# MANUEL DE CALCUL NUMÉRIQUE APPLIQUÉ

Christian Guilpin

*Maître de Conférence, responsable de l'enseignement des mathématiques appliquées en maîtrise de Physique et Applications à l'université Paris VII-Denis Diderot*


7, avenue du Hoggar
Parc d'Activité de Courtabœuf, BP 112
91944 Les Ulis Cedex A, France

ISBN : 2-86883-406-X

# Avant-propos

Ce livre de calcul appliqué trouve son origine dans un cours dispensé aux étudiants de maîtrise de physique et applications (MPA) à l'université Paris VII depuis 1984, ainsi que dans quelques préoccupations de recherche au laboratoire.

Ce manuel ne constitue pas à proprement parler un cours au sens usuel du mot, et si certains chapitres ont des liens évidents et s'enchaînent naturellement, d'autres, plus isolés, ont été réunis sous forme d'annexes pour préserver au mieux l'unité, mais leur importance n'est pas secondaire.

L'organisation de cet ouvrage vise à une présentation suffisamment concise et assimilable des algorithmes numériques fondamentaux, développés jusqu'à leur mise en œuvre, de telle sorte qu'ils soient susceptibles d'aider l'étudiant, le chercheur et l'ingénieur dans l'exercice quotidien de leur art : il s'agit de pouvoir obtenir des résultats numériques convenables chaque fois qu'une méthode analytique fait défaut.

Pour ce qui concerne certains théorèmes très importants, nous nous sommes parfois borné à les énoncer sans les démontrer, le contraire eut risqué de nous éloigner de notre préoccupation majeure : le résultat numérique ; cependant, les indications bibliographiques permettent d'obtenir aisément ces démonstrations qui sont classiques.

Les objectifs poursuivis se situent sur deux plans que l'on a coutume de séparer mais qui sont indissociables de notre point de vue : l'acquisition d'algorithmes numériques indispensable à la résolution de problèmes usuels et la maîtrise du traitement des données expérimentales selon la méthode statistique. Traiter les données de l'expérience impose l'usage de techniques numériques appropriées, et l'examen des résultats entachés d'erreur et d'incertitude impose l'usage de la statistique. La propagation des erreurs à travers les algorithmes relève d'une analyse subtile qui est éternellement omise tant elle est délicate. Nous avons tenté de l'effleurer et c'est une des raisons qui nous a poussé à développer l'étude des lois de distribution ainsi que leurs fondements dans une partie qui est davantage dévolue aux statistiques.

De même qu'il est impensable de vouloir apprendre à jouer du piano la veille de donner un concert, de même il est impensable de vouloir apprendre l'algorithmique numérique le jour où le besoin s'impose. Dans les deux cas, il convient de recourir aux gammes afin d'acquérir une solide expérience. En calcul numérique il n'y a pas de voie royale, et aucun algorithme n'est capable de fournir de résultats corrects quelles que soient les données fournies. Il est toujours possible de mettre en défaut une procédure et d'obtenir des résultats abérrants pourvu que l'on s'en donne la peine... Un très bel exemple est étudié à l'occasion de la résolution des systèmes linéaires dépendant d'une matrice de Hilbert.

L'expérience pratique prend alors toute sa valeur, et c'est ainsi que notre enseignement comporte une séance hebdomadaire de trois heures sur calculateur arithmétique. Peu importe

le langage et la manière, seul le « résultat correct » compte et ce n'est pas une mince affaire que de se faire une opinion sur les erreurs qui entachent les résultats finals. Ensuite viendront éventuellement se greffer les problèmes d'élégance et d'optimisation.

En aucun cas, cet enseignement n'a pour but d'explorer et de recenser tous les algorithmes ayant trait à un type de problèmes. Nous avons voulu présenter ceux qui se sont montrés paradigmatiques soit sous l'angle de la simplicité soit sous l'angle de l'efficacité. Il s'agit de construire des programmes que nous aurons soigneusement testés, dont nous connaîtrons les limites et qui rempliront peu à peu notre boîte à outils.

Pour simplifier, nous dirons que ce livre peut se subdiviser en trois parties à savoir :

1. Études d'algorithmes numériques et leur mise en oeuvre.
2. Analyse statistique des résultats d'expériences.
3. Annexes, problèmes et corrigés.

Nous l'avons déjà dit, les deux premières parties interfèrent partiellement, et c'est une des raisons pour laquelle nous avons renoncé à présenter un ouvrage où tout ce qui est étudié dans un chapitre s'appuie nécessairement sur ce qui a été établi précédemment. Par souci d'unité nous avons préféré regrouper les titres par centre d'intérêt. Ainsi, il nous est apparu plus intéressant d'avoir rassemblé l'étude des polynômes orthogonaux plutôt que d'avoir dispersé l'information dans différents chapitres concernant l'interpolation et l'intégration numérique.

Il aurait été dommage de ne pas avoir abordé, ne serait-ce que rapidement, les méthodes de Monte-Carlo d'une part, et les problèmes mal posés d'autre part. Ces domaines illustrent bien la synthèse des deux premières parties, d'autant plus qu'ils s'intègrent remarquablement dans les préoccupations des chercheurs et des ingénieurs. Qui, en physique, n'a pas eu à résoudre numériquement une équation de convolution? Qui n'a pas tenté la résolution d'un problème au moyen d'une simulation?

Pour terminer nous proposons un avant-dernier chapitre constitué d'un ensemble de problèmes et d'exercices qui illustrent quelques usages des méthodes qui ont été présentées ; ils servent également à éclairer quelques points de théorie qui seraient venus alourdir le cours s'ils avaient été intégrés dans les divers chapitres : on montre par exemple que le coefficient de conformité de Pearson obéit bien à une loi du $\chi^2$. Le dernier chapitre donne les solutions des problèmes présentés.

La plupart des chapitres font l'objet d'une illustration et se terminent par des programmes écrits dans le langage C : il s'agit du langage de base qui assure la portabilité. Ce point de vue s'explique par la facilité qu'il y a à changer de langage : Fortran, Pascal, etc., sans avoir grand chose à modifier dans le programme source. On n'est pas obligé de partager ces vues, mais il est très facile de modifier les programmes proposés pour qu'ils apparaissent moins « archaïques ».

Pour en finir avec les algorithmes choisis et les programmes présentés, nous dirons qu'ils sont fournis **sans garantie** d'aucune sorte malgré le grand soin porté à ce travail. Ils peuvent comporter des imprécisions voire des imperfections, à ceci s'ajoute le fait qu'aucun algorithme n'est irréprochable dans la mesure où il est toujours possible de trouver des valeurs numériques qui le mette en défaut.

Bien sûr, nous formons le vœu que cet ouvrage puisse apporter une aide solide aux étudiants, ingénieurs et chercheurs pour lesquels il constituera un outil dont le rôle favorise la réalisation de sa propre boîte à outils.

La rédaction d'un ouvrage ne se réalise jamais dans l'isolement, et il m'a fallu bien des oreilles attentives, bien des lecteurs vigilants, bien des conseillers éclairés. L'instant est venu de remercier tous ceux qui, à quelque titre que ce soit, m'ont apporté une aide inconditionnelle, je citerai par ordre alphabétique : Claude Bardos, Jean Bornarel, Jacques Gacougnolle, Patricia Guilpin,

Michel Jacques, Claude Marti, Yvan Simon ainsi que l'équipe de physique théorique de Chaouqui Misbah.

Pour terminer, j'ajouterai une mention particulière à EDP Sciences qui m'a offert un contexte de travail optimum afin d'obtenir la meilleure réalisation possible.

Christian GUILPIN

## PROGRAMMES SOURCES ACCOMPAGNANT CE MANUEL

Les programmes sources cités dans ce livre sont disponibles sur le site Web d'EDP Sciences (adresse : `http://www.edpsciences.com/guilpin/`).

**Chapitre 2**
- aitken_0.c
- epsilon_0.c
- richard.h
- richar_0.c
- epsilon_1.c
- epsilon_2.c
- epsilon_3.c
- epsilon_4.c
- aitken_2.c
- kacmarz1.c
- fredholm.c
- newton_1.c

**Chapitre 4**
- dichot0.c
- itera.c
- newton1d.c
- newtonpp.c
- kacmarz.c
- newton2d.c
- bairstow.c
- bairstow.h
- racine.h

**Chapitre 5**
- systlin.c
- triangle.c
- trianlin.c
- hilbert.c
- trianinv.c
- leverier.c
- givens.c
- rutisacc.c
- danilev.c
- dsyslin.h

**Chapitre 6**
- lagpoly.c
- ascend.c
- descend.c
- lispline.c
- spline.h
- sudeter.c
- multma.h
- transpos.h
- dsyslin.h
- invers.h

**Chapitre 7**
- legendre.c
- decomleg.c
- r_legend.c
- getname.h

**Chapitre 9**
- laguerre.c
- r_laguer.c

**Chapitre 10**
- hermite.c
- r_hermit.c

**Chapitre 11**
- rn_x.txt
- sn_x.txt
- un_x.txt
- vn_x.txt

**Chapitre 12**
- argent.c
- cotes_1.c
- cotes_2.c

**Chapitre 13**
- epsilon.h
- combina.h
- retrogra.c
- runge.c
- pendule.c
- moulton.c
- bashfort.c
- taylor.c

**Chapitre 14**
- triode.c
- chaleur.c
- corde.c

**Chapitre 15**
- echantil.c
- gibbs.c
- dirac.c
- filtre.c

**Chapitre 16**
- dfft0.h
- dfftinv.h
- tf_image.c
- inv_imag.c

**Chapitre 17**
- fredh_1.c
- gaussien.h
- dconvol.c
- dconvol.h
- intmonte.h
- xaleamen.h

**Chapitre 18**
- calculpi.c
- matmonte.c
- intmonte.c
- recuit.c

**Chapitre 20**
- gaussleg.h

**Chapitre 22**
- khi2.h
- student.h

**Chapitre 24**
- histogra.c
- kolmogor.c

**Chapitre 25**
- teststat.c
- varian_1.c
- varian_2.c

**Chapitre 26**
- regres.c

**Annexe A**
- sturm.c

**Annexe F**
- bessel1n.c
- bessel1f.c
- bessel2n.c
- bessel2f.c
- besseljf.h
- besseljn.h
- bessel2f.h

**Annexe I**
- dzeta0.c
- dzeta1.c
- grosys1.c
- jacobi0.c
- jacobi1.c
- pendule0.c
- predcor0.c
- souriau0.c
- sudeter.c
- syslinit.c
- tangent2.c
- gradconj.c
- refrig1.c
- mathieu9.c
- vanderp0.c
- card0.c

# Sommaire

# 1 Généralités sur le calcul numérique

Le calcul numérique est une branche des mathématiques appliquées qui étudie les méthodes pratiques destinées à fournir des solutions numériques aux problèmes formalisés dans le langage des mathématiques pures. La plupart du temps, ce sont les ingénieurs et les chercheurs qui se trouvent concernés par l'usage de ces méthodes pratiques, car ce sont en définitive les nombres qui vont retenir leur attention. Face à un jeu d'équations plus ou moins complexes, ils devront obtenir un ensemble de valeurs numériques qui pourra servir par exemple soit à la réalisation d'un édifice ou d'un prototype, soit à la confrontation des résultats expérimentaux et théoriques etc.

## 1. La notion d'algorithme en calcul numérique

Un algorithme est un procédé de calcul qui ne met en œuvre que des opérations arithmétiques et logiques. L'étymologie de ce mot est arabe et c'est une altération du nom du mathématicien Al-Khwârizmi (†812?), probablement sous l'influence du mot grec (repris par les latins) ὁ αριθμός, le nombre.

Quoi qu'il en soit, c'est une dénomination commode qui sert à désigner l'ensemble des opérations qui interviennent au cours d'une démonstration conduisant à l'énoncé d'un théorème. Cependant sa portée ne dépasse pas celle de la « prose que chacun fait sans le savoir », et en aucun cas cette définition ne permet de donner une manière de construction des algorithmes. Il s'agit donc d'un concept commode mais peu fécond qui sert à désigner un certain type d'organisation de propositions à caractère mathématique. Sans que rien ne soit changé, il est tout à fait possible de remplacer ce mot par procédé de calcul, technique de calcul, procédure...

Pour illustrer ce concept, on peut évoquer, par exemple, la technique de résolution des équations du deuxième degré à coefficients réels à condition toutefois d'admettre que l'on dispose outre les quatre opérations fondamentales (addition, soustraction, multiplication et division) de l'opération racine carrée. L'examen des mathématiques montre que le nombre d'opérations proposées dans un algorithme peut être infini (mais dénombrable), c'est le cas des développements en série de fonctions : série entière, série de Fourier, etc.

Le concept d'algorithme en calcul numérique est quelque peu plus restrictif : le nombre d'opérations élémentaires arithmétiques et logiques est obligatoirement fini. En outre, cela implique que l'on ne peut manipuler que des nombres admettant une représentation finie ce qui conduit à effectuer des troncatures et des arrondis au cours des opérations successives. Cette remarque en apparence triviale doit pourtant être présente à l'esprit lorsque l'on fait usage d'une machine arithmétique (mais aussi du calcul manuel...) : la première division venue, la première

racine carrée venue ont toutes les chances d'introduire un nombre dont la représentation impose une infinité de chiffres significatifs, il faudra les tronquer...

La connaissance de la manière dont les nombres sont représentés dans la machine utilisée est fondamentale en ce sens qu'elle permet l'estimation de la précision attachée aux résultats d'un calcul. Dès lors, on comprend très bien que la recherche d'une grande précision imposera une grande taille de mots-machine et par conséquent un grand temps de calcul. En revanche, une faible précision n'imposera qu'une petite taille du mot-machine ainsi qu'un faible temps de calcul. Le choix retenu est en général un compromis entre ces deux situations extrêmes.

Cette dernière étape intervient lors de l'évaluation de l'incertitude qui entache les résultats d'un calcul, mais ce n'est pas la seule source. Le problème général de l'évaluation des incertitudes et erreurs est délicat qui plus est lorsque l'on fait usage d'une machine ; nous l'aborderons un peu plus en détail dans quelques paragraphes.

## 2. Le calcul numérique ne concerne que les nombres entiers

Ce n'est pas une boutade, et il faut bien admettre que l'on ne peut manipuler (au sens étymologique) que des représentations finies. La représentation dite en virgule flottante éclaire ce point de vue car en fait elle traite de nombres ayant une certaine quantité fixe de chiffres significatifs (agrémenté d'un facteur de cadrage appelé exposant) et dans la réalité, à chaque opération élémentaire sur les nombres il y a une opération de cadrage suivie d'une opération arithmétique sur les nombres entiers, elle-même suivie éventuellement d'une opération de troncature ou d'arrondi.

Cette conception n'est pas tellement restrictive en soi, car les seules opérations arithmétiques pratiques que l'Homme est susceptible de réaliser ne peuvent que concerner les nombres admettant une représentation finie de symboles (chiffres significatifs) donc en fait des entiers. Pour ce qui concerne les machines arithmétiques, les deux représentations, entière et flottante, ne doivent pas masquer la réalité et il faut voir là uniquement deux représentations commodes. En fin de chapitre, on trouvera un paragraphe traitant de quelques représentations assez générales.

## 3. Le calcul numérique traite du problème pratique de l'approximation de fonctions explicites ou implicites

Un problème d'analyse se trouvant modélisé dans le langage des mathématiques pures, la première préoccupation consiste à rechercher la solution sous forme littérale à l'aide des fonctions connues qui sont les polynômes et les fonctions transcendantes élémentaires. On peut se poser la question de savoir si cette façon de concevoir est toujours possible.

Tous les problèmes ne sont pas algorithmiquement solubles, et force est de répondre non à la question posée. Pour s'en convaincre, il suffit de considérer un exemple, sous l'angle de l'histoire, qui illustre cette difficulté : la recherche des racines d'un polynôme de degré quelconque à coefficients réels.

Si l'équation du deuxième degré est connue depuis l'Antiquité, on peut dire qu'il revient à Al-Khwârizmi (fin VIIIe, début IXe) et à Luca Pacioli (1445?–1514?) le fait d'avoir raffiné les solutions.

L'école italienne de Bologne s'attaque à l'équation du troisième degré ; on retiendra à son propos les noms de Tartaglia (1500?–1557), de Cardan (1501–1576) qui parvinrent à la solution au travers de défis et de provocations qui semblaient être coutumiers à cette époque. Signalons toutefois que les bases de l'étude sont dues à Del Fero (1465?–1526).

Ensuite l'équation du quatrième degré fut abordée et résolue par Ferrari (1522–1565) qui fut un élève de Cardan. Plus tard, les mathématiciens s'intéressèrent tout naturellement à l'équation du cinquième degré et les recherches furent très riches d'enseignements dans la mesure où, selon toute vraisemblance pour la première fois dans l'histoire des sciences, le travail conduisit à l'énoncé d'un résultat négatif.

1. En 1608, Rothe (?–1617) émit une proposition selon laquelle toute équation algébrique de degré $n$ possédait $n$ racines réelles ou complexes, et Gauss (1777–1855) en effectua la démonstration rigoureuse en 1799.
2. Cette même année Ruffini (1765–1822) affirma l'impossibilité formelle d'obtenir la résolution des équations algébriques de degré supérieur à quatre. Il fallut attendre la publication des travaux d'Abel (1802–1829) pour obtenir la démonstration définitive de la proposition de Ruffini (1826).

Ce résumé de l'histoire des équations algébriques nous mène à deux types de remarques :

***Remarque 1*** : Le fait de poser un problème dont la solution existe n'est pas suffisant pour permettre d'exhiber effectivement ladite solution selon des moyens donnés. On peut se rappeler à ce propos le problème de la trisection de l'angle qui n'est pas algorithmiquement soluble si les moyens de construction autorisés sont la règle et le compas.

Aujourd'hui, on sait que les problèmes que l'on peut se poser sont de trois types à savoir :

a. les problèmes algorithmiquement solubles, l'équation du deuxième degré par exemple ;
b. les problèmes algorithmiquement non solubles, par exemple la quadrature du cercle à l'aide de la règle et du compas ;
c. les problèmes indécidables pour lesquels on ne peut rien démontrer. Du reste, dans une axiomatique donnée, on sait qu'il existe des propositions vraies que l'on ne peut pas démontrer, à condition de considérer toutefois que les axiomes de l'arithmétique ne sont pas contradictoires (*cf.* Gödel (1906–1978)).

***Remarque 2*** : Ce n'est pas parce qu'un problème est algorithmiquement insoluble que l'on n'est pas en mesure de proposer une solution approchée — généralement avec une précision fixée à l'avance — et c'est justement la raison d'être du calcul numérique que de fournir de telles solutions ; par exemple, on verra comment calculer les racines d'un polynôme de degré quelconque à coefficients réels.

En terminant ce paragraphe, nous remarquerons que les problèmes de calcul numérique sont uniquement des problèmes d'approximation de fonctions au sens le plus large. Souvenons-nous d'une des premières équations différentielles que nous avons rencontrée en Physique, il s'agit de l'équation du pendule pesant assujetti à osciller dans un plan. On écrit traditionnellement l'équation à laquelle obéit son mouvement :

$$\frac{\mathrm{d}^2\theta(t)}{\mathrm{d}t^2} + \omega^2 \sin\theta(t) = 0,$$

où $\theta(t)$ est l'angle que le pendule fait avec la verticale à un instant donné, $\omega$ étant une constante dépendant de la longueur du pendule et de l'accélération. Alors, on apprend que l'on ne peut pas obtenir la solution sous forme littérale ($t$, $\theta_0$, $\theta'_0$, $\omega^2$) à l'aide des transcendances usuelles ($\theta_0$ et $\theta'_0$ sont les conditions initiales sur l'angle et la vitesse angulaire). Pourtant, le théorème de Cauchy-Lipschitz nous permet d'affirmer l'existence et l'unicité d'une telle solution qui est de surcroît une fonction analytique.

La tradition veut également que l'on ne s'intéresse qu'au cas des petits angles pour lesquels on peut écrire le développement de $\sin\theta(t)$ au premier ordre. Alors, dans ce cas particulier, on sait

obtenir la solution sous forme littérale. Cependant, chacun sait que les pendules sont capricieux, et tous n'ont pas le bon goût de vouloir limiter l'amplitude de leurs oscillations. C'est le rôle du calcul numérique que d'apporter une réponse à ces types de problèmes et l'on verra chaque fois de quelle manière.

## 4. Solutions littérales et solutions analytiques

À propos du pendule, nous avons dit que la solution $\theta(t, \theta'_0, \omega^2)$ ne pouvait être exprimée sous forme littérale c'est-à-dire que l'on ne pouvait pas donner une expression formelle en fonction des transcendances usuelles et des polynômes. Cela ne signifie nullement que la solution n'est pas une fonction analytique. À ce sujet, rappelons qu'une fonction est dite analytique lorsqu'elle est développable en série entière, que la fonction soit réelle ou complexe, que les arguments soient réels ou complexes. D'ailleurs, sans préjuger de la suite, on peut ajouter que le calcul numérique fait amplement appel aux développements en série dans de nombreuses méthodes, et d'une façon tout à fait générale, on peut affirmer que toutes les solutions numériques sont des valeurs numériques de fonctions analytiques. En résumé, pour éviter tout abus et toute ambiguïté, il convient de distinguer l'expression formelle d'une fonction avec son expression analytique.

## 5. Que sait-on calculer rigoureusement ?

Comme nous ne pouvons traiter que de nombres admettant une représentation finie de chiffres significatifs, seules l'addition, la soustraction et la multiplication donneront des résultats rigoureux lors de l'exécution des différentes opérations (encore avons-nous fait abstraction des réels problèmes de la taille réservée pour représenter les nombres). La première division venue a statistiquement toutes les chances d'introduire un résultat comportant un nombre infini de chiffres significatifs.

Sur le plan du calcul des fonctions, seuls les polynômes (dont le degré est évidemment fini) sont susceptibles d'être calculés rigoureusement (ou avec une précision souhaitée à l'avance dans la mesure où le calcul peut faire intervenir la division). Là repose le grand intérêt des développements en série (de MacLaurin (1698–1746) et Taylor (1685–1731)) qui fournit une expression calculable sachant que l'on se fixe à l'avance une précision donnée et sachant que l'on peut majorer convenablement l'expression du reste de la série. Au passage, il est bon de noter que l'on sait en général réaliser une opération formelle importante : la dérivation, indispensable au calcul des coefficients du développement en série.

Considérons une fonction $f(x)$ régulière c'est-à-dire continue, dérivable autant de fois que l'on veut dans un domaine $D$ contenant $a$ et $a + h$, son développement en série de Taylor s'écrit :

$$f(a+h) = f(a) + \frac{f'(a)}{1!}h + \frac{f''(a)}{2!}h^2 + \cdots + \frac{f^{(n)}(a)}{n!}h^n + R_{n+1}$$

où $R_{n+1}$ est l'expression du reste lorsque l'on tronque par nécessité le développement à l'ordre $n$. Ce reste s'écrit :

$$R_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}$$

où $\xi$ est un nombre compris entre $a$ et $(a + h)$.

Il est quasiment impossible de connaître $\xi$ et, conformément à l'usage, on cherche un majorant de $R_{n+1}$ dans l'intervalle $(a, a + h)$. À ce propos, rappelons que si $R_{n+1}$ tend vers zéro quand

$n$ tend vers l'infini, pour une fonction pourvue d'une infinité de dérivées, le développement de Taylor devient une **série entière** appelée série de MacLaurin.

En définitive, nous ne savons calculer que peu de choses avec une extrême rigueur, mais les polynômes en font partie, et cela leur confère un rôle particulièrement intéressant en calcul numérique. À bien y réfléchir, cela met en relief une autre facette du théorème d'approximation de Weierstrass (1815–1897) qui date de 1885 et sur lequel nous aurons le loisir d'insister ultérieurement à propos de l'interpolation (chapitre 6).

## 6. Les erreurs et les incertitudes

Un algorithme donné permet d'obtenir des résultats numériques qui, malheureusement, sont entachés d'erreurs dont les sources proviennent d'origines différentes.

Tout d'abord, il faut évoquer l'erreur au sens trivial du terme, c'est-à-dire la méprise ; elle mérite quelques commentaires tout simplement parce qu'elle est polymorphe et qu'elle n'est pas toujours évidente à détecter. Cela peut aller de l'erreur de méthode ou d'algorithme à la simple faute de copie ou de transcription. Que d'ennuis liés à la transmission de mauvaises données, alors que le programme est tout à fait convenable. C'est la raison pour laquelle, derrière toute instruction de lecture, il est impératif de réécrire la donnée communiquée et c'est du reste une bonne façon de la voir figurer sur la feuille de résultats.

Ici, comme ailleurs, il n'y a pas de règles universelles qui permettent d'éviter la bévue ; cependant, lors de la phase de mise au point, évitons les compilateurs tolérants, vérifions que nous retrouvons les résultats connus liés à certains types de problèmes, sollicitons le jugement éclairé d'un collègue...

Maintenant, dans l'hypothèse où l'algorithme retenu est adéquat, et que sa programmation apparaît sans faille (ce qui implique de surcroît que l'algorithme soit adapté aux données traitées), on peut envisager l'étude des erreurs entachant les résultats numériques obtenus.

### 6.1. Erreurs à caractère mathématique

La différence entre la solution strictement mathématique et la solution approchée fournit un terme d'erreur : c'est le cas fréquemment rencontré lors de l'étude des séries infinies, des suites infinies, des produits infinis, etc. Dans tous les cas classiques, on arrive à connaître un majorant raisonnable des restes qui sont abandonnés.

### 6.2. Erreurs liées à l'exécution des calculs

Les données numériques confiées à une machine ont deux origines : ce sont des nombres d'origine strictement mathématique (calcul d'une intégrale par exemple) ou des données provenant de mesures lesquelles sont déterminées dans un intervalle (autrement dit à une certaine précision près), c'est ce que le physicien dénomme **incertitude**. Quelle que soit leur origine, les données comportent une erreur ou une incertitude, car même dans le premier cas, les opérations de conversion et d'arrondi introduisent statistiquement une erreur.

La machine traite ces données selon un certain algorithme et procède à chaque étape élémentaire à une opération d'arrondi. L'arrondi est une opération qui consiste à ajouter à la mantisse du nombre la valeur 0,5 puis à tronquer le résultat, ce qui permet de diviser les erreurs machine par deux.

Après l'exécution de tous les calculs, on doit se demander inévitablement quelle est la précision du résultat final. Il s'agit de voir comment l'incertitude finale a été propagée au cours de toutes

les opérations d'arrondi. Le problème n'est pas si simple que cela et dépend essentiellement de la manière dont l'algorithme propage les erreurs indépendamment des calculs effectués.

## 7. Un problème difficile : la propagation des erreurs en calcul automatique

Si les calculs font intervenir des fonctions transcendantes usuelles que l'on trouve sur toutes les machines, il faut alors avoir recours au manuel du fabricant de logiciels pour connaître l'incertitude attachée à l'usage de telle ou telle fonction, incertitudes qui dépendent, en règle générale, de la taille de l'argument. Nous faisons allusion plus spécialement aux fonctions dites **fonctions de bibliothèque** : sinus, cosinus, tangente, arc sinus, arc cosinus, arc tangente, exponentielle, logarithme, racine carrée (annexe E)... Toutefois les erreurs affectant les résultats (fournies par le fabricant) ne sont pas la majoration de l'erreur relative ou absolue mais plus simplement l'écart quadratique moyen ; cela fournit une valeur bien plus raisonnable de l'erreur qui serait sans cela toujours exagérément surestimées. Cette conception est tout à fait convenable dans la mesure où l'on peut supposer que les erreurs obéissent à la loi de Gauss-Laplace, mais cette hypothèse admet des limites qu'il convient de ne pas franchir et nous en verrons un exemple un peu plus loin.

En résumé, les calculs sont réalisés au moyen d'opérations introduisant des arrondis et de fonctions de bibliothèque affectées d'une précision limitée. Que peut-on conclure quant à la précision des résultats finals ? Il n'est pas possible d'apporter une réponse à la question ainsi posée car les erreurs dépendent de la manière dont l'algorithme les propage et la limite supérieure de l'erreur finale n'est pas nécessairement un majorant de la somme de la borne supérieure du module des erreurs évaluées à chaque opération élémentaire. Pour s'en convaincre, il suffit de procéder selon cette technique et l'on s'apercevra bien vite que cette façon de concevoir les incertitudes se révèle très exagérée et très surestimée ; en effet, il n'y a aucune chance pour que toutes les opérations du calcul soient systématiquement l'objet d'une erreur maximum.

L'algorithme propage des erreurs, mais pour bien situer le problème disons que, de ce point de vue, on rencontre deux types d'algorithmes : ceux qui font appel à des calculs cumulatifs reposant sur l'addition (au sens large) et les calculs itératifs reposant sur la répétition du calcul avec chaque fois une nouvelle valeur donnée par le précédent tour.

### 7.1. Les calculs itératifs

Comme nous l'avons dit ce sont des calculs répétitifs que l'on limite nécessairement à un certain ordre, et qui consistent à réintroduire dans le calcul la dernière valeur calculée. Le plus souvent, la procédure n'a d'intérêt que dans la mesure où la suite générée est convergente.

La limite obtenue ne dépend alors que de la précision de la machine utilisée (représentation des nombres, fonctions de bibliothèque, ...). Cependant, il convient d'ajouter que le cumul des erreurs au cours d'un tour de calcul interviendra au niveau de la vitesse de convergence du processus.

***Exemple de calcul répétitif*** – Calcul de la racine carrée d'un nombre positif $N$.

Désignons par $a_0$ une première approximation de $\sqrt{N}$ et par $e_0$ l'erreur liée à cette estimation. On a la relation :

$$\sqrt{N} = a_0 + e_0.$$

Nous allons essayer d'obtenir une approximation de $e_0$. Pour cela développons au premier ordre l'expression :

$$N = (a_0 + e_0)^2.$$

On obtient :

$$N = a_0(a_0 + 2e'_0),$$

expression dans laquelle $e_0$ est remplacée par $e'_0$ qui en est une approximation. On déduit :

$$e'_0 = (N/a_0 - a_0)/2,$$

soit encore :

$$a_1 = a_0 + e'_0 = (N/a_0 + a_0)/2.$$

La somme $a_0 + e'_0$ constitue une meilleure approximation que $a_0$ sous réserve toutefois que le développement soit convergent. Dans ce dernier cas, rien ne nous empêche de recommencer les calculs avec la dernière approximation évaluée $a_1$ (itération), et ceci jusqu'à ce que l'on ait obtenu le meilleur résultat que la précision de la machine puisse permettre d'obtenir. De toute façon, il faudra choisir un critère tel que le nombre d'itérations soit fini, encore faut-il que la suite de nombres générée soit convergente.

Pour s'assurer de cette convergence, il faut et il suffit que :

$$|e_j| < |a_j|.$$

Donc pour débuter les itérations on peut prendre $a_0 = N$ par exemple. L'erreur sur le calcul dépend de la limite de l'erreur mathématique $e_j$ qui ne tend pas vers zéro du point de vue numérique. $e'_\infty$ admet une limite non nulle $e_\infty$ parce que nous avons affaire à des erreurs de troncature. Cela conditionne la précision de la valeur finale, mais cela peut avoir pour effet de donner une suite $(a_k)$ qui, lorsque $k$ est grand, fournit alternativement deux valeurs limites. Nous aurons l'occasion d'examiner plus en détail ce problème en étudiant les racines des équations.

Dans cette classe de problèmes où les erreurs ne sont pas cumulées dès le début de l'algorithme, on peut intégrer d'autres types d'algorithmes qui relèvent de la même analyse parmi lesquels on peut citer la dichotomie (*cf.* chapitre 4).

## 7.2. Les calculs cumulatifs

Il s'agit de calculs dont le résultat dépend de tout l'ensemble des calculs intermédiaires. Les exemples les plus immédiats concernent la somme d'une série numérique ou encore le calcul d'une intégrale. Le résultat définitif dépend évidemment, toujours du point de vue de la précision, du nombre de termes calculés et du nombre de chiffres significatifs dont la machine dispose. À cet égard, une méfiance toute particulière doit être manifestée lorsque l'on traite des séries lentement convergentes ou lentement divergentes, car il est possible d'obtenir à peu près n'importe quoi, l'erreur pouvant devenir quasiment infinie. L'exemple le plus connu qui illustre parfaitement ces propos repose sur la série harmonique qui est très lentement divergente.

Le calcul effectif de la série divergente $\sum_{n=1}^{\infty} 1/n$ donne toujours un résultat fini en faisant usage des machines usuelles, car la somme partielle $S_p$ obtenue par sommation des $p$ premiers termes qui apportent effectivement une contribution à la somme est toujours très inférieure à la taille du plus grand nombre représentable dans lesdites machines. Pour fixer les idées, si la précision relative fournie par la machine est $10^{-k}$, le nombre $p$ sera obtenu lorsque :

$$1/(p-1)/S_p < 10^{-k}.$$

Autrement dit le terme $(p+1)$ ne peut plus modifier la somme partielle déjà calculée, il en sera de même des termes suivants qui sont encore plus petits.

Si la série avait divergé plus rapidement, nous aurions pu être alerté par un dépassement de capacité qui aurait arrêté l'exécution des calculs c'est-à-dire qu'à partir d'un certain rang la somme partielle aurait dépassé le plus grand nombre représentable en machine, malheureusement il n'en a rien été. Ces considérations sur les séries attirent deux autres remarques :

1. Un dépassement de capacité signalé ne signifie pas obligatoirement que l'algorithme utilisé est divergent, mais tout simplement qu'au cours du calcul la somme partielle s'est montrée supérieure au plus grand nombre représentable. Il est facile de donner un tel exemple en considérant le développement de $\exp(x)$ :

$$\exp(x) = x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

Utilisons ce développement toujours convergent (rayon de convergence infini) pour calculer $\exp(-1\,000)$. Nous savons que cette valeur est très petite, mais bien que le développement de $\exp(x)$ soit absolument convergent le calcul de la série alternée va poser des problèmes quasiment insurmontables. *Grosso modo* la somme partielle va croître jusqu'à ce que $n!$ l'emporte sur $x^n$, donc lorsque $1\,000^n \sim n!$. En passant aux logarithmes puis en faisant usage de la formule de Stirling $\log_e(n!) = n\log_e(n) - n$, on aboutit au résultat suivant :

$$n\log_e(1\,000) = n\log_e(n) - n,$$

d'où

$$\log_e\left(\frac{n}{1\,000}\right) = 1.$$

On obtient $n$ en écrivant que $n/1\,000 = e$ ($e$ est la base des logarithmes népériens). Soit $n = 2\,718$. Il n'est peut-être pas inutile d'insister sur la taille immense des nombres intermédiaires qui doivent être calculés ($2\,718! \sim 10^{8\,154}$) et qui dépasse de très loin la capacité des mots-mémoire les plus optimistes. Cet exemple montre à l'évidence que les fonctions de bibliothèque ne sont certainement pas calculées au moyen de développements de ce type. Nous avons consacré les annexes C, D, E, à la technique de calcul des fonctions de bibliothèque.

2. Le calcul numérique n'exclut pas de ses méthodes l'usage de certaines séries divergentes (encore appelées séries semi-convergentes) : on verra des applications lors de l'étude des développements asymptotiques et lors de l'étude de l'epsilon-algorithme appliqué aux fonctions admettant un prolongement analytique (chapitre 2).

Revenons un instant sur les séries lentement convergentes ; elles constituent un piège redoutable car rien ne permet de se défier du résultat si ce n'est justement un calcul d'erreur. Nous pensons plus particulièrement aux séries de Fourier (1768–1830) lorsque les coefficients décroissent comme $1/n$. Dans ce domaine, rien n'est simple car la majoration abusive des erreurs conduit tout aussi inévitablement à pénaliser voire à rejeter des résultats qui pourraient être acceptables.

***Un exemple de calcul d'erreur sur la somme d'une série convergente*** – Dans le simple but de supprimer les cadrages des nombres intermédiaires, et donc de raisonner plus facilement sur les erreurs absolues, nous allons examiner en détail le calcul d'une série banale et connue dont la somme vaut l'unité, soit :

$$S = \frac{1}{1\times 2} + \frac{1}{2\times 3} + \frac{1}{3\times 4} + \frac{1}{4\times 5} + \cdots + \frac{1}{n\times(n+1)} + \cdots$$

Comme

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{(n+1)},$$

on voit que la somme partielle limitée au $m^{\text{e}}$ terme s'écrit :

$$S_m = 1 - \frac{1}{(m+1)}.$$

En effectuant un passage à la limite quand $m$ devient infini, on en conclut que la limite de la série est égale à 1. Ce n'est pas l'examen mathématique de cette série qui va retenir notre attention, mais uniquement le calcul numérique effectif ainsi que l'erreur qui en résulte.

Quand on limite la somme aux $m$ premiers termes, on peut majorer l'erreur de troncature de cette série, c'est une erreur mathématique qui est majorée par le premier terme abandonné de la série convergente alternée :

$$e_m = \frac{1}{(m+1)}.$$

À présent supposons que les calculs soient réalisés sur une machine qui travaille avec des nombres représentés sur 5 octets. Chaque étape du calcul introduit une erreur de l'ordre de (*cf.* § 9) :

$$e_a = 10^{-9{,}332}.$$

Comme il n'y a pas d'erreur sur les multiplications (pas de troncature), seules sont prises en compte les erreurs sur les inversions et sur les additions : soient au total $2m$ opérations susceptibles d'apporter une contribution à ce que nous appellerons l'erreur globale $E_g$. On peut majorer $E_g$ :

$$E_g = 2me_a.$$

Nous donnons ci-dessous un tableau (Tab. 1.1) où sont portés $m$, $S_m$, l'erreur globale, l'erreur statistique et $\left|1 - S - \frac{1}{m+1}\right|$.

$|1 - S - \frac{1}{m+1}|$ est un terme qui représente les erreurs cumulées au cours des opérations réalisées en machine, ce terme est désigné par erreur réelle dans le tableau. Nous avons choisi de faire varier $m$ de 10 000 en 10 000 jusqu'à 50 000, puis de passer à la valeur 100 000.

**Tableau 1.1.**

| m | S | err. globa. $10^8$ | err. stat. $10^6$ | err. réelle $10^6$ |
|---|---|---|---|---|
| 10 000 | 0,999 900 | 0,931 | 0,475 | 0,329 |
| 20 000 | 0,999 950 | 1,86 | 0,672 | 0,440 |
| 30 000 | 0,999 967 | 2,79 | 0,822 | 0,255 |
| 40 000 | 0,999 975 | 3,72 | 0,950 | 2,88 |
| 50 000 | 0,999 980 | 4,66 | 1,06 | 2,30 |
| 100 000 | 0,999 991 | 9,31 | 1,502 | 77,0 |

Il est facile de constater que les erreurs globales $E_g$ majorées *a priori* ne représentent pas du tout la réalité, et pour $m = 50\,000$, $E_g$ est environ 500 fois trop grand. On en conclut que cette manière de concevoir les erreurs n'apporte rien et surtout n'est pas réaliste. Il est bien préférable de raisonner en termes de probabilité. Du reste, la probabilité pour que l'erreur soit effectivement égale à $E_g$ est pratiquement nulle.

# 8. Réexamen des erreurs du point de vue statistique

Supposons que l'on utilise une machine qui travaille sur des mantisses de $n$ chiffres significatifs. On va alors raisonner sur la valeur absolue de la mantisse considérée comme un nombre entier.

Faisons l'hypothèse que l'erreur de troncature puisse être considérée comme une variable aléatoire continue que l'on désigne par $X$ à valeur sur $(0,1)$.

## 8.1. Cas où la distribution des X peut être considérée comme rectangulaire

On peut aisément calculer les caractéristiques de la variable aléatoire $X$. On calcule d'abord la moyenne :

$$m = \langle X \rangle = \int_0^1 X \mathrm{d}X = 0{,}5,$$

puis l'écart quadratique moyen $\sigma^2$ :

$$\sigma^2 = \int_0^1 (X-m)^2 \mathrm{d}X = 0{,}083\,3,$$

d'où l'on tire $\sigma = 0{,}288\,7$.

À chaque opération, on peut associer à l'erreur une variable aléatoire $X$. S'il y a $N$ opérations dans la procédure globale et si l'on suppose que les erreurs sont indépendantes, l'erreur totale sera la variable aléatoire :

$$Y = \sum_{j=1}^{N} X_j.$$

D'après le théorème central limite, $Y$ est une variable aléatoire gaussienne de moyenne $m$ et d'écart quadratique moyen $\sigma$ :

$$\sigma_y^2 = N\sigma^2$$

ce qui donne en définitive une mesure de l'erreur donnée par l'écart type :

$$\sigma_y = 0{,}288\,7\sqrt{N}.$$

$\sigma_y$ constitue en général une bonne estimation de l'erreur et nous rappelons à ce sujet que la probabilité pour que l'erreur soit inférieure en module à $\sigma_y$ est 68%, à $2\sigma_y$ est 95% et à $3\sigma_y$ est 99,7%.

En réalité la variable aléatoire $X^m$ prend ses valeurs sur l'intervalle $(0,\ 0,5)$ puisque la machine effectue non pas des troncatures mais des arrondis, ce qui divise chaque erreur élémentaire par 2. Il s'ensuit que : $m = \langle X \rangle = 0{,}25$, $\sigma^2 = 1{,}042\,10^{-2}$, soit $\sigma = 0{,}102$, et $\sigma_y = 0{,}102\sqrt{N}$.

## 8.2. Cas où X n'est plus à distribution rectangulaire

Reprenons l'exemple précédent et cherchons à calculer la somme de la série « à la précision de la machine », c'est-à-dire que l'on va arrêter les calculs lorsque $S_n/(n+1)/(n+2)$ sera plus grand que $10^{9{,}332}$ si l'on dispose de 4 octets pour représenter la mantisse.

À partir d'un certain rang $K < n$, la division va introduire une erreur systématique qui va garder le même sens très longtemps. La distribution ne sera plus du tout rectangulaire et les estimations effectuées au moyen des erreurs gaussiennes deviennent caduques.

*Grosso modo*, la précision optimum est obtenue pour une valeur de $n$ voisine de 100 000 que nous avons portée dans le tableau. On voit très bien que, pour ce type de problèmes, les grandeurs statistiques ne sont plus des estimateurs acceptables de l'erreur, cependant elles le demeurent pour $n$ allant jusqu'à 50 000 environ.

# 9. Sur la représentation des nombres en machine

## 9.1. Les nombres entiers

Pour les entiers positifs la représentation retenue est la représentation binaire pure et pour les entiers négatifs la représentation binaire en complément à deux. Nous allons montrer de quelle façon se réalisent ces représentations sur une machine travaillant sur des mots-mémoire de $k$ bits (bit est la contraction de binary digit signifiant chiffre binaire).

Le nombre de configurations différentes susceptibles d'être obtenues est alors $2^k$, on peut donc représenter $2^k$ nombres entiers en binaire. La plupart du temps il est convenu d'utiliser le premier bit à gauche pour exprimer le signe du nombre et l'on adopte la valeur zéro pour désigner un nombre positif et la valeur un pour désigner un nombre négatif.

En ce qui concerne les nombres négatifs, on préfère utiliser la représentation en complément à deux qui permet alors d'effectuer l'addition des nombres et non la soustraction. Ajoutons que l'on gagne un chiffre dans la représentation car il n'y a qu'une seule configuration pour zéro alors qu'il y en aurait deux en signant les nombres négatifs : une pour les positifs et une pour les négatifs.

Si l'on se limite au point de vue opératoire, la représentation en complément à deux consiste à écrire le nombre en binaire pur, puis à changer le symbole zéro en symbole un et le symbole un en symbole zéro, et enfin à ajouter un. Ainsi, de cette façon, on représente les nombres sur l'intervalle fini $(-2^{k-1}, +2^{k-1} - 1)$.

***Exemple*** – On suppose que le mot-mémoire a la taille d'un octet (8 bits), donc on peut représenter 256 configurations différentes, soit encore les nombres de $-128$ à $+127$.

On désire réaliser l'opération $c = a + b$ avec $a = 57$ et $b = -36$. Les représentations binaires sont les suivantes :

| | | | |
|---|---|---|---|
| $a$ | 00111001 | $\|b\|$ | 00100100 |
| $b$ | 11011100 | $c$ | 100010101 |
| | | | report |

On s'aperçoit alors, qu'à l'exécution, il y a un dépassement de la capacité du mot-mémoire. La plupart des compilateurs masquent ce dépassement de capacité (report) et l'on génère des nombres entiers modulo $2^k$. Cette propriété sera exploitée pour générer des nombres pseudo-aléatoires (*cf.* chapitre 19).

## 9.2. Les nombres décimaux ou flottants ou réels

La représentation entière ne permet pas une large dynamique et se trouve mal adaptée à la représentation des nombres de grande taille (grand module de l'exposant). Pour fixer les idées, considérons un nombre de l'ordre de $10^{100}$. Il faudrait des mots-mémoire constitués de 300 à 350 bits pour le représenter en binaire pur ce qui est prohibitif dans la mesure où statistiquement peu de nombres aussi grands sont utilisés, alors que toutes les opérations arithmétiques devront porter sur tous les bits sans exception ; cela montre que la machine passerait le plus clair de son temps à travailler pour rien... Tout bien considéré, même si la taille de la mémoire devait être considérable, le problème des très petits nombres fractionnaires resterait entièrement posé.

En définitive, il s'agit de trouver un compromis acceptable entre la précision de la représentation et la taille du mot-machine, ce qui a une incidence directe sur le temps d'exécution des opérations et la taille de la mémoire centrale. Avant d'envisager la manière de stocker un

nombre en machine, il nous faut parler des nombres normalisés qui permettent d'optimiser la représentation. Un nombre, quel qu'il soit, s'exprime soit dans le vocabulaire écrit courant, soit au moyen d'une représentation symbolique écrite dans une base donnée (à l'aide de chiffres) ; c'est évidemment la seconde option qui est utilisée pour représenter l'information codée. Ceci peut apparaître comme une évidence, il n'en est rien ; pour s'en convaincre il suffit d'écrire deux nombres en chiffres romains et d'en effectuer la division. On s'aperçoit sans difficulté que l'écriture en chiffres romains n'est pas bien adaptée à la réalisation effective de cette opération.

Un nombre peut toujours être représenté dans une base $B$ quelconque selon l'expression :

$$N = \sum_{n=m_1}^{m_2} \beta_n B^n \quad \text{avec} \quad \{\beta_k\} = \big(0,\, 1,\, 2, \ldots (B-1)\big).$$

On remarque que $m_1$ et $m_2$ sont des nombres entiers finis dès qu'il s'agit de calculs effectifs (positifs, négatifs ou nuls). On peut encore écrire :

$$N = \left( \sum_{n=m_1}^{m_2} \beta_n B^{n-m_2-1} \right) B^{m_2+1}$$

Cette dernière représentation est dite normalisée à condition toutefois d'avoir $\beta_{m_2} \neq 0$. Comme la base n'a pas besoin d'être explicitement exprimée, on écrira $N$ sous la forme :

$$N = \{0, \beta_{m-2}\, \beta_{m_2-1} \ldots \beta_{m_1}\, ;\, m_2 + 1\},$$

la suite ordonnée des $\beta_m$ s'appelle la **mantisse** tandis que la puissance de la base s'appelle l'**exposant**. Il s'agit donc d'une représentation semi-logarithmique particulière dans la mesure où le premier chiffre significatif est différent de zéro. Sa raison d'être s'explique par la recherche d'une représentation optimum qui permet d'occuper la place mémoire la plus petite. Pour ce faire, on stocke uniquement la mantisse et l'exposant ; la base est une donnée implicite qui n'est pas stockée avec le nombre mais qui est évidemment indispensable pour effectuer les calculs.

***a – Une représentation des nombres flottants sur quatre octets*** – C'est une représentation qui a été fréquemment retenue par les constructeurs de grosses machines dans les années 60, aussi mérite-t-elle qu'on s'y attarde un peu. Le mot-machine est constitué de quatre octets numérotés de gauche à droite de 1 à 4, et la base implicite est 16.

Les chiffres utilisés sont donc les symboles hexadécimaux 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F ; chacun de ces symboles peut être codé en binaire sur un demi-octet, car un chiffre hexadécimal se code sur quatre bits.

Par convention, la mantisse s'exprime sur les octets 2, 3 et 4, ce qui limite la représentation à 6 chiffres hexadécimaux. Comme les nombres sont normalisés, le premier demi-octet de l'octet 2 doit être différent de zéro.

L'octet 1 contient le signe du nombre dans le premier bit et l'exposant dans les sept bits suivants. Comme précédemment le signe plus est codé par un 0 et le signe moins par un 1 (premier bit).

Pour ce qui concerne l'exposant, on pourrait concevoir une représentation entière affectée d'un signe puisque les exposants peuvent être négatifs ou positifs. Ce n'est généralement pas la solution retenue. Comme on peut représenter 128 configurations possibles sur les 7 bits affectés à l'exposant, la convention veut que les configurations comprises entre 0 et 64 soient attribuées aux exposants négatifs (et l'exposant nul) et les configurations comprises entre 65 et 127 aux exposants positifs. En définitive, il suffit d'ajouter 64 à l'exposant avant de procéder au stockage.

***b – Précision de la représentation*** – La troncature du nombre porte sur la mantisse seulement. Dans le cas le plus défavorable on aura négligé une suite infinie de F en notation hexadécimale (ce serait une suite infinie de 9 en notation décimale, ou une suite infinie de 1 en notation binaire). Toujours est-il que cette suite infinie, quelle que soit sa base, a pour limite l'unité, et par conséquent, représente une erreur absolue de 1 unité sur le dernier chiffre significatif de la mantisse du nombre stocké ; autrement dit, on aura une erreur absolue de 1 unité sur la mantisse considérée comme un nombre entier. Cette façon de concevoir n'est pas toujours la mieux adaptée au calcul des erreurs, aussi préfère-t-on raisonner en terme d'erreur relative sur la représentation du nombre, qui se réduit à l'erreur relative sur la mantisse :

$$e_r = \frac{1}{\text{mantisse}}.$$

Ici encore le cas le plus défavorable se rencontre lorsque la mantisse est la plus petite possible. Puisque la représentation est hexadécimale et normalisée, la plus petite mantisse vaut $16^5$, il s'ensuit que l'erreur relative la plus défavorable vaut :

$$e_r = 9{,}54\,10^{-7},$$

soit environ $10^{-6}$.

Compte tenu de la présence d'un digit supplémentaire utilisé en mémoire centrale lors de l'exécution, on ne réalise pas des troncatures mais des arrondis ce qui a pour conséquence de diviser les erreurs par deux, soit :

$$e_r = 0{,}5\,10^{-6}.$$

Ce type de représentation permet d'obtenir une dynamique s'étendant de $10^{-78}$ à $10^{-75}$ approximativement.

Il existe également une double précision qui occupe huit octets. Les quatre octets supplémentaires sont uniquement affectés à la mantisse. La dynamique n'est pas affectée tandis que la précision relative de la représentation est passée à $10^{-16}$ environ.

***c – Représentation usuelle des nombres flottants dans les Basics*** – La plupart du temps la représentation s'effectue sur 5 octets et la base implicite est 2. Cette conception conduit, comme on va le voir, à une précision plus grande mais aussi à une dynamique plus faible que celle précédemment étudiée.

L'octet 1 est utilisé pour représenter l'exposant tandis que les 4 derniers octets servent au stockage de la mantisse. Comme le nombre est normalisé en notation binaire, le premier bit de l'octet 2 est inévitablement 1. Cela constitue une information sans grand intérêt et certains constructeurs utilisent ce bit pour y stocker le signe du nombre, soit 0 si le nombre est positif et 1 si le nombre est négatif. Le premier octet permet de représenter 256 configurations qui se décomposent de la façon suivante : les exposants négatifs ou nuls sont représentés de 0 à 127 et les exposants positifs de 128 à 255. La précision relative est majorée par :

$$e_r = 1/2^{31} = 4{,}7\,10^{-10}.$$

Le plus petit nombre représentable est de l'ordre de $10^{-39}$, tandis que le plus grand est de l'ordre de $10^{38}$.

***d – Représentation des entiers et des flottants dans certains langages utilisés avec les micro-ordinateurs*** – Selon les logiciels et les ordinateurs utilisés, la représentation entière s'effectue sur 2, 4 voire 8 octets : en binaire pur pour les entiers positifs et en complément à deux pour les entiers négatifs. Dans le cas de deux octets, on dispose de nombres compris dans l'intervalle (−32 768, +32 767).

***Remarque*** : En général, en arithmétique entière, au cours des calculs les dépassements de capacité sont automatiquement masqués et les calculs sont effectués modulo $2m$, avec $m = 8k-1$ où $k$ est le nombre d'octets retenus pour la représentation. Cette propriété sera utilisée pour générer des nombres pseudo-aléatoires.

La représentation des nombres flottants s'effectue sur des doubles mots soit 4 octets. Le système est identique à celui utilisé pour les Basics à la différence près toutefois qu'on ne dispose plus que de 3 octets au lieu de 4 pour écrire la mantisse. En conséquence de quoi la dynamique est évidemment la même, tandis que la précision se trouve un peu réduite, elle se situe au voisinage de :

$$e_r = 1/2^{23} = 1{,}2\,10^{-7}.$$

***e – La double précision*** – Par exemple, en C, dans les micro-ordinateurs usuels, la représentation des nombres en double précision s'effectue sur 8 octets, soit 16 demi-octets. Les trois premiers demi-octets contiennent le bit de signe du nombre suivi de l'exposant, ce qui permet une dynamique de l'ordre de $10^{-308}$ à $10^{308}$. La mantisse est représentée sur les treize demi-octets restants. La plus petite mantisse est donc $8 \times 16^{12}$ sur laquelle l'erreur d'arrondi est 0,5. La précision relative la plus défavorable est donc de l'ordre de $0{,}2\,10^{-16}$.

## 10. Éléments de bibliographie


N. BAKHVALOV (1976) *Méthodes Numériques*, Éditions MIR.

C. BERNARDI, B. METIVET et R. VERFURTH (1996) *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Wiley.

H. BESTOUGEFF, GUILPIN Ch. et M. JACQUES (1975) *La Technique Informatique*, Tomes I et II, Masson.

B. DÉMIDOVITCH et L. MARON (1979) *Éléments de Calcul Numérique*, Éditions MIR.

P. HENRICI (1964) *Elements of Numerical Analysis*, Wiley.

F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.

D. MC CRACKEN et W. DORN (1964) *Numerical Methods and Fortran Programming*, Wiley.

E. NAGEL, J. NEWMAN, K. GÖDEL et J.-Y. GIRARD (1989) *Le théorème de Gödel*, Seuil.

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

H. RUTISHAUSER (1990) *Lectures on numerical mathematics*, Springer-Verlag.

J. STOER et R. BULIRSCH (1993) *Introduction to numerical analysis*, Springer-Verlag.

# 2 Quelques algorithmes accélérateurs de la convergence des suites

Au cours de ce chapitre, nous allons présenter quelques algorithmes très puissants qui ont pour but d'accélérer la vitesse de convergence des suites. Il s'agit de l'algorithme d'Aitken, du procédé d'extrapolation de Richardson et de l'epsilon-algorithme, ce dernier ayant vu le jour en 1956 à la suite des travaux de P. Wynn. Nous nous attacherons essentiellement à leur présentation et à leur mise en œuvre dans la mesure où nous ferons souvent appel à eux au cours des différents chapitres.

Sans entrer dans des considérations théoriques de haut niveau, il est possible de dire que d'une part l'epsilon-algorithme est une généralisation du procédé d'Aitken, et que d'autre part il est relié aux approximants de Padé et aux fractions continues (*cf.* annexes C et D). Il n'est pas question d'effectuer ici les justifications théoriques relatives à ces algorithmes et nous limiterons nos ambitions à en apprendre le maniement. Cependant, le lecteur intéressé par les fondements des techniques d'accélération de convergence pourra avoir recours à l'excellent ouvrage de C. Brézinski : Algorithmes d'Accélération de la Convergence. Étude Numérique, aux Éditions Technip, Paris (1978). Dans cet ouvrage figurent d'autres procédures d'accélération de la convergence, de nombreux exemples d'applications et des programmes rédigés en Fortran. Pour tout ce qui a trait aux procédures d'accélération de la convergence, il est nécessaire d'insister sur le fait que les résultats proposés ne concernent que les nombres susceptibles de pouvoir être effectivement calculés avec une précision suffisante.

## 1. L'algorithme $\Delta^2$ d'Aitken (1895–1967)

Supposons que nous connaissions une suite numérique convergente $S_0, S_1, \ldots, S_n$. Par exemple les $S_k$ peuvent être constitués par les sommes partielles d'une série, ou encore par la suite des approximations successives de la racine d'une équation $f(x) = 0$ obtenues par une méthode itérative. À partir de cette suite de $(n+1)$ termes, nous allons construire une autre suite de $n$ termes et ainsi de suite. Nous définissons une procédure récurrente, et pour les besoins de la cause, nous allons numéroter des suites ainsi formées au moyen d'un exposant placé entre parenthèses. Ainsi, la suite initiale s'écrit :

$$S_0^{(0)}, S_1^{(0)}, \ldots, S_n^{(0)}.$$

### 1.1. Présentation de la méthode

Considérons une suite numérique $S_0^{(0)}, S_1^{(0)}, S_2^{(0)} \ldots, S_n^{(0)}$ dont la convergence est lente. À partir de la suite des $S_k^{(0)}$, on forme une nouvelle suite $S_j^{(1)}$ obtenue de la manière suivante :

$$S_k^{(1)} = S_{k+2}^{(0)} - \frac{\left[S_{k+2}^{(0)} - S_{k+1}^{(0)}\right]^2}{S_k^{(0)} + S_{k+2}^{(0)} - 2S_{k+1}^{(0)}}.$$

À partir de la dernière suite formée, celle des $S_k^{(1)}$, on obtient une autre suite au moyen du même procédé. On note qu'à chaque construction d'une nouvelle suite, le nombre d'éléments de la suite diminue de deux unités. L'application successive de ce processus nous conduit à l'obtention de la suite qui n'est plus constituée que d'un seul terme à condition toutefois que le nombre de termes de la suite initiale soit impair. Le terme général s'écrit alors :

$$S_k^{(q+1)} = S_{k+2}^{(q)} - \frac{\left[S_{k+2}^{(q)} - S_{k+1}^{(q)}\right]^2}{S_k^{(q)} + S_{k+2}^{(q)} - 2S_{k+1}^{(q)}}. \tag{2.1}$$

### 1.2. Un exemple numérique

On se propose d'appliquer cet algorithme à la suite des sommes partielles obtenue par la sommation de la série :

$$S = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \frac{\pi}{4}$$

qui est bien connue pour sa convergence lente. En retenant les cinq premiers termes de la suite on trouve $S = 0{,}834\,920$. L'usage de l'algorithme d'Aitken, avec les mêmes nombres, permet de trouver $S = 0{,}785\,526$. Autrement dit, l'erreur relative sur $\pi$ passe de $6{,}3\,10^{-2}$ à $5{,}1\,10^{-4}$.

On trouvera sur le Web (*) un programme `aitken0.c` mettant en œuvre cet algorithme.

## 2. Le procédé d'extrapolation de Richardson (1881–1953)

On désigne toujours par $S_j^{(0)}$, avec $j = 0, 1, 2 \ldots n$, la suite initiale que l'on désire voir converger plus rapidement, et par $x_k$ une suite auxiliaire « choisie convenablement ». Nous aborderons un peu plus loin le problème du choix de la suite auxiliaire. La méthode de Richardson consiste à former une suite de « vecteurs » au moyen de la relation suivante :

$$S_m^{(k+1)} = \frac{x_m S_{m+1}^{(k)} - x_{m+k+1} S_m^{(k)}}{x_m - x_{m+k+1}}, \tag{2.2}$$

avec $m, k = 0, 1, 2, \ldots n$. La succession des opérations est symbolisée par le tableau 2.1, page ci-contre.

Dans la mesure où les $x_m$ sont indépendants des $S_m^{(k)}$, il est aisé de s'apercevoir, qu'il s'agit d'un algorithme en triangle et que $S_m^{(k+1)}$ est une combinaison linéaire de $S_{m+1}^{(k)}$ et $S_m^{(k)}$, il s'ensuit, alors, que le procédé de Richardson est une transformation linéaire de suites.

***Remarque*** : Dans le cas particulier où l'on choisit comme suite auxiliaire la suite :

$$x_m = S_{k+1}^{(0)} - S_k^{(0)} = \Delta S_k^{(0)},$$

* `http://www.edpsciences.com/guilpin/`

**Tableau 2.1.**

| | | | | |
|---|---|---|---|---|
| $S_0^{(0)}$ | | | | |
| | $S_0^{(1)}$ | | | |
| $S_1^{(0)}$ | | $S_0^{(2)}$ | | |
| | $S_1^{(1)}$ | | $S_0^{(3)}$ | |
| $S_2^{(0)}$ | | $S_1^{(2)}$ | | $S_0^{(4)}$ |
| | $S_2^{(1)}$ | | $S_1^{(3)}$ | |
| $S_2^{(0)}$ | | $S_2^{(2)}$ | | |
| | $S_3^{(1)}$ | | | |
| $S_4^{(0)}$ | | | | |

on obtient alors un algorithme qui n'est plus en triangle et qui n'est plus linéaire. La relation générale est donnée par l'expression :

$$S_m^{(k+1)} = \frac{\Delta S_m^{(0)} S_{m+1}^{(k)} - \Delta S_{m+k+1}^{(0)} S_m^{(k)}}{\Delta S_m^{(0)} - \Delta S_{m+k+1}^{(0)}} . \tag{2.3}$$

## 2.1. Quelques éléments de théorie

Revenons à la suite initiale $S_j^{(0)}$ qui converge vers $S$ ainsi qu'à la suite auxiliaire $x_k$ indépendante des $S_j^{(0)}$ que l'on suppose d'une part strictement décroissante et d'autre part tendre vers zéro lorsque $n$ tend vers l'infini. Le procédé d'extrapolation de Richardson est fondé sur la formule de Neville-Aitken qui donne une façon de construire les polynômes d'interpolation pour une valeur particulière de la variable (*cf.* chapitre 6 sur l'interpolation). En effet, $S_j^{(k)}$ est la valeur en zéro du polynôme d'interpolation de degré $k$, lequel, aux abscisses $x_p$, prend les valeurs $S_p^{(0)}$ pour $p = j, j+1, \ldots, j+k$.

À présent nous allons énoncer quelques théorèmes importants, mais dont nous ne donnons pas la démonstration.

***a* – *Théorème I*** – Pour que $S_j^{(k)}$ tende vers $S$ quel que soit $j > N$, il faut et il suffit que $S_m^{(0)} = S + \sum_{i=1}^{k} a_i x_m^i$ quel que soit $m > N$.

Autrement dit, les $S_j^{(k)}$ peuvent s'exprimer sous la forme d'un rapport de deux déterminants, ce qui constitue une autre expression de ce théorème :

$$S_j^{(k)} = \frac{\begin{vmatrix} S_j^{(0)} & \ldots & S_{j+k}^{(0)} \\ x_j & \ldots & x_{j+k} \\ \ldots & \ldots & \ldots \\ x_j^k & \ldots & x_{j+k}^k \end{vmatrix}}{\begin{vmatrix} 1 & \ldots & 1 \\ x_j & \ldots & x_{j+k} \\ \ldots & \ldots & \ldots \\ x_j^k & \ldots & x_{j+k}^k \end{vmatrix}}$$

***b – Théorème II*** – Si la suite $x_k$ (formée de nombres réels positifs, décroissante et tendant vers zéro lorsque $n$ tend vers l'infini) vérifie la relation :

$$\frac{x_n}{x_{n+1}} \geq a > 1$$

pour tout $n$, alors la limite de la suite $S_j^{(k)}$ tend vers $S$ quand $k$ tend vers l'infini. La condition énoncée est nécessaire et suffisante. Ce théorème nous donne des indications sur les choix possibles de la suite des $x_j$. Par exemple, il est raisonnable de retenir la suite $x_m = \alpha^m$ et de l'assortir de la condition $0 < \alpha < 1$.

***c – Théorème III*** – Dans le cas où $S_j^{(0)} = \Phi(x_j)$ et que $\Phi(x)$ est suffisamment dérivable, et que le module de la dérivée $(k+1)^{\text{e}}$ demeure borné sur l'intervalle $I$ ($I$ est le plus petit intervalle contenant le point zéro et toute la suite $x_n$) par un nombre $M$ indépendant de $x$, alors $S_l^{(k)}$ tend vers $S$ quand $l$ tend vers l'infini. Si de plus $M$ est indépendant de $k$, alors $S_l^{(k)}$ tend vers $S$ lorsque $k$ tend vers l'infini.

***d – Théorème IV*** – (Ce théorème concerne l'accélération de la convergence.) Dans le cas où les conditions du théorème II sont satisfaites, pour que la suite $S_j^{(k+1)}$ converge plus vite que la suite $S_j^{(k)}$, il faut et il suffit que :

$$\lim \frac{S_m^{(k+1)} - S}{S_m^{(k)} - S} = \lim \frac{x_{m+k+1}}{x_m} \quad \text{quand } m \text{ tend vers l'infini.}$$

Cependant, force nous est de reconnaître que cette condition n'est pas aisément vérifiable dans les cas pratiques.

## 2.2. Généralisation du procédé de Richardson

On considère une fonction $\Psi(x)$ définie, continue et strictement croissante sur l'intervalle $(0, \beta)$ avec $\beta > 0$. Alors, dans la relation (2.2), on peut remplacer $x_q$ par $[\Psi(x_q) - \Psi(0)]$ sans que rien de ce qui vient d'être établi ne soit modifié. On obtient alors la relation :

$$S_j^{(k+1)} = \frac{[\Psi(x_j) - \Psi(0)]\, S_{j+1}^{(k)} - [\Psi(x_{j+k+1}) - \Psi(0)]\, S_j^{(k)}}{\Psi(x_j) - \Psi(x_{j+k+1})}. \tag{2.4}$$

Le théorème I s'énonce alors de la façon suivante :

***Théorème V*** – Pour que $S_j^{(k)}$ tende vers $S$ quel que soit $j > N$, il faut et il suffit que :

$$S_m^{(0)} = S + \sum_{i=1}^{k} a_i \Big( \Psi(x_i) - \Psi(0) \Big).$$

Voici quelques choix possibles pour les fonctions $\Psi(x)$ :

$$\begin{aligned} &\Psi(x) = x^q && \text{avec } q \geq 1 \\ &\Psi(x) = a^x && \text{avec } a > 1 \\ &\Psi(x) = \log_e(1 + x). \end{aligned}$$

***Remarque*** : Si $\Psi_1(x)$ et $\Psi_2(x)$ vérifient les conditions de la généralisation, alors les fonctions :

$$g_1(x) = a_1\Psi_1(x) + a_2\Psi_2(x)$$
$$g_2(x) = \Psi_1(x) \cdot \Psi_2(x)$$
$$g_3(x) = \Psi_1\left[\Psi_2(x)\right],$$

vérifient aussi les conditions de la généralisation.

### 2.3. Relation entre le procédé $\Delta^2$ d'Aitken et celui de Richardson

Revenons à l'expression (2.3) obtenue en remplaçant $x_n$ par $\Delta S_n^{(0)}$. Elle vérifie les conditions des théorèmes précédents, et si l'on fait $k = 0$, nous obtenons :

$$S_j^{(1)} = \frac{S_j^{(0)} S_{j+2}^{(0)} - S_{j+1}^{(0)} S_{j+1}^{(0)}}{S_{j+2}^{(0)} - 2S_{j+1}^{(0)} + S_j^{(0)}}.$$

On reconnaît le procédé d'Aitken en réduisant au même dénominateur l'expression (2.3), et du reste on verra un peu plus loin que ce vecteur de composantes $S_j^{(2k+1)}$ est également la deuxième colonne du tableau de l'epsilon-algorithme.

***Un exemple numérique*** – Nous allons appliquer l'algorithme de Richardson à la suite des sommes partielles de la série :

$$S = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \cdots = \frac{\pi^2}{6} = 1{,}644\,934\,066\,8\ldots$$

Ici il nous faut en plus choisir une suite auxiliaire, et nous avons retenu arbitrairement la suite classique $x_m = \frac{1}{m}$ convergente et qu'il ne faut pas confondre avec la série harmonique qui diverge.

Nous avons effectué les calculs avec les 25 premières sommes, et voici les résultats trouvés :

$$S(25) = 1{,}605\,723\,403$$
$$S(\text{Richardson}) = 1{,}644\,941.$$

En comparant avec le résultat connu on voit que l'erreur relative sur la somme est de $5{,}0\,10^{-6}$.

On trouvera sur le Web(*) un sous-programme `richar0.c` qui réalise cet algorithme.

## 3. Présentation de l'epsilon-algorithme scalaire

Il s'agit sans doute, dans l'état actuel des connaissances, du plus puissant algorithme permettant d'accélérer la convergence des suites. Comme précédemment on conserve les notations introduites au cours de ce chapitre, à l'exception toutefois de la suite initiale qui prend le rang 1 au lieu du rang zéro. Autrement dit, la suite initiale s'écrit : $S_k^{(1)}$.

Les termes de la deuxième suite s'expriment au moyen des relations suivantes :

$$S_k^{(2)} = S_{k+1}^{(0)} + \frac{1}{S_{k+1}^{(1)} - S_k^{(1)}}$$

* `http://www.edpsciences.com/guilpin/`

où $k = 0, 1, \ldots, (n-1)$, et avec comme convention $S_k^{(0)} = 0$ quel que soit $k$. On forme la troisième suite au moyen de la relation :

$$S_k^{(3)} = S_{k+1}^{(1)} + \frac{1}{S_{k+1}^{(2)} - S_k^{(2)}} ;$$

et d'une manière tout à fait générale, la $p^e$ suite sera donnée par :

$$S_k^{(p)} = S_{k+1}^{(p-2)} + \frac{1}{S_{k+1}^{(p-1)} - S_k^{(p-1)}}$$

où $k = 0, 1, \ldots (n - p + 1)$. On poursuit les calculs jusqu'à ce que l'on n'obtienne plus qu'un seul terme qui est $S_0^{(n)}$. Il s'agit maintenant non plus d'un algorithme en triangle mais d'un algorithme en losange. On se persuade facilement que les résultats intéressants figurent dans les suites de rang impair, il faut donc choisir un nombre impair de données au départ, soit $(n+1)$ impair, pour que le terme $S_0^{(n)}$ soit une approximation de la limite de la suite $S_0^{(1)}$. Les suites de rang pair ne constituent que des intermédiaires de calcul.

## 3.1. Exemples numériques

On désire calculer la somme de la série alternée déjà évoquée au cours du premier paragraphe :

$$S = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \frac{\pi}{4} = 0{,}785\,398\,2.$$

Sur le tableau 2.2, nous avons donné les suites successives $S^{(q)}$ en limitant la suite initiale à 5 termes :

**Tableau 2.2.**

| $S^{(1)}$ | $S^{(2)}$ | $S^{(3)}$ | $S^{(4)}$ | $S^{(5)}$ |
|---|---|---|---|---|
| 1 | | | | |
| | −3 | | | |
| 0,666 666 6 | | 0,791 666 6 | | |
| | +5 | | −115 | |
| 0,866 666 6 | | 0,783 333 3 | | 0,785 585 |
| | −7 | | +329 | |
| 0,723 809 5 | | 0,786 309 5 | | |
| | +9 | | | |
| 0,834 920 6 | | | | |

L'erreur relative sur la valeur donnée par $S_0^{(5)}$ est $2{,}4\,10^{-4}$.

## 3.2. Calcul de la somme d'une série de Fourier

La fonction qui vaut $-1$ entre $-\pi$ et 0 et $+1$ entre 0 et $\pi$ admet le développement en série de Fourier :

$$f(x) = \frac{4}{\pi}\left(\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \cdots + \frac{\sin\left[(2p+1)x\right]}{2p+1} + \cdots\right).$$

Cette série sera étudiée chapitre 16 et elle servira d'introduction à l'étude du phénomène de Gibbs. Du fait de la discontinuité de première espèce de la fonction, la série converge lentement et, qui plus est, nous sommes en présence du phénomène de Gibbs. En appliquant l'epsilon-algorithme à la suite des onze sommes partielles formées entre le terme d'ordre 40 et le terme d'ordre 50 dont voici les valeurs :

$$S_{40}^{(1)} = 1{,}002\,194\,232\,20 \qquad S_{41}^{(1)} = 1{,}017\,439\,936\,43 \qquad S_{42}^{(1)} = 1{,}031\,279\,464\,80$$
$$S_{43}^{(1)} = 1{,}043\,240\,233\,83 \qquad S_{44}^{(1)} = 1{,}052\,942\,746\,94 \qquad S_{45}^{(1)} = 1{,}060\,110\,382\,53$$
$$S_{46}^{(1)} = 1{,}064\,575\,093\,01 \qquad S_{47}^{(1)} = 1{,}066\,278\,967\,14 \qquad S_{48}^{(1)} = 1{,}065\,271\,752\,62$$
$$S_{49}^{(1)} = 1{,}061\,704\,573\,49 \qquad S_{50}^{(1)} = 1{,}055\,820\,201\,84.$$

on obtient la valeur : $S = 1{,}000\,689$ pour la valeur $x = 0,1$.

Conclusion : l'epsilon-algorithme fait disparaître le phénomène de Gibbs, c'est-à-dire les lentes oscillations au voisinage de la discontinuité.

## 4. L'epsilon-algorithme vectoriel

Les suites $S^{(q)}$ ne sont plus des grandeurs scalaires mais des grandeurs vectorielles. Il est donc nécessaire de préciser la technique de calcul des opérations proposées notamment en ce qui concerne la grandeur :

$$1/\left(S_{k+1}^{(p)} - S_k^{(p)}\right).$$

La différence de deux vecteurs ne pose pas de problème, et il nous reste à définir l'inverse $V^{-1}$ $(= 1/V)$ d'un vecteur $V$. Cette opération n'est pas définie habituellement, mais pour le cas qui nous préoccupe, on admettra qu'elle est réalisée en multipliant numérateur et dénominateur par le vecteur $V^*$ dont les composantes sont les composantes complexes conjuguées du vecteur $V$. Ainsi nous avons :

$$V^{-1} = \frac{V^*}{V \cdot V^*} = \frac{V^*}{V^2}.$$

Cette précision apportée, rien n'est modifié dans la conduite du calcul.

Si la suite à traiter est formée de nombres complexes, rien ne nous empêche de considérer un nombre complexe comme un vecteur à deux dimensions et d'appliquer l'epsilon-algorithme vectoriel.

Cet algorithme vectoriel trouvera son emploi chaque fois que le résultat d'un calcul itératif est donné par un ensemble de valeurs $(U_j)$ que l'on peut alors considérer comme les composantes d'un vecteur. Par exemple, nous étudierons la méthode de Picard qui permet d'obtenir par itérations un échantillon d'une certaine fonction solution d'une équation différentielle. Cette méthode converge très lentement et il va de soi que nous pourrons l'accélérer au moyen de l'epsilon-algorithme vectoriel.

## 5. L'epsilon-algorithme matriciel

Ici, il n'y a aucune difficulté à définir la différence de deux matrices carrées de même ordre et l'inverse d'une matrice, pourvu que cette dernière soit régulière. La technique de calcul se trouve

inchangée. Il est bon de remarquer que l'epsilon-algorithme ne s'applique qu'à des matrices carrées ; il s'ensuit qu'une suite convergente de matrices rectangulaires ne peut pas être accélérée par ce procédé puisque ces matrices ne possèdent pas d'inverse mais seulement deux pseudo-inverses qui ne peuvent pas faire l'affaire.

En revanche, sur le plan pratique, il est tout à fait possible d'appliquer l'epsilon-algorithme à une suite de matrices rectangulaires ou carrées pour en accélérer la convergence et cela en faisant usage de l'epsilon-algorithme vectoriel appliqué à chacun des vecteurs-lignes ou chacun des vecteurs-colonnes. Cette remarque prend toute sa signification lorsque la suite des matrices est constituée de ce que l'on appelle des matrices mal conditionnées. Dans ce cas, il est extrêmement fréquent d'obtenir des erreurs de calcul très importantes qui se répercutent sur l'ensemble des éléments de la matrice notamment lors de l'opération d'inversion. Cet aspect négatif des choses se trouve alors réduit par l'usage de l'epsilon-algorithme vectoriel.

L'epsilon-algorithme matriciel peut trouver une application lors des problèmes de résolution de l'équation de Laplace (1749–1827) (ou de Poisson) par la méthode itérative de Gauss-Seidel dans le cas particulier où la représentation du maillage conduit à une matrice carrée. Toutefois, quels que soient la géométrie et le maillage retenu pour intégrer l'équation de Laplace, dès lors que l'on utilise une méthode itérative, il est toujours possible d'utiliser ou bien l'epsilon-algorithme scalaire ou bien l'epsilon-algorithme vectoriel.

## 6. Remarques et propriétés de l'epsilon-algorithme

1. Nous venons de voir la forme scalaire de l'epsilon-algorithme, la forme vectorielle et la forme matricielle. Pour information, il existe aussi deux formes topologiques que nous nous contentons de mentionner.

2. Lorsqu'on applique l'epsilon-algorithme à une suite, il n'est pas nécessaire de débuter les calculs à partir du premier terme de la suite et l'on peut très bien prendre les termes de $q$ en $q$ à partir du rang $k$ pour former la suite initiale (*cf.* l'exemple concernant le phénomène de Gibbs).

3. L'epsilon-algorithme ne peut donner que ce qu'il a, c'est-à-dire qu'il peut accélérer la convergence à condition toutefois que la précision des termes retenus soit suffisante. En aucun cas il ne peut fournir une précision supérieure à celle qui a servi à faire les calculs des termes initiaux. Autrement dit, si le dernier élément $S_n^{(1)}$ de la suite initiale a été obtenu à la « précision de la machine », c'est-à-dire que l'addition du terme suivant de la série laisse inchangée la valeur de $S_n^{(1)}$, il est inutile de vouloir appliquer l'epsilon-algorithme quand bien même bénéficierait-on d'un accroissement de précision pour réaliser cette dernière opération. Par exemple, si nous avons obtenu $S_n^{(1)}$ en simple précision, l'application de l'epsilon-algorithme en double précision n'apportera strictement rien. Il n'est pas possible d'extraire une information qui n'est pas contenue dans les données.

4. À l'occasion de l'étude des séries de Fourier, on rencontre un phénomène gênant qui se manifeste au voisinage des points de discontinuité de première espèce (lorsque la fonction développée présente ces discontinuités) : c'est le phénomène de Gibbs (1839–1903). L'epsilon-algorithme supprime le phénomène de Gibbs qui se traduit normalement par des oscillations de grande amplitude lesquelles s'amortissent très lentement avec le nombre de termes de la série de Fourier.

5. Il est possible qu'à un moment donné la quantité $\left(S_{k+1}^{(p)} - S_k^{(p)}\right)$ soit nulle ce qui fait tomber en défaut le mécanisme d'accélération de la convergence puisque l'inverse de cette grandeur produit un dépassement de capacité en machine. Dans la plupart des cas cette difficulté se rencontre lorsque les $S_k^{(p)}$ ont été calculés à un moment donné « à la précision de la machine » et il s'ensuit que l'on retiendra ce résultat. S'il n'en est pas ainsi, bien qu'il existe des règles particulières qui peuvent être utilisées, il est souvent préférable d'abandonner la procédure. En conséquence de quoi, il est recommandé d'effectuer un test sur les différences $(S_{k+1}^{(p)} - S_k^{(p)})$ dans le programme et de procéder à l'arrêt des calculs lorsque l'une de ces différences est nulle. Au préalable, on aura pris soin d'enregistrer ou de faire imprimer les résultats partiels qui demeurent susceptibles d'être exploités.

***Réalisations pratiques*** – Sur le Web (*), on trouvera deux programmes réalisant l'un l'epsilon-algorithme scalaire `epsilon.h` et l'autre l'epsilon-algorithme vectoriel `evectr0.h`.

#### *L'epsilon-algorithme dans les autres chapitres*

1. Calcul des limites de suites, application au calcul numérique des intégrales.
2. Calcul des séries de Fourier, suppression du phénomène de Gibbs, calcul des dérivées ayant un développement divergent.
3. Résolution des équations $f(x) = 0$.
4. Accélération de la convergence des méthodes itératives utilisées lors de l'intégration des équations différentielles.
5. Accélération de la convergence lors de la recherche des valeurs propres des matrices.
6. Accélération de la convergence lors de la résolution des systèmes linéaires et non linéaires.
7. Intégration des équations aux dérivées partielles de type elliptique.
8. Résolution des équations intégrales de Fredholm de deuxième espèce (série de Liouville-Neumann).

On trouvera sur le Web (*) des illustrations de ces différents problèmes : `aitken2.c`, `epsil0.c`, `epsil1.c`, `epsil2.c`, `epsil4.c`, `kacmarz1.c` et `newton0.c`.

## 7. Propriétés remarquables du procédé $\Delta^2$ d'Aitken et de l'epsilon-algorithme

### 7.1. Le prolongement analytique

Si une fonction admet un développement en série de puissances convergent dans le disque $D$, on peut en général effectuer le prolongement analytique de cette série en dehors du disque de convergence. Alors il est possible d'appliquer l'un des algorithmes à la suite des sommes partielles

* `http://www.edpsciences.com/guilpin/`

même si celle-ci est divergente, c'est-à-dire hors du cercle de convergence. Considérons par exemple la série :

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 + \cdots$$

Cette série a un rayon de convergence strictement plus petit que 1. Appliquons les algorithmes à la valeur $x = 99$. En limitant le nombre de termes à 5, on obtient les résultats suivants

$$\begin{aligned} S_0^1 &= 1{,}000\,000\,000\,0\ E + 00 \\ S_1^1 &= -9{,}800\,000\,000\,0\ E + 01 \\ S_2^1 &= 9{,}703\,000\,000\,0\ E + 03 \\ S_3^1 &= -9{,}605\,960\,000\,0\ E + 05 \\ S_4^1 &= 9{,}509\,900\,500\,0\ E + 07. \end{aligned}$$

Il est évident que cette suite diverge ; cependant l'application des algorithmes donne comme limites :

| | |
|---|---|
| Aitken | $S = 1{,}000\,000\,536\,4\ E - 02$ |
| Epsilon-algorithme | $S = 1{,}000\,000\,000\,0\ E - 02$ |

alors que le calcul direct de la fraction donne $S = 10^{-2}$.

Sur le Web (*), on donne les programmes `aitken1.c` et `epsil3.c` qui réalisent ces opérations.

## 7.2. Les développements asymptotiques

Il est une autre série divergente extrêmement intéressante qui concerne les développements asymptotiques sur lesquels les algorithmes d'accélération de la convergence donnent de remarquables résultats. Sur le plan théorique, il est aisé de rattacher ces développements aux prolongements analytiques, car on passe de l'un à l'autre par une inversion, c'est-à-dire en posant $x = 1/x$. Il n'est plus nécessaire de choisir $x$ « grand » ni de limiter strictement la somme partielle à calculer à un ordre bien précis $N$ donné par le calcul des erreurs.

Voici quelques exemples de développements asymptotiques et les différents résultats obtenus au moyen des procédés d'accélération de la convergence.

$$f(x) = \int_x^\infty t^{-1} \exp(x - t)\, dt \text{ avec } x > 0.$$

admet comme développement asymptotique l'expression (*cf.* chapitre 3) :

$$f(x) = \frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \cdots + (-1)^n \frac{n!}{x^{n+1}} + \cdots$$

* `http://www.edpsciences.com/guilpin/`

Pour $x = 1$, $f(1) = 0{,}596\,336$, tandis que les quinze premières sommes partielles sont :

$$\begin{aligned}
S_0^1 &= 1{,}0\\
S_1^1 &= 0{,}0\\
S_2^1 &= 2{,}0\\
S_3^1 &= -4{,}0\\
S_4^1 &= 20{,}0\\
S_5^1 &= -100{,}0\\
S_6^1 &= 62{,}0\\
S_7^1 &= -442{,}0\\
S_8^1 &= 3\,590{,}0\\
S_9^1 &= -3{,}269\,80\ E + 05\\
S_{10}^1 &= 3{,}301\,820\ E + 06\\
S_{11}^1 &= -3{,}661\,498\ E + 07\\
S_{12}^1 &= 4{,}423\,866\,20\ E + 08\\
S_{13}^1 &= -5{,}784\,634\,18\ E + 09\\
S_{14}^1 &= 8{,}139\,365\,702\ E + 10.
\end{aligned}$$

Voici les résultats obtenus :

| $\Delta^2$d'Aitken | Epsilon-algorithme |
|---|---|
| 0,596 347 | 0,596 572 |

## 7.3. La série de Liouville (1809–1882) - Neumann (1832–1925)

Considérons une équation intégrale de Fredholm (1866–1927) de deuxième espèce :

$$x(t) - \rho \int_0^1 K(t,s)x(s)\,\mathrm{d}s = y(t),$$

que l'on écrira de façon plus concise :

$$x - \rho Ax = y$$

où $A$ est l'opérateur « intégrale » qui est évidemment linéaire.

Sous réserve de convergence, la série de Liouville-Neumann donne la solution du problème :

$$x = y + \rho Ay + \rho^2 A^2 y + \rho^3 A^3 y + \cdots + \rho^n A^n y + \cdots$$

On reconnaît la série géométrique dont la convergence sera assurée si

$$\frac{1}{|\rho|} > |\lambda_k|,$$

où $|\lambda_k|$ est la plus grande valeur propre de $A$ en module (*cf.* Jean Bass, Cours de mathématiques, Masson, 1968 et Lichnerowicz, Algèbre et Analyse linéaires, Masson, 1955).

Ici encore, l'application de l'epsilon-algorithme vectoriel à une suite divergente conduit au bon résultat.

***Exemple***

$$U(x) = \exp x^2 + 5 \int_0^1 \Big(\cos(\pi x) + 2\sqrt{3}\sin(\pi x)(y - 0{,}5)\Big)\, U(y)\, \mathrm{d}y.$$

On a retenu neuf vecteurs, c'est-à-dire les sommes partielles jusqu'à $\rho^8 A^8 y$. Chacune des ces fonctions a été échantillonnée sur vingt et un points. Nous avons obtenu pour les éléments du dernier vecteur l'ordre de grandeur de $10^4$ car la série de Liouville-Neumann est divergente.

La technique de résolution de l'équation de Fredholm est une méthode self-consistante qui permet d'obtenir aisément l'erreur. L'epsilon-algorithme donne une erreur inférieure à $10^{-12}$ au sens du sup.

On trouvera sur le Web(*) le programme `fredholm.c` qui réalise ce calcul.

## 8. Éléments de bibliographie

C. BRÉZINSKI (1978) *Algorithmes d'Accélération de la Convergence, Étude Numérique*, Éditions Technip, Paris.

J.P. DELAHAY (1988) *Sequence transformations*, Springer-Verlag.

A. RALSTON et R.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Éditions Dunod.

* `http://www.edpsciences.com/guilpin/`

# 3 Les développements asymptotiques

Les développements asymptotiques, encore appelés séries semi-convergentes, constituent une fort belle application numérique des séries divergentes. Le but est d'associer à une certaine fonction $f(x)$ un développement du type :

$$a_0 + \sum_{n=1}^{\infty} \frac{a_n}{x^n} \cdot$$

Sous certaines conditions que nous allons évoquer, le développement divergent, tronqué à un ordre convenable, permet de calculer numériquement la fonction $f(x)$ avec une précision d'autant plus grande que le module de l'argument est plus élevé. Par exemple, on rencontre en analyse classique un tel développement quand on veut calculer la constante d'Euler (1707–1783) avec une grande économie de moyen (*cf.* annexe H, exercice 1–4).

## 1. Un exemple de développement asymptotique

Considérons la fonction cosinus intégral :

$$Ci(x) = -\int_x^{\infty} \frac{\cos(t)}{t}\, \mathrm{d}t \quad \text{avec } x > 0.$$

Intégrons successivement par parties, nous obtenons :

$$Ci(x) = \frac{\sin(x)}{x} - \frac{\cos(x)}{x^2} - \frac{2\sin(x)}{x^3} + \frac{6\cos(x)}{x^4} + \cdots$$
$$+ (-1)^n \frac{(2n)!\sin(x)}{x^{2n+1}} - (-1)^n \frac{(2n+1)!\cos(x)}{x^{2n+2}} + (2k+1)! \int_x^{\infty} \frac{\cos(t)}{t^{2k+2}}\, \mathrm{d}t.$$

Soit encore :

$$Ci(x) = \frac{\sin(x)}{x} \sum_{n=0}^{\infty} (-1)^n \frac{(2n)!}{x^{2n}} - \frac{\cos(x)}{x} \sum_{n=0}^{\infty} (-1)^n \frac{(2n+1)!}{x^{2n+1}} \cdot$$

Ce développement est divergent comme le montre le critère de d'Alembert. Pour la série ayant $\sin(x)/x$ en facteur, on obtient :

$$\frac{(2n+2)!x^{2n}}{(2n)!x^{2n+2}} \approx \frac{4n^2}{x^2}$$

quel que soit $x$. On peut trouver $n$ suffisamment grand pour que $4n^2 > x^2$ ce qui montre le caractère divergent de la série. On a le même résultat avec la seconde série ayant $\cos(x)/x$ comme coefficient, mais une seule série suffit... pour montrer la divergence. On aurait pu dire aussi que le terme général de chacune des séries ne tend pas vers zéro quand $n$ tend vers l'infini, ce qui suffit à assurer la divergence.

Désignons par $S_{2k}(x)$ la somme constituée des deux séries qui sont tronquées chacune à l'ordre $k$ :

$$S_{2k}(x) = \frac{\sin(x)}{x}\sum_{n=0}^{k}(-1)^n\frac{(2n)!}{x^{2n}} - \frac{\cos(x)}{x}\sum_{n=0}^{k}(-1)^n\frac{(2n+1)!}{x^{2n+1}},$$

il est alors facile de calculer l'erreur liée à l'approximation :

$$E_{2k}(x) = |Ci(x) - S_{2k}(x)|$$

que l'on peut écrire encore :

$$E_{2k}(x) = (2k+1)!\left|\int_x^\infty \frac{\cos(t)}{t^{2k+2}}\,\mathrm{d}t\right| \leq (2k+1)!\left|\int_x^\infty \frac{\mathrm{d}t}{t^{2k+2}}\right|.$$

On s'aperçoit que $E_{2k}(x)$ est une fonction croissante en $k$ et décroissante en $x$. Pour connaître l'ordre $k$ du développement pour lequel l'erreur est minimum, il suffit d'écrire que :

$$E_{2k}(x) \approx E_{2(k+1)}(x) \quad (\text{ou encore} \approx E_{2(k-1)}(x)).$$

On en déduit immédiatement, dans le cadre de notre exemple particulier, que :

$$\frac{(2k)!}{x^{2k+1}} = \frac{(2k+2)!}{x^{2k+3}}$$

d'où l'on tire : $x \approx 2k+1$.

Donc, pour $x$ fixé, il suffit de calculer $2k$ termes de la série asymptotique en choisissant $k$ donné par l'expression :

$$k = \frac{x-1}{2}, \quad \text{soit} \quad k \approx \frac{x}{2}\cdot$$

Considérons un exemple numérique en supposant que $x = 10$, on en déduit immédiatement que $k = 5$. On peut alors évaluer :

$$E_{10}(x) = \frac{10!}{10^{11}} = 0{,}36\,10^{-4}.$$

Rappelons que cette erreur est liée strictement à la troncature des deux séries ; lors de l'exécution des calculs, il faudra ajouter les erreurs effectuées sur chacun des termes constituant les sommes.

Il suffit donc de calculer la somme des cinq premiers termes des deux séries pour connaître $Ci(10)$ avec une précision supérieure à $0{,}36\,10^{-4}$. La somme $S_{2k}(10)$ donne : $S_{2k}(10) = -0{,}454\,81\,10^{-2}$ et l'on peut donc voir que l'algorithme proposé permet d'obtenir les trois premiers chiffres significatifs exacts. En définitive :

$$Ci(10) = 4{,}548\,10^{-2} \pm 0{,}003\,6\,10^{-2}.$$

Pour la petite histoire, il semble que ce soit Stokes G.G. (1819–1903) qui ait pressenti le premier l'intérêt d'une telle procédure, et c'est à Poincaré H. (1854–1912) que revient le mérite d'avoir développé la théorie (1886), essentiellement dans le but de résoudre quelques équations différentielles apparaissant en mécanique céleste et en mécanique analytique.

Ajoutons que ces développements servent principalement à approcher les fonctions dites « spéciales » pour des arguments dont le module est élevé — disons de plusieurs unités ou dizaines d'unités.

Les travaux de Poincaré permettent d'apporter quelques précisions sur la notion de développement asymptotique et ce sont ses propres idées que nous allons rappeler succinctement.

### Définition

On dira que la série $S_n(x)$ représente un développement asymptotique de la fonction $f(x)$ au voisinage de l'infini si, sur un certain intervalle $I]a, \infty[$ auquel appartient $x$, les conditions suivantes sont remplies :

En posant $R_n(x) = x^n\big(f(x) - S_n(x)\big)$, on doit avoir :

$$\lim_{x\to\infty} R_n(x) = 0 \text{ quel que soit } n \text{ même lorsque}$$
$$\lim_{n\to\infty} R_n(x) > \infty \text{ quel que soit } x \text{ fixé.}$$

Il en résulte que l'on pourra toujours prendre $x$ suffisamment grand de telle sorte que l'inégalité suivante soit vérifiée :

$$\Big|x^n\,[f(x) - S_n(x)]\Big| < \varepsilon,$$

où $\varepsilon$ est un infiniment petit positif fixé à l'avance.

Pour utiliser au mieux le développement asymptotique, on choisit $n$ de telle façon que l'erreur commise sur l'approximation soit la plus petite possible.

## 2. Quelques propriétés utiles des développements asymptotiques

***Unicité du développement asymptotique*** – Quand une fonction $f(x)$ admet un développement asymptotique, alors ce développement est unique.

L'affirmation inverse est fausse car deux fonctions différentes peuvent avoir le même développement asymptotique. En effet, supposons que les deux fonctions diffèrent de $\exp(-\alpha x)$ avec la partie réelle de $\alpha$ positive, alors, les coefficients du développement asymptotique de cette quantité sont identiquement nuls.

Toutes les fonctions $f(x)$ n'admettent pas nécessairement un développement asymptotique, cependant, il se peut qu'il existe une fonction $\Phi(x)$ telle que la fonction :

$$\Psi(x) = \frac{f(x)}{\phi(x)}$$

admette un développement asymptotique $\sum_{j=0}^{\infty} \frac{a_j}{x^j}$. Alors on peut écrire :

$$f(x) = \phi(x) \sum_{j=0}^{\infty} \frac{a_j}{x^j} \cdot$$

***Intégration d'un développement asymptotique*** – Supposons que la fonction $f(x)$ admette un développement asymptotique que nous écrirons :

$$a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} + \cdots + \frac{a_n}{x^n} + \cdots$$

avec $a_0 \neq 0$ ; alors on peut intégrer le développement sur l'intervalle $x \leq t \leq \infty$ pour les grandes valeurs de $x$. Nous obtenons donc :

$$\int_x^{\infty} \left[ f(t) - a_0 - \frac{a_1}{t} \right] \mathrm{d}t \leq \sum_{k=2}^{\infty} \frac{a_k}{(k-1)x^{k-1}} \cdot$$

On comprendra immédiatement la nécessité de retrancher les deux premiers termes qui conduiraient inévitablement à des divergences.

***Différentiation d'un développement asymptotique*** – La différentiation des séries en général est une opération délicate et il faut des conditions assez fortes sur les fonctions développées pour qu'il y ait égalité entre la dérivée de $f(x)$ et la dérivée du développement asymptotique. Pour commencer, il faut que $f(x)$ admette une dérivée $f'(x)$, de plus, il faut que $f'(x)$ soit continue et admette un développement asymptotique ; alors l'opération de dérivation devient légitime.

***Combinaison linéaire de développements asymptotiques*** – Si deux fonctions $f(x)$ et $g(x)$ admettent chacune un développement asymptotique que nous écrirons sous la forme :

$$f(x) = \sum_{n=0}^{\infty} \frac{a_n}{x^n} \quad \text{et} \quad g(x) = \sum_{n=0}^{\infty} \frac{b_n}{x^n},$$

alors la fonction $\Phi(x) = \lambda f(x) + \mu g(x)$, où $\lambda$ et $\mu$ sont deux nombres, admet pour développement asymptotique :

$$\Phi(x) = \sum_{k=0}^{\infty} \frac{\lambda a_k + \mu b_k}{x^k} \cdot$$

***Produit de deux développements asymptotiques*** – Si deux fonctions $f(x)$ et $g(x)$ admettent chacune un développement asymptotique alors la fonction $\Psi(x) = f(x) \cdot g(x)$ a pour développement asymptotique :

$$\Psi(x) = \sum_{n=0}^{\infty} \frac{1}{x^n} \sum_{k=0}^{\infty} a_k b_{n-k}.$$

***Rapport de deux développements asymptotiques*** – Si deux fonctions $f(x)$ et $g(x)$ admettent chacune un développement asymptotique alors la fonction $\Omega(x) = f(x)/g(x)$ a pour développement asymptotique :

$$\Omega(x) = \frac{\sum_{k=0}^{\infty} \frac{a_k}{x^k}}{\sum_{k=0}^{\infty} \frac{b_k}{x^k}},$$

à condition toutefois que $b_0$ soit différent de zéro.

– Si une fonction $f(x)$ admet un développement asymptotique et si une fonction $\Theta[f(x)]$ possède un développement en série entière convergent dans le cercle de rayon $|f| < \rho$, alors le développement asymptotique de $\Theta$ s'obtient en portant le développement asymptotique de $f$ dans le développement en série entière de $\Theta$ .

# 3. Développement asymptotique de quelques fonctions spéciales

## 3.1. Sinus intégral

$$Si(x) = \int_0^x \frac{\sin(t)}{t}\,\mathrm{d}t$$

$$Si(x) = \frac{\pi}{2} - \frac{\cos(x)}{x}\sum_{n=0}^{\infty}(-1)^n\frac{(2n)!}{x^{2n}} - \frac{\sin(x)}{x}\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{x^{2n+1}}\,.$$

## 3.2. Cosinus intégral

$$Ci(x) = -\int_x^{\infty} \frac{\cos(t)}{t}\,\mathrm{d}t$$

$$Ci(x) = \frac{\sin(x)}{x}\sum_{n=0}^{\infty}(-1)^n\frac{(2n)!}{x^{2n}} - \frac{\cos(x)}{x}\sum_{n=0}^{\infty}(-1)^n\frac{(2n+1)!}{x^{2n+1}}\,.$$

## 3.3. Exponentielle intégrale

$$I(x) = \int_x^{\infty} \frac{\exp(x-t)}{t}\,\mathrm{d}t \quad \text{pour} \quad x > 0$$

$$I(x) = \sum_{n=0}^{\infty}(-1)^n\frac{n!}{x^{n+1}} \cdots + (-1)^{n+1}(n+1)!\int_x^{\infty}\frac{\exp(x-t)}{t^{n+2}}\,\mathrm{d}t.$$

## 3.4. Les fonctions erf(x) et cerf(x)

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,\mathrm{d}t$$

$$\begin{aligned}\mathrm{cerf}(x) &= 1 - \mathrm{erf}(x)\\ &= \frac{\exp(-x^2)}{x\sqrt{\pi}}\left(1 - \frac{1}{2x^2} + \frac{1\cdot 3}{2^2x^4} - \frac{1\cdot 3\cdot 5}{2^3x^6} + \cdots + (-1)^n\frac{1\cdot 3\cdot 5\cdots(2n-3)}{2^{n-1}x^{2n-2}}\cdots\right)\end{aligned}$$

## 3.5. Les intégrales de Fresnel (1788–1827)

$$C(x) = \int_0^x \cos(\pi t^2/2)\,\mathrm{d}t \quad \text{et} \quad S(x) = \int_0^x \sin(\pi t^2/2)\,\mathrm{d}t$$

$$C(x) = \frac{1}{2} + \frac{\cos(\pi x^2/2)}{\pi x} \sum_{n=0}^{\infty} (-1)^{n+1} \frac{1 \cdot 5 \cdot 9 \cdots (4n+1)}{(\pi x^2)^{2n+1}} - \frac{\sin(\pi x^2/2)}{\pi x} \sum_{n=0}^{\infty} (-1)^{n+1} \frac{3 \cdot 7 \cdot 11 \cdots (4n-1)}{(\pi x^2)^{2n+1}}.$$

$$S(x) = \frac{1}{2} + \frac{\sin(\pi x^2/2)}{\pi x} \sum_{n=0}^{\infty} (-1)^{n+1} \frac{1 \cdot 5 \cdot 9 \cdots (4n+1)}{(\pi x^2)^{2n}} + \frac{\cos(\pi x^2/2)}{\pi x} \sum_{n=0}^{\infty} (-1)^{n+1} \frac{3 \cdot 7 \cdot 11 \cdots (4n-1)}{(\pi x^2)^{2n}}.$$

## 3.6. Les fonctions de Bessel (1784–1846)

Le calcul numérique des fonctions de Bessel fait l'objet d'un chapitre à part (annexe F), cependant le calcul de ces fonctions pour les grandes valeurs de l'argument trouve sa place dans ce chapitre.

Les fonctions de Bessel de première et de seconde espèces notées respectivement $J_n(x)$ et $N_n(x)$ sont solutions de l'équation différentielle :

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + (1 - \frac{n^2}{x^2})y = 0.$$

Les développements asymptotiques sont donnés par les expressions :

$$J_n(x) = \sqrt{\frac{2}{\pi x}}\Big(A_n(x)\cos(U) - B_n(x)\sin(U)\Big)$$
$$N_n(x) = \sqrt{\frac{2}{\pi x}}\Big(A_n(x)\sin(U) - B_n(x)\cos(U)\Big)$$

dans lesquelles on a noté :

$$U = x - (n + \frac{1}{2})\frac{\pi}{2},$$

$$A_n = 1 - \frac{(4n^2-1^2)(4n^2-3^2)}{2!(8x)^2} + \frac{(4n^2-1^2)(4n^2-3^2)(4n^2-5^2)(4n^2-7^2)}{4!(8x)^4} - \cdots$$
$$B_n = \frac{(4n^2-1^2)}{1!(8x)} - \frac{(4n^2-1^2)(4n^2-3^2)(4n^2-5^2)}{3!(8x)^3} + \cdots$$

## 4. Éléments de bibliographie

A. ANGOT (1972) *Complément de Mathématiques*, Masson, Paris.
K.W. BREITUNG (1994) *Asymptotic approximations for probability integrals*, Springer-Verlag.
R.B. DINGLE (1973) *Asymptotic expansions*, Academic Press.
F.W.J. OLVER (1974) *Asymptotics and special functions*, Academic Press.
E.T. WHITTAKER et C.N. WATSON (1927) *A course of modern analysis*, Cambridge University Press, 4$^{e}$ édition.
R. WONG (1989) *Asymptotic approximations of integrals*, Academic Press.

# 4 Résolution des équations numériques

Dans ce chapitre, nous étudions quelques méthodes de résolution des équations du type $f(x) = 0$ suivi de quelques méthodes de résolution de systèmes non linéaires d'équations, puis nous terminons par l'étude d'une méthode spécifique concernant les équations algébriques. Nous limitons notre étude aux fonctions réelles, mais, comme nous le verrons, bien des méthodes pourront être étendues sans grande difficulté au cas de la variable complexe ; et dans la mesure où les calculateurs utilisés permettent d'effectuer les calculs en complexes, il n'y aura que peu de changements à apporter aux algorithmes et programmes proposés.

On a l'habitude de classer les équations $f(x) = 0$ en deux types qui sont les équations algébriques (polynômes) et les équations transcendantes c'est-à-dire celles qui ne sont pas réductibles à un polynôme. Toutes les méthodes qui s'appliquent à la résolution d'équations transcendantes s'appliquent également aux équations algébriques et cela sans préjuger des ennuis possibles de calcul numérique. En revanche, pour ce qui concerne les polynômes, il existe quelques méthodes spécifiques dont certaines sont très utiles à connaître, c'est la raison pour laquelle nous étudions d'abord les méthodes générales avant d'aborder les méthodes spécifiques.

## 1. Généralités sur la résolution des équations f(x) = 0

### 1.1. Localisation des racines des équations

Avant d'entreprendre le calcul à proprement parler d'une ou de plusieurs racines, il convient de localiser soigneusement ces racines pour ne pas avoir à rencontrer de surprises désagréables, telles que l'arrêt inopiné de la machine... Plusieurs possibilités s'offrent à nous pour localiser les racines, en voici quelques-unes.

**a.** On peut avoir recours à certains théorèmes de mathématiques nous fournissant ces renseignements. Ce sont par exemple les théorèmes généraux concernant les polynômes orthogonaux relativement à une fonction poids $\omega(x)$ donnée.

**b.** On peut, dans le même ordre d'idée, effectuer une étude particulière de la fonction $f(x)$ et en effectuer la représentation graphique.

**c.** On peut aussi tabuler la fonction $f(x)$ avec un pas variable pour tenter d'obtenir le maximum de renseignements.

**d.** Dans certains cas, on pourra essayer d'utiliser les suites de Sturm (1803–1855), mais il faut bien dire que l'intérêt de telles suites est essentiellement théorique, et l'utilisation devient délicate même pour les polynômes à coefficients entiers (de quelques unités cependant) dont

le degré ne dépasse pas la dizaine. Nous examinerons en détail cette méthode dans l'annexe A. La localisation préliminaire des racines peut être plus difficile qu'il n'y paraît en première analyse notamment lorsque $f(x)$ présente des extremums au voisinage de l'axe des $x$ et cela sans préjuger qu'il y ait des racines ou non. Ces valeurs approchées des racines vont servir de point de départ à tout un ensemble de méthodes, le plus souvent itératives, qui convenablement utilisées vont permettre d'obtenir des valeurs précises des racines.

## 1.2. La dichotomie

Étymologiquement ce mot d'origine grecque signifie action de couper en deux. Le principe de la méthode est le suivant : supposons que l'on sache qu'une racine ait été localisée dans l'intervalle $(a, b)$ et que cette racine soit unique dans cet intervalle. Nous allons étudier le signe de $f(x)$ dans $(a, b)$. Pour cela nous comparons les signes de $f(a)$ et de $f\left(\frac{a+b}{2}\right)$. S'ils sont identiques cela signifie que la racine est comprise dans l'intervalle $\left(\frac{a+b}{2}, b\right)$ ; s'ils sont différents cela veut dire que la racine appartient à l'intervalle $\left(a, \frac{a+b}{2}\right)$. On recommence la procédure dans l'intervalle où se situe la racine, mais on s'aperçoit que la taille de l'intervalle a été divisée par deux. On va poursuivre ainsi jusqu'à ce que l'on obtienne la précision souhaitée. Il est facile de connaître le nombre de tours que doit comporter la procédure à partir du moment où l'on sait quelle est la précision de la machine utilisée et que la racine a fait l'objet d'une localisation raisonnable.

***Combien de fois devons-nous répéter la division de l'intervalle par deux ?*** – Désignons par $x_0$ la racine : $f(x_0) = 0$, et par $(a, b)$ l'intervalle sur lequel on exécute la dichotomie, $x_0$ appartenant à $(a, b)$.

Après $n$ tours, la dichotomie définit l'intervalle minimum $\varepsilon_1$ qui sera le dernier susceptible de modifier effectivement la valeur $x_0$, c'est-à-dire :

$$\varepsilon_1 = (b - a)\left(\frac{1}{2}\right)^n$$

soit en valeur relative :

$$\left|\frac{\varepsilon_1}{x_0}\right| = \left|\frac{b-a}{x_0}\right|\left(\frac{1}{2}\right)^n \leq 10^{-p}.$$

L'inégalité $(a)$ définit le nombre de tours à effectuer pour obtenir la précision optimum, soit en passant aux logarithmes :

$$0,301n - \log_{10}\left|\frac{b-a}{x_0}\right| \leq p,$$

$$\text{soit encore : } n = \frac{1}{0{,}301}\left(p + \log_{10}\left|\frac{b-a}{x_0}\right|\right).$$

Comme on a effectué une localisation de $x_0$ telle que $\left|\frac{b-a}{x_0}\right|$ soit de l'ordre de quelques unités, disons 10 pour fixer les idées (ce qui résout la plupart des cas réels rencontrés), $n$ sera de l'ordre de :

$$n = \frac{1}{0{,}301}(p + 1).$$

Si $p$ vaut 9, on trouve alors 33 tours de calcul, si $p$ vaut 15 on obtient 53 tours de calcul.

Comme le montre l'expression donnant $n$, quand bien même la dynamique de l'intervalle serait multipliée par 10, cela ne ferait qu'ajouter une unité à $p+1$, ce qui ne changerait pas grand-chose, et, dans l'exemple précédent, on trouverait alors 36 et 56 tours.

Quoi qu'il en soit, la prudence la plus élémentaire exige toujours de reporter la prétendue racine dans l'équation et d'afficher le résultat... À ce sujet, encore une remarque : **quel doit être l'ordre de grandeur de $f(x_0)$?**

On s'attend à trouver zéro, mais en général il n'en sera rien. Cela dépend de la fonction $f(x)$ dont le calcul est agrémenté d'une erreur congénitale liée d'une part aux approximations utilisées par l'algorithme, d'autre part aux propagations des erreurs de troncature entre autres. Dans les centre de calcul, il est possible de consulter une documentation fournie par le fabricant de logiciel qui informe l'utilisateur sur ce type d'erreur. Pour des valeurs de l'argument appartenant à un certain domaine, un écart type est fourni, car il s'agit dans ce cas d'une erreur statistique.

Comme $f(x_0) \simeq 0$ et $f(x_0+\varepsilon_1) \simeq 0$, un développement au premier ordre donne l'ordre de grandeur de $f(x_0+\varepsilon_1)$ :

$$f(x_0+\varepsilon_1) = f(x_0) + \varepsilon_1 f'(x_0),$$
$$\text{d'où} \quad f(x_0+\varepsilon_1) - f(x_0) \simeq \varepsilon_1 f'(x_0) \simeq 10^{-p} x_0 f'(x_0).$$

Cette technique est simple à mettre en œuvre car elle ne fait appel qu'au calcul de la fonction elle-même, en outre elle offre l'avantage de conserver la suite des solutions approchées dans le voisinage retenu ce qui n'est pas le cas avec d'autres méthodes comme celle de Newton par exemple. Sur le Web (*) on trouvera le programme `dichot.c` réalisant la recherche de racine par dichotomie.

## 1.3. Méthode d'approximations successives ne faisant intervenir que la fonction

Le problème qui consiste à rechercher la racine d'une équation $f(x) = 0$ dans un certain domaine peut être modifié en ajoutant simplement $x$ à chacun des deux membres, soit : $f(x) + x = x$. Par goût de la simplicité, on peut poser $f(x) + x = g(x)$ sans changer quoi que ce soit.

Nous allons décrire une méthode itérative qui ne fait apparaître que la seule fonction $f(x)$ (ou $g(x)$) à condition toutefois que la fonction $f(x)$ obéisse à certaines règles de régularité telles que la continuité et la dérivabilité (dérivée première continue). À partir de la première approximation $x_0$ obtenue par un des procédés proposés précédemment, on forme la suite :

$$x_1 = g(x_0)$$
$$x_2 = g(x_1)$$
$$\ldots\ldots\ldots\ldots$$
$$x_n = g(x_{n-1})$$
$$\ldots\ldots\ldots\ldots$$

Pour que cette technique soit efficace, il est indispensable que la suite des $x_k$ soit convergente. Dans l'hypothèse où $f(x)$ est continue et où la suite des $x_k$ converge, alors la limite de la suite que nous désignerons par $x^*$ est racine de $f(x) = 0$. En effet, quand $k$ tend vers l'infini $x_k$ et $x_{k+1}$ tendent vers $x^*$. La continuité de $f(x)$ entraîne celle de $g(x)$ qui entraîne à son tour

* `http://www.edpsciences.com/guilpin/`

que $\lim_{k\to\infty} x_{k+1} = \lim g(x_k) = x^* = g(x^*)$. Maintenant il nous faut examiner quelles sont les conditions qui assurent la convergence des $x_k$.

À la suite des $x_k$ associons la série des $u_k$ définie ainsi :

$$u_n = x_n - x_{n-1} \qquad \text{avec } u_0 = x_0.$$

On peut d'ores et déjà remarquer que la convergence de la série $u_n$ entraîne celle de la suite $x_n$ et que la suite $x_n$ et la série associée $u_n$ admettent la même limite $x^*$. En effet si nous appelons $S_p$ la somme des $(p+1)$ termes de la série, nous avons :

$$S_p = \sum_{k=0}^{p} u_k = x_p,$$

ce qui démontre la dernière proposition.

Développons le terme général $u_k$ de la série :

$$u_k = x_k - x_{k-1} = g(x_{k-1}) - g(x_{k-2}).$$

Appliquons le théorème de Rolle à la dernière expression, cela donne :

$$u_k = g(x_{k-1}) - g(x_{k-2}) = (x_{k-1} - x_{k-2})g'(\xi)$$

avec $\xi$ compris dans l'intervalle $(x_{k-1} - x_{k-2})$. En utilisant la relation de définition de la série, nous pouvons écrire :

$$u_k = u_{k-1}g'(\xi).$$

La règle de d'Alembert (1717–1783) va nous donner la condition de convergence de la série $u_k$ :

$$\left|\frac{u_k}{u_{k-1}}\right| < 1.$$

On en déduit que l'algorithme converge lorsque la condition $|g'(\xi)| < 1$ est vérifiée dans tout le domaine voisin de la racine dans lequel on effectue les itérations.

## 1.4. Généralisation de la méthode

Au cas où $|g'(\xi)| > 1$, la suite des itérations est divergente. Pour pallier cet inconvénient, au lieu d'ajouter $x$ aux deux membres de l'équation $f(x) = 0$, il suffit d'ajouter $mx$ et de choisir convenablement $m$ pour que la convergence soit assurée. Nous pouvons écrire : $f(x)+mx = mx$, puis l'on pose $f(x) + mx = g(x)$ ; d'où la nouvelle suite des $x_k$ :

$$x_k = \frac{g(x_{k-1})}{m}.$$

La condition de convergence de la suite des $x_k$ est :

$$\left|\frac{g'(x_{k-1})}{m}\right| < 1, \quad \text{soit encore :} \quad -1 < \frac{f'(x)+m}{m} < +1.$$

Il est toujours possible de choisir m positif ou négatif, de taille suffisante, pour assurer cette double inégalité à condition que la dérivée soit bornée et que $f(x^*) \neq f'(x^*)$ si $x^*$ est une racine.

***a – Représentation géométrique de la méthode*** – On porte sur le même graphe les fonctions $y = mx$ et $y = g(x) = f(x) + mx$.

Cela va nous permettre de localiser non seulement la ou les racines mais aussi de connaître la dérivée au voisinage des racines donc éventuellement d'ajuster le coefficient $m$ selon le cas qui se présente. Sur le graphe de la figure 4.1, on peut suivre la suite des points générée à partir de la première approximation $x_0$ :

| | |
|---|---|
| $x_0$ | $g(x_0)$ |
| $x_1 = g(x_0)$ | $g(x_1)$ |
| $x_2 = g(x_1)$ | $g(x_2)$ |
| ..... | ..... |
| $x_n = g(x_{n-1})$ | $g(x_n)$ |

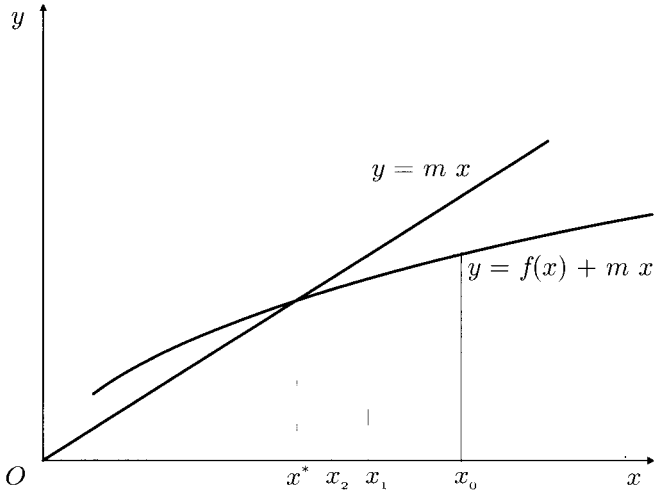


**Figure 4.1.** *Méthode par itérations.*

***b – Précision sur la détermination de la racine*** – On désigne par $M$ la limite supérieure de la valeur absolue de la dérivée dans le domaine où l'on effectue les itérations. Nous pouvons donc écrire :

$$|u_n| = M\,|u_{n-1}| \text{ avec } M < 1.$$

Si nous arrêtons les itérations après $n$ tours, nous aurons :

$$x_n = \sum_{p=0}^{n} u_p = S_n$$

et l'on déduit que l'erreur $e$ sur la racine $x^*$, en adoptant $x^*$ comme approximation, est :

$$e = \sum_{p=n+1}^{\infty} u_p.$$

On en conclut que :

$$e \leq \sum_{p=n+1}^{\infty} |u_p| = |u_n| \left| K + K^2 + K^3 + \cdots \right| .$$

On reconnaît une progression géométrique de raison $K$ et de premier terme $K$. Donc :

$$e \leq |u_n| \frac{K}{1-K} .$$

Comme en général on arrête les itérations lorsque deux valeurs consécutives sont égales à la précision relative de la machine on peut alors écrire :

$$\left| \frac{u_n}{x_n} \right| = \left| \frac{x_n - x_{n-1}}{x_n} \right| = 10^{-p}$$

d'où

$$e \leq |x_n| \frac{K}{1-K} 10^{-p}.$$

En pratique, on peut estimer $K$ de la façon suivante :

$$K = \left| \frac{g(x_n) - g(x_{n-1})}{x_n - x_{n-1}} \right| .$$

Sur le Web[*], on trouvera le programme `itera.c` qui calcule les racines par cette méthode itérative.

## 1.5. La méthode de Newton (1643–1727)

Nous avons déjà rencontré la méthode de Newton à propos du calcul de la racine carrée d'un nombre et il nous reste donc à généraliser cette technique à la recherche d'un zéro d'une fonction quelconque $f(x)$.

Supposons qu'un travail préliminaire nous ait permis de localiser une racine dans l'intervalle $(a, b)$. On désigne par $x_0$ la première approximation de cette racine. Comme $x_0$ n'est qu'une approximation, la racine $x^*$ que l'on recherche s'exprime de la façon suivante : $x^* = x_0 + h_0$. Nous allons chercher à calculer $h_0$ ou plus exactement une approximation de $h_0$. Nous pouvons écrire :

$$f(x^*) = f(x_0 + h_0) = 0.$$

Linéarisons le problème, c'est-à-dire que nous développons la fonction $f(x)$ au premier ordre au voisinage de $x_0$. Nous obtenons :

$$f(x_0 + h_0) = 0 = f(x_0) + h_0 f'(x_0)$$

ce qui va permettre d'obtenir la valeur approchée de $h_0$ :

$$h_0 = -\frac{f(x_0)}{f'(x_0)} .$$

* `http://www.edpsciences.com/guilpin/`

Cette valeur nous permet d'obtenir, sous certaines réserves, une meilleure approximation que nous écrivons :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} .$$

Rien ne nous empêche, excepté toutefois l'absence éventuelle de convergence, d'itérer successivement la procédure jusqu'à ce que l'on obtienne la racine $x^*$ avec la précision souhaitée. On aura formé la suite :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} .$$

Dans la mesure où l'on dispose non seulement de la fonction mais aussi de sa dérivée on peut espérer obtenir une vitesse de convergence rapide du moins dans le cas où les racines sont simples. Au reste, l'interprétation géométrique est très parlante. Représentons la courbe $y = f(x)$ sur un diagramme (*cf.* Fig. 4.2) ainsi que la première approximation $x_0$. La linéarisation du problème consiste à remplacer un arc de courbe de $f(x)$ par un segment de droite qui est tangent à la courbe dans le cas général. Traçons la tangente au point $[x_0, f(x_0)]$. Elle coupe l'axe des $x$ au point $x_1$ donné par l'expression :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} .$$

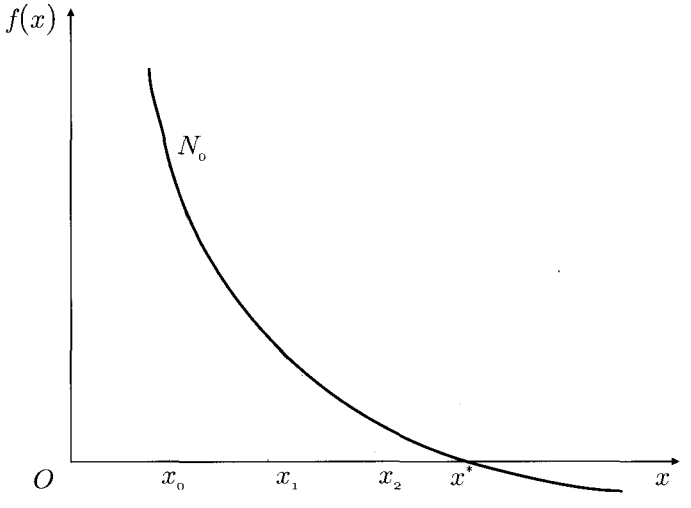


**Figure 4.2.** *Méthode de Newton.*

La répétition du processus nous permet d'obtenir sans problème les points suivants qui sont $x_2, x_3, \ldots$

***a – Ennuis possibles avec la méthode de Newton*** – En l'absence de difficultés, la méthode de Newton assure une convergence très rapide de la suite de $x_k$. Dans le cas de racines simples, la convergence est même quadratique ce qui confère à cette technique un intérêt particulier quand on dispose déjà d'un chiffre significatif. Nous utiliserons cet algorithme pour réaliser une partie de la fonction de bibliothèque calculant la racine carrée, et nous verrons ultérieurement dans quelles conditions.

Malheureusement des problèmes peuvent surgir qui sont liés à la nature de la fonction, de sa dérivée ainsi que du point $x_0$ à partir duquel s'effectuent les itérations. En effet dans le domaine voisin de la racine où l'on travaille, la courbe peut présenter des changements de concavité ou encore présenter des extremums. Si un point $x_k$ tombe au voisinage d'un extremum, alors $f'(x)$ est proche de zéro, et la technique de Newton nous donnera un point $x_{k+1}$ qui pourra être très loin de la solution recherchée. Pour peu qu'il y ait une racine dans le voisinage de $x_{k+1}$, on aura toutes les chances de la calculer en croyant être dans un autre voisinage. Cela laisse la porte ouverte à quelques désastres possibles... Lorsque l'on se trouve dans un cas analogue, on doit obligatoirement le savoir avant d'entamer les calculs puisqu'on a pris les précautions d'usage. Pour éviter qu'au départ la suite générée présente des instabilités ou des divergences locales, on combine la méthode de Newton avec celle des parties proportionnelles.

***b – Estimation de la précision de la méthode de Newton*** – La plus grande sagesse veut que l'on reporte la valeur $x_k$ que l'on estime être une approximation raisonnable de la racine dans la fonction $f(x)$. On note : $e_k = f(x_k)$. Si $e_k$ est de l'ordre de grandeur de l'incertitude introduite par le calcul de $f(x_k)$, cela ne signifiera certainement pas que $x_k$ soit une racine, mais qu'il n'y a pas de contradiction avec le fait qu'elle puisse en être une. En revanche, si $e_k$ est trop grand, il y a lieu de s'inquiéter et d'examiner de nouveau le problème. C'est une attitude identique à celle qu'on adopte lorsque l'on fait usage de la preuve par 9 ou par 11. Seul un résultat négatif donne la certitude d'une erreur, tandis qu'un résultat positif ne donne pas la certitude d'une exactitude mais seulement une probabilité plus ou moins élevée pour que se présente cette éventualité.

Deux cas peuvent se présenter : ou bien la suite générée est une suite encadrante, ou bien la suite générée est une suite monotone croissante ou monotone décroissante du moins à partir d'un certain rang.

• ***La suite est encadrante*** – Il faut bien avouer que c'est un cas d'école qui se présente rarement, car cela impose que la racine soit aussi un point d'inflexion. On retiendra les deux dernières valeurs de $x_k$ et $x_{k-1}$ car on sait que la racine se situe entre ces deux dernières valeurs entachées toutefois des inévitables erreurs d'arrondi et des erreurs dues à l'approximation des fonctions de bibliothèque. Du fait de ces erreurs, il est fort possible que l'on trouve deux limites en machine. Bien entendu, l'estimation de l'incertitude est liée intimement à la différence $|x_k - x_{k-1}|$ sans oublier les erreurs entachant les calculs de $f(x_{k-1})$ et $f(x_k)$. À partir de là, on retombe dans le calcul standard des erreurs.

• ***La suite est monotone à partir d'un certain rang*** – À partir d'un certain rang $r$, tous les $x_q$ générés sont donc classés par ordre croissant ou décroissant. Il faut s'assurer que lorsque deux valeurs consécutives $x_k$ et $x_{k+1}$ sont égales à la précision de la machine, le reste des opérations négligées n'apporterait plus de contribution à la précision du calcul, autrement dit il faut que :

$$\sum_{j=k+1}^{\infty} |x_{k+1} - x_k| = \sum_{j=k+1}^{\infty} \left| \frac{f(x_k)}{f'(x_k)} \right|$$

soit négligeable.

Comme il n'y a aucune raison particulière pour que la tangente soit verticale, la seule condition de convergence est que $f(x_k)$ tende effectivement vers zéro quand $k$ tend vers l'infini. C'est alors que la précision de la machine utilisée intervient ainsi que la nature de la fonction $f(x)$. Pour obtenir une estimation raisonnable de l'erreur, il faudra chaque fois faire une étude particulière dont il est possible de donner les grandes lignes : il est facile de donner à la valeur approchée $x^*$ de la racine une suite d'accroissements $\pm\ \mathrm{d}x, \pm 2\ \mathrm{d}x, \pm 3\ \mathrm{d}x \ldots, \pm k\ \mathrm{d}x$ avec $\mathrm{d}x = 10^{-p} x^*$

où $10^{-p}$ est la précision relative de la représentation en machine. Peu de valeurs suffisent. On tabulera alors la fonction $f(x)$ pour ces différentes valeurs et l'on notera l'intervalle dans lequel s'effectue le changement de signe. Compte tenu de la précision de la représentation on pourra estimer un intervalle dans lequel se trouve effectivement la racine. C'est une évaluation directe de l'erreur qui constitue un garde-fou contre les mauvaises surprises.

On trouvera sur le Web(*) le programme `newton1d.c` qui calcule les racines par la méthode de Newton.

### 1.6. Méthode de Newton et des parties proportionnelles

En général, cette méthode est utilisée lorsque dans l'intervalle où l'on recherche la racine, peut se trouver un point d'inflexion ou un extremum. Donc, pour éviter que la valeur de la dérivée apparaissant dans la méthode de Newton notamment durant les premiers tours d'itération, ne provoque des instabilités préjudiciables à la bonne conduite du calcul, on lui associe la méthode des parties proportionnelles qui ramène les valeurs des approximations dans un domaine que l'on espère généralement plus raisonnable. Soit $x_0$ la première approximation et $[x_0, f(x_0)]$ le point $A$ sur la courbe $y = f(x)$. On effectue une deuxième approximation :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \cdot$$

à laquelle correspond le point $B$ sur la courbe.

La méthode des parties proportionnelles consiste à choisir comme approximation suivante le point d'intersection de l'axe des $x$ et de la droite joignant les points $A$ et $B$. Désignons par $x_2$ l'abscisse de ce point, nous obtenons :

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} \cdot$$

À nouveau on calcule le point $x_3$ par la méthode de Newton puis le point $x_4$ par la méthode des parties proportionnelles et ainsi de suite (*cf.* Fig. 4.3, page suivante).

On trouvera sur le Web(*) le programme `newtonpp.c` qui calcule les racines au moyen de la méthode de Newton et des parties proportionnelles.

***Remarque*** : Les ennuis de stabilité liés à l'usage de la méthode de Newton naissent essentiellement du choix de la première approximation qui parfois est trop éloignée de la racine que l'on recherche ou encore quand une abscisse de la suite formée se situe trop près d'un extremum. Quand on tient un certain nombre de chiffres significatifs, il faut rappeler qu'en l'absence de racines multiples, la méthode converge quadratiquement et pratiquement seules les erreurs de troncature viennent altérer la précision du résultat.

## 2. Résolution d'un système non linéaire de deux équations à deux inconnues

Nous nous proposons de résoudre un système non linéaire de deux équations à deux inconnues que nous écrivons :

$$f(x, y) = 0$$
$$g(x, y) = 0.$$

* `http://www.edpsciences.com/guilpin/`

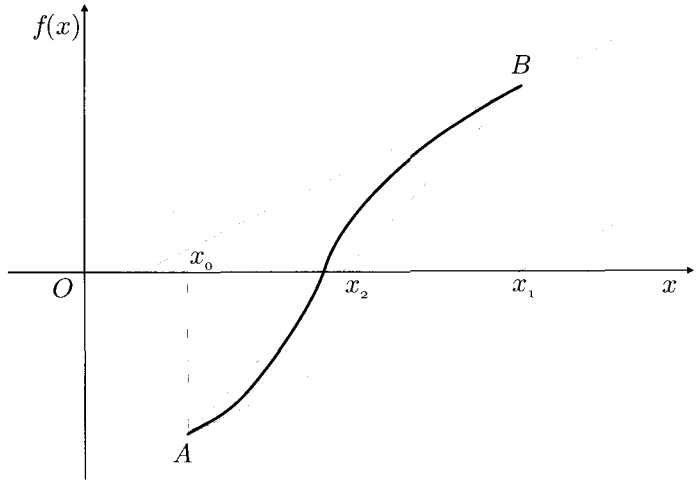


**Figure 4.3.** *Méthode de Newton et des parties proportionnelles.*

Comme chaque fois en pareil cas, il sera nécessaire de commencer l'étude par une représentation graphique laquelle permettra de se faire une idée des éventuelles difficultés qui peuvent se présenter, et de déterminer une première approximation $(x_0, y_0)$. Le système étant non linéaire, nous allons commencer par le linéariser puis par effectuer des itérations sur le système linéarisé. Nous admettrons que les fonctions sont deux fois continûment différentiables dans un domaine $D$ du plan $(x, y)$ qui contient la racine recherchée et que le déterminant fonctionnel (ou encore le jacobien) :

$$J(x,y) = \frac{\partial f}{\partial x}\frac{\partial g}{\partial y} - \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}$$

est différent de zéro pour la valeur de la racine.

## 2.1. La méthode de Newton

Nous généralisons sans difficulté la méthode étudiée lors de l'étude de l'équation $f(x) = 0$. Soit $x_0$ et $y_0$ une première approximation. On note par $\mathrm{d}x$ et $\mathrm{d}y$ les écarts à la solution exacte c'est-à-dire : $x^* = x_0 + \mathrm{d}x_0$ et $y^* = y_0 + \mathrm{d}y_0$.

Le système s'écrit :

$$f(x^*, y^*) = 0 = f(x_0 + \mathrm{d}x_0, y_0 + \mathrm{d}y_0)$$
$$g(x^*, y^*) = 0 = g(x_0 + \mathrm{d}x_0, y_0 + \mathrm{d}y_0).$$

Pour linéariser le problème il suffit d'effectuer un développement au premier ordre des fonctions $f$ et $g$ au voisinage de $x_0$ et $y_0$. Nous obtenons :

$$f(x_0, y_0) + \mathrm{d}x_0\frac{\partial f}{\partial x} + \mathrm{d}y_0\frac{\partial f}{\partial y} = 0.$$
$$g(x_0, y_0) + \mathrm{d}x_0\frac{\partial g}{\partial x} + \mathrm{d}y_0\frac{\partial g}{\partial y} = 0.$$

Il nous reste à résoudre le système linéaire de deux équations à deux inconnues pour obtenir $\mathrm{d}x_0$ et $\mathrm{d}y_0$. On trouve alors :

$$\mathrm{d}x_0 = \frac{g\dfrac{\partial f}{\partial y} - f\dfrac{\partial g}{\partial y}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

$$\mathrm{d}y_0 = \frac{f\dfrac{\partial g}{\partial x} - g\dfrac{\partial f}{\partial x}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

expressions que l'on calcule au point $(x_0, y_0)$.

Bien entendu on obtient dans des conditions convenables une meilleure approximation :

$$x_1 = x_0 + \mathrm{d}x_0$$
$$y_1 = y_0 + \mathrm{d}y_0$$

à partir de laquelle il va être facile de réitérer la procédure. Ici encore la méthode convergera quadratiquement en l'absence de racines multiples. Cependant il existe des méthodes à convergence simple qui font apparaître elles aussi les dérivées. Au premier examen on peut s'interroger sur l'intérêt présenté par de telles méthodes puisque, avec des moyens semblables, les méthodes quadratiques sont beaucoup plus performantes. Ceci n'est vrai qu'en apparence car dans certains cas les instabilités de calcul rendent inopérantes les méthodes quadratiques tant qu'un certain degré de précision n'a pas été atteint. En revanche ces problèmes disparaissent avec les méthodes à convergence simple. Du reste, comme la vitesse de convergence est éventuellement une source de préoccupation, il est intéressant d'accélérer la convergence de la suite formée en se servant d'un algorithme accélérateur de la convergence des suites : l'epsilon-algorithme par exemple. C'est par ce procédé que nous calculons des diagrammes de phase théoriques lorsque les conditions deviennent délicates.

Le lecteur trouvera le programme `newton2d.c` sur le Web(*).

## 2.2. La méthode de Kacmarz (1937)

Nous allons présenter la méthode, d'abord élaborée pour les systèmes multilinéaires, sur un système linéaire de deux équations à deux inconnues à partir duquel on pourra effectuer toutes les généralisations souhaitables : soit la transposition au cas d'un système multilinéaire, soit la transposition au cas non linéaire que l'on traitera au moyen d'une procédure itérative s'appliquant sur le système linéarisé.

Soit le système

$$a_1x + b_1y + c_1 = 0 \quad (\text{droite } D_1)$$
$$a_2x + b_2y + c_2 = 0 \quad (\text{droite } D_2) \qquad \text{avec } a_1b_2 \neq a_2b_1.$$

Ce système admet une solution désignée par $(x^*, y^*)$. Partant d'une première approximation $A$ de coordonnées $(x_0, y_0)$, on projette ce point sur la droite $D_1$, on obtient le point $A_1$ de

* `http://www.edpsciences.com/guilpin/`

coordonnées $(x_1, y_1)$. Ensuite, on projette le point $A_1$ sur la droite $D_2$ ce qui nous donnera le point $A_2$. On poursuit en projetant le point $A_2$ sur la droite $D_1$ et ainsi de suite jusqu'à ce que l'on obtienne la précision désirée. À présent écrivons les coordonnées de la suite des points ainsi formée. Nous avons :

$$x_1 = x_0 - \frac{a_1 R_1}{a_1^2 + b_1^2}$$
$$y_1 = y_0 - \frac{b_1 R_1}{a_1^2 + b_1^2} \quad \text{avec} \quad R_1 = a_1 x_0 + b_1 y_0 + c_1.$$

Le point $(x_1, y_1)$ est projeté sur la droite $D_2$, à partir duquel on calcule le point $(x_2, y_2)$ :

$$x_2 = x_1 - \frac{a_2 R_2}{a_2^2 + b_2^2}$$
$$y_2 = y_1 - \frac{b_2 R_2}{a_2^2 + b_2^2} \quad \text{avec} \quad R_2 = a_2 x_1 + b_2 y_1 + c_2,$$

et ainsi de suite. Il est possible de montrer que cette méthode est toujours convergente mais qu'elle est seulement à convergence linéaire. L'intérêt de cette méthode repose avant tout sur le fait qu'elle est transposable aux équations non linéaires.

***Application au cas des équations non linéaires*** – Le système des équations linéarisées que nous avons obtenu au paragraphe traitant de la méthode de Newton sera simplement résolu par la méthode de Kacmarz. Nous avions :

$$f(x_0, y_0) + \mathrm{d}x_0 \frac{\partial f}{\partial x} + \mathrm{d}y_0 \frac{\partial f}{\partial y} = 0$$
$$g(x_0, y_0) + \mathrm{d}x_0 \frac{\partial g}{\partial x} + \mathrm{d}y_0 \frac{\partial g}{\partial y} = 0.$$

Si l'on pose :

$$a_1 = \frac{\partial f}{\partial x}$$
$$a_2 = \frac{\partial g}{\partial x}$$
$$b_1 = \frac{\partial f}{\partial y}$$
$$b_2 = \frac{\partial g}{\partial y}$$
$$R_1 = f(x, y)$$
$$R_2 = g(x, y),$$

on obtient les suites des $x_k$ et des $y_k$ données par les expressions :

$$x_k = x_{k-1} - \frac{a_1 R_1}{a_1^2 + b_1^2}$$
$$y_k = y_{k-1} - \frac{b_1 R_1}{a_1^2 + b_1^2}$$

expressions calculées au point $(x_{k-1}, y_{k-1})$

$$x_{k+1} = x_k - \frac{a_2 R_2}{a_2^2 + b_2^2}$$

$$y_{k+1} = y_k - \frac{b_2 R_2}{a_2^2 + b_2^2}$$

expressions calculées au point $(x_k, y_k)$. La convergence n'est plus systématiquement assurée comme dans le cas linéaire. Dans certains calculs, elle a retenu notre attention pour sa grande stabilité, et nous avons pu accélérer la vitesse de convergence au moyen de l'epsilon-algorithme.

Le lecteur trouvera sur le Web(*) le programme `kacmarz.c` permettant l'usage de cette procédure.

# 3. Racines d'un polynôme

## 3.1. Méthode de Bairstow (1914)

Ici encore, nous allons limiter nos préoccupations à l'étude des polynômes à coefficients réels que l'on notera de la façon suivante :

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n$$

avec comme condition quasi évidente $a_0 \neq 0$. Nous savons qu'un polynôme de degré $n$ admet $n$ racines réelles ou complexes et nous allons étudier une méthode très élégante pour obtenir lesdites racines.

L'idée générale du calcul repose sur le schéma suivant : dans le cas où n est supérieur à deux on peut diviser $P_n(x)$ par un trinôme $T(x)$ de la forme :

$$T(x) = x^2 + ux + v$$

ce qui nous permet d'obtenir une expression nouvelle de $P_n(x)$ à savoir :

$$P_n(x) = T(x) Q_{n-2}(x) + Ax + B$$

où $Q_{n-2}(x)$ est un polynôme de degré $(n-2)$ et $(Ax+B)$ un monôme exprimant le reste de la division. Nous allons faire en sorte que $A$ et $B$ qui sont des fonctions implicites de $u$ et $v$ soient nulles ; on réalisera cela en ajustant convenablement les valeurs de $u$ et $v$ apparaissant dans le trinôme. Lorsque nous y serons parvenus, le polynôme s'écrira sous la forme d'un produit strict d'un trinôme et d'un polynôme de degré $(n-2)$. Ainsi, les racines du trinôme seront également les racines du polynôme $P_n(x)$, et l'on sait qu'il n'y a pas de difficultés particulières à calculer les racines d'un trinôme.

Rien ne nous empêche de faire subir le même traitement au polynôme $Q_{n-2}(x)$, et ainsi de suite jusqu'à ce que $Q_p(x)$ soit un polynôme de degré 1 ou 2. En définitive, tout le problème repose sur la détermination des paramètres $u$ et $v$ qui annulent simultanément $A$ et $B$. Nous avons donc à résoudre un système non linéaire de deux équations à deux inconnues que nous écrivons :

$$A(u, v) = 0$$
$$B(u, v) = 0$$

* `http://www.edpsciences.com/guilpin/`

système que l'on résoudra par la méthode de Newton ou de Kacmarz, de toute façon au moyen d'itérations. Si $u_0$ et $v_0$ constituent une première approximation, la méthode de Newton nous permet de calculer les corrections $\mathrm{d}u$ et $\mathrm{d}v$ qu'il convient d'apporter, soit :

$$\mathrm{d}u = \frac{B\dfrac{\partial A}{\partial v} - A\dfrac{\partial B}{\partial v}}{D}$$

$$\mathrm{d}v = \frac{A\dfrac{\partial B}{\partial u} - B\dfrac{\partial A}{\partial u}}{D},$$

$$\text{avec } D = \frac{\partial A}{\partial u}\frac{\partial B}{\partial v} - \frac{\partial A}{\partial v}\frac{\partial B}{\partial u}.$$

les fonctions étant calculées au point $(u_0, v_0)$.

Nous serons effectivement capable de calculer $\mathrm{d}u$ et $\mathrm{d}v$ dans la mesure où nous serons capable d'exprimer $A$ et $B$ ainsi que leurs dérivées partielles et c'est sous cet aspect que nous allons envisager les calculs. Explicitons le polynôme $Q_{n-2}(x)$ :

$$Q_{n-2}(x) = b_0 x^{n-2} + b_1 x^{n-3} + \cdots + b_{n-2}.$$

Par identification de $P_n(x)$ et de $T(x)Q_{n-2}(x) + Ax + B$ on obtient les relations de récurrence suivantes :

$$\begin{array}{lcl}
b_0 & = & a_0 \\
b_1 & = & a_1 - ub_0 \\
b_2 & = & a_2 - ub_1 - vb_0 \\
\cdots\cdots & & \cdots\cdots \\
b_j & = & a_j - ub_{j-1} - vb_{j-2} \\
\cdots\cdots & & \cdots\cdots \\
b_{n-2} & = & a_{n-2} - ub_{n-3} - vb_{n-4} \\
A & = & a_{n-1} - ub_{n-2} - vb_{n-3} \\
B & = & a_n - vb_{n-2}
\end{array}$$

À présent nous savons calculer $A$ et $B$ en fonction des coefficients du polynôme $P_n(x)$ et des paramètres $u$ et $v$. Calculons maintenant les dérivées partielles de $A$ et $B$ par rapport à $u$ et $v$ lesquels sont — rappelons-le — des paramètres indépendants. Pour ce faire, dérivons la forme générale de la relation de récurrence par rapport à $u$ :

$$\frac{\partial b_j}{\partial u} = -b_{j-1} - u\frac{\partial b_{j-1}}{\partial u} - v\frac{\partial b_{j-2}}{\partial u}$$

et posons

$$c_{j-1} = \frac{\partial b_j}{\partial u}$$

nous obtenons ainsi la relation de récurrence suivante :

$$c_{j-1} = -b_{j-1} - uc_{j-2} - vc_{j-3}.$$

Nous explicitons à nouveau l'ensemble des relations :

$$\begin{aligned}
c_0 &= -b_0 \\
c_1 &= -b_1 - uc_0 \\
c_2 &= -b_2 - uc_1 - vc_0 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
c_{n-3} &= -b_{n-3} - uc_{n-4} - vc_{n-5} \\
\frac{\partial A}{\partial u} &= -b_{n-2} - uc_{n-3} - vc_{n-4} \\
\frac{\partial B}{\partial u} &= -vc_{n-3}
\end{aligned}$$

Recommençons le calcul en tout point semblable en dérivant la relation de récurrence initiale par rapport à $v$.

$$\frac{\partial b_j}{\partial v} = -b_{j-1} - u\frac{\partial b_{j-1}}{\partial v} - v\frac{\partial b_{j-2}}{\partial v}$$

Posons

$$d_{j-2} = \frac{\partial b_j}{\partial v}$$

ce qui nous permet d'écrire la relation de récurrence :

$$d_{j-2} = -b_{j-2} - ud_{j-3} - vd_{j-4}.$$

qu'à nouveau nous explicitons complètement :

$$\begin{aligned}
d_0 &= -b_0 \\
d_1 &= -b_1 - ud_0 \\
d_2 &= -b_2 - ud_1 - vd_0 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
d_{n-4} &= -b_{n-4} - ud_{n-5} - vd_{n-6} \\
\frac{\partial A}{\partial v} &= -b_{n-3} - ud_{n-4} - vd_{n-5} \\
\frac{\partial B}{\partial v} &= -b_{n-2} - ud_{n-4}.
\end{aligned}$$

Remarquons immédiatement que les suites $c_j$ et $d_j$ sont identiques, il n'est donc pas nécessaire de calculer la suite $d_j$, et l'on obtiendra les dérivées partielles par rapport à $v$ à partir de la suite des $c_j$ :

$$\begin{aligned}
\frac{\partial A}{\partial v} &= -b_{n-3} - ud_{n-4} - vd_{n-5} = c_{n-3} \\
\frac{\partial B}{\partial v} &= -b_{n-2} - vc_{n-4}.
\end{aligned}$$

## 3.2. Conduite du calcul

Il nous faut donc calculer les fonctions $A$ et $B$ ainsi que leurs dérivées partielles. Connaissant une première approximation $(u_0, v_0)$ nous sommes donc en mesure de déterminer les termes

correctifs $du$ et $dv$ à chaque tour d'itération (on utilisera les relations de Newton par exemple). Bien sûr nous recommençons le même calcul à partir des nouvelles valeurs :

$$u_1 = u_0 + \mathrm{d}u$$
$$v_1 = v_0 + \mathrm{d}v$$

et ainsi de suite. Nous avons défini une procédure itérative que nous arrêterons en machine lorsque $\mathrm{d}u/u$ et $\mathrm{d}v/v$ seront inférieurs à la précision de la représentation en machine.

Regroupons la liste des opérations utiles (*cf.* Tab. 4.1.).

**Tableau 4.1.** *Liste des opérations utiles.*

| | | | | | |
|---|---|---|---|---|---|
| $b_0$ | $=$ | $a_0$ | $c_0$ | $=$ | $-b_0$ |
| $b_1$ | $=$ | $a_1 - ub_0$ | $c_1$ | $=$ | $-b_1 - uc_0$ |
| $b_2$ | $=$ | $a_2 - ub_1 - vb_0$ | $c_2$ | $=$ | $-b_2 - uc_1 - vc_0$ |
| ... | ... | ..... | ... | ... | ..... |
| $b_j$ | $=$ | $a_j - ub_{j-1} - vb_{j-2}$ | $c_j$ | $=$ | $-b_j - uc_{j-1} - vc_{j-2}$ |
| ... | ... | ..... | ... | ... | ..... |
| $b_{n-2}$ | $=$ | $a_{n-2} - ub_{n-3} - vb_{n-4}$ | $c_{n-3}$ | $=$ | $-b_{n-3} - uc_{n-4} - vc_{n-5}$ |
| $A$ | $=$ | $a_{n-1} - ub_{n-2} - vb_{n-3}$ | $\dfrac{\partial A}{\partial u}$ | $=$ | $-b_{n-2} - uc_{n-3} - vc_{n-4}$ |
| $B$ | $=$ | $a_n - vb_{n-2}$ | $\dfrac{\partial B}{\partial u}$ | $=$ | $-vc_{n-3}$ |
| | | | $\dfrac{\partial A}{\partial v}$ | $=$ | $c_{n-3}$ |
| | | | $\dfrac{\partial B}{\partial v}$ | $=$ | $-b_{n-2} - vc_{n-4}$ |

Comment choisir les valeurs de départ $u_0$ et $v_0$ ?

Une bonne méthode pour débuter les itérations consiste à choisir les valeurs suivantes :

$$u_0 = \frac{a_{n-1}}{a_{n-2}} \quad \text{et} \quad v_0 = \frac{a_n}{a_{n-2}} \cdot$$

On calcule la suite des $u_k$ et des $v_k$ jusqu'à obtenir la précision souhaitée. Pour réaliser ces opérations il ne faut surtout pas utiliser de tableaux pour stocker les $b_k$ et les $c_k$, car nous n'avons besoin que des trois dernières valeurs calculées. Il suffira alors de réaliser des décalages sur seulement trois variables. Cette remarque permet de gagner un temps considérable en calcul ; ensuite quand on a obtenu l'annulation du reste $Ax + B$, on calcule la suite des $b_k$ dans le même tableau que celui qui a servi à stocker les valeurs des $a_k$. Alors il est possible de pouvoir chercher deux nouvelles racines grâce à la même procédure. Dans le but de simplifier l'usage des relations de récurrence, on programme les relations générales données par $b_j$ et $c_j$, et l'on démarre les calculs avec :

$$b_{-1} = b_{-2} = 0 \quad \text{et} \quad c_{-1} = c_{k-2} = 0.$$

***Mise en œuvre de la méthode*** – On trouvera sur le Web(*) les programmes `bairstow.c` et `bairstow.h` réalisant l'algorithme de Bairstow. Toutefois ces programmes appellent plusieurs remarques :

1. Le programme peut être mis en défaut lorsque le polynôme présente des racines multiples et notamment des racines doubles car le jeu des erreurs peut nous faire basculer vers des valeurs négatives du discriminant et peut nous donner l'illusion de deux racines complexes conjuguées. Cela mérite évidemment un examen attentif. En fait l'erreur absolue sur le discriminant est de l'ordre de $10^{-16}$ et par conséquent la partie imaginaire sera de l'ordre de $10^{-8}$.
2. Hormis l'ennui précédemment cité, il n'y a pas lieu de procéder à une détermination locale préalable des racines qui peuvent du reste être complexes.
3. On pourrait mettre en œuvre une méthode qui utilise un binôme au lieu d'un trinôme, mais seules seraient calculées les racines réelles du polynôme à coefficients réels, ce qui ôte tout attrait à ce procédé. Cependant, **dans le cas où les coefficients du polynôme sont complexes**, la réécriture de la méthode de Bairstow appliquée à la division par un binôme donne une excellente manière d'opérer (voir les problèmes).

## 3.3. Racines d'un polynôme et valeurs propres d'une matrice

L'usage des mathématiques veut que la détermination des valeurs propres des matrices passe par l'intermédiaire de l'équation caractéristique dont on recherche ensuite les racines. L'équation caractéristique est en fait un polynôme dont les zéros sont les valeurs propres de la matrice. Ainsi, quand nous aurons obtenu l'équation caractéristique, la méthode de Bairstow nous donnera les valeurs propres de la matrice.

Ce n'est pas la meilleure technique pour calculer les valeurs propres d'une matrice et l'on étudiera d'autres méthodes directes qui nous permettront de les obtenir. Cependant, réciproquement, si l'on sait associer une matrice carrée d'ordre $n$ à un polynôme de degré $n$ qui est le polynôme caractéristique de ladite matrice, on pourra utiliser les méthodes directes de calcul des valeurs propres pour obtenir les zéros des polynômes.

Pour simplifier le problème sans pour autant nuire à la généralité, on s'intéressera aux polynômes à coefficient principal réduit que nous écrirons donc :

$$\Pi_n(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1}x + a_n = 0.$$

Il suffit alors d'associer la matrice suivante (une des formes canoniques de Frobenius (1849–1917)) :

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & -a_{n-3} & -a_{n-4} & & \dots & -a_1 \end{pmatrix}.$$

La première surdiagonale de cette matrice est constituée de 1 et la dernière ligne est formée par les opposés des coefficients du polynômes caractéristique ordonnés selon les puissances croissantes, partout ailleurs il y a des zéros. On montre que $(-1)^n \Pi_n(x)$ est l'équation caractéristique

* `http://www.edpsciences.com/guilpin/`

de cette matrice. Il est facile de former cette matrice puis de chercher ses valeurs propres par une méthode directe, on obtiendra alors les zéros de $\Pi_n(x)$.

On donne sur le Web(*) `froben.c` un programme qui forme la matrice $A$ à partir des coefficients du polynôme. Ceci appelle une remarque importante. Si la matrice est mal conditionnée, nous aurons beaucoup de difficultés à calculer des valeurs propres correctes. Le lien entre l'équation caractéristique et la matrice associée montre également que les coefficients des polynômes peuvent aussi être mal conditionnés et dans ce cas les algorithmes peuvent donner des résultats aberrants.

* `http://www.edpsciences.com/guilpin/`

# 5 | Éléments de calcul matriciel

Ce chapitre est constitué de deux rubriques consacrées l'une à la résolution de systèmes linéaires, l'autre au calcul des valeurs propres des matrices carrées. Nous allons voir comment résoudre ce premier problème selon deux méthodes générales, puis nous examinerons une solution propre aux « matrices » de Vandermonde rencontrée en interpolation. La résolution d'un système linéaire nous conduira directement au calcul d'un déterminant et à l'inversion des matrices (carrées). Pour terminer, nous aborderons quelques méthodes de calcul des valeurs propres des matrices. Auparavant, nous traiterons des opérations élémentaires sur les matrices.

Si l'on se réfère au calcul d'un déterminant d'ordre $n$ selon la méthode de Cramer (1704–1752), on voit que ce procédé exige environ $n!$ opérations ce qui devient rapidement irréalisable du point de vue pratique. Il nous faut utiliser des méthodes beaucoup plus économiques et celles que nous présentons utilisent grossièrement $n^3$ voire $n^4$ opérations. Du reste le dénombrement exact est facile à obtenir en modifiant très légèrement les programmes proposés : il suffit d'ajouter un compteur d'opérations.

Tout au long de ce chapitre, on note par des lettres majuscules les matrices et les vecteurs, et par des lettres minuscules (généralement indicées) les éléments des matrices et les composantes des vecteurs.

## 1. Multiplication de deux matrices

Soient deux matrices $A(n,m)$ et $B(m,l)$ expressions dans lesquelles le premier terme de la parenthèse représente le nombre de lignes et le second le nombre de colonnes. $C(n,l)$ le produit des matrices $A$ et $B$ s'écrit :

$$C(n,l) = A(n,m)B(m,l),$$

avec

$$c_{jk} = \sum_{i=1}^{m} a_{ji} b_{ik}.$$

Le nombre d'opérations est environ $2m \times n \times l$. On trouvera le sous-programme `multma.h` calculant le produit de deux matrices sur le Web(*).

* `http://www.edpsciences.com/guilpin/`

## 2. Résolution d'un système linéaire

On considère un système linéaire de $n$ équations à $n$ inconnues que l'on écrit sous une forme matricielle :

$$AX = B.$$

Les coefficients de la matrice $A$ (d'ordre $n$) sont notés selon l'habitude classique : $a_{lk}$ (dans l'ordre ligne colonne). $X$ (de composantes $x_j$) est le vecteur des inconnues et $B$ (de composantes $b_j$)le vecteur second membre. La notation du vecteur inconnu est sans intérêt dans la manière dont on résout le problème et seules les grandeurs que nous allons manipuler sont les $a_{lk}$ et les $b_j$.

### 2.1. Méthode des pivots

Nous allons traiter simultanément l'ensemble des données du problème et pour ce faire nous formons un tableau $T(A, B)$ en juxtaposant les éléments de la matrice $A$ et les composantes du vecteur $B$ selon l'expression ci-dessous :

$$T(A,B) = \left|\begin{array}{cccccc} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ a_{41} & a_{42} & a_{43} & \dots & a_{4n} & b_4 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right|$$

On voit que le tableau $T(A, B)$ représente une forme multilinéaire avec les propriétés suivantes :

**a.** La permutation de deux lignes ne change rien au tableau.
**b.** La multiplication d'une ligne par un nombre ne change rien au tableau.
**c.** L'addition terme à terme de deux lignes ne change rien au tableau,
**d.** La combinaison linéaire de deux lignes du tableau ne change rien au tableau.

Quand le système linéaire est soluble et résolu, le tableau $T(A, B)$ se présente sous la forme :

$$T(I,S) = \left|\begin{array}{cccccc} 1 & 0 & 0 & \dots & 0 & S_1 \\ 0 & 1 & 0 & \dots & 0 & S_2 \\ 0 & 0 & 1 & \dots & 0 & S_3 \\ 0 & 0 & 0 & \dots & 0 & S_4 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & S_n \end{array}\right|$$

où $I$ est la matrice unité d'ordre $n$ et $S$ le vecteur solution.

Comme $T(A, B) = T(I, S)$, nous allons voir comment passer du tableau $T(A, B)$ au tableau $T(I, S)$ par les opérations linéaires décrites en $a$, $b$, $c$ et $d$ au moyen de la méthode des pivots.

Nous allons transformer le tableau $T(A, B)$ de telle sorte que nous obtenions le vecteur colonne $(1, 0, 0, \dots, 0)$. Pour y parvenir il faut d'abord diviser la première ligne du tableau par l'élément $a_{11}$ que l'on suppose différent de zéro. La première ligne s'écrit alors :

$$1 \quad a_{12}^1 \quad a_{13}^1 \quad \dots \quad a_{1n}^1 \quad b_1^1.$$

Pour annuler les autres termes de la première colonne, il suffit d'opérer ainsi :

les éléments de la première ligne multipliés par $a_{21}$ sont retranchés des éléments correspondants de la deuxième ligne. Ensuite, d'une façon tout à fait générale, les éléments de la première ligne multipliés par $a_{k1}$ sont soustraits des éléments correspondants de la $k^{\text{e}}$ ligne. Donc, pour $k > 1$ :

$$a^1_{ki} = a_{ki} - a_{k1}a^1_{1i} \quad \text{et} \quad b^1_k = b_k - a_{k1}b^1_1.$$

Les éléments de la première colonne du tableau transformé équivalent sont nuls pour $k > 1$, seul le premier élément vaut 1. Après cette première phase, le tableau s'écrit :

$$T(A,B) = \begin{vmatrix} 1 & a^1_{12} & a^1_{13} & \dots & a^1_{1n} & b^1_1 \\ 0 & a^1_{22} & a^1_{23} & \dots & a^1_{2n} & b^1_2 \\ 0 & a^1_{32} & a^1_{33} & \dots & a^1_{3n} & b^1_3 \\ 0 & a^1_{42} & a^1_{43} & \dots & a^1_{4n} & b^1_4 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & a^1_{n2} & a^1_{n3} & \dots & a^1_{nn} & b^1_n \end{vmatrix}.$$

À présent, nous allons traiter le deuxième vecteur colonne de la matrice $A$. Nous divisons toute la deuxième ligne par l'élément $a^1_{22}$ à condition qu'il soit différent de zéro. Cette deuxième ligne devient :

$$0 \quad a^2_{23} \quad \dots \quad a^2_{2n} \quad b^2_2.$$

ensuite, les éléments de la deuxième ligne multipliés par $a^1_{k2}$ sont soustraits des éléments correspondants de la $k^{\text{e}}$ ligne, et cela pour $k \neq 2$. Ainsi on aura :

$$a^2_{ki} = a^1_{ki} - a^1_{k2}a^2_{2i} \quad \text{et} \quad b^2_k = b^1_k - a^1_{k2}b^2_2.$$

À présent la deuxième colonne du tableau a été transformée selon nos vœux.

Il n'y a pas de difficultés à poursuivre de la même façon les transformations de la troisième ligne puis des lignes suivantes, et ainsi de suite jusqu'à la $n^{\text{e}}$ ligne. À la fin du calcul, nous obtenons la configuration souhaitée :

$$T(I,b^n) = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 & b^n_1 \\ 0 & 1 & 0 & \dots & 0 & b^n_2 \\ 0 & 0 & 1 & \dots & 0 & b^n_3 \\ 0 & 0 & 0 & \dots & 0 & b^n_4 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & b^n_n \end{vmatrix}.$$

Les $b^n_k$ sont les solutions recherchées du système linéaire.

L'élément $a^m_{kk}$ qui se trouve sur la diagonale principale a vu sa valeur changer dynamiquement au cours des calculs tant que $m$ est inférieur à $k$. Quand $m = k$, on procède à la division de la ligne $k$ par cet élément qui porte le nom de **pivot** d'où le nom de la méthode.

Au cours des calculs, nous avons supposé que le pivot rencontré n'était jamais nul. **Qu'arrive-t-il si l'on rencontre un pivot nul?** De deux choses l'une, ou bien la matrice n'est pas inversible et son déterminant est nul, ou bien elle est inversible et son déterminant est différent de zéro.

Il nous est toujours possible de permuter la ligne $k$ avec une des lignes qui suivent, donc numérotées de $k+1$ à $n$. Il n'y aura d'intérêt à réaliser cette permutation que si au moins un

élément de la colonne $k$ est différent de zéro. Si tel est le cas, on échange les deux lignes, rien n'est changé dans le tableau $T$ et l'on poursuit les calculs. Si la permutation ne permet pas d'amener un pivot non nul, c'est que la matrice est singulière. Si, à partir de la $k^e$ ligne on ne peut plus permuter de lignes, la matrice est dégénérée d'ordre $n - k$. Seules $k$ lignes ou $k$ colonnes sont linéairement indépendantes.

***Remarque 1*** : La prudence la plus élémentaire exige que l'on reporte la solution trouvée dans les équations d'origine afin de déceler éventuellement un écart sensible avec le second membre.

***Remarque 2*** : Une matrice dégénérée possède une infinité de solutions. À bien y réfléchir, il est peu probable qu'un calcul de pivot amène la valeur zéro, tout simplement parce que le jeu du cumul des erreurs et des troncatures a peu de chances de réaliser un tel événement. Donc on risque fortement à un moment donné de diviser une ligne par un pivot qui aurait dû être nul sans qu'il le soit dans la réalité. Nous allons obtenir une solution parmi l'infinité de solutions possibles dans un tel cas.

**Comment prévenir une telle infortune?** Il est nécessaire de détecter un tel dysfonctionnement, et, on peut y parvenir en exécutant plusieurs fois les calculs sur des présentations différentes du tableau obtenues en intervertissant les lignes. Ainsi on fera apparaître plusieurs solutions différentes du même système linéaire et notre attention sera attirée par sa singularité.

Dans le cas où l'on ne rencontre pas un pivot nul, on peut évaluer de façon approchée le nombre d'opérations : $n^3$. Sur le Web[*], on trouvera le programme `systlin.c` qui réalise le calcul selon la méthode des pivots.

## 2.2. Méthode de la décomposition de la matrice A en un produit de deux matrices triangulaires

Nous nous proposons de décomposer la matrice $A$ en un produit de deux matrices triangulaires notées $D$ et $G$ où $D$ est une matrice triangulaire inférieure et $G$ une matrice triangulaire supérieure telles que $A = DG$, puis d'appliquer cette technique de décomposition à la résolution d'un système linéaire. Notons que ce type de décomposition sera encore employé pour inverser les matrices et pour calculer les valeurs propres.

La matrice $A$ étant d'ordre $n$ il en sera de même des matrices $D$ et $G$, et l'écriture de $A = DG$ conduit à écrire $n^2$ équations à $(n^2+n)$ inconnues. Nous sommes donc libre de choisir $n$ inconnues à notre convenance : choisissons d'écrire les éléments de la diagonale principale de $D$ égaux à l'unité. Les deux matrices $D$ et $G$ s'écrivent :

$$D = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ d_{21} & 1 & 0 & \dots & 0 \\ d_{31} & d_{32} & 1 & \dots & 0 \\ d_{41} & d_{42} & d_{43} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & d_{n3} & \dots & 1 \end{vmatrix}$$

$$G = \begin{vmatrix} g_{11} & g_{12} & g_{13} & \cdots & g_{1n} \\ 0 & g_{22} & g_{23} & \cdots & g_{2n} \\ 0 & 0 & g_{33} & \cdots & g_{3n} \\ 0 & 0 & 0 & \dots & g_{4n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & g_{nn} \end{vmatrix}.$$

* `http://www.edpsciences.com/guilpin/`

Par identification, on trouve :

1. la première ligne de $G$ :

$$g_{11} = a_{11} \qquad g_{12} = a_{12} \qquad \ldots \qquad g_{1n} = a_{1n}$$

2. la deuxième ligne de $D$, soit l'élément $d_{21}$ qui s'obtient à partir de :

$$a_{21} = d_{21} g_{11}$$

3. la deuxième ligne de $G$ s'obtient au moyen des relations :

$$a_{22} = d_{21} g_{12} + g_{22} \qquad a_{23} = d_{21} g_{13} + g_{23} \qquad \ldots \qquad a_{2n} = d_{21} g_{1n} + g_{2n}$$

4. la troisième ligne de $D$ est donnée par :

$$a_{31} = d_{31} g_{11} \qquad a_{32} = d_{31} g_{12} + d_{32} g_{22}$$

5. la troisième ligne de $G$ est obtenue au moyen de :

$$a_{33} = d_{31} g_{13} + d_{32} g_{23} \qquad \ldots \qquad a_{3n} = d_{31} g_{1n} + d_{32} g_{2n} + g_{3n}$$

6. On poursuit en calculant alternativement la ligne suivante de $D$ puis la même ligne de $G$ jusqu'à la ligne $n$.

On trouvera sur le Web(*) le programme `triangle.c` qui réalise une telle décomposition.

### 2.3. Application à la résolution du système linéaire AX = B

Nous pouvons donc écrire :

$$AX = DGX = B$$

il est possible de poser $Z = GX$, et donc d'obtenir aisément le vecteur $Z$ dont les composantes sont données par le système d'équations $DZ = B$ qui se décompose de la manière suivante :

$$y_1 = d_1 \quad d_{21} y_1 + y_2 = d_2 \quad d_{31} y_1 + d_{32} y_2 + y_3 = d_3 \quad \ldots \quad d_{n1} y_1 + d_{n2} y_2 + \cdots + y_n = d_n.$$

Cet ensemble d'équations permet de déterminer les composantes de $Z$. Connaissant $Z$, il est aisé de calculer le vecteur $X$ au moyen de l'équation $Z = GX$ dont les composantes donnent :

$$\begin{aligned}
&g_{nn} x_n = z_n \\
&g_{n-1n} x_n + g_{n-1n-1} x_{n-1} = z_{n-1} \\
&\ldots\ldots\ldots\ldots \\
&g_{1n} x_n + g_{1n-1} x_{n-1} + \cdots + g_{12} x_2 + g_{12} x_2 + g_{11} x_1 = z_1.
\end{aligned}$$

Donc à partir des récurrences ainsi définies nous sommes en mesure de résoudre un système linéaire.

***Remarque*** : Si la matrice est singulière, elle ne peut pas être décomposée en un produit de deux matrices triangulaires.

On trouvera sur le Web(*) le programme `trianlin.c` mettant en œuvre cet algorithme.

* `http://www.edpsciences.com/guilpin/`

## 2.4. Calcul d'un déterminant

Après la présentation des algorithmes de résolution de systèmes linéaires, il résulte que le déterminant $\Delta$ d'une matrice carrée $W$ est égal au produit des pivots multiplié par $(-1)^p$, $p$ étant le nombre de permutations de lignes effectuées au cours du calcul. Il est aussi égal au produit des éléments de la diagonale principale de la matrice triangulaire $G$.

***Remarque à propos de la méthode des pivots*** – On rencontre souvent l'idée selon laquelle la méthode des pivots peut être améliorée si, à chaque fois que l'on traite une ligne, on amène celle qui possède le plus grand pivot en valeur absolue.

Cette façon de voir est complètement erronée pour la raison qui suit. Le calcul propage les erreurs cumulées sur chacun des pivots obtenus avec une ultime soustraction (susceptible de fournir une redoutable erreur relative). Le produit des pivots $\pi$ étant constant et égal au déterminant $\Delta$, le fait d'utiliser les grands pivots au début du calcul impose les petits pivots à la fin... Donc le déterminant, le système linéaire etc. verront croître les erreurs au fur et à mesure que se déroule le calcul, et cela d'une manière plus rapide que la simple proportionnalité au nombre des opérations arithmétiques réalisées.

• ***Que faut-il faire alors ?*** – Comme nous avons : $\Delta = \prod_{k=0}^{n} p_k$, l'erreur relative $\delta\Delta$ sur $\Delta$ est donnée par l'expression :

$$\left|\frac{\delta\Delta}{\Delta}\right| \leq \sum_{k=0}^{n}\left|\frac{\delta_{pi}}{p_k}\right| = |\delta_e| \sum_{k=0}^{n}\left|\frac{1}{p_k}\right|$$

car les erreurs absolues $\delta p_i$ sont *grosso modo* les mêmes. On peut encore écrire :

$$\left|\frac{\delta\Delta}{\Delta}\right| \leq \left|\frac{\delta_e}{\Delta}\right| \sum_{k=0}^{n} |p_k| .$$

Le membre de gauche sera majoré par la somme des modules des pivots (à une constante multiplicative près). C'est lorsque les pivots sont égaux en module que la somme des modules est minimum puisque le produit des pivots doit être constant.

En conséquence, la seule méthode raisonnable consiste à choisir le pivot qui est toujours le plus proche en module de $\sqrt[n]{|\Delta|}$. Pour cela il faut obtenir une première valeur de $\Delta$ et procéder par itérations jusqu'à ce que deux valeurs consécutives du déterminant soient égales à la précision de la machine.

## 2.5. A est une matrice de Vandermonde (1735–1796)

Nous rencontrerons à nouveau ce type de problème à propos du polynôme d'interpolation de Lagrange ; on cherche à calculer les coefficients du polynôme de degré $n$ qui passe par les $(n+1)$ points d'un échantillon $(\alpha_j, \beta_j)$ avec $j = 0,1,2,\ldots,n$. La seule restriction consiste à supposer que les abscisses $\alpha_j$ sont toutes différentes les unes des autres.

Le polynôme s'écrit sous la forme :

$$P_n(x) = \sum_{k=O}^{n} a_k x^k.$$

Nous obtenons ainsi le système linéaire suivant :

$$\begin{vmatrix} 1 & \alpha_0^1 & \alpha_0^2 & \alpha_0^3 & \dots & \alpha_0^n \\ 1 & \alpha_1^1 & \alpha_1^2 & \alpha_1^3 & \dots & \alpha_1^n \\ 1 & \alpha_2^1 & \alpha_2^2 & \alpha_2^3 & \dots & \alpha_2^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \alpha_n^1 & \alpha_n^2 & \alpha_n^3 & \dots & \alpha_n^n \end{vmatrix} \cdot \begin{vmatrix} a_0 \\ a_1 \\ a_2 \\ \\ a_n \end{vmatrix} = \begin{vmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \\ \beta_n \end{vmatrix}$$

On se propose de résoudre ce système d'ordre $n+1$ au moyen des déterminants de Vandermonde dont nous rappelons la propriété essentielle suivante :

$$\begin{vmatrix} 1 & \alpha_0^1 & \alpha_0^2 & \alpha_0^3 & \dots & \alpha_0^n \\ 1 & \alpha_1^1 & \alpha_1^2 & \alpha_1^3 & \dots & \alpha_1^n \\ 1 & \alpha_2^1 & \alpha_2^2 & \alpha_2^3 & \dots & \alpha_2^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \alpha_n^1 & \alpha_n^2 & \alpha_n^3 & \dots & \alpha_n^n \end{vmatrix} = \prod_{i=0}^{n} \left\{ \prod_{j>i}^{n} [\alpha_j - \alpha_i] \right\} = \Delta^{(n+1)}(\alpha).$$

Ici, l'ordre du déterminant de Vandermonde est $n+1$, d'où la notation $\Delta^{(n+1)}(\alpha)$. On exprime alors $a_0$ sous la forme du rapport classique de deux déterminants que l'on note :

$$a_0 = \frac{\Delta_{a_0}}{\Delta^{(n+1)}(\alpha)},$$

$$\text{avec } \Delta_{a_0} = \begin{vmatrix} \beta_0 & \alpha_0^1 & \alpha_0^2 & \alpha_0^3 & \dots & \alpha_0^n \\ \beta_1 & \alpha_1^1 & \alpha_1^2 & \alpha_1^3 & \dots & \alpha_1^n \\ \beta_2 & \alpha_2^1 & \alpha_2^2 & \alpha_2^3 & \dots & \alpha_2^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \beta_n & \alpha_n^1 & \alpha_n^2 & \alpha_n^3 & \dots & \alpha_n^n \end{vmatrix}$$

Divisons la première ligne de $\Delta_{a_0}$ par $\alpha_0^1$, puis la deuxième ligne par $\alpha_1^1$ et ainsi de suite, nous obtenons :

$$\Delta_{a_0} = \prod_{k=0}^{n} \alpha_k \begin{vmatrix} \frac{\beta_0}{\alpha_0} & 1 & \alpha_0^1 & \alpha_0^2 & \alpha_0^3 & \dots & \alpha_0^{n-1} \\ \frac{\beta_1}{\alpha_1} & 1 & \alpha_1^1 & \alpha_1^2 & \alpha_1^3 & \dots & \alpha_1^{n-1} \\ \frac{\beta_2}{\alpha_2} & 1 & \alpha_2^1 & \alpha_2^2 & \alpha_2^3 & \dots & \alpha_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{\beta_n}{\alpha_n} & 1 & \alpha_n^1 & \alpha_n^2 & \alpha_n^3 & \dots & \alpha_n^{n-1} \end{vmatrix}$$

En développant ce déterminant par rapport à la première colonne, on voit que $\Delta_{a_0}$ est une combinaison linéaire de déterminants de Vandermonde d'ordre $n$ notés :

$$\Delta_i^n(\alpha) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \left[ \prod_{\substack{j>k \\ j \neq i}}^{n} (\alpha_j - \alpha_k) \right],$$

soit encore :

$$\Delta_{a_0} = \left(\prod_{k=0}^{n} \alpha_k\right) \cdot \left(\sum_{j=0}^{n} (-1)^j \frac{\beta_j}{\alpha_j} \Delta_j^n(\alpha)\right).$$

À présent, on sait calculer $a_0$ ; on peut donc reporter sa valeur dans les équations de départ et faire disparaître la première ligne qui devient inutile puisque nous connaissons effectivement $a_0$. Le calcul de $a_1, a_2, \ldots, a_n$ se ramène à la résolution du système linéaire d'ordre $n$ suivant :

$$\begin{vmatrix} 1 & \alpha_1^1 & \ldots & \alpha_1^{n-1} \\ 1 & \alpha_2^1 & \ldots & \alpha_2^{n-1} \\ \ldots\ldots & & & \\ 1 & \alpha_n^1 & \ldots & \alpha_n^{n-1} \end{vmatrix} \cdot \begin{vmatrix} a_1 \\ a_2 \\ \\ a_n \end{vmatrix} = \frac{1}{C} \begin{vmatrix} \beta_1 - a_0 \\ \beta_2 - a_0 \\ \\ \beta_n - a_0 \end{vmatrix} \qquad \text{avec } C = \prod_{i=1}^{n} \alpha_i.$$

L'opération qui consiste à calculer une valeur $a_j$ (d'abord $a_0$) puis à éliminer une ligne s'appelle **déflation**. À présent, nous sommes en mesure de calculer $a_1$ par le même procédé, toutefois, l'ordre de la matrice a baissé d'une unité, $a_0$ ayant disparu. En poursuivant les opérations, on calcule de proche en proche toutes les valeurs $a_k$.

***Cas où un des $\alpha_i$ est nul*** – L'algorithme semble tomber en défaut si l'un des $\alpha_i$ est nul. Il n'en est rien. En effet, si la valeur $\alpha_j$ est nulle, c'est la seule par hypothèse, donc on peut permuter la ligne $j$ avec la ligne zéro sans que rien ne soit changé. Dans ces conditions, on connaît immédiatement la valeur $a_0 = \beta_0$ puisque la première ligne de la matrice ne contient que des zéros à l'exception du premier élément qui vaut un. Après déflation, nous revenons au problème précédemment traité.

Nous donnons sur le Web (*) le programme `vdmonde.c` qui réalise ces opérations. Remarquons que ce programme utilise très peu de place mémoire : seuls les vecteurs y sont utilisés.

## 2.6. A est une matrice de Hilbert (1862–1943)

Une matrice de Hilbert $A$ d'ordre $n$ est une matrice symétrique dont les éléments $a_{lk}$ sont donnés par l'expression suivante :

$$a_{lk} = \frac{1}{l+k+1} \quad \text{avec } l, k = 0, 1, 2, \ldots, n-1.$$

C'est une matrice bien connue pour son mauvais conditionnement. Pour s'en convaincre, il suffit de choisir le vecteur inconnu $X$ et de calculer le second membre $B$. Prenons pour composantes du vecteur $X$ d'ordre $n$ les entiers successifs de 1 à $n$, soit $x_0 = 1, x_1 = 2, \ldots, x_{n-1} = n$. Le produit $AX$ donne le second membre $B$. À partir de $A$ et de $B$ proposons-nous de déterminer $X$ pour $n = 5, 10, 15$ et 20. Rapidement les résultats deviennent aberrants. C'est la raison pour laquelle on dit que la matrice $A$ est mal conditionnée. Il est facile de calculer chaque fois le déterminant de la matrice $A$ qui tend rapidement vers zéro avec l'ordre $n$. Pour $n = 15$, on trouve un déterminant de l'ordre de $10^{-78}$... on comprend ainsi où se situent les pertes de signification.

* `http://www.edpsciences.com/guilpin/`

On donne sur le Web(*) le programme `hilbert.c` qui fournit les résultats de ces calculs.

## 2.7. A est une matrice creuse

On considère à présent un système linéaire $AX = D$ dont la matrice $A$ d'ordre $N$ est constituée de trois bandes tridiagonales de sous-matrices carrées d'ordre $n$ et notées $A_k$, $B_k$ et $C_k$ comme le montre la figure ci-dessous :

$$\begin{vmatrix} B_1 & C_1 & & & & \\ A_2 & B_2 & C_2 & & & \\ & A_3 & B_3 & C_3 & & \\ \ldots\ldots & & \ldots\ldots & & \ldots\ldots & \\ & & & & C_{m-1} & \\ & & & A_m & B_m & \end{vmatrix} \cdot \begin{vmatrix} X_1 \\ X_2 \\ X_3 \\ \ldots\ldots \\ X_{m-1} \\ X_m \end{vmatrix} = \begin{vmatrix} D_1 \\ D_2 \\ D_3 \\ \ldots\ldots \\ D_{m-1} \\ D_m \end{vmatrix}$$

le vecteur inconnu $X$ et le vecteur second membre $D$ étant partitionnés en sous-vecteurs de $n$ composantes. On dit que la matrice $A$ est creuse. Comme l'inverse d'une matrice creuse est une matrice pleine, on va mettre en œuvre une méthode spécifique qui n'est rien d'autre que la généralisation de la méthode exposée lors de l'étude des fonctions-spline (*cf.* chapitre 6). Au moyen d'opérations linéaires, le but est de faire disparaître les matrices $A_k$ et de remplacer les matrices $B_k$ par la matrice unitéd'ordre $n$ notée $I_k$ ; on obtient alors la disposition suivante :

$$\begin{vmatrix} I_1 & W_1 & & & & \\ & I_2 & W_2 & & & \\ & & I_3 & W_3 & & \\ \ldots\ldots & & & \ldots\ldots & & \\ & & & & W_{m-1} & \\ & & & & I_m & \end{vmatrix} \cdot \begin{vmatrix} X_1 \\ X_2 \\ X_3 \\ \ldots\ldots \\ X_{m-1} \\ X_m \end{vmatrix} = \begin{vmatrix} G_1 \\ G_2 \\ G_3 \\ \ldots\ldots \\ G_{m-1} \\ G_m \end{vmatrix}$$

On obtient aisément les expressions suivantes :

$$\begin{aligned} W_1 &= B_1^{-1} C_1 \\ W_2 &= [B_2 - A_2 W_1]^{-1} C_2 \\ &\ldots\ldots \\ W_k &= [B_k - A_k W_{k-1}]^{-1} C_k \qquad \text{pour } k = 2, 3, \ldots m-1, \\ G_1 &= B_1^{-1} D_1 \\ G_2 &= [B_2 - A_2 W_1]^{-1} [D_2 - A_2 G_1] \\ &\ldots\ldots \\ G_k &= [B_k - A_k W_{k-1}]^{-1} [D_k - A_k G_{k-1}] \\ &\ldots\ldots \\ G_m &= [B_m - A_m W_{m-1}]^{-1} [D_m - A_m G_{m-1}] . \end{aligned}$$

On calcule alors facilement $X_m = G_m$, puis en remontant on calcule par récurrence :

$$X_k = G_k - W_k X_{k+1} \text{ pour } k = m-1, m-2, \ldots 1.$$

* `http://www.edpsciences.com/guilpin/`

# 3. Inversion d'une matrice carrée d'ordre n

## 3.1. Méthode des pivots

Soit une matrice carrée $A$ d'ordre $n$ que l'on suppose inversible. Le problème consiste à déterminer la matrice $A^{-1}$ telle que :

$$AA^{-1} = A^{-1}A = I$$

expression dans laquelle $I$ est la matrice unité d'ordre $n$.

Dans un premier temps, portons notre attention sur le système linéaire suivant :

$$AX = E_i$$

où $E_i$ est le vecteur colonne unité qui possède la valeur zéro partout sauf sur la ligne $i$ où la valeur est un. Nous pouvons écrire :

$$X = A^{-1}E_i = A_i^{-1}$$

autrement dit, la résolution de ce système linéaire fournit la $i^{e}$ colonne de la matrice inverse $A^{-1}$.

Nous savons résoudre ces $n$ systèmes linéaires, mais il est plus habile de les résoudre simultanément car, sinon, nous effectuerions inutilement $n$ fois la transformation de la matrice. Au lieu d'accoler à la matrice $A$ le vecteur second membre, il suffit de lui accoler la matrice unité (les $n$ vecteurs unité), puis de procéder à la même transformation que celle qui a été proposée pour résoudre un système linéaire, à la différence près que l'on traitera $n$ seconds membres au lieu d'un seul.

Sur le Web(*), il est proposé le programme `matinv.c` qui inverse une matrice carrée.

## 3.2. Méthode par triangularisation

L'inverse d'une matrice triangulaire supérieure est une matrice triangulaire supérieure, et l'inverse d'une matrice triangulaire inférieure est une matrice triangulaire inférieure. Commençons par traiter la matrice $D$ dont la matrice inverse est $\Delta$. Nous avons :

$$\Delta D = I$$

qui fournit les équations :

$$\delta_{kk}d_{kk} = 1$$

et, pour $l > k$

$$\delta_{lk} = -\frac{\delta_{l,k+1}d_{l+1,k} + \delta_{l,k+2}d_{l+2,k} + \cdots + \delta_{ll}d_{lk}}{d_{kk}}$$

soit encore :

$$\delta_{lk} = -(\delta_{l,k+1}d_{l+1,k} + \delta_{l,k+2}d_{l+2,k} + \cdots + \delta_{ll}d_{lk})\delta_{kk}.$$

* `http://www.edpsciences.com/guilpin/`

Après avoir calculé les $d_{kk}$, on calcule ligne à ligne à partir de la dernière, et dans chaque ligne les éléments colonne par colonne à partir de la dernière colonne.

On procède d'une façon tout à fait identique pour la matrice $G$ dont la matrice inverse est $\Gamma$. Nous écrivons :

$$\Gamma G = I$$

qui fournit les équations :

$$\gamma_{kk} g_{kk} = 1$$

et, pour $l < k$

$$\gamma_{lk} = -\frac{\gamma_{ll} g_{lk} + \gamma_{l,l+1} g_{l+1,k} + \gamma_{l,l+2} g_{l+2,k} + \cdots}{g_{kk}}$$

soit encore :

$$\gamma_{lk} = -(\gamma_{ll} g_{lk} + \gamma_{l,l+1} g_{l+1,k} + \gamma_{l,l+2} g_{l+2,k} + \cdots)\gamma_{kk}.$$

Après avoir calculé les $\gamma_{kk}$, on calcule ligne à ligne à partir de la première, et dans chaque ligne les éléments colonne par colonne à partir de la première colonne.

Sur le Web (*), on trouvera le programme `trianinv.c` qui inverse une matrice carrée au moyen de cette procédure.

***Remarque*** : Quelle que soit la méthode utilisée, il est prudent d'effectuer toujours les vérifications usuelles. Il sera bien venu de procéder ou bien à la multiplication de $A$ et de $A^{-1}$ qui doit redonner — aux erreurs près — la matrice unité, ou bien à l'inversion de la matrice inverse qui doit fournir la matrice de départ — toujours aux erreurs près.

## 4. Calcul des valeurs propres

Il s'agit de calculer les valeurs propres de matrices carrées d'ordre $n$ à coefficients réels ; cependant, les méthodes proposées peuvent s'appliquer aux matrices à éléments complexes. Nous allons examiner diverses méthodes qui donnent des résultats plus ou moins intéressants selon la taille de la matrice $A$ ou son conditionnement. Le problème est donc la résolution de l'équation caractéristique ou séculaire :

$$AX = \lambda X$$

où $X$ s'appelle vecteur propre et $\lambda$ valeur propre.

### 4.1. La méthode de Le Verrier (1811–1877)

L'équation caractéristique d'une matrice d'ordre $n$ est un polynôme de degré $n$, et les racines de ce polynôme sont les valeurs propres de la matrice. Si nous pouvons obtenir les coefficients du polynôme caractéristique, la méthode de Bairstow nous permettra d'en calculer les racines. La solution repose sur les relations de Newton qui établissent une expression entre les coefficients d'un polynôme et les $S^q$ (somme des racines du polynôme élevées à la puissance $q$).

* `http://www.edpsciences.com/guilpin/`

***Relations de Newton*** – Soit $P_n(x)$ un polynôme de degré $n$ que nous écrivons :

$$P_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_n, \text{ avec } a \neq 0.$$

Désignons par $\rho_i$ les racines réelles ou complexes de ce polynôme qui se met alors sous la forme :

$$P_n(x) = a_0 \prod_{i=1}^{n} (x - \rho_i).$$

À présent, nous allons dériver (par rapport à $x$) de deux façons différentes $P_n(x)$ et identifier les résultats obtenus. Nous obtenons :

$$P'_{n-1}(x) = P_n(x) \sum_{i=1}^{n} \frac{1}{x - \rho_i}$$

expression dans laquelle nous développons $\frac{1}{x-\rho_i}$ soit :

$$\frac{1}{x - \rho_i} = \frac{1}{x}\left(1 + \left(\frac{\rho_i}{x}\right) + \left(\frac{\rho_i}{x}\right)^2 + \left(\frac{\rho_i}{x}\right)^3 + \cdots + \left(\frac{\rho_i}{x}\right)^p + \cdots\right)$$

d'où nous déduisons :

$$\sum_{i=1}^{n} \frac{1}{x - \rho_i} = \frac{1}{x}\left(n + \sum_{i=1}^{n}\left(\frac{\rho_i}{x}\right) + \sum_{i=1}^{n}\left(\frac{\rho_i}{x}\right)^2 + \sum_{i=1}^{n}\left(\frac{\rho_i}{x}\right)^3 + \cdots + \sum_{i=1}^{n}\left(\frac{\rho_i}{x}\right)^p + \cdots\right)$$

comme :

$$\sum_{i=1}^{n} \rho_i = \text{ somme des racines,}$$

$$\sum_{i=1}^{n} \rho_i^2 = \text{ somme des carrés des racines,}$$

on peut donc poser :

$$S^q = \sum_{i=1}^{n} \rho_i^q \text{ avec } q = 0, 1, 2, \ldots \text{( en remarquant que } S^0 = n).$$

Nous obtenons alors :

$$\sum_{i=1}^{n} \frac{1}{x - \rho_i} = \frac{1}{x}\left(S^0 + \frac{S^1}{x} + \frac{S^2}{x^2} + \frac{S^3}{x^3} + \cdots + \frac{S^p}{x^p} + \cdots\right)$$

Par ailleurs la dérivation directe nous donne :

$$P'_{n-1}(x) = a_0nx^{n-1} + a_1(n-1)x^{n-2} + a_2x^{n-2} + \cdots + a_{n-1}$$

ce qui permet d'écrire l'identité :

$$a_0nx^{n-1} + a_1(n-1)x^{n-2} + a_2x^{n-2} + \cdots + a_{n-1}$$
$$= (a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_n)\frac{1}{x}(S^0 + \frac{S^1}{x} + \frac{S^2}{x^2} + \frac{S^3}{x^3} + \cdots + \frac{S^p}{x^p} + \cdots).$$

Il reste à identifier les coefficients des termes de même degré en $x$, ce qui nous fournit les relations de Newton :

$$\begin{aligned}
&a_0S^1 + a_1 = 0\\
&a_0S^2 + a_1S^1 + 2a_2 = 0\\
&a_0S^3 + a_1S^2 + a_2S^1 + 3a_3 = 0\\
&\cdots\cdots\cdots\cdots\\
&a_0S^q + a_1S^{q-1} + \cdots + a_{q-1}S^1 + qa_3 = 0\\
&\cdots\cdots\cdots\cdots\\
&a_0S^n + a_1S^{n-1} + \cdots + a_{n-1}S^1 + na_n = 0
\end{aligned}$$

pour $p \geq n$, nous obtenons :

$$a_0S^p + a_1S^{p-1} + \cdots + a_{n-1}S^{p+1-n} + a_nS^{p-n} = 0.$$

Le problème qu'il reste à résoudre consiste à calculer effectivement les valeurs $S^q$ qui, ici, seront les sommes des valeurs propres à la puissance $q$, pour $q = 1, 2, \ldots, n$. Revenons à l'équation aux valeurs propres :

$$AX = \lambda X$$

et multiplions successivement $(n-1)$ fois à gauche les deux membres de cette équation par $A$ :

$$A^2X = A\lambda X = \lambda AX = \lambda^2 X$$

et ainsi de suite, on obtient d'une façon générale :

$$A^qX = \lambda^q X \text{ avec } q = 1, 2, 3, \ldots, n \ldots$$

Nous savons que la trace de la matrice $A$ est égale à la somme de ses valeurs propres ; de même, la trace de la matrice $A^q$ est égale à la somme des valeurs propres de $A$ élevées à la puissance $q$. Par conséquent, la trace de la matrice $A^q$ est égale à $S^q$. (On rappelle que la trace d'une matrice est la somme de ses éléments sur la diagonale principale.) Il suffit de calculer la trace de $A$ puis de multiplier $A$ par $A$ soit $B = AA$ et de calculer la trace de $B$. Ensuite, on multipliera $B$ par $A$ puis on calculera la trace de $B$, et ainsi de suite jusqu'à obtenir la valeur de $S^n$.

On trouvera sur le Web(*) le programme `leverier.c` qui calcule les valeurs propres (réelles ou complexes) de matrices réelles par la méthode de Le Verrier.

## 4.2. Méthode de Rutishauser (1918–1970)

Nous décomposons la matrice $A$ dont on cherche les valeurs propres en un produit de deux matrices triangulaires l'une, $D_1$, triangulaire inférieure et l'autre, $G_1$, triangulaire supérieure telles que $A = D_1G_1$.

On calcule la matrice $B_1$ qui est le produit de $G_1$ et $D_1$ dans l'ordre, soit :

$$B_1 = G_1D_1.$$

* `http://www.edpsciences.com/guilpin/`

On décompose $B_1$ en un produit de deux matrices triangulaires $D_2$ et $G_2$ telles que :

$$B_2 = D_2 G_2,$$

puis on calcule le produit de $G_2$ et $D_2$ que l'on appelle $B_2$, on poursuit ce type de décomposition et l'on obtient en définitive les résultats suivants :

$$\begin{aligned}
B_0 &= A = D_1 G_1 \\
B_1 &= G_1 D_1 = D_2 G_2 \\
B_2 &= G_2 D_2 = D_3 G_3 \\
&\cdots\cdots\cdots\cdots \\
B_n &= G_n D_n = D_{n+1} G_{n+1} \\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

Lorsque la décomposition triangulaire est possible, on montre que la matrice $B_n$ a les mêmes valeurs propres que la matrice $A$.

En général, $B_n$ tend vers une matrice triangulaire et les valeurs propres se situent alors sur la diagonale. De plus si celles-ci sont réelles et distinctes, elles apparaissent dans l'ordre des modules décroissants en descendant le long de la diagonale principale.

On trouvera sur le Web(*) le programme `rutishau.c` qui calcule les valeurs propres selon la méthode de Rutishauser.

Sans entrer dans les détails liés aux valeurs propres multiples, aux valeurs propres complexes, etc., il est utile de mentionner que cette méthode, dans sa version accélérée, sert de fondement à l'établissement de la méthode de Givens.

On trouvera sur le Web(*) le programme `rutishac.c` qui calcule les valeurs propres selon la méthode accélérée de Rutishauser.

## 4.3. Méthode de Givens (1912– )

La méthode de Givens consiste à réduire la matrice $A$ en une forme quasi triangulaire au moyen d'une suite de **transformations unitaires**, puis d'appliquer une des méthodes de Rutishauser à convergence quadratique. Nous allons exposer la suite de ces opérations.

Désignons par $A$ la matrice dont on cherche les valeurs propres et par $a_{ij}$ ses éléments. Pour réaliser la transformation unitaire notée $U$, $(U^T)^* U = I$, on pose :

$$\rho = \frac{a_{p-1,q}}{a_{p-1,p}} \qquad r = \frac{1}{\sqrt{1+\rho\rho^*}} \qquad s = \frac{\rho}{\sqrt{1+\rho\rho^*}},$$

l'astérisque désignant la valeur conjuguée et $U^T$ la transposée de $U$.

Les colonnes $p$ et $q$ sont transformées selon les expressions :

$$\begin{aligned}
b_{p-1,p} &= a_{p-1,p}\sqrt{1+\rho\rho^*} \qquad b_{p-1,q} = 0 \\
b_{kp} &= r\, a_{kp} + s^*\, a_{kq} \\
b_{kq} &= -s\, a_{kp} + r\, a_{kq} \qquad \text{avec } k = p, \ldots, n.
\end{aligned}$$

* `http://www.edpsciences.com/guilpin/`

Ensuite, les quatre valeurs $b_{qq}$, $b_{pp}$, $b_{qp}$ et $b_{pq}$ sont remplacées par les quantités :

$$c_{pp} = r\,b_{pp} + s\,b_{qp} \qquad c_{qq} = -s^*\,b_{pq} + r\,b_{qq}$$
$$c_{pq} = r\,b_{pq} + s\,b_{qq} \qquad c_{qp} = c_{pq}^*.$$

Les lignes $p$ et $q$ se transforment par des expressions analogues :

$$c_{pl} = r\,b_{pl} + s\,b_{ql} \qquad c_{ql} = -s^*\,b_{pl} + r\,b_{ql} \qquad \text{avec } l = 1,2,\ldots,n.$$

Toutes ces expressions viennent se substituer aux éléments correspondants dans la matrice $A$.

Dans le cas où $a_{p-1,p} = 0$, alors on prend $b_{p-1,p} = a_{p-1,q}$ puis $r = 0$ et $s = 1$, ce qui correspond à une rotation de $\pi/2$.

Cette méthode nous conduit donc à obtenir une matrice quasi triangulaire inférieure. Une telle matrice se décompose aisément selon la méthode de Rutishauser en une matrice triangulaire inférieure $D$ et une matrice triangulaire supérieure $G$.

Pour des raisons de commodité d'écriture, on note la matrice $A$ au moyen des $a_{ij}$ pour la partie triangulaire et $\alpha_1$ pour la diagonale située immédiatement au-dessus de la diagonale principale. On écrit donc :

$$A = \begin{vmatrix} a_{11} & \alpha_2 & 0 & \ldots & 0 \\ a_{21} & a_{22} & \alpha_3 & \ldots & 0 \\ a_{31} & a_{32} & a_{33} & \ldots & 0 \\ a_{41} & a_{42} & a_{43} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & \ldots & \alpha_n \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{vmatrix}$$

$$D = \begin{vmatrix} b_{11} & 0 & 0 & \ldots & 0 \\ b_{21} & b_{22} & 0 & \ldots & 0 \\ b_{31} & b_{32} & b_{33} & \ldots & 0 \\ b_{41} & b_{42} & b_{43} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{n-1,1} & b_{n-1,2} & b_{n-1,3} & \ldots & 0 \\ b_{n1} & b_{n2} & b_{n3} & \ldots & b_{nn} \end{vmatrix}$$

$$G = \begin{vmatrix} 1 & \beta_2 & 0 & \ldots & 0 \\ 0 & 1 & \beta_3 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \beta_n \\ 0 & 0 & 0 & \ldots & 1 \end{vmatrix}$$

Les $b_{ij}$ et les $\beta_j$ sont donnés par les expressions :

$$\beta_q = \frac{\alpha_q}{b_{q-1,q-1}} \qquad \text{pour } q = 2,\ldots,n,$$
$$b_{pq} = a_{pq} - \beta_q\,b_{p,q-1} \qquad \text{pour } p = 1,\ldots,n,$$

lesquelles viennent remplacer les valeurs correspondantes dans le tableau des $a_{ij}$.

Le calcul du produit $C = G \cdot D$ conduit aux nouvelles expressions :

$$\begin{aligned} c_{p1} &= b_{p1} + \beta_{p+1}\, b_{p+1,1} \qquad \text{pour } p = 1, \ldots, n-1, \\ c_{n1} &= b_{n1} \end{aligned}$$

puis en notant par $\gamma_k$ les éléments de $C$ au-dessus de la diagonale principale :

$$\begin{aligned} \gamma_q &= \beta_q\, b_{qq} \\ c_{pq} &= b_{pq} + \beta_{p+1}\, b_{p+1,q} \qquad \text{pour } p = q, \ldots, n-1, \\ c_{nq} &= b_{nq}. \end{aligned}$$

À la fin de ce premier tour de calcul, on note $c_{nn}^{(1)}$ la dernière valeur figurant sur la diagonale de la matrice $C$. On retranche $c_{nn}^{(1)}$ de tous les éléments diagonaux et l'on réitère la procédure jusqu'à ce que la valeur

$$\lambda_k = \sum_{i=1}^{k} c_{nn}^{(i)}$$

ne soit plus modifiée par un tour d'itération supplémentaire, c'est-à-dire lorsque l'on a obtenu la meilleure précision avec la machine utilisée. $\lambda_k$ est alors une des valeurs propres recherchées.

Une fois obtenu une valeur propre, on réduit l'ordre de la matrice de 1 unité en supprimant la dernière ligne et la dernière colonne. Ces opérations s'appellent déflation. Ensuite, par le même procédé, on passe au calcul de la valeur propre suivante. **Cette méthode de Rutishauser concernant les matrices quasi triangulaires converge quadratiquement.** (Pratiquement cela signifie que le nombre de chiffres significatifs exacts est multiplié par un coefficient proche de deux à chaque tour d'itération.)

Rappelons que les calculs s'effectuent en arithmétique complexe dans le cas général.

On trouvera sur le Web (*) le programme `givens.c` qui calcule les valeurs propres réelles des matrices réelles.

## 4.4. Méthode de Danilevski (1937)

Cette méthode consiste à transformer la matrice $A$ ou plutôt le déterminant caractéristique en sa forme canonique de Frobenius au moyen de transformations linéaires qui laissent le déterminant inchangé. Une fois la forme canonique obtenue, on écrit directement l'équation caractéristique dont on cherche ensuite les racines au moyen d'une méthode appropriée, la méthode de Bairstow par exemple.

Écrivons l'équation caractéristique $A(\lambda)$ associée à la matrice $A$ :

$$A(\lambda) = \begin{vmatrix} a_{11}-\lambda & -a_{12} & a_{13} & \ldots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \ldots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33}-\lambda & \ldots & a_{3n-1} & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn-1} & a_{nn}-\lambda \end{vmatrix}$$

* `http://www.edpsciences.com/guilpin/`

On désire obtenir la forme suivante :

$$B(\lambda) = \begin{vmatrix} b_1 - \lambda & -b_2 & -b_3 & \dots & -b_{n-1} & -b_n \\ 1 & -\lambda & 0 & \dots & 0 & 0 \\ 0 & 1 & -\lambda & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -\lambda \end{vmatrix}$$

On passe du déterminant $A(\lambda)$ au déterminant $B(\lambda)$ au moyen de combinaisons linéaires soit de lignes, soit de colonnes (*cf.* la méthode des pivots concernant la résolution des systèmes linéaires). La transformation s'effectue de bas en haut, à savoir :

1. On divise l'avant-dernière colonne par $a_{n,n-1}$ pour amener 1 à la place de $a_{n,n-1}$ (les pivots de la transformation se situent sur la première sous-diagonale). Pour conserver le coefficient 1 devant $\lambda$, on est amené à multiplier l'avant-dernière ligne par $a_{n,n-1}$. On désigne par $a^1_{nq}$ les valeurs obtenues sur la dernière ligne.
2. Ensuite, on combine linéairement les colonnes pour que le déterminant ait sa dernière ligne selon la forme canonique :

$$0 \quad 0 \quad \dots \quad 0 \quad 1 \quad -\lambda.$$

   Pour cela on multiple l'avant dernière colonne par $a^1_{nq}$ et l'on retranche le résultat de la colonne $q$, avec $q = 0, 1, 2, \dots, n-2, n$. Cette opération a amené des termes en $\lambda$ sur l'avant dernière ligne avec les coefficients $a^1_{nq}$, $q = 1, 2, \dots, n-2, n$, qu'il faut à présent éliminer sauf sur la diagonale principale.
3. En retranchant terme à terme la première ligne multipliée par $a^1_{nq}$, on fait disparaître $\lambda$ du premier élément $a_{n-1,1}$ de l'avant-dernière ligne. On opère de la même façon avec la deuxième ligne multipliée par $a_{n2}$ pour éliminer le terme en $\lambda$ de $a_{n-1,2}$ et ainsi de suite pour obtenir l'avant dernière ligne privée de termes en $\lambda$ excepté l'élément diagonal.

On passe ensuite au pivot suivant qui est l'élément $a_{n-1,n-2}$, on divise alors la ligne $(n-1)$ par cet élément et on multiplie la colonne $(n-1)$ par ce même élément, puis on poursuit les calculs de la même manière que pour la ligne $n$.

***Cas où un pivot est nul*** – Si un pivot est nul la méthode tombe en défaut. Supposons qu'il s'agisse du pivot $a_{k,k-1}$ alors on regarde dans la ligne $k$ si un des termes compris entre les colonnes 1 et $k-2$ est différent de zéro. Supposons qu'il en soit ainsi et que le terme $a_{kq}$ soit différent de zéro. On ajoute alors les éléments de la colonne $q$ à ceux de la colonne $k-1$. Cette opération fait apparaître un terme en $\lambda$ sur la ligne $q$ et la colonne $k-1$. On le fait disparaître en retranchant à la ligne $q$ la ligne $q+1$.

Le cas où il n'y a que des zéros sur la ligne $k$ dont le pivot est nul montre que les calculs ne peuvent pas se poursuivre de cette manière, mais le déterminant est alors le produit de deux déterminants dont l'un a la forme voulue de Frobenius et l'autre la forme ordinaire. À ce dernier déterminant, on peut alors faire subir à nouveau le traitement de Danilevski. Dans le cas le plus défavorable, on aura un produit de déterminants chacun écrit sous la forme ordinaire.

On trouvera sur le Web (*) le programme `danilev.c` qui calcule les valeurs propres réelles des matrices réelles.

* `http://www.edpsciences.com/guilpin/`

## 4.5. La méthode de Krylov (1879–1955)

Elle repose sur le théorème de Cayley-Hamilton (toute matrice $A$ d'ordre $n$ est solution de son équation caractéristique), soit :

$$A^n + \sum_{q=1}^{n} a_q A^{n-q} = 0,$$

les $a_k$ étant les coefficients du polynôme caractéristique. On multiplie les deux membres de cette expression par un vecteur arbitraire d'ordre $n$ que l'on appelle $X$, on obtient :

$$\sum_{q=1}^{n} a_q A^{n-q} X = -A^n X.$$

On commence par calculer $B_0 = X$, $B_1 = AB_0, \ldots, B_{k+1} = AB_k, \ldots$, et enfin $B_n = AB_{n-1}$. Les vecteurs $B_k$ sont les colonnes de la matrice du système linéaire à résoudre pour obtenir les coefficients $a_k$ du polynôme caractéristique. Les dispositions du calcul sont les suivantes :

$$\begin{vmatrix} b_{1n-1} & b_{1n-2} & b_{1n-3} & \ldots & b_{11} & b_{10} \\ b_{2n-1} & b_{2n-2} & b_{2n-3} & \ldots & b_{21} & b_{20} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{nn-1} & b_{nn-2} & b_{nn-3} & \ldots & b_{n1} & b_{n0} \end{vmatrix} \cdot \begin{vmatrix} a_1 \\ a_2 \\ \\ a_n \end{vmatrix} = - \begin{vmatrix} b_{1n} \\ b_{2n} \\ \\ b_{nn} \end{vmatrix}.$$

Le vecteur $X$ est arbitraire, et on le choisit très souvent avec toutes ses composantes égales à l'unité. Reste à déterminer $a_0 = (-1)^n$.

Cette méthode n'exigeant que peu de calculs est fort attrayante, malheureusement elle conduit à un système mal conditionné lorsque $n$ atteint et dépasse quelques unités de l'ordre de 8 à 10.

On donne sur le Web(*) le programme `krylov.c` qui réalise cet algorithme.

## 4.6. Cas des matrices symétriques

Les matrices symétriques sont hermitiennes et par conséquent ne possèdent que des valeurs propres réelles. Quand on calcule les zéros d'un polynôme caractéristique dont les coefficients sont mal conditionnés on court le risque de voir apparaître presque sûrement des racines complexes dont on se passerait bien. La méthode de Givens-Rutishauser permet de contourner cette difficulté. On trouvera dans les exercices un problème mettant en œuvre la méthode de Jacobi adaptée au cas des matrices symétriques et qui donne de très bons résultats, on fournit également le programme (*cf.* annexe H, problème 5.8).

## 4.7. Retour sur les matrices de Hilbert

Les valeurs propres des matrices de Hilbert sont réelles puisque la matrice est symétrique, cependant, leur calcul est une source de difficultés puisque le déterminant de la matrice d'ordre $n$ tend vers zéro quand $n$ tend vers l'infini. En effet, le produit des valeurs propres est égal au déterminant de la matrice. Par ailleurs, la somme des valeurs propres est égale à la trace de la matrice. Toutes les conditions sont réunies pour avoir des valeurs propres très grandes et à coup sûr d'autres très petites.

* `http://www.edpsciences.com/guilpin/`

## 5. Éléments de bibliographie

A. ANGOT (1972) *Compléments de mathématiques*, Éditions Masson.

P.G. CIARLET (1982) *Introduction à l'analyse numérique matricielle et à l'optimisation*, Éditions Masson.

E. DURAND (1961) *Solutions numériques des équations algébriques*, Tome II, Éditions Masson.

C.T. KELLEY (1995) *Iterative methods for linear and non linear equations*, Society for Industrial and Applied Mathematics, Philadelphie.

R. KRESS (1998) *Numerical analysis*, Springer.

O. NEVANLINNA (1993) *Convergence of iterations for linear equations*, Birkhauser.

R.S. VARGA (1962) *Matrix iterative analysis*, Éditions Prentice Hall.

# 6 | L'interpolation

Avant de développer le thème de l'interpolation, il est sans doute utile de préciser la nature du problème proposé, car, très souvent, les opérations d'interpolation et les opérations de lissage sont confondues dans l'esprit de bien des utilisateurs. Que ce soient des résultats expérimentaux ou que ce soient des tables, la plupart des fonctions, au sens large, ne nous sont données que pour un ensemble discret de points qui constitue un échantillon. Toutes les fonctions transcendantes dites spéciales ont fait l'objet de tabulation pour des valeurs de la variable en progression arithmétique. Bien entendu il est rare que la valeur de la fonction figure dans la table pour la valeur dont on a réellement besoin et l'on est obligé de se livrer à une opération d'interpolation — voire d'extrapolation — pour obtenir le résultat désiré.

Un problème identique apparaît lorsque l'on veut traiter numériquement un ensemble de résultats expérimentaux quand bien même le procédé utilisé pour les recueillir eut été réalisé par voie analogique.

Bref, c'est l'éternel problème : on ne peut avoir accès qu'à un nombre fini de valeurs. Le problème posé par l'interpolation ne doit pas être dissocié à proprement parler du problème de l'extrapolation (dans la mesure où il s'agit d'effectuer une extrapolation qui a effectivement un sens). Quel que soit le problème auquel on s'attache, on dispose toujours au départ d'un ensemble fini d'ordonnées correspondant à un ensemble fini d'abscisses, ces deux ensembles appartenant nécessairement à des intervalles finis. Appelons $I$ l'intervalle de définition des abscisses. Ces ensembles de valeurs constituent un échantillon de la fonction $f(x)$ à laquelle nous nous intéressons.

Quand nous cherchons à connaître $f(x)$ pour une quelconque valeur appartenant à $I$, hormis pour les points de l'échantillon, nous dirons que nous effectuons une interpolation. En revanche, quand nous cherchons à obtenir $f(x)$ pour une valeur de $x$ située à l'extérieur de $I$, nous dirons que nous réalisons une extrapolation. Ces deux problèmes s'abordent de la même façon, mais c'est le calcul d'erreur qui va introduire une différence entre les deux. Au passage, nous noterons que l'extrapolation dans les voisinages immédiats des bornes de $I$ fournit d'excellents résultats.

Pour ce qui concerne les opérations de lissage, il faut dire qu'elles ont un autre but : il s'agit avant tout d'effectuer un filtrage sur un ensemble de données entachées d'erreur. On cherche alors à rendre compte d'une allure globale des données expérimentales au moyen de fonctions plus ou moins arbitraires dont on ajustera les paramètres — le plus souvent en utilisant la méthode des moindres carrés — mais, ce qui importera en priorité, sera l'élimination de ces erreurs qui constituent un bruit de fond. Nous aurons amplement l'occasion d'approfondir ce problème, et nous examinerons un certain nombre de méthodes qui réalisent ces opérations de filtrage

éventuellement après qu'auront été abordés quelques éléments de statistique, la méthode des moindres carrés et la transformée de Fourier.

## 1. De la légitimité de l'interpolation

En règle générale, la fonction $f(x)$ connue au moyen d'un échantillon de points est continue sur l'intervalle $I$ que nous venons de définir. Il s'ensuit que sur cet intervalle elle peut être approchée uniformément par un polynôme et cela en vertu du théorème de Weierstrass (1815–1897). Du reste il n'y a pas de difficulté à généraliser cette proposition au cas où l'approximation est réalisée par un système quelconque de fonctions pourvu qu'il soit complet.

Les systèmes complets usuels sont : le système des puissances (les polynômes orthogonaux usuels), le système des sinus et des cosinus, les fonctions de Bessel, les fonctions du cylindre parabolique etc.

L'idée d'interpoler par un polynôme est bien antérieure à la publication du théorème de Weierstrass (1815–1897) et elle a été exploitée par de nombreux savants mais c'est le nom de Lagrange qui est resté attaché à cette technique. Comme elle présente un grand intérêt théorique, nous allons nous y attarder quelque peu.

## 2. Le polynôme de Lagrange (1736–1813)

Le problème consiste à obtenir le polynôme de degré $n$ qui passe par les $(n+1)$ points de l'échantillon. Les abscisses $a_0, a_1, \ldots a_n$ appartiennent à l'intervalle $I$ qui est le plus petit possible, et la fonction réelle $f(x)$ prend les valeurs $b_0, b_1, \ldots b_n$ correspondant aux abscisses respectives.

Aucune hypothèse n'est faite sur les échantillons à cette réserve près toutefois que, pour une valeur donnée de $a_k$, il ne peut correspondre qu'une seule valeur $b_k$. Cela implique que tous les $a_k$ sont différents. Si tel n'était pas le cas, il conviendrait d'effectuer une partition (ou plusieurs) pour réaliser cette condition, et l'on associerait à chacune des partitions obtenues un polynôme de Lagrange unique. Supposons que pour la valeur $a_k$ il y ait deux valeurs $y_1$ et $y_2$, cela veut dire que l'on découpe l'intervalle $I$ en deux sous-intervalles $I_1$ et $I_2$ ayant $a_k$ comme intersection commune. $y_1$ et $y_2$ sont attribués chacun à l'intervalle correspondant $I_1$ ou $I_2$ donné par l'allure de la courbe représentative.

Tout cela précisé, nous allons établir l'existence de ce polynôme unique de degré $n$ que nous désignons par $P_n(x)$ et qui s'explicite de la façon suivante :

$$P_n(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1} x^{n-1} + \alpha_n x^n = \sum_{k=0}^{k=n} \alpha_k x^k.$$

En utilisant le fait que ce polynôme $P_n(x)$ doit prendre la valeur $b_k$ lorsque la variable $x$ prend la valeur $a_k$, nous obtenons le système linéaire suivant :

$$\begin{vmatrix} 1 & a_0 & a_0^2 & a_0^3 & \ldots & a_0^n \\ 1 & a_1 & a_1^2 & a_1^3 & \ldots & a_1^n \\ 1 & a_2 & a_2^2 & a_2^3 & \ldots & a_2^n \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & a_n & a_n^2 & a_n^3 & \ldots & a_n^n \end{vmatrix} \cdot \begin{vmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \\ \alpha_n \end{vmatrix} = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \\ \\ b_n \end{vmatrix}$$

que nous écrirons d'une manière plus concise sous la forme :

$$A\alpha = B.$$

Le déterminant de la matrice $A$ est un déterminant de Vandermonde pour lequel tous les $a_k$ sont différents :

$$\det[A] = \prod_{j=0}^{j=n} \prod_{k>j}^{k=n} (a_j - a_k).$$

Comme conséquence immédiate il s'ensuit que la solution existe et qu'elle est unique.

Si l'on se situe dans un contexte historique, à l'époque de Lagrange (1736–1813), le calcul des coefficients ne pouvait pas s'effectuer par la résolution d'un système linéaire dès que l'ordre de la matrice dépassait quelques unités, et Lagrange proposa une technique qui a permis de réaliser directement l'interpolation sans calculer explicitement les coefficients du polynôme. C'est cette méthode que nous exposons maintenant.

Considérons un polynôme de degré $(n+1)$ appelé $Q_{n+1}(x)$ et défini par la relation :

$$Q_{n+1}(x) = \prod_{j=0}^{n} (x - a_j).$$

À présent, décomposons le rapport $P_n(x)/Q_{n+1}(x)$ en éléments simples, opération possible puisque, par hypothèse, tous les $a_k$ sont différents ; nous pouvons écrire :

$$\frac{P_n(x)}{Q_{n+1}(x)} = \sum_{i=0}^{n} \frac{A_i}{(x - a_i)}$$

d'où l'expression prise par le polynôme $P_n(x)$ :

$$P_n(x) = Q_{n+1}(x) \left( \sum_{i=0}^{n} \frac{A_i}{(x - a_i)} \right) = \sum_{i=0}^{n} \frac{A_i Q_{n+1}(x)}{(x - a_i)} .$$

Comme il est possible d'écrire que :

$$\frac{Q_{n+1}(x)}{(x - a_i)} = \prod_{\substack{j=0 \\ j \neq i}}^{n} (x - a_j)$$

nous obtenons alors :

$$P_n(x) = \sum_{i=0}^{n} A_i \cdot \prod_{\substack{j=0 \\ j \neq i}}^{n} (x - a_j).$$

Maintenant il nous faut chercher à expliciter les $A_i$ en tenant compte du fait que le polynôme doit passer par les points d'échantillonnage. Cela donne les relations suivantes :

$$P_n(a_k) = b_k = \sum_{i=0}^{n} A_i \cdot \prod_{\substack{j=0 \\ j \neq i}}^{n} (a_k - a_j).$$

Comme l'indice $j$ prend toutes les valeurs de 0 à $n$, excepté la valeur $i$, les produits seront toujours tous nuls sauf celui pour lequel nous aurons $i = k$. On en déduit l'expression suivante :

$$b_k = A_k \cdot \prod_{\substack{j=0 \\ j \neq i}}^{n} (a_k - a_j).$$

Nous tirons alors la valeur de $A_k$ que nous reportons dans l'expression de $P_n(x)$ :

$$P_n(x) = \sum_{i=0}^{n} b_i \cdot \prod_{\substack{s=0 \\ s \neq i}}^{n} \frac{(x - a_s)}{(a_s - a_i)} \cdot \tag{6.1}$$

$P_n(x)$ est donc le polynôme passant par tous les points de l'échantillonnage, et nous en avons obtenu deux expressions. Il convient d'ajouter que ce polynôme peut être utilisé aussi bien pour réaliser une interpolation qu'une extrapolation.

Comme par expérience on sait que l'extrapolation est une opération peu précise loin du domaine $I$, il est nécessaire dès à présent de se faire une opinion sérieuse sur l'erreur commise en remplaçant une fonction $f(x)$ échantillonnée par le polynôme de $P_n(x)$, et c'est l'expression due à Lagrange qui va nous permettre de résoudre le problème.

## 3. Évaluation de l'erreur

Supposons que $f(x)$, remplacée par $P_n(x)$, soit $(n + 1)$ fois dérivable sur l'intervalle $I$, et considérons la fonction auxiliaire $\Phi(u)$ définie comme suit :

$$\Phi(u) = f(u) - P_n(u) - \left[f(x) - P_n(x)\right] \frac{\prod_{i=0}^{n} (u - a_i)}{\prod_{i=0}^{n} (x - a_i)} \cdot$$

Cette fonction $\Phi(u)$ qui apparaîtra sans doute un peu artificielle va nous permettre d'exprimer aisément la valeur absolue de l'erreur $E(x)$ commise en effectuant le remplacement de $f(x)$ par $P_n(x)$ :

$$E(x) = f(x) - P_n(x),$$

ce qui, par parenthèses, explique la structure de $\Phi(u)$. Pour parvenir à notre fin, dérivons $(n+1)$ fois la fonction $\Phi(u)$ par rapport à $u$, ce qui donne :

$$\Phi^{(n+1)}(u) = f^{(n+1)}(u) - \left[f(x) - P_n(x)\right] \frac{(n+1)!}{\prod_{i=0}^{n} (x - a_i)} \cdot$$

Remarquons que la fonction $\Phi(u)$ s'annule pour toutes les valeurs $u = a_i$ mais aussi pour la valeur $u = x$. L'application successive du théorème de Rolle (1652–1719) à $\Phi(u)$ permet de voir que la dérivée $\Phi^{(n+1)}(u)$ s'annule pour une valeur $u = \xi$ appartenant au plus petit

intervalle contenant $x$ et les $a_i$ (dans le cas de l'extrapolation $x$ n'appartient pas à l'intervalle $I$ précédemment défini). Nous obtenons alors l'expression suivante :

$$\Phi^{(n+1)}(\xi) = f^{(n+1)}(\xi) - E(x)\frac{(n+1)!}{\prod_{i=0}^{n}(x-a_i)} = 0,$$

de là nous tirons l'expression de l'erreur $E(x)$ :

$$E(x) = \prod_{i=0}^{n}(x-a_i)\frac{f^{(n+1)}(\xi)}{(n+1)!}\cdot$$

Comme chaque fois en pareil cas, on procède à la majoration de $f^{(n+1)}(x)$ dans l'intervalle adéquat. Rappelons que cet intervalle sera plus grand que $I$ s'il s'agit d'une extrapolation, c'est-à-dire que $x$ n'appartient pas à l'intervalle $I$. Désignons par $M_{n+1}$ la limite supérieure de $f^{(n+1)}(x)$ dans l'intervalle concerné, l'expression de l'erreur devient :

$$E(x) \leq \frac{M_{n+1}}{(n+1)!}\left|\prod_{i=0}^{n}(x-a_i)\right|.$$

On voit immédiatement que l'erreur est nulle sur les points d'échantillonnage et qu'elle est très petite dans le voisinage de ces points. De plus il est important de noter que l'expression de $E(x)$ varie en fonction de $x$ et que, de plus, elle dépend des points $a_k$. On peut alors se poser la question de savoir comment choisir les $a_k$ pour que l'erreur $E(x)$ soit minimum.

## 4. Comment minimiser E(x)

La seule façon possible de réaliser l'opération est de rendre minimum l'expression $\prod_{i=0}^{n}(x-a_i)$ qui n'est rien d'autre qu'un polynôme de degré $(n+1)$. Nous montrerons au chapitre 9 le théorème suivant : parmi tous les polynômes à coefficient principal réduit de degré $(n+1)$, c'est le polynôme de Tchebycheff (1821–1894) qui s'écarte le moins de l'axe des $x$ sur l'intervalle $(-1,+1)$. Ce polynôme a pour expression :

$$T_{n+1}(z) = \frac{1}{2^n}\cos[(n+1)\mathrm{Arcos}(z)].$$

Les zéros de ce polynôme sont donnés par :

$$(n+1)\mathrm{Arcos}(z_k) = \frac{\pi}{2}(2k+1),$$

soit :

$$z_k = \cos\frac{\pi(2k+1)}{2(n+1)} \quad \text{avec} \quad k = 0,1,2,\ldots,n.$$

En réalité, on ne travaille pas sur l'intervalle $(-1,+1)$, mais sur l'intervalle $I$ dont la borne inférieure est $a$ et la borne supérieure $b$. Le changement de variable :

$$z = \frac{2x-a-b}{b-a}$$

permet de passer de l'intervalle $(-1,+1)$ à l'intervalle $(a,b)$. On en déduit que le polynôme de Tchebycheff de degré $(n+1)$ à coefficient principal réduit défini sur $(a,b)$ a pour forme :

$$\frac{(b-a)^{n+1}}{2^{2n+1}} T_{n+1}\left(\frac{2x-a-b}{b-a}\right),$$

ce changement de variable donne également les zéros de ce polynôme :

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\frac{\pi(2k+1)}{2(n+1)} \quad \text{avec} \quad k = 0,1,2,\ldots,n. \tag{6.2}$$

Pour minimiser l'erreur, il faut choisir les $a_k$ identiques aux $x_k$ qui sont donnés par l'expression (6.2).

***Remarque*** : Il est à noter que l'erreur est donc minimum pour un partage de l'intervalle $I$ selon les $x_k$, et cela est indépendant de la fonction à interpoler. Cette règle demeure dans le cas où la fonction $f(x)$ est connue empiriquement, et la meilleure façon d'obtenir des valeurs exploitables avec la plus petite erreur repose sur le choix des $x_k$ donnés par l'expression (6.2). On insiste donc sur la généralité de ce théorème qui ne préjuge en rien de la nature de la fonction à interpoler si ce n'est l'hypothèse traditionnelle de continuité.

## 5. Autre disposition pratique du calcul du polynôme de Lagrange

Il faut bien reconnaître que le calcul effectif du polynôme de Lagrange pour un ensemble de valeurs de $x$ demande un travail assez laborieux, aussi préfère-t-on employer une autre expression qui conduit à des calculs plus économiques et surtout qui permettra d'établir un certain nombre d'autres formes réputées canoniques. Pour ce faire, on se propose d'adopter la disposition suivante :

$$P_n(x) = B_0 + B_1(x-a_0) + B_2(x-a_0)(x-a_1) + \cdots + B_n(x-a_0)(x-a_1)\ldots(x-a_{n-1}), \tag{6.3}$$

dont il faut déterminer les $B_k$. D'abord on a $P_n(a_0) = b_0 = B_0$, ensuite on peut poser :

$$Q_{n-1}(x) = \frac{P_n(x)-b_0}{x-a_0} = B_1 + B_2(x-a_1) + B_3(x-a_1)(x-a_2) + \cdots + B_n(x-a_1)(x-a_2)\ldots(x-a_{n-1}),$$

ce qui permet d'écrire

$$Q_{n-1}(a_1) = \frac{b_1-b_0}{a_1-a_0} = B_1.$$

Considérons à présent :

$$Q_{n-2}(x) = \frac{Q_{n-1}(x)-B_1}{x-a_1} = B_2 + B_3(x-a_2) + B_4(x-a_2)(x-a_3) + \cdots + B_n(x-a_2)(x-a_3)\ldots(x-a_{n-1}),$$

expression dans laquelle on fait $x = a_2$, ce qui donne :

$$Q_{n-2}(a_2) = \frac{Q_{n-1}(a_2) - B_1}{a_2 - a_1} = B_2,$$

soit encore :

$$B_2 = \frac{b_2 - b_0 - B_1(a_2 - a_0)}{(a_2 - a_0)(a_2 - a_1)}.$$

On poursuit de la même façon le calcul pour obtenir tous les $B_k$.

### Récapitulation de la conduite du calcul

Sur le tableau ci-dessous on a représenté la suite des opérations qui permet de calculer tous les $B_j$.

$$\begin{aligned}
&a_0 b_0 = B_0 \\
&a_1 b_1 \frac{b_1 - b_0}{a_1 - a_0} = q^1_{n-1} = B_1 \\
&a_2 b_2 \frac{b_2 - b_0}{a_2 - a_0} = q^2_{n-1} \frac{q^3_{n-1} - q^2_{n-1}}{a_2 - a_1} = q^2_{n-2} = B_2 \\
&a_3 b_3 \frac{b_3 - b_0}{a_3 - a_0} = q^3_{n-1} \frac{q^3_{n-1} - q^1_{n-1}}{a_3 - a_1} = q^3_{n-2} \frac{q^3_{n-2} - q^2_{n-2}}{a_3 - a_2} = q^3_{n-3} = B_3 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&a_n b_n \frac{b_n - b_0}{a_n - a_0} = q^n_{n-1} \frac{q^n_{n-1} - q^{n-1}_{n-1}}{a_n - a_1} = q^n_{n-2} \frac{q^n_{n-2} - q^{n-1}_{n-2}}{a_n - a_2} = q^n_{n-3} = B_n
\end{aligned}$$

## 6. Cas où les abscisses sont en progression arithmétique

Désignons par $h$ la raison de la progression ; nous avons donc la suite des abscisses donnée par : $a_0$, $a_1 = a_0 + h$, $a_2 = a_0 + 2h, \ldots, a_n = a_0 + nh$.

C'est un cas particulier, certes, mais il est très répandu, et à ce titre il mérite d'être examiné car il conduit à des expressions très connues et très simples d'emploi. Pour aborder cette étude il nous faut au préalable définir les opérateurs de différence.

On appelle différence première d'ordre $k$ la valeur $\Delta b_k$ donnée par l'expression :

$$\Delta^1 b_k = b_{k+1} - b_k = f[a_0 + (k+1)h] - f(a_0 + kh).$$

Ensuite, on définit les différences deuxièmes d'ordre $k$ :

$$\Delta^2 b_k = \Delta^1 b_{k+1} - \Delta^1 b_k$$

et ainsi de suite pour les différences troisièmes, quatrièmes, etc. D'une façon générale on définira les différences $p^{\text{e}}$ d'ordre $k$ au moyen de la relation :

$$\Delta^p b_k = \Delta^{p-1} b_{k+1} - \Delta^{p-1} b_k \quad \text{avec} \quad p \leq n.$$

***Remarque*** : Si les valeurs $(a_i, b_i)$ sont données par un polynôme de degré $m$ $(f(x) = P_m(x))$, il va de soi que les différences d'ordre $k$ supérieures à $m$ sont nulles et que toutes les différences $m^{\text{e}}$ sont égales.

La signification des différences est simple, ce sont les expressions numériques approchées des valeurs de la dérivée première, de la dérivée deuxième, etc. au point $(a_k, b_k)$.

Ici la valeur $a_0$ est la valeur initiale de la suite, mais on peut très bien écrire les formules d'approximation en choisissant pour valeur initiale $a_0$ un point d'intérêt particulier dans le tableau des données ; dans ce cas, les abscisses précédant $a_0$ sont notées par des indices négatifs :

$$a_{-k} = a_0 - kh.$$

Cette dernière remarque conduit à l'écriture de différentes formes du polynôme de Lagrange qui offrent de grands avantages lorsque l'on ne désire employer qu'une partie restreinte du tableau initial des données. Ces différentes formes portent le nom du savant qui s'est attaché à une de ces études, on étudiera donc les polynômes ascendant et descendant de Newton, les polynômes de Stirling et de Bessel.

Dans le cas où l'on utilise toutes les données du tableau, soit $(n+1)$ points, tous les polynômes que nous venons d'évoquer donnent alors les mêmes résultats puisque ce ne sont que des expressions différentes du même polynôme de Lagrange. Toutefois nous verrons une petite nuance concernant les polynômes de Stirling et de Bessel qui sont respectivement l'un pair l'autre impair, mais cela ne restreint pas la généralité de notre propos.

## 7. Les polynômes d'interpolation de Newton (1643–1727)

Reprenons l'équation (6.3) dans laquelle nous tenons compte du fait que les abscisses sont en progression arithmétique :

$$P_n(x) = N_0 + N_1(x - a_0) + N_2(x - a_0)(x - a_0 - h) + \cdots + N_n(x - a_0)\ldots[x - a_0 - (n-1)h]. \quad (6.4)$$

Pour obtenir les valeurs des $(n-1)$ coefficients $N_k$, il suffit de donner à $x$ la suite des valeurs $a_0 + kh$ avec $k = 0, 1, 2, \ldots, n$. À l'aide des différences premières, deuxièmes etc., le tableau précédent se transforme ainsi :

| $x$ | $b$ | $\Delta^1 b$ | $\Delta^2 b$ | $\Delta^3 b$ | $\Delta^4 b$ | $\ldots$ |
|---|---|---|---|---|---|---|
| $a_0$ | $b_0$ | | | | | |
| $a_0 + h$ | $b_1$ | $\Delta^1 b_0$ | | | | |
| $a_0 + 2h$ | $b_2$ | $\Delta^1 b_1$ | $\Delta^2 b_0$ | | | |
| $a_0 + 3h$ | $b_3$ | $\Delta^1 b_2$ | $\Delta^2 b_1$ | $\Delta^3 b_0$ | | |
| $a_0 + 4h$ | $b_4$ | $\Delta^1 b_3$ | $\Delta^2 b_2$ | $\Delta^3 b_1$ | $\Delta^4 b_1$ | |
| ..... | | | | | | |

On obtient alors :

$$\begin{aligned}
P_n(a_0) &= b_0 = N_0 \\
P_n(a_0 + h) &= b_1 = N_0 + N_1 h \\
P_n(a_0 + 2h) &= b_2 = N_0 + 2N_1 h + 2!N_2 h^2 \\
&\ldots\ldots\ldots \\
P_n(a_0 + qh) &= b_q = N_0 + qN_1 h + q(q-1)N_2 h^2 + q(q-1)(q-2)N_3 h^3 + \cdots + q!N_q h^q \\
&\ldots\ldots\ldots \\
P_n(a_0 + nh) &= b_n = N_0 + nN_1 h + n(n-1)N_2 h^2 + n(n-1)(n-2)N_3 h^3 + \cdots + n!N_q h^n.
\end{aligned}$$

À présent, nous allons exprimer de proche en proche les coefficients $N_k$ au moyen des opérateurs de différence définis au paragraphe précédent. On obtient donc :

$$
\begin{aligned}
N_0 &= b_0 \\
N_1 &= \frac{b_1 - b_0}{h} = \frac{\Delta^1 b_0}{h} \\
N_2 &= \frac{(b_2 - b_1) - (b_1 - b_0)}{2!h^2} = \frac{\Delta^2 b_0}{2!h^2} \\
&\cdots\cdots\cdots\cdots\cdots \\
N_k &= \frac{\Delta^k b_0}{k!h^k} \\
&\cdots\cdots\cdots\cdots\cdots \\
N_n &= \frac{\Delta^n b_0}{n!h^n}
\end{aligned}
$$

Pour des raisons de concision d'écriture il est commode de poser :

$$z = \frac{x - a_0}{h}$$

Ainsi l'expression devient :

$$P_n(z) = b_0 + z\Delta^1 b_0 + \frac{z(z-1)}{2!}\Delta^2 b_0 + \cdots + \frac{z(z-1)\ldots[z-(n-1)]}{n!}\Delta^n b_0. \qquad (6.5)$$

Cette formule s'appelle **polynôme de Newton descendant**. Elle est utilisée pour effectuer des interpolations en tête de tableau (ce qui signifie que l'on réalise l'interpolation avec le début du tableau et non pas avec tout le tableau).

Pour obtenir les interpolations en queue de tableau — disons avec le dernier quart des données — il suffit d'écrire le polynôme de Newton en considérant un pas négatif, c'est-à-dire en changeant de signe la raison de la progression arithmétique. L'expression (6.4) devient :

$$
\begin{aligned}
P_n(x) = M_0 + M_1(x - a_0) + M_2(x - a_0)(x - a_0 + h) \\
+ \cdots + M_n(x - a_0)\ldots[x - a_0 + (n-1)h]. \qquad (6.6)
\end{aligned}
$$

Comme précédemment, on donnera à $x$ les valeurs $a_0 - kh$ avec $k = 0, 1, 2, \ldots n$, mais, pour éviter toute confusion, on affectera un indice aux valeurs correspondantes des ordonnées avec un signe négatif, soit $b_{-k}$. On obtient alors :

$$
\begin{aligned}
P_n(a_0) &= b_0 = M_0 \\
P_n(a_0 - h) &= b_{-1} = M_0 - M_1 h \\
P_n(a_0 - 2h) &= b_{-2} = M_0 - 2M_1 h + 2!M_2 h^2 \\
&\cdots\cdots\cdots\cdots\cdots \\
P_n(a_0 - qh) &= b_{-q} = M_0 - qM_1 h + q(q-1)M_2 h^2 - q(q-1)(q-2)M_3 h^3 \\
&\qquad + \cdots + (-1)^q q! M_q h^q \\
&\cdots\cdots\cdots\cdots\cdots \\
P_n(a_0 - nh) &= b_{-n} = M_0 - nM_1 h + n(n-1)M_2 h^2 - n(n-1)(n-2)M_3 h^3 \\
&\qquad + \cdots + (-1)^n n! M_q h^n.
\end{aligned}
$$

À partir de ces expressions, on tire la suite des $M_k$ données en fonction des opérateurs de différence :

$$
\begin{aligned}
M_0 &= b_0 \\
M_1 &= \frac{b_0 - b_1}{h} = \frac{\Delta^1 b_{-1}}{h} \\
M_2 &= \frac{\Delta^2 b_{-2}}{2!h^2} \\
&\cdots\cdots\cdots \\
M_k &= \frac{\Delta^k b_{-k}}{k!h^k} \\
&\cdots\cdots\cdots \\
M_n &= \frac{\Delta^n b_{-n}}{n!h^n}
\end{aligned}
$$

Toujours en posant $z = (x - a_0)/h$, nous obtenons la forme canonique du **polynôme de Newton ascendant** :

$$P_n(z) = b_0 + z\Delta^1 b_{-1} + \frac{z(z+1)}{2!}\Delta^2 b_{-2} + \cdots + \frac{z(z+1)\ldots(z+n-1)}{n!}\Delta^n b_{-n}\cdot \qquad (6.7)$$

En résumé, les deux polynômes de Newton permettent d'obtenir des expressions adaptées à l'interpolation des panneaux de tête et de queue des tableaux lorsque l'on n'utilise qu'une partie des données. Il reste donc la partie centrale que l'on interpole par les polynômes de Stirling et de Bessel.

## 8. Le polynôme d'interpolation de Stirling (1692–1770)

On obtient ce polynôme en faisant jouer un rôle symétrique à l'ensemble des données par rapport à $a_0$, pour cela on écrit le polynôme (6.4) sous la forme suivante :

$$
\begin{aligned}
&P_n(x) = S_0 + S_1(x - a_0) + S_2(x - a_0 + h)(x - a_0) + S_3(x - a_0 + h)(x - a_0)(x - a_0 - h) \\
&+ \cdots + S_{2q}(x - a_0 + qh)\ldots[x - a_0 - (q-1)h] + S_{2q+1}(x - a_0 + qh)\ldots(x - a_0 - qh). \qquad (6.8)
\end{aligned}
$$

Pour obtenir les valeurs de $S_k$, il suffit de donner à $x$ la suite des valeurs $a_0$, $a_0 = h$, $a_0 + h$, $a_0 - 2h$, etc., ce qui permet d'écrire :

$$
\begin{aligned}
P_n(a_0) &= b_0 &&= S_0 \\
P_n(a_0 - h) &= b_{-1} &&= S_0 - S_1 h \\
P_n(a_0 + h) &= b_1 &&= S_0 + S_1 h - 2S^2 h^2 \\
P_n(a_0 - 2h) &= b_{-2} &&= S_0 - 2S_1 h + 2!S_2 h^2 - 3!S^3 h^3 \\
&\cdots\cdots\cdots\cdots \\
P_n(a_0 - qh) &= b_{-q} &&= S_0 - qS_1 h + q(q-1)S_2 h^2 - q(q-1)(q+1)S_3 h^3 \\
&&&\quad + \cdots + (-1)^q(2q-1)!S_{2q-1}h^{2q-1} \\
&\cdots\cdots\cdots\cdots \\
P_n(a_0 + qh) &= b_n &&= S_0 + qS_1 h - q(q-1)S_2 h^2 + q(q+1)(q-2)S_3 h^3 + \cdots + (2q)!S_{2q}h^{2q}.
\end{aligned}
$$

Tout comme précédemment, on tire les valeurs des coefficients $S_k$ exprimés en fonction des opérateurs de différence :

$$\begin{aligned}
S_0 &= b_0 \\
S_1 &= \frac{\Delta^1 b_{-1}}{h} \\
S_2 &= \frac{\Delta^2 b_{-2}}{2!h^2} \\
&\cdots\cdots\cdots\cdots \\
S_{2q} &= \frac{\Delta^{2q} b_{-q}}{(2q)!h^{2q}} \\
S_{2q+1} &= \frac{\Delta^{2q+1} b_{-q-1}}{(2q+1)!h^{2q+1}}
\end{aligned}$$

En utilisant le même changement de variable que celui déjà utilisé à propos des polynômes de Newton, nous obtenons une forme plus concise :

$$\begin{aligned}
P_n(z) = b_0 + z\Delta^1 b_{-1} + \frac{z(z+1)}{2!}\Delta^2 b_{-1} + \frac{z(z^2-1)}{3!}\Delta^3 b_{-2} \\
+ \frac{z(z^2-1)(z+2)}{4!}\Delta^4 b_{-2} + \frac{z(z^2-1)(z^2-4)}{5!}\Delta^5 b_{-3} + \cdots
\end{aligned} \tag{6.9}$$

À présent, écrivons le même polynôme (6.8) en fonction d'abscisses données pour une raison négative :

$$\begin{aligned}
P_n(x) = R_0 + R_1(x-a_0) + R_2(x-a_0)(x-a_0-h) \\
+ R_3(x-a_0+h)(x-a_0)(x-a_0-h) \\
+ \cdots + R_{2q}[x-a_0+(q-1)h]\ldots(x-a_0-qh) \\
+ R_{2q+1}(x-a_0+qh)\ldots(x-a_0-qh)
\end{aligned} \tag{6.10}$$

Sans entrer dans le détail puisque les opérations ont été explicitées à plusieurs reprises, on obtient les coefficients $R_k$ au moyen des expressions :

$$\begin{aligned}
R_0 &= b_0 \\
R_1 &= \frac{\Delta^1 b_0}{h} \\
R_2 &= \frac{\Delta^2 b_{-1}}{2!h^2} \\
&\cdots\cdots\cdots\cdots \\
R_{2q} &= \frac{\Delta^{2q} b_{-q}}{(2q)!h^{2q}} \\
R_{2q+1} &= \frac{\Delta^{2q+1} b_{-q}}{(2q+1)!h^{2q+1}}
\end{aligned}$$

En procédant au changement de variable en $z$ on peut écrire :

$$\begin{aligned}
P_n(z) = b_0 + z\Delta^1 b_0 + \frac{z(z-1)}{2!}\Delta^2 b_{-1} + \frac{z(z^2-1)}{3!}\Delta^3 b_{-1} \\
+ \frac{z(z^2-1)(z-2)}{4!}\Delta^4 b_{-2} + \frac{z(z^2-1)(z^2-4)}{5!}\Delta^5 b_{-2} + \cdots
\end{aligned} \tag{6.11}$$

Par définition le polynôme de Stirling est la demi-somme des expressions données par (6.9) et (6.10), lequel s'écrit simplement :

$$P_n(z) = b_0 + z\frac{\Delta^1 b_0 + \Delta^1 b_{-1}}{2} + \frac{z^2}{2!}\Delta^2 b_{-1} + \frac{z(z^2-1)}{3!}\frac{(\Delta^3 b_{-1} + \Delta^3 b_{-2})}{2} + \frac{z^2(z^2-1)}{4!}\Delta^4 b_{-2} + \frac{z(z^2-1)(z^2-4)}{5!}\frac{(\Delta^5 b_{-2} + \Delta^5 b_{-3})}{2} \cdots \quad (6.12)$$

Il convient de remarquer que l'on arrête l'expression du polynôme de Stirling au terme $\Delta^{2q} b_{-q}$ et non au terme précédent :

$$\frac{(\Delta^{2q-1} b_{-q+1} + \Delta^{2q-1} b_{-q})}{2},$$

tout simplement parce que si l'on connaît $\Delta^{2q-1} b_{-q+1}$ et $\Delta^{2q-1} b_{-q}$, on connaît $\Delta^{2q} b_{-q}$. On en déduit que le polynôme de Stirling est toujours un polynôme de degré pair et qu'il faut connaître $(2k+1)$ points pour un polynôme de degré $2k$.

Le polynôme de Stirling est utilisé pour effectuer les interpolations situées dans le deuxième quart du tableau lorsque, évidemment, on utilise une fraction seulement du tableau.

Dans les mêmes conditions, il nous reste à s'intéresser au troisième quart du tableau, ce sera chose faite après l'étude du polynôme de Bessel.

## 9. Le polynôme d'interpolation de Bessel (1784–1846)

Pour obtenir le polynôme de Bessel il suffit d'effectuer la demi-somme du polynôme (6.9) et du polynôme (6.11). Au préalable on réalise dans la relation (6.9) le changement de variable $z$ en $(z-1)$ qui correspond à un décalage des indices d'une unité, nous pouvons écrire :

$$P_n(z) = b_1 + (z-1)\Delta^1 b_0 + \frac{z(z-1)}{2!}\Delta^2 b_0 + \frac{(z-1)(z-2)}{3!}\Delta^3 b_{-1} + \frac{z(z^2-1)(z-2)}{4!}\Delta^4 b_{-1} + \cdots \quad (6.13)$$

le polynôme d'interpolation prend alors la forme :

$$P_n(z) = \frac{b_0 + b_1}{2} + (z-1/2)\Delta^1 b_0 + \frac{z(z-1)}{2!}\frac{(\Delta^2 b_0 + \Delta^2 b_{-1})}{2} + \frac{z(z-1)(z-1/2)}{3!}\Delta^3 b_{-1} + \frac{z(z^2-1)(z-2)}{4!}\frac{(\Delta^4 b_{-1} + \Delta^4 b_{-2})}{2} + \cdots$$

Ajoutons que nous devons arrêter le polynôme de Bessel à l'ordre $(2k+1)$, c'est en effet un polynôme impair qui demande $(2k+2)$ points pour sa définition.

## 10. Erreurs commises en utilisant les polynômes d'interpolation

Il suffit de réaménager la formule établie lors de l'étude du polynôme de Lagrange pour chacun des cas envisagés ici, c'est-à-dire l'expression de $E(x)$. Nous obtenons alors les expressions suivantes :

### 10.1. Le polynôme de Newton ascendant

$$|E_n(z)| < |z(z-1)\dots(z-n)| \frac{h^{n+1}}{(n+1)!} M_{n+1}. \tag{6.14}$$

### 10.2. Le polynôme de Newton descendant

$$|E_n(z)| < |z(z+1)\dots(z+n)| \frac{h^{n+1}}{(n+1)!} M_{n+1}. \tag{6.15}$$

### 10.3. Le polynôme de Stirling (polynôme pair)

$$|E_{2n+1}(z)| < \left|z(z^2-1)\dots(z^2-n^2)\right| \frac{h^{2n+1}}{(2n+1)!} M_{2n+1}. \tag{6.16}$$

### 10.4. Le polynôme de Bessel (polynôme impair)

$$|E_{2n+2}(z)| < \left|z(z^2-1)\dots(z^2-n)(z^2-n-1)\right| \frac{h^{2n+2}}{(2n+2)!} M_{2n+2}. \tag{6.17}$$

Dans ces quatre expressions $M_n$ conserve la même définition que celle qui a été explicitée au cours du paragraphe 3 de ce chapitre.

## 11. Programmes déterminant les polynômes d'interpolation

Nous avons donné sur le Web(*) les programmes `lagpoly.c`, `ascend.c` et `descend.c` qui permettent d'obtenir les formes de quelques polynômes rencontrés dans ce chapitre, soient le polynôme de Lagrange et les deux polynômes de Newton qui sont les formes les plus usitées (*cf.* la méthode d'Adams concernant l'intégration des équations différentielles).

## 12. Interpolation par les fonctions-spline

Replaçons-nous dans les conditions générales du début du chapitre du moins pour ce qui concerne les données. L'interpolation dans un tableau de $(n+1)$ points réalisée au moyen du polynôme de Lagrange de degré $n$ ou d'une des variantes due à Newton, Stirling ou Bessel, peut souvent conduire à des calculs assez longs et dont l'intérêt n'est pas toujours vraiment immédiat. Alors, on pourra préférer de réaliser les interpolations, par exemple, au moyen des fonctions-spline qui ne sont rien d'autre que $n$ polynômes de degré $q$ représentant les données expérimentales dans chacun des $n$ intervalles. Un polynôme et un seul représente la fonction dans un intervalle formé par deux abscisses consécutives $(a_k, a_{k+1})$.

Chaque polynôme ayant $(q+1)$ coefficients, on doit au total calculer $n(q+1)$ termes. Pour réaliser cette opération on impose les conditions suivantes :

* `http://www.edpsciences.com/guilpin/`

**a.** Chaque polynôme $Q_q^k(x)$ réalisant l'interpolation entre les points $a_k$ et $a_{k+1}$ passe par ces deux points.

**b.** Les dérivées premières, deuxièmes et ainsi de suite jusqu'à l'ordre $(q-1)$ des polynômes $Q_q^{k-1}(x)$ et $Q_q^k(x)$ se raccordent au point $a_k$, c'est-à-dire qu'elles sont égales chacune à chacune.

**c.** Comme nous avons imposé $[2n+(n-1)](q-1)$ conditions, il manque encore $(q-1)$ conditions pour fermer le système. Alors, on impose arbitrairement que les dérivées d'ordre le plus élevé concernant les deux polynômes extrêmes $Q_q^1(x)$ et $Q_q^n(x)$ s'annulent respectivement au point $a_0$ et au point $a_n$. Si $q$ est impair, les $(q-2)$ dernières dérivées de $Q_q^1(x)$ et $Q_q^n(x)$ s'annuleront respectivement au point $a_0$ et au point $a_n$. Si $q$ est pair le problème devient dyssymétrique et l'on choisira les $(q-2)/2$ dernières dérivées pour un polynôme et $q/2$ pour l'autre.

Ces généralités étant énoncées, nous allons fixer notre attention sur les fonctions-spline du troisième degré qui constituent de loin celles qui sont le plus fréquemment employées.

## 13. Les fonctions-spline du troisième degré

Il n'y a dans ce cas que deux conditions à ajouter pour que le système des équations linéaires soit fermé. Nous avons :

$$\begin{aligned} Q_k(x) &= \alpha_{k0} + \alpha_{k1}x + \alpha_{k2}x^2 + \alpha_{k3}x^3 \\ Q_k(a_{k-1}) &= b_{k-1} \\ Q_k(a_k) &= b_k \\ Q'_k(a_k) &= Q'_{k+1}(a_k) \\ Q''_k(a_k) &= Q''_{k+1}(a_k) \end{aligned}$$

plus les deux conditions arbitraires :

$$Q''_1(a_0) = 0 \qquad Q''_n(a_n) = 0.$$

En toute rigueur, il nous faut résoudre un système linéaire de $4n$ équations à $4n$ inconnues. Comme la matrice des coefficients est une matrice-bande, nous allons développer une méthode spécifique pour obtenir les paramètres inconnus. Il s'agit en fait de réduire le plus possible le nombre d'inconnues en exploitant convenablement les conditions que nous venons d'expliciter. Pour cela posons :

$$\begin{aligned} h_k &= a_k - a_{k-1} \\ M_q &= Q''_q(a_q) \quad \text{pour } q = 1, 2, \ldots (n-1). \end{aligned}$$

Nous allons exprimer les polynômes en fonction des $M_q$ qui sont pour l'instant des inconnues. Ces quantités vont être déterminées par un système linéaire que nous allons définir. Comme nous avons :

$$\begin{aligned} Q''_k(x) &= 2\alpha_{k2} + 6\alpha_{k3}x \\ Q''_k(a_k) &= M_k = 2\alpha_{k2} + 6\alpha_{k3}a_k \\ Q''_k(a_{k-1}) &= Q''_{k-1}(a_{k-1}) = M_{k-1=}2\alpha_{k2} + 6\alpha_{k3}a_{k-1} \end{aligned}$$

on en déduit que :

$$6\alpha_{k3} = \frac{M_k - M_{k-1}}{a_k - a_{k-1}}$$

et

$$2\alpha_{k2} = M_k - \frac{M_k - M_{k-1}}{a_k - a_{k-1}} a_k$$

puis

$$Q_k''(x) = M_k - \frac{M_k - M_{k-1}}{h_k} a_k + \frac{M_k - M_{k-1}}{h_k} x\,;$$

cette dernière expression se transforme en :

$$Q_k''(x) = \left[M_k a_k - M_k a_{k-1} - M_k a_k + M_{k-1} a_k + M_k x - M_{k-1} x\right] / h_k$$

d'où nous tirons

$$Q_k''(x) = \left[M_k (x - a_{k-1}) - M_{k-1}(x - a_k)\right] / h_k.$$

Maintenant, nous allons intégrer successivement deux fois cette expression puis déterminer les deux constantes d'intégration au moyen des deux conditions :

$$Q_k(a_{k-1}) = b_{k-1} \qquad \text{et} \qquad Q_k(a_k) = b_k.$$

Nous pouvons alors écrire :

$$h_k Q_k(x) = M_k \frac{(x - a_{k-1})^3}{6} - M_{k-1} \left[\frac{(x - a_k)^3}{6}\right] + Kx + L,$$

d'où

$$K = b_k - b_{k-1} - (M_k - M_{k-1}) \frac{h_k^2}{6}$$

et

$$L = h_k b_k - M_k \frac{h_k^3}{6} - (b_k - b_{k-1}) a_k + (M_k - M_{k-1}) \frac{h^{2k}}{6} a_k.$$

Nous obtenons alors :

$$Kx + L = b_k x - b_{k-1} x - M_k \frac{h_k^2}{6} x + M_{k-1} \frac{h_k^2}{6} x + h_k b_k - M_k \frac{h_k^3}{6} - a_k b_k + b_{k-1} a_k + M_k \frac{h_k^2}{6} a_k - M_{k-1} \frac{h_k^2}{6} a_k.$$

En regroupant les termes et en tenant compte du fait que $h_k$ est donné par l'expression $h_k = a_k - a_{k-1}$, on obtient :

$$Kx + L = \left(b_k - M_k \frac{h^{2k}}{6}\right)(x - a_{k-1}) + \left(b_{k-1} - M_{k-1} \frac{h_k^2}{6}\right)(a_k - x)\,;$$

d'où nous pouvons écrire :

$$h_k Q_k(x) = M_k \frac{(x-a_{k-1})^3}{6} - M_{k-1}\frac{(x-a_k)^3}{6} + \left(b_k - M_k\frac{h^{2_k}}{6}\right)(x-a_{k-1}) + \left(b_{k-1} - M_{k-1}\frac{h^{2_k}}{6}\right)(a_k - x)\,;$$

Par cette succession de transformations nous avons réduit le nombre d'inconnues du système linéaire, et ce sont les $M_k$ qui deviennent les nouvelles inconnues au lieu des coefficients $\alpha_{k0}$, $\alpha_{k1}$, $\alpha_{k2}$ et $\alpha_{k3}$. Pour calculer effectivement les $M_k$, nous devons utiliser le fait que les polynômes $Q_k(x)$ et $Q_{k+1}(x)$ doivent avoir leurs dérivées premières qui se raccordent au point $a_k$. Nous obtenons ainsi :

$$M_k\frac{h_k}{2} - \frac{b_{k-1}}{h_k} + h_k\frac{M_{k-1}}{6} + \frac{b_k}{h_k} - h_k\frac{M_k}{6} = M_k\frac{h_{k+1}}{2} - \frac{b_k}{h_{k+1}} + h_{k+1}\frac{M_k}{6} + \frac{b_{k+1}}{h_{k+1}} - h_{k+1}\frac{M_{k+1}}{6}\,.$$

Regroupons les termes en $M_k$ ordonnés selon les indices croissants :

$$\frac{h_k}{6}M_{k-1} + \frac{h_k + h_{k+1}}{3}M_k + \frac{h_{k+1}}{6}M_{k+1} = \frac{b_{k+1}-b_k}{h_{k+1}} - \frac{b_k - b_{k-1}}{h_k}\,.$$

Puisque les polynômes $Q_0(x)$ et $Q_1(x)$ sont déterminés par $M_0 = M_n = 0$, on obtient donc un système de $(n-1)$ inconnues en faisant varier $k$ de 1 à $(n-1)$.

À présent, il reste à résoudre un système linéaire dont le premier membre dépend d'une matrice tridiagonale.

## 14. Résolution d'un système linéaire dépendant d'une matrice tridiagonale

Comme la plupart des éléments de la matrice sont nuls, il est préférable d'utiliser une méthode adéquate pour traiter ce système linéaire. En effet l'utilisation d'un programme général serait mal venue car la plus grande partie du temps de calcul serait consacrée à l'obtention de résultats nuls. Donc nous allons présenter une méthode adaptée au problème, méthode d'autant plus intéressante qu'elle est généralisable au cas d'une matrice pentadiagonale par exemple.

D'une façon tout à fait formelle, écrivons un système linéaire qui dépend d'une matrice tridiagonale :

$$\begin{vmatrix} m_{11} & m_{12} & 0 & 0 & 0 & \dots & 0 \\ m_{21} & m_{22} & m_{23} & 0 & 0 & \dots & 0 \\ 0 & m_{32} & m_{33} & m_{34} & 0 & \dots & 0 \\ \dots & & & & & & \\ 0 & 0 & 0 & m_{n-1,n-2} & m_{n-1,n-1} & & m_{n-1,n} \\ 0 & 0 & 0 & 0 & m_{n,n-1} & & m_{n,n} \end{vmatrix} \cdot \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ \\ x_{n-1} \\ x_n \end{vmatrix} = \begin{vmatrix} c_1 \\ c_2 \\ c_3 \\ \\ c_{n-1} \\ c_n \end{vmatrix}$$

système que l'on peut écrire sous la forme condensée :

$$MX = C.$$

On traitera plus avantageusement le tableau $T$ à $n$ lignes et $(n+1)$ colonnes obtenu en juxtaposant à la matrice $M$ le vecteur $C$.

Une transformation linéaire tout à fait analogue à celle utilisée dans la méthode des pivots permet de passer du tableau $T$ au tableau $T^*$ par suppression d'une diagonale de la matrice $M$, tous les éléments de la diagonale principale prenant la valeur 1 (*cf.* chapitre 5). Pour parvenir à notre fin, nous allons diviser la première ligne du tableau $T$ par l'élément diagonal $m_{11}$. Ensuite, il suffit de multiplier la première ligne par $m_{21}$ et de retrancher le résultat de la deuxième ligne. On divise alors cette deuxième ligne par l'élément situé sur la diagonale. On pourrait penser qu'il vaut mieux commencer par effectuer la division par le pivot avant d'effectuer la combinaison linéaire qui élimine $m_{21}$, mais dans ce dernier cas, on réalise une opération en trop qui est $m_{21}/m_{22}$. En effet, il est inutile de diviser $m^*_{21}$ par $m^*_{22}$ après la réalisation de la combinaison linéaire car $m^*_{21}$ est nul. Cette remarque fait gagner $(n-1)$ opérations ce qui n'est pas négligeable lors de procédures répétitives où le temps d'exécution peut devenir important.

En poursuivant de ligne en ligne la même façon d'opérer, on aboutira nécessairement au tableau $T^*$ qui aura la forme suivante :

$$\left|\begin{array}{ccccccc} 1 & p_{12} & 0 & 0 & & & d_1 \\ 0 & 1 & p_{23} & 0 & & & d_2 \\ 0 & 0 & 1 & p_{34} & & & d_3 \\ \cdots & \cdots & \cdots & & & & \\ 0 & 0 & 0 & 0 & 1 & p_{n-1,n} & d_{n-1} \\ 0 & 0 & 0 & 0 & 0 & 1 & d_n \end{array}\right| .$$

La dernière ligne donne directement la valeur de la dernière inconnue. La ligne $(n-1)$ permet d'obtenir alors l'inconnue $(n-1)$ par substitution et ainsi de suite en remontant jusqu'à la première ligne du tableau. En considérant le problème de résolution d'un système linéaire sous cet angle particulier, on aboutit à une économie considérable de mémoire et de temps de calcul.

Sur le Web(*), on trouvera le programme `lispline.c` qui réalise l'interpolation par des fonctions-spline du troisième degré.

## 15. Une application simple des polynômes d'interpolation

L'intégration formelle d'un polynôme est une des rares opérations que l'on sache réaliser rigoureusement. Il n'y a pas de difficulté particulière donc à utiliser les polynômes d'interpolation pour réaliser les opérations de quadrature.

En particulier, si l'on utilise les fonctions-spline du troisième degré,on aboutira à une valeur numérique dont la précision est analogue à celle donnée par la méthode de Simpson (1710–1761) puisque, dans les deux cas, il s'agit d'interpoler une fonction par une parabole cubique.

Nous fournissons sur le Web(*) le sous-programme `aire.h` qui assure cette intégration. Nous donnons les étapes du calcul de l'expression qui est programmée. Partons de l'expression générale du polynôme d'ordre $k$ :

$$P_k(x) = M_k\frac{(x-a_{k-1})^3}{6h_k} - M_{k-1}\frac{(x-a_k)^3}{6h_k} + \left(b_k - M_k\frac{h^{2_k}}{6}\right)\frac{(x-a_{k-1})}{h_k} + \left(b_{k-1} - M_{k-1}\frac{h^{2_k}}{6}\right)\frac{(a_k-x)}{h_k} .$$

* `http://www.edpsciences.com/guilpin/`

Désignons par $S_k$ la surface comprise entre $a_{k-1}$ et $a_k$, il vient :

$$S_k = \int_{a_{k-1}}^{a_k} P_k(x)\,\mathrm{d}x = \left\{ M_{k-1}\frac{(x-a_{k-1})^4}{24h_k} - M_{k-1}\frac{(a_k-x)^4}{24h_k} \right.$$
$$\left. - \left(b_{k-1} - M_{k-1}\frac{h^{2k}}{6}\right)\frac{(a_k-x)^2}{2h_k} + \left(b_k - M_k\frac{h_k^2}{6}\right)\frac{(x-a_{k-1})^2}{2h_k} \right\}_{a_{k-1}}^{a_k}.$$

Le développement de cette dernière expression conduit à la formule simple :

$$S_k = -\frac{h_k^3}{24}(M_k + M_{k-1}) + \frac{h_k}{2}(b_k + b_{k-1})$$

là on tire :

$$S = \sum_{k=1}^{n} S_k,$$

expression dans laquelle $S$ est l'intégrale totale. Dans le cas fréquent où les abscisses sont en progression arithmétique, nous pouvons poser $h = h_k$ quel que soit l'indice $k$, on arrive alors à une expression encore plus simple de $S$ :

$$S = h\left(\sum_{k=0}^{n} b_k - \frac{b_0 + b_n}{2}\right) - \frac{h^3}{12}\left(\sum_{k=0}^{n} M_k + \frac{M_n + M_0}{2}\right).$$

## 16. L'algorithme d'interpolation d'Aitken (1932)

Il s'agit d'une procédure — encore appelée algorithme de Neville-Aitken — destinée au calcul d'un polynôme pour une valeur particulière alors que le dit polynôme est connu par $(n+1)$ points $(x_k, y_k)$ avec $k = 0, 1, 2, \ldots, n$. C'est une méthode directe qui permet d'obtenir donc une valeur unique du polynôme, en ce sens elle est analogue au polynôme de Lagrange.

Dans ce paragraphe, on désigne par $P_{p\ldots q}$ le polynôme de degré $(q-p-1)$ qui passe par les points $(x_p, y_p)\ldots(x_q, y_q)$ et par $P_{p,\ldots,\langle\pi\rangle,\ldots q}$ le polynôme de degré $(p-q-2)$ qui passe par les points $(x_p, y_p)\ldots(x_{kq}, y_q)$ à l'exception du point $(x_\pi, y_\pi)$.

Avant même de rentrer dans le cœur du sujet, il est possible d'ores et déjà de remarquer que l'ordre des points ne joue aucun rôle quant à la nature des polynômes.

### 16.1. Relation de récurrence entre les polynômes P

Entre les polynômes $P_{p,\ldots,q}$, $P_{p,\ldots,\langle\pi\rangle,\ldots q}$ et $P_{p,\ldots,\langle\sigma\rangle,\ldots q}$ il existe la relation de récurrence suivante :

$$(x_\pi - x_\sigma)P_{p,\ldots q} = \begin{vmatrix} (x_\sigma - x) & P_{p,\ldots,\langle\pi\rangle,\ldots q} \\ (x_\pi - x) & P_{p,\ldots,\langle\sigma\rangle,\ldots q} \end{vmatrix}; \tag{6.18}$$

où, pour plus de clarté, nous précisons que le second membre est un déterminant.

La démonstration de cette relation repose sur le fait que, pour une valeur $x_\nu$ différente de $x_\pi$ et $x_\sigma$, on a :

$$P_{p\ldots q}(x_\nu) = P_{p,\ldots,\langle\sigma\rangle,\ldots q^{(x_\nu)}} = P_{p,\ldots,\langle\pi\rangle,\ldots q^{(x_\nu)}} = y_\nu.$$

La relation (6.18) peut donc s'écrire :

$$(x_\pi - x_\sigma)y_\nu = (x_\pi - x_\nu)y_\nu - (x_\sigma - x_\nu)y_\nu = (x_\pi - x_\sigma)y_\nu.$$

Pour le cas où $x_\pi = x_\nu$, nous pouvons écrire :

$$(x_\pi - x_\sigma)y_\pi = -(x_\sigma - x_\pi)y_\pi.$$

et dans le cas où $x_\nu = x_\sigma$, nous obtenons encore l'identité des deux membres de (6.18) :

$$(x_\pi - x_\sigma)y_\sigma = -(x_\pi - x_\sigma)y_\sigma.$$

On en conclut que l'égalité (6.18) est une relation qui permet de calculer le polynôme $P_{p,\ldots,q}$ de degré $(q-p-1)$ à partir des deux polynômes de degré $(q-p-1)$ qui sont $P_{p,\ldots,\langle\pi\rangle,\ldots q}$ et $P_{p,\ldots,\langle\sigma\rangle,\ldots q}$.

## 16.2. Calcul numérique de $y_\mu$

En faisant usage de la relation (6.18), nous allons calculer le polynôme $P_{0,1,\ldots n}$ pour la valeur $x_\mu$ en formant une suite convenable de polynômes intermédiaires $P_{p,\ldots,q}$ et $P_{p,\ldots,\langle\pi\rangle,\ldots q}$. À ce propos nous étudierons deux méthodes, l'une due à Aitken l'autre due à Neville laquelle n'est qu'une modification de la méthode d'Aitken.

***a – Méthode d'Aitken*** – Pour résoudre le problème des polynômes intermédiaires, Aitken a proposé la disposition suivante :

$$\begin{array}{cccccccccc}
 & & & x_0 - x_\mu & & y_0 & & & & \\
 & & & x_0 - x_1 & x_1 - x_\mu & y_1 & P_{01} & & & \\
 & & x_1 - x_2 & x_0 - x_2 & x_2 - x_\mu & y_2 & P_{02} & P_{012} & & \\
 & x_2 - x_3 & x_1 - x_3 & x_0 - x_3 & x_3 - x_\mu & y_3 & P_{03} & P_{013} & P_{0123} & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
x_{\nu-1} - x_\nu & \ldots & x_1 - x_\nu & x_0 - x_\nu & x_\nu - x_\mu & y_\nu & P_{0n} & P_{01n} & P_{012n} & P_{01\ldots n}
\end{array}$$

À partir de la disposition des $y_k$ et des $(x_\alpha - x_\beta)$ on va calculer les polynômes $P_{01}, P_{02}, \ldots, P_{0n}$ à l'aide de la relation :

$$P_{0j} = \frac{1}{(x_0 - x_j)}[(x_0 - x_\mu)y_j - (x_j - x_\mu)y_0] \text{ avec } j = 1, 2, 3, \ldots n.$$

Ensuite on calculera la colonne suivante dans le tableau, c'est-à-dire $P_{012}, P_{013}, \ldots, P_{01n}$ au moyen de l'expression :

$$P_{01j} = \frac{1}{(x_0 - x_j)}[(x_1 - x_\mu)P_{0j} - (x_j - x_\mu)P_{01}] \text{ avec } j = 1, 2, 3, \ldots n.$$

On procédera ainsi de suite jusqu'à l'obtention de l'élément $P_{01\ldots n}$ qui constitue, en définitive, la valeur recherchée $y_\mu$.

***Remarque 1*** : Pour obtenir une précision optimum, on peut se poser la question de savoir quelle doit être la distribution initiale des points car nous sommes libre de numéroter les points comme bon nous semble.

Dans le cas d'une extrapolation, on adoptera l'ordre des points allant du plus près de $x_m$ au plus éloigné.

S'il s'agit d'une interpolation, on choisira une distribution initiale qui correspond à un encadrement du point $x_m$ de telle sorte que l'on réalisera la disposition suivante :

$$\ldots < x_{2k} < \ldots < x_2 < x_0 < x_m < x_1 < x_3 < \ldots < x_{2k+1} \ldots$$

Autrement dit, comme le montre le calcul concernant l'interpolation, ce sont les points les plus proches voisins de $x_m$ qui jouent un rôle prépondérant, donc il convient tout simplement de faire jouer un rôle prioritaire à ces points dans les calculs.

***Remarque 2*** : Il n'est donc pas nécessaire d'utiliser les points très éloignés de $x_m$ dans le calcul de $y_m$ car ils n'auront que très peu d'influence sur le résultat.

Du reste, si au cours du calcul, on rencontre deux termes $P_{01\ldots k}$ et $P_{01\ldots k+1}$ qui sont chacun le dernier élément de deux lignes consécutives du tableau proposé par Aitken et qui sont égaux « *à la précision de la machine* », il n'est pas utile de poursuivre les calculs jusqu'à l'obtention de $P_{01\ldots n}$.

***b – Méthode de Neville*** – Neville a proposé une autre disposition du tableau d'Aitken qui est reportée ci-dessous :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $x_0 - x_\mu$ | $y_0$ | | | |
| $x_0 - x_1$ | | | | $x_1 - x_\mu$ | $y_1$ | $P_{01}$ | | |
| $x_0 - x_2$ | | $x_1 - x_2$ | | $x_2 - x_\mu$ | $y_2$ | $P_{12}$ | $P_{012}$ | |
| $x_0 - x_3$ | | $x_1 - x_3$ | $x_2 - x_3$ | $x_3 - x_\mu$ | $y_3$ | $P_{23}$ | $P_{123}$ | $P_{0123}$ |
| ..... | | | | | | | | |
| $x_0 - x_n$ | ... | $x_{n-2} - x_n$ | $x_{n-1} - x_n$ | $x_n - x_\mu$ | $y_n$ | $P_{n-1n}$ | $P_{012n}$ | $P_{01\ldots n}$ |

## 16.3. Exemple numérique d'interpolation par la méthode d'Aitken

On se propose d'effectuer une interpolation dans la table suivante :

| $x$ | $y$ |
|---|---|
| 0,0 | 1,000 0 |
| 0,5 | 0,938 5 |
| 1,0 | 0,765 2 |
| 1,5 | 0,511 8 |
| 2,0 | 0,223 9 |
| 2,5 | −0,048 4 |
| 3,0 | −0,260 1 |
| 3,5 | −0,380 1 |

On reconnaîtra là un extrait de la table de la fonction de Bessel de première espèce $J_0(x)$. On désire effectuer l'interpolation pour la valeur $x = 1,70$. La table donne $y = 0,398\,0$. On trouvera sur le Web (*) le programme `p_aitken.c` réalisant la procédure d'Aitken, lequel fournit pour la valeur d'interpolation 0,397 97.

* `http://www.edpsciences.com/guilpin/`

## 17. Approximation par une combinaison linéaire de fonctions

Les polynômes de degré $n$ constituent une première méthode pour approcher une fonction $f(x)$ lorsque nous en connaissons $(n+1)$ points. Cependant, dans bien des situations concrètes, l'échantillonnage à traiter est obtenu à partir de données expérimentales entachées d'erreur. Il devient illusoire de vouloir trouver un polynôme passant par chacun des points obtenus, car il n'y a en général aucun intérêt à vouloir rendre compte de l'aspect aléatoire des phénomènes étudiés... Dans ce cas précis, on dispose de considérations à caractère théorique qui permettent de pressentir une fonction ou une combinaison linéaire de fonctions qui passe au voisinage de tous les points d'échantillonnage. Bien entendu, il convient de définir ce qu'on entend par voisinage, mais au préalable, il faut distinguer les types d'approximation que nous venons d'évoquer : soit nous désirons connaître les coefficients figurant dans une certaine fonction, plus ou moins compliquée, censée rendre compte d'une expérience, soit nous cherchons à déterminer les valeurs des coefficients d'une forme linéaire destinée à « lisser » les points expérimentaux.

Le premier problème est généralement non linéaire et pose un certain nombre des difficultés techniques sur le plan du calcul numérique. Le second se résout tout naturellement au moyen du calcul matriciel que nous avons évoqué au début de ce chapitre. C'est cet aspect qui va retenir notre attention, et nous nous proposons de « passer le plus près possible de tous les points » au moyen d'une combinaison linéaire de fonctions $\Phi_j(x)$ dont le nombre $(m+1)$ est inférieur (ou égal) au nombre $(n+1)$ de points.

*A priori*, il y a bien des moyens de « passer le plus près possible de tous les points », et il est aisé de subodorer qu'il existe une infinité de solutions au sens mathématique du terme. À ce stade, il est nécessaire de préciser ses désirs, et c'est la méthode des moindres carrés qui se présente en excellente position pour résoudre le problème. Désignons par :

$$\Theta(x) = \sum_{j=0}^{m} \alpha_j \Phi_j(x)$$

notre combinaison linéaire. La méthode des moindres carrés consiste à rendre minimum la forme quadratique :

$$\sum_{k=0}^{n} \left( \sum_{j=0}^{m} \alpha_j \Phi_j(a_k) - b_k \right)^2 .$$

On dit aussi que l'on rend la somme des carrés des résidus la plus petite possible, car la différence $\varepsilon_k = \sum_{j=0}^{m} \alpha_j \Phi_j(a_k) - b_k$ s'appelle résidu d'observation. La dérivation par rapport aux coefficients $\alpha_m$ qui rend minimum la forme quadratique donne immédiatement une forme linéaire qui détermine précisément les coefficients $\alpha_j$.

Nous cherchons donc à rendre minimum la quantité :

$$E^2 = \sum_{j=0}^{m} \varepsilon_k^2 = \sum_{k=0}^{n} \left( \sum_{j=0}^{m} \alpha_j \Phi_j(a_k) - b_k \right)^2 ,$$

pour cela on écrit que les $m$ équations s'annulent, soit :

$$\frac{\partial E^2}{\partial \alpha_l} = 0 = 2 \sum_{k=0}^{n} \Phi_l(a_k) \left( \sum_{j=0}^{m} \alpha_j \Phi_j(a_k) - b_k \right),$$

cette dernière expression peut encore se transformer avantageusement de la façon suivante :

$$0 = \sum_{k=0}^{n} \left( \sum_{j=0}^{m} \alpha_j \Phi_l(a_k) \Phi_j(a_k) - \Phi_l(a_k) b_k \right) \quad \text{avec } l = 0, 1, 2, \ldots, m.$$

$$\sum_{j=0}^{m} \alpha_j \sum_{k=0}^{n} \Phi_l(a_k) \Phi_j(a_k) = \sum_{k=0}^{m} \Phi_l(a_k) b_k \quad \text{avec } l = 0, 1, 2, \ldots, m.$$

Nous obtenons alors un système linéaire de $(m + 1)$ équations à $(m + 1)$ inconnues dont la résolution nous fournira les coefficients $\alpha_j$.

***Remarque*** : Dans le cas où la fonction $f(x)$ que l'on approche par la fonction $\Theta(x)$ est connue analytiquement, alors les sommes discrètes sont remplacées par des intégrales sur l'intervalle d'interpolation $I$. Il s'ensuit que

$$E^2 = \int_a^b \varepsilon^2(x)\, \mathrm{d}x \qquad \text{avec } \varepsilon(x) = f(x) - \sum_{j=1}^{m} \alpha_j \Phi_j(x).$$

Les coefficients $\alpha_j$ sont alors déterminés par les $(m + 1)$ équations suivantes :

$$\frac{\partial E^2}{\partial \alpha_l} = \sum_{j=1}^{m} \alpha_j \int_a^b \Phi_l(x) \Phi_j(x)\, \mathrm{d}x - \int_a^b \Phi_l(x) f(x)\, \mathrm{d}x = 0,$$

avec $l = 0, 1, 2, \ldots, m$. Le calcul de ces intégrales peut être éventuellement conduit numériquement dans le cas où les quadratures n'existent pas, ou pour le moins ne sont pas évidentes.

## Applications immédiates

***a*** – Il est aisé de rechercher un polynôme de degré strictement inférieur à $n$ qui passe le plus près possible de tous les points au sens des moindres carrés. Généralement on parle de régression parabolique au moyen d'un polynôme de degré $m$ ; le cas le plus connu étant le polynôme de degré un qui est encore appelé **droite de régression** en hommage au travaux de Galton (1822–1911) qui introduisit cette dénomination à l'occasion d'études statistiques concernant la biologie et l'eugénique.

***b*** – Système linéaire surdéterminé de $(m+1)$ équations à $(n+1)$ inconnues $m$ étant plus grand ou égal à $n$. On écrit le système :

$$\sum_{j=0}^{m} a_{lk} x_k + \alpha_k = 0, \qquad \text{pour } l = 0, 1, 2, \ldots, n.$$

Sans trop de difficulté on obtient le **système des équations normales** associé au système surdéterminé de $(n + 1)$ équations à $(m + 1)$ inconnues. Il suffit de remplacer $\Phi_l(x)$ par $x_l$ dans la relation précédemment établie ; on obtient alors :

$$\sum_{j=0}^{m} \alpha_j \sum_{k=0}^{n} a_{lk} a_{jk} = \sum_{j=0}^{m} a_{lk} b_k \qquad \text{avec } l = 0, 1, 2, \ldots, m.$$

Il est facile de montrer que si le système linéaire surdéterminé s'écrit au moyen de la notation matricielle :

$$\mathbf{A}X = b,$$

$\mathbf{A}$ étant une matrice de $n$ lignes et de $m$ colonnes (avec $m \leq n$), on obtient le système des équations normales en effectuant l'opération suivante :

$$\left[\mathbf{A}^T\mathbf{A}\right] X = \left[\mathbf{A}^T\right] b,$$

expression dans laquelle $\mathbf{A}^T$ est la matrice transposée de la matrice $\mathbf{A}$. Au passage, on notera que la matrice obtenue $\left[\mathbf{A}^T\mathbf{A}\right]$ est une forme quadratique qui est définie positive — la matrice est alors symétrique, ce qui est intéressant lors de la mise au point du programme correspondant.

Sur le Web (*), on trouvera le programme `surdeter.c` qui réalise cet algorithme lequel fournit un calcul d'erreur au sens statistique du terme : il s'agit de l'écart type de chacune des inconnues. Ce dernier calcul est explicité au chapitre de la régression linéaire.

## 18. Éléments de bibliographie

A. ANGOT (1972) *Compléments de Mathématiques*, Éditions Masson.
N. BAKHVALOV (1976) *Méthodes Numériques*, Éditions MIR.
J.Ch. FIOROT et P. JEANNIN (1992) *Courbes splines rationnelles*, Masson.
P. HENRICI (1964) *Elements of Numerical Analysis*, Wiley.
F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.
R.A. LORENTZ (1992) *Multivariate Birkhoff interpolation*, Springer-Verlag.
D. MC CRACKEN et W. DORN (1964) *Numerical Methods and Fortran Programming*, Wiley.
H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.
H. MINEUR (1938) *Technique de la méthode des moindres carrés*, Éditions Gauthier-Villars.
A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.
L.A. SAKHNOVICH (1997) *Interpolation theory and its applications*, Kluwer Academic Publishers.

* `http://www.edpsciences.com/guilpin/`

# 7 Les polynômes de Legendre. Méthode d'intégration de Gauss-Legendre

Ces polynômes sont nés d'une étude que Legendre (1752–1833) entreprit à propos de l'étude de l'attraction newtonienne en $1/r^2$. MacLaurin et Clairault (1713–1765) ont démontré que la figure d'équilibre d'une masse fluide en rotation dont les particules s'attirent mutuellement selon la loi de Newton est un ellipsoïde de révolution. MacLaurin détermina aussi la valeur de l'attraction exercée par un ellipsoïde homogène sur un point situé en son intérieur et sur sa surface. En 1783, Legendre étendit le calcul de l'attraction exercée sur un point situé à l'extérieur de l'ellipsoïde. La solution repose sur l'introduction d'un ellipsoïde homofocal et d'une fonction génératrice des fameux polynômes de Legendre. Clairaut introduisit le premier la notion de potentiel pour ce genre de problèmes et en 1785 Laplace montra que le potentiel obéissait à l'équation aux dérivées partielles qui porte aujourd'hui son nom. Les polynômes de Legendre rentrent, comme cas particulier, dans la classe des fonctions sphériques introduites par Laplace en 1785.

Aujourd'hui, c'est en électrostatique que l'on se familiarise avec les polynômes de Legendre, mais il faut rappeler que la loi de Coulomb (1736–1806) est postérieure de deux ans à leur avènement.

Nous connaissons l'intérêt qu'il y a de définir une base (complète) de fonctions orthogonales dans le problème d'approximation de fonctions par des combinaisons linéaires. Les polynômes de Legendre ont la propriété d'être orthogonaux et de constituer un système complet sur $(-1, +1)$ pour les fonctions généralement continues, ainsi s'ils sont utilisés comme base de décomposition, les coefficients de la combinaison linéaire ne sont pas corrélés, c'est-à-dire qu'ils sont calculés une fois pour toutes, comme le sont les coefficients des séries de Fourier avec lesquels nous sommes davantage familiarisés.

Les polynômes de Legendre ont connu rapidement une autre fortune entre les mains de Gauss (1777–1855) qui a utilisé leur propriété d'orthogonalité en proposant une méthode d'intégration numérique particulièrement précise. La technique opératoire est d'une très grande simplicité, en revanche, l'établissement des relations exige quelques calculs, et c'est le but de ce chapitre que de proposer les enchaînements mathématiques propres à éclairer la méthode.

## 1. Les polynômes de Legendre

### 1.1. Définition

Les polynômes de Legendre sont des polynômes orthogonaux sur l'intervalle fondamental $(-1, +1)$ qui prennent la valeur $+1$ pour la valeur $+1$ de la variable.

Désignons par $P_n(x)$ le polynôme de Legendre de degré $n$, la définition impose que :

$$\int_{-1}^{+1} P_n(x)P_m(x)\,\mathrm{d}x = 0 \tag{7.1}$$

si $P_m(x)$ est le polynôme de Legendre de degré $m$ tel que $m \neq n$.

Si l'on remarque qu'un polynôme quelconque $Q_p(x)$ de degré $p$ peut s'exprimer selon une combinaison linéaire de polynômes de Legendre de degré égal et inférieur à $p$ (*cf.* paragraphe 1.14), alors, pour $p < n$ on aura également :

$$\int_{-1}^{+1} P_n(x)Q_p(x)\,\mathrm{d}x = 0. \tag{7.2}$$

Cette dernière relation va nous permettre d'exprimer les polynômes de Legendre d'une autre façon en procédant à une intégration par parties. Pour cela, on posera :

$$P_n(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n} R_{2n}(x)$$

où $R_{2n}(x)$ est un polynôme de degré $2n$ que nous allons déterminer. On obtient alors :

$$\int_{-1}^{+1} P_n(x)\cdot Q_p(x)\,\mathrm{d}x = \Big[Q_p(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}R_{2n}(x) - \frac{\mathrm{d}}{\mathrm{d}x}Q_p(x)\frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}}R_{2n}(x) + \cdots + (-1)^n \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}Q_p(x)R_{2n}(x)\Big]_{-1}^{+1}$$

Il convient de remarquer que le polynôme $R_{2n}(x)$ est défini à un polynôme de degré $(n-1)$ près ; autrement dit, il nous est possible de choisir arbitrairement $n$ constantes définissant $R_{2n}(x)$, et l'on peut par conséquent convenir que les $(n-1)$ premières dérivées de $R_{2n}(x)$ ainsi que $R_{2n}(x)$ s'annulent pour $x = -1$. Cela nous conduit à écrire que :

$$\Big[Q_p(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}R_{2n}(x) - \frac{\mathrm{d}}{\mathrm{d}x}Q_p(x)\frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}}R_{2n}(x) + \cdots + (-1)^n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}Q_p(x)R_{2n}(x)\Big]_{x=-1} = 0,$$

cependant pour que l'intégrale (7.2) soit nulle, il faut encore que :

$$\Big[Q_p(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}R_{2n}(x) - \frac{\mathrm{d}}{\mathrm{d}x}Q_p(x)\frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}}R_{2n}(x) + \cdots + (-1)^n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}Q_p(x)R_{2n}(x)\Big]_{x=1} = 0.$$

Cette condition ne peut être satisfaite, quel que soit le polynôme arbitraire $Q_p(x)$, que lorsque $R_{2n}(x)$ et ses $(n-1)$ premières dérivées s'annulent pour $x = 1$.

À présent nous avons imposé $n$ conditions au polynôme $R_{2n}(x)$ ; il reste cependant défini à une constante multiplicative près que l'on désigne par $K$. On peut donc l'exprimer sous la forme :

$$R_{2n}(x) = K(x-1)^n(x+1)^n = K(x^2-1)^n.$$

Il y a plusieurs façons « canoniques » de choisir $K$. On peut adopter les polynômes orthonormés, on peut préférer la condition $P_n(1) = 1$, mais l'on peut tout aussi bien s'intéresser aux polynômes à coefficient principal réduit, c'est-à-dire ceux dont le coefficient du terme de degré

le plus élevé est égal à un. Cela est une affaire de circonstances et nous aurons l'occasion de revenir sur ce point. Quoi qu'il en soit, puisque

$$P_n(x) = K\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x-1)^n(x+1)^n,$$

on peut calculer $K$ au moyen de la formule de Leibnitz (1646–1716) qui fournit l'expression de la dérivée $n^{\mathrm{e}}$ d'un produit $(u \cdot v)$ :

$$\frac{\mathrm{d}^n(uv)}{\mathrm{d}x^n} = v\frac{\mathrm{d}^n u}{\mathrm{d}x^n} + \cdots C_n^q \frac{\mathrm{d}^{n-q}u}{\mathrm{d}x^{n-q}}\frac{\mathrm{d}^q v}{\mathrm{d}x^q} + \cdots + u\frac{\mathrm{d}^n v}{\mathrm{d}x^n},$$

il suffit de poser $u = (x-1)^n$ et $v = (x+1)^n$ pour obtenir la valeur de $K$. En effet, pour $x = 1$, tous les termes sont nuls sauf le premier et l'on obtient alors :

$$P_n(1) = 1 = Kv\frac{\mathrm{d}^n}{\mathrm{d}x^n}u = Kn!2^n$$

ce qui donne :

$$K = \frac{1}{2^n n!} = \frac{1}{2 \cdot 4 \cdot 6 \ldots (2n)},$$

et l'on obtient alors l'expression générale des polynômes de Legendre connue sous le nom de formule de Rodriguès (1815) :

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n. \tag{7.3}$$

On déduit sans peine les premiers polynômes :

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= (3x^2-1)/2 \\ &\ldots \end{aligned}$$

## 1.2. Relation de récurrence entre trois polynômes consécutifs

Nous allons montrer qu'il existe une relation de récurrence entre trois polynômes consécutifs $P_{n+1}(x)$, $P_n(x)$ et $P_{n-1}(x)$. Pour cela partons de la formule de Rodriguès (1794–1851) écrite pour le degré $(n+1)$ :

$$P_{n+1}(x) = \frac{1}{2^{n+1}}\frac{1}{(n+1)!}\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}}(x^2-1)^{n+1}.$$

On dérive une fois l'expression $(x^2-1)^{n+1}$, et l'on obtient :

$$P_{n+1}(x) = \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{x(x^2-1)^n\right\}$$

On peut encore écrire :

$$P_{n+1}(x) = \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{(x^2-1)^n + 2nx^2(x^2-1)^{n-1}\right\}$$

puis on ajoute et l'on retranche $2n(x^2-1)^{n-1}$ dans la dérivée du second terme :

$$\begin{aligned} P_{n+1}(x) &= \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[(x^2-1)^n + 2nx^2(x^2-1)^{n-1} - 2n(x^2-1)^{n-1} + 2n(x^2-1)^{n-1}\right] \\ &= \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[(2n+1)(x^2-1)^n + 2n(x^2-1)^{n-1}\right], \end{aligned}$$

soit encore :

$$P_{n+1}(x) = \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[(2n+1)(x^2-1)^n\right] + P_{n-1}(x).$$

Calculons le premier terme du second membre en utilisant la relation découlant de la règle de Leibnitz, cette quantité est égale à :

$$\begin{aligned} \frac{2n+1}{2^n n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[(x^2-1)^n\right] &= \frac{2n+1}{n2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[x(x^2-1)^n\right] - \frac{(2n+1)x}{n2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[(x^2-1)^n\right] \\ &= \frac{2n+1}{n}P_{n+1}(x) - \frac{2n+1}{n}xP_n(x). \end{aligned}$$

En définitive, il vient :

$$P_{n+1}(x) = \frac{2n+1}{n}P_{n+1}(x) - \frac{2n+1}{n}xP_n(x) + P_{n-1}(x).$$

On aboutit finalement à la relation de récurrence désirée :

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0. \tag{7.4}$$

Cette expression jointe à $P_0(x) = 1$ et $P_1(x) = x$ permet donc de calculer aisément tous les coefficients de chaque polynôme de Legendre $P_n(x)$ quel que soit $n$.

### 1.3. Relations de récurrence faisant intervenir des dérivées

Dérivons la formule de Rodriguès par rapport à $x$, cela donne :

$$P'_n(x) = \frac{1}{2^n}\frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{2nx(x^2-1)^{n-1}\right\},$$

soit encore, en utilisant la relation de Leibnitz :

$$P'_n(x) = \frac{1}{2^{n-1}}\frac{1}{(n-1)!}\left\{x\frac{\mathrm{d}^n}{\mathrm{d}x^n} + n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\right\}(x^2-1)^{n-1},$$

d'où une première relation de récurrence entre polynômes et dérivées :

$$P'_n(x) = xP'_{n-1}(x) + nP_{n-1}(x). \tag{7.5}$$

La relation (7.5) peut encore s'écrire

$$(n+1)P'_{n+1}(x) = (n+1)\left\{xP'_n(x) + (n+1)P_n(x)\right\}.$$

D'autre part, la dérivation de la relation de récurrence (7.4) donne :

$$(n+1)P'_{n+1}(x) = (2n+1)P_n(x) + (2n+1)xP'_n(x) - nP'_{n-1}(x).$$

La soustraction de ces deux dernières expressions conduit à la seconde relation de récurrence entre les polynômes et les dérivées :

$$nP_n(x) = xP'_n(x) - P'_{n-1}(x) \quad \text{avec} \quad n \neq 0. \tag{7.6}$$

Ces deux dernières formules sont utiles dans le cadre de l'étude de la méthode d'intégration de Gauss-Legendre.

## 1.4. Équation différentielle dont les $P_n(x)$ sont solutions

Dérivons une fois par rapport à $x$ la première relation de récurrence (7.5) faisant intervenir les dérivées :

$$P''_n(x) = P'_{n-1}(x) + xP''_{n-1}(x) + nP'_{n-1}(x) = (n+1)P'_{n-1}(x) + xP''_{n-1}(x).$$

De même, la dérivation de la relation (7.6) donne :

$$nP'_n(x) = xP''_n(x) + P'_n(x) - P''_{n-1}(x),$$

d'où l'on tire :

$$P''_{n-1}(x) = -(n-1)P'_n(x) + xP''_n(x).$$

Éliminons $P''_{n-1}(x)$ entre ces deux relations :

$$P''_n(x) = (n-1)P'_{n-1}(x) + x\left\{xP''_n(x) - (n-1)P'_n(x)\right\}.$$

Reste à éliminer $P'_{n-1}(x)$ l'aide de la relation (7.6) :

$$P''_n(x) = (n-1)[xP'_n(x) - nP_n(x)] + x^2P''_n(x) - (n-1)xP'_n(x).$$

Cette expression ne dépend plus que de $P_n(x)$ et de ses deux premières dérivées :

$$(1-x^2)P''_n(x) - 2xP'_n(x) + n(n+1)P_n(x) = 0. \tag{7.7}$$

Cette dernière relation est vraie quel que soit $n$, c'est une équation différentielle linéaire du deuxième ordre appelée équation différentielle de Legendre.

## 1.5. Propriétés des racines des polynômes de Legendre

Un théorème permet d'affirmer que les zéros de chaque polynôme de Legendre sont simples.

Revenons à la fonction $R_{2n}(x) = K(x^2-1)^n$. Elle possède $2n$ racines égales à $-1$ et $+1$. Le théorème de Rolle montre que $R'_{2n}(x)$ a une racine réelle entre $-1$ et $+1$, et l'on peut ajouter que cette fonction a $(n-1)$ racines égales à $-1$ et $(n-1)$ racines égales à $+1$. En poursuivant le calcul jusqu'à la dérivée $n$ ième, on voit que $R^{(n)}_{2n}(x)$ est un polynôme de degré $n$ qui possède $n$ racines réelles comprises entre $-1$ et $+1$. Ce sont, en vertu de la formule de Rodriguès, les racines de $P_n(x)$.

## 1.6. Norme des polynômes de Legendre

Calculons à présent l'expression :

$$J = \int_{-1}^{+1} P_n^2(x)\,\mathrm{d}x = K^2 \int_{-1}^{+1} \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n\,\mathrm{d}x \qquad \text{avec } K = \frac{1}{2^n n!}\cdot$$

Posons

$$\mathrm{d}u = \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n\,\mathrm{d}x \qquad \text{et} \quad v = \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n,$$

puis intégrons par parties ce qui donne :

$$J = K^2\left[\left\{\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(x^2-1)^n\right\}_{-1}^{+1} - \int_{-1}^{+1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}(x^2-1)^n\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}}(x^2-1)^n\,\mathrm{d}x\right].$$

Le premier terme à l'intérieur du crochet est nul puisque $-1$ et $+1$ sont racines, et de proche en proche on obtient :

$$J = K^2(-1)^n\int_{-1}^{+1}(x^2-1)^n\frac{\mathrm{d}^{2n}}{\mathrm{d}x^{2n}}(x^2-1)^n\,\mathrm{d}x.$$

Comme $(x^2-1)^n$ est un polynôme de degré $2n$ (à coefficient principal réduit) que l'on dérive $2n$ fois, le résultat est une constante qui vaut $(2n)!$. Alors, $J$ prend l'expression suivante :

$$J = (-1)^n K^2(2n)!\int_{-1}^{+1}(x^2-1)^n\,\mathrm{d}x.$$

Il n'est pas très difficile de calculer cette dernière intégrale, pour cela on pose :

$$L = \int_{-1}^{+1}(x^2-1)^n\,\mathrm{d}x,$$

expression que l'on intègre par parties. Il suffit d'écrire :

$$u = (x^2-1)^n \text{ et donc } \mathrm{d}u = 2x\,\mathrm{d}x\,n(x^2-1)^{n-1},$$

ainsi que $\mathrm{d}v = \mathrm{d}x$ et $v = x$ ; on obtient alors :

$$L = \left[x(x^2-1)^n\right]_{-1}^{+1} - 2n\int_{-1}^{+1}x^2(x^2-1)^{n-1}\,\mathrm{d}x.$$

La première quantité (toute intégrée) a une valeur nulle. Une autre intégration par parties permet d'écrire :

$$L = \frac{2}{3}n(n-1)^2\int_{-1}^{+1}x^4(x^2-1)^{n-2}\,\mathrm{d}x.$$

On poursuit ainsi de suite les calculs et, en notant par $\varepsilon$ le signe que l'on déterminera plus tard, on arrive à l'expression :

$$L = \varepsilon 2n\frac{1}{3}2(n-1)\frac{1}{5}2(n-2)\ldots\frac{1}{2n-1}2(1)\int_{-1}^{+1}x^{2n-2}(x^2-1)^1\,\mathrm{d}x,$$

qui conduit à l'expression suivante :

$$L = \varepsilon\frac{2\cdot4\cdot6\ldots2n}{1\cdot3\cdot5\ldots(2n+1)}2\cdot$$

Le problème du signe $\varepsilon$ que nous avons délibérément abandonné se trouve résolu en remarquant que $(x^2-1)^n$ est une fonction qui a le signe de $(-1)^n$ sur l'intervalle $(-1,+1)$ ainsi $L$ s'écrit :

$$L = (-1)^n \frac{2 \cdot 4 \cdot 6 \ldots 2n}{1 \cdot 3 \cdot 5 \ldots (2n+1)} 2.$$

Ensuite, nous pouvons écrire :

$$J = \frac{(2n)!}{n!} \frac{1}{2^{n-1}} \frac{1}{1 \cdot 3 \cdot 5 \ldots (2n+1)} = \frac{2}{(2n+1)} \cdot$$

En définitive, on obtient le résultat suivant :

$$\int_{-1}^{+1} P_n^2(x)\, \mathrm{d}x = \frac{2}{(2n+1)} \cdot \tag{7.8}$$

## 1.7. Les premiers polynômes de Legendre

$$\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x \\
P_2(x) &= (3x^2-1)/2 \\
P_3(x) &= (5x^3-3x)/2 \\
P_4(x) &= (35x^4-30x^2+3)/8 \\
P_5(x) &= (63x^5-70x^3+15x)/8 \\
P_6(x) &= (231x^6-315x^4+105x^2-5)/16 \\
P_7(x) &= (429x^7-693x^5+315x^3-35x)/16.
\end{aligned}$$

## 1.8. Calcul des coefficients des polynômes de Legendre

À l'aide de la relation de récurrence (7.4), de $P_0(x)$ et de $P_1(x)$, on peut calculer les coefficients selon les puissances croissantes ou décroissantes de la variable $x$ des $k$ premiers polynômes de Legendre. Dans le cas où l'on utilise un calculateur arithmétique, pour des raisons d'économie de place mémoire, on entasse tous les coefficients des polynômes les uns à la suite des autres dans un tableau unique à une seule dimension.

On trouvera sur le Web (*) le programme `legendre.c` permettant de calculer ces coefficients.

## 1.9. Variantes concernant les polynômes de Legendre

Il n'est pas bien difficile de reprendre cette étude en changeant l'intervalle fondamental $(-1,+1)$ qui est remplacé par l'intervalle fini $(a,b)$. Rien d'essentiel n'est changé pourvu que $a$ et $b$ restent finis. Il est courant de définir d'autres polynômes de Legendre sur l'intervalle $(0,1)$, et ils sont appelés parfois polynômes de Legendre modifiés. On les rencontre lors de problèmes de lissage sur l'intervalle de définition $(0,1)$. On en trouve une application intéressante lors du calcul de diagrammes de phase.

* `http://www.edpsciences.com/guilpin/`

## 1.10. Une propriété importante des polynômes de Legendre

Par définition, on dit qu'un polynôme de degré $n$ est à coefficient principal réduit lorsque le coefficient du terme de degré $n$ est égal à l'unité. Cette notion sert essentiellement à comparer les polynômes de degré $n$ entre eux, sinon ils seraient définis à une constante multiplicative près qui ne permettrait plus de comparaison.

***Théorème*** – Parmi tous les polynômes de degré $n$ à coefficient principal réduit, celui dont la norme, évaluée sur l'intervalle $(-1, +1)$, est la plus petite est le polynôme de Legendre, la norme envisagée étant celle du produit scalaire (la fonction poids étant égale à l'unité).

Désignons par $G_n(x)$ un polynôme répondant à la question et décomposons-le selon une combinaison linéaire de polynômes de Legendre, opération toujours possible. On note par un astérisque les polynômes de Legendre à coefficient principal réduit et l'on écrit par conséquent :

$$G_n(x) = P_p^*(x) + \sum_{k=1}^{n} \alpha_k P_{n-k}^*(x).$$

où les $\alpha_k$ sont des coefficients constants. La norme de $G_n(x)$ s'écrit :

$$N[G_n(x)] = \int_{-1}^{+1} \left[ P_n^*(x) + \sum_{k=1}^{n} \alpha_k P_{n-k}^*(x) \right]^2 \mathrm{d}x$$

Soit encore :

$$N[G_n(x)] = N\left[P_n^*(x)\right] + \sum_{k=1}^{n} \alpha_k^2 N\left[P_{n-k}^*(x)\right].$$

Ce second membre est minimum lorsque tous les $\alpha_k$ sont égaux à zéro. Il s'ensuit que $N[G_n(x)]$ est minimum quand $G_n(x) = P_n^*(x)$, ce qui démontre le théorème.

## 1.11. Fonction génératrice des polynômes de Legendre

Nous admettrons le résultat suivant sans démonstration :

$$\frac{1}{\sqrt{1 - 2hx + h^2}} = P_0(x) + hP_1(x) + h^2 P_2(x) + \cdots + h^n P_n(x) + \cdots$$

ainsi que cette autre expression :

$$P_n(x) = \frac{1}{\pi} \int_0^{\pi} \left( x \pm \sqrt{(x^2 - 1)} \cos \Phi \right)^n \mathrm{d}\Phi.$$

## 1.12. Décomposition d'une fonction f(x) en série de polynômes de Legendre

On considère une fonction $f(x)$ continue sur l'intervalle $(-1, +1)$ ou éventuellement ayant des points de discontinuité de première espèce en quantité dénombrable. D'une façon tout à fait analogue à l'analyse de Fourier, on se propose de rechercher un développement de $f(x)$ sous forme d'une combinaison linéaire de polynômes de Legendre :

$$f(x) = \sum_{n=0}^{\infty} a_n P_n(x).$$

En multipliant chacun des membres par $P_k(x)$ puis en intégrant sur l'intervalle $(-1, +1)$, on obtient :

$$\int_{-1}^{+1} f(x)P_k(x)\,\mathrm{d}x = \int_{-1}^{+1} \sum_{n=0}^{\infty} a_n P_n(x)P_k(x)\,\mathrm{d}x.$$

Soit encore :

$$\int_{-1}^{+1} f(x)P_k(x)\,\mathrm{d}x = \sum_{n=0}^{\infty} a_n \int_{-1}^{+1} P_n(x)P_k(x)\,\mathrm{d}x,$$

et compte tenu des relations d'orthogonalité :

$$\int_{-1}^{+1} f(x)P_k(x)\,\mathrm{d}x = a_k \int_{-1}^{+1} P_k(x)P_k(x)\,\mathrm{d}x = a_k \frac{2}{2k+1},$$

ce qui donne :

$$a_k = \frac{2k+1}{2} \int_{-1}^{+1} f(x)P_k(x)\,\mathrm{d}x.$$

En toute rigueur, il faudrait justifier l'égalité entre $f(x)$ et son développement selon une étude analogue à celle effectuée à propos des séries de Fourier : il faudrait associer la série à la fonction $f(x)$ et déterminer à quelles conditions l'égalité peut être obtenue.

## 1.13. Décomposition d'un polynôme quelconque en polynômes de Legendre

Considérons un polynôme quelconque de degré $n$ :

$$Q_n(x) = \sum_{k=0}^{n} a_k x^{n-k} \qquad \text{avec } a_0 \neq 0.$$

que l'on se propose de décomposer sur la base des polynômes orthogonaux de Legendre $P_j(x)$ que l'on écrit :

$$P_j(x) = \sum_{m=0}^{j} \alpha_{jm} x^{j-m}.$$

Le principe de la décomposition est très simple, il suffit de remarquer que le polynôme

$$Q_n(x) - \frac{a_0}{\alpha_{n0}} P_n(x)$$

est un polynôme de degré $n-1$, désignons-le par $\Pi_{n-1}(x)$. Ses coefficients sont donnés par :

$$\Pi_{n-1}(x) = Q_n(x) - \frac{a_0}{\alpha_{n0}} P_n(x) = (a_1 - \frac{a_0}{\alpha_{n0}}\alpha_{n1})x^{n-1} + (a_2 - \frac{a_0}{\alpha_{n0}}\alpha_{n2})x^{n-2}$$
$$+ (a_3 - \frac{a_0}{\alpha_{n0}}\alpha_{n3})x^{n-3} + \ldots = \sum_{k=0}^{n-1} (a_k - \frac{a_0}{\alpha_{n0}}\alpha_{nk})x^{n-1-k} = \sum_{k=0}^{n-1} \beta_k x^{n-1-k}.$$

Il n'y a pas de difficulté à calculer les coefficients $\beta$, donc à partir de ce moment nous sommes ramené au problème précédent à cette différence près que les polynômes mis en jeu ont perdu un degré : nous sommes passé du degré $n$ au degré $(n-1)$. Par ce procédé, nous arriverons au degré zéro, et il est très aisé d'effectuer cette décomposition en calcul automatique, et l'on trouvera sur le Web(*) le programme `dcomleg.c` réalisant cet algorithme.

### 1.14. Calcul de $\int_{-1}^{+1} x^n P_n dx$

On peut toujours décomposer $x^n$ en polynômes de Legendre selon la manière qui vient d'être décrite, on obtient :

$$x^n = \sum_{k=0}^{n} B_k P_{n-k}(x),$$

et comme :

$$P_n(x) = \sum_{j=0}^{n} \alpha_{nj} x^{n-j},$$

on peut écrire sous la forme :

$$x^n = \frac{1}{\alpha_{n0}} P_n(x) + \sum_{k=1}^{n} B_k P_{n-k}(x).$$

Il est aisé de déduire :

$$\int_{-1}^{+1} x^n P_n(x)\, \mathrm{d}x = \frac{1}{\alpha_{n0}} \frac{2}{2n+1},$$

en tenant compte des relations d'orthogonalité. Cette relation est utile lors de l'étude de la méthode d'intégration de Gauss.

## 2. Méthode d'intégration de Gauss-Legendre

### 2.1. Position du problème

On se propose de calculer numériquement l'intégrale :

$$I = \int_a^b F(t)\, \mathrm{d}t \tag{7.9}$$

lorsque celle-ci, bien entendu, a un sens et que les bornes d'intégration $a$ et $b$ demeurent finies, en outre $F(t)$ ne comporte pas de singularités notamment aux bornes du domaine d'intégration c'est-à-dire en $a$ et $b$. Avant d'entamer ce calcul, remarquons que le changement de variable :

$$t = \frac{a+b}{2} + \frac{b-a}{2} x$$

* `http://www.edpsciences.com/guilpin/`

ramène le domaine d'intégration à l'intervalle fondamental $(-1, +1)$, ce qui donne :

$$I = \int_a^b F(t)\,\mathrm{d}t = \int_{-1}^{+1} f(x)\,\mathrm{d}x. \tag{7.10}$$

La méthode de Gauss pour évaluer numériquement l'intégrale $I$ consiste à :

**a.** adopter pour valeur approchée de $I$ l'expression suivante :

$$J = \sum_{k=0}^{n} H_k f(x_k)\,; \tag{7.11}$$

**b.** choisir un critère pour déterminer les $x_k$ et les $H_k$.

L'examen de la formule (7.11) montre qu'il y a $(n+1)$ valeurs $x_k$ à déterminer ainsi que les $(n+1)$ valeurs $H_k$ qui leur correspondent. Pour obtenir ces valeurs, on retient comme critère l'égalité rigoureuse des expressions (7.10) et (7.11) dans le cas particulier où $f(x)$ est un polynôme, et on fera en sorte que le degré du polynôme soit le plus élevé possible pour $n$ fixé arbitrairement ($n$ figure dans la relation (7.11)). Comme il y a $2(n+1)$ coefficients arbitraires, le polynôme sera au plus de degré $(2n+1)$.

Dans le cas général où $f(x)$ n'est pas un polynôme, on peut espérer une bonne approximation en vertu du théorème de Weierstrass selon lequel toute fonction continue sur un compact $(a, b)$ peut être approchée uniformément par un polynôme. Bien sûr, le théorème de Weierstrass est très postérieur à la méthode de Gauss – un petit siècle les sépare – néanmoins on peut remarquer que, dans le cas où $f(x)$ est continue et possède des dérivées successives continues sur le compact $(a, b)$, le développement en série de Taylor montre que $f(x)$ peut se développer selon un polynôme de degré $n$, à une approximation près qui est donnée par le reste de la formule de Lagrange. Quoi qu'il en soit, un calcul d'erreur fondé sur cette remarque viendra apporter tous les éclaircissements utiles.

On trouvera à la page 46 de l'ouvrage de Crouzeix et Mignot cité en bibliographie la démonstration de l'unicité d'une telle formule d'approximation qui vaut pour toutes les méthodes de Gauss que l'on va rencontrer par la suite.

## 2.2. Calcul des abscisses $x_k$

Supposons que $f(x)$ soit un polynôme $Q_n(x)$ de degré $n$. Ce polynôme peut toujours s'exprimer au moyen d'une combinaison linéaire unique de polynômes de Legendre de degré égal et inférieur à $n$. Soit :

$$f(x) = Q_n(x) = \sum_{j=0}^{n} A_j P_j(x)$$

expression dans laquelle $P_j(x)$ est le polynôme de Legendre de degré $j$ et $A_j$ le coefficient numérique de la décomposition. En vertu des relations d'orthogonalité, nous pouvons écrire :

$$M = \int_{-1}^{+1} Q_n(x) P_{n+1}(x)\,\mathrm{d}x = \int_{-1}^{+1} P_{n+1}(x) \sum_{j=0}^{n} A_j P_j(x)\,\mathrm{d}x = \sum_{j=0}^{n} A_j \int_{-1}^{+1} P_{n+1}(x) P_j(x)\,\mathrm{d}x = 0.$$

En introduisant la relation (7.11), nous obtenons l'expression suivante :

$$M = 0 = \sum_{j=0}^{n} H_j Q_n(x_j) P_{n+1}(x_j). \tag{7.12}$$

Comme cette relation doit être vérifiée quel que soit le polynôme $Q_n(x)$, de degré égal ou inférieur à $n$, et comme par ailleurs les $H_k$ ne peuvent être nuls sinon le nombre des abscisses $x_k$ s'abaisserait, on en conclut nécessairement que $P_{n+1}(x_j)$ est nul. Il s'ensuit que les $x_k$ sont les $(n+1)$ racines du polynôme de Legendre de degré $(n+1)$. À ce propos, nous savons que toutes les racines de $P_{n+1}(x)$ sont réelles, distinctes et comprises dans l'intervalle $(-1, +1)$.

***Remarque 1*** : Avant de poursuivre plus avant les calculs, d'ores et déjà, il est possible de voir que la précision de la méthode de Gauss sera d'autant plus grande que le polynôme $Q_n(x)$ qui approche la fonction $f(x)$ sera de degré plus élevé.

***Remarque 2*** : Puisque toutes les racines de tous les polynômes de Legendre appartiennent à l'intervalle $(-1, +1)$, la méthode de Bairstow va se révéler particulièrement pratique et efficace pour obtenir les valeurs numériques des racines.

***Remarque 3*** : Il est important de constater que les valeurs $x_k$ ne dépendent pas de la fonction à intégrer et sont donc des constantes universelles, elles pourront être rentrées une fois pour toute dans un calculateur.

## 2.3. Calcul numérique des racines $x_k$ des premiers polynômes

À l'aide de la méthode de Bairstow, nous allons calculer les racines des $k$ premiers polynômes de Legendre. Comme ces valeurs sont universelles, il est bon de les obtenir avec une bonne précision, ne serait-ce que pour les employer lors de calculs réalisés sur les petites machines portatives. Bien entendu, si l'on dispose d'une possibilité d'effectuer les calculs en double précision, il est recommandé de l'utiliser. On a reporté sur le tableau 7.3, page 131, les valeurs numériques des racines du polynôme de degré 13. Cette limitation est due au fait que seulement 16 chiffres significatifs ont été utilisés et qu'il en faut davantage pour calculer les racines des polynômes de degré supérieur à 13.

## 2.4. Calcul des coefficients $H_k$

Puisque le polynôme de Legendre $P_{n+1}(x)$ ne possède que des racines réelles et simples, nous pouvons l'exprimer en fonction directe d'un produit de monômes :

$$P_{n+1}(x) = a_{0,n+1} \prod_{j=0}^{n} (x - x_j)$$

$a_{0,n+1}$ étant le coefficient du terme de degré le plus élevé dans le développement de $P_{n+1}(x)$. À présent considérons le polynôme $Q_{nk}(x)$ de degré $n$, ainsi défini :

$$Q_{nk}(x) = \frac{P_{n+1}(x)}{(x - x_k)} = a_{0,n+1} \prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i).$$

Puisque l'expression (7.11) doit donner une valeur exacte de l'intégrale de $Q_{nk}(x)$, nous pouvons écrire :

$$L = \int_{-1}^{+1} Q_{nk}(x)\, \mathrm{d}x = \sum_{j=0}^{n} H_j Q_{nk}(x_j) = a_{0,n+1} \sum_{j=0}^{n} H_j \prod_{\substack{i=0 \\ i \neq k}}^{n} (x_j - x_i). \tag{7.13}$$

Le second membre a toutes ses sommes nulles sauf une seule lorsque $j = k$. Nous obtenons :

$$L = \int_{-1}^{+1} Q_{nk}(x)\,\mathrm{d}x = a_{0,n+1} H_k \prod_{\substack{i=0 \\ i\neq k}}^{n} (x_k - x_i) = H_k Q_{nk}(x_k),$$

expression à partir de laquelle on tire :

$$H_k = \frac{1}{Q_{nk}(x_k)} \int_{-1}^{+1} Q_{nk}(x)\,\mathrm{d}x.$$

Une expression de $Q_{nk}(x_k)$ plus commode à exploiter est souhaitable, et pour cela calculons la dérivée de $P_{n+1}(x)$ ; nous avons :

$$P'_{n+1}(x) = a_{0,n+1} \sum_{j=0}^{n} \prod_{\substack{i=0 \\ i\neq j}}^{n} (x - x_i)$$

et pour $x = x_k$ nous obtenons :

$$P'_{n+1}(x_k) = a_{0,n+1} \prod_{\substack{i=0 \\ i\neq k}}^{n} (x_k - x_i) = Q_{nk}(x_k).$$

L'expression (7.13) se transforme alors en une expression ne faisant intervenir que le polynôme de Legendre de degré $(n+1)$ ainsi que son polynôme dérivé :

$$L = H_k P'_{n+1}(x_k) = \int_{-1}^{+1} \frac{P_{n+1}(x)}{x - x_k}\,\mathrm{d}x.$$

Cette dernière expression permet en principe de calculer les coefficients $H_k$ qui ne dépendent que des polynômes de Legendre et de leurs zéros. Insistons sur le fait qu'ils ne dépendent pas de la fonction dont on cherche à calculer l'intégrale de façon approchée. On obtient en définitive :

$$H_k = \frac{1}{P'_{n+1}(x_k)} \int_{-1}^{+1} \frac{P_{n+1}(x)}{x - x_k}\,\mathrm{d}x. \tag{7.14}$$

Nous allons procéder à la transformation de l'expression (7.14) qui n'est pas encore d'un emploi très commode. Pour ce faire, considérons l'expression suivante :

$$\int_{-1}^{+1} P_{n+1}(x) \frac{P_n(x) - P_n(x_k)}{x - x_k}\,\mathrm{d}x. \tag{7.15}$$

On voit immédiatement que $(x - x_k)$ est une racine de $P_n(x) - P_n(x_k)$ et que par conséquent $[P_n(x) - P_n(x_k)]/(x - x_k)$ est un polynôme de degré $(n-1)$. En vertu de la relation d'orthogonalité et du théorème de décomposition, l'intégrale (7.15) est nulle. Développons cette expression :

$$0 = \int_{-1}^{+1} \frac{P_{n+1}(x) P_n(x)}{x - x_k}\,\mathrm{d}x - P_n(x_k) \int_{-1}^{+1} \frac{P_{n+1}(x)}{x - x_k}\,\mathrm{d}x. \tag{7.16}$$

Évaluons le premier terme :

$$T = \int_{-1}^{+1} \frac{P_{n+1}(x)P_n(x)}{x - x_k}\,\mathrm{d}x.$$

Comme $P_{n+1}(x)$ admet un développement du genre :

$$P_{n+1}(x) = a_{0,n+1}x^{n+1} + a_{1,n+1}x^n + \cdots + a_{n+1,n+1}$$

et comme d'autre part $x_k$ est une racine du polynôme, on peut écrire :

$$P_{n+1}(x) = (x - x_k)(b_{0,n+1}x^n + b_{1,n+1}x^{n-1} + \cdots + b_{n,n+1}),$$

par identification, on note que :

$$a_{0,n+1} = b_{0,n+1}.$$

En tenant compte des relations d'orthogonalité, $T$ s'exprime alors de la façon suivante :

$$T = \int_{-1}^{+1} a_{0,n+1}x^n P_n(x)\,\mathrm{d}x = \frac{a_{0,n+1}}{a_{0,n}}\frac{2}{2n+1}\cdot$$

En revenant à l'expression (7.16), on peut alors écrire :

$$\int_{-1}^{+1} \frac{P_{n+1}(x)}{x - x_k}\,\mathrm{d}x = \frac{1}{P_n(x_k)}\frac{a_{0,n+1}}{a_{0,n}}\frac{2}{2n+1},$$

en remplaçant dans l'expression (7.14), on obtient :

$$H_k = \frac{1}{P'_{n+1}(x_k)}\frac{1}{P_n(x_k)}\frac{a_{0,n+1}}{a_{0,n}}\frac{2}{2n+1}\cdot$$

La relation de récurrence entre trois polynômes consécutifs permet d'obtenir directement $a_{0,n+1}/a_{0,n}$ en identifiant les termes de degré $(n+1)$ ; on obtient la valeur $(2n+1)/(n+1)$. En définitive, $H_k$ prend une forme très simple :

$$H_k = \frac{1}{P'_{n+1}(x_k)}\frac{1}{P_n(x_k)}\frac{2}{n+1}\cdot \tag{7.17}$$

Il existe une autre forme de $H_k$ très utile à connaître parce que très facile d'emploi. Éliminons $P'_{n-1}(x)$ entre les relations de récurrence (7.13) et (7.14) ; on obtient :

$$(1 - x^2)P'_n(x) - nP_{n-1}(x) - nxP_n(x) = 0.$$

Remplaçons $n$ par $(n+1)$ et $x$ par $x_k$ dans cette expression, nous obtenons :

$$(1 - x_{k^2})P'_{n+1}(x_k) = (n+1)P_n(x_k).$$

Il suffit de reporter la valeur de $(n+1)P_n(x_k)$ dans la relation (7.17) pour écrire :

$$H_k = \frac{2}{(1 - x_k^2)}P'^2_{n+1}(x_k)\cdot \tag{7.18}$$

Cette formule est d'un emploi plus intéressant que la relation (7.17) non seulement sur le plan de la programmation mais aussi sur le plan du temps de calcul.

Il n'y a pas de difficulté particulière pour obtenir ces valeurs numériques dans la mesure où les coefficients des polynômes ainsi que les racines ont été convenablement placés dans un ou des fichiers.

On trouvera sur le Web (*) le programme `r_legend.c` calculant les $H_k$.

## 2.5. Retour sur le calcul de l'intégrale définie sur (a, b)

Nous avons montré que le changement de variable :

$$y = \frac{b+a}{2} + \frac{b-a}{2}x$$

nous ramène à l'intervalle $(-1, +1)$. Par conséquent, il suffit alors de calculer :

$$y_k = \frac{b+a}{2} + \frac{b-a}{2}x_k.$$

Les $H_k$ n'étant pas modifiés, on peut écrire les deux relations :

$$I = \frac{b-a}{2}\sum_{k=0}^{n} H_k F(y_k) = \sum_{k=0}^{n} H_k f(x_k).$$

Le cas des intégrales continues par morceaux — c'est-à-dire présentant des discontinuités de première espèce — est justiciable de la méthode de Gauss : on calcule autant d'intégrales qu'il y a d'intervalles sur lesquels la fonction est continue, et chacune des intégrales se ramène au cas qui vient d'être présenté.

## 2.6. Calcul de l'erreur commise lors de l'approximation

Supposons que $f(x)$ soit une fonction continûment différentiable sur l'intervalle fini (canonique) $(-1, +1)$. Le développement de Taylor-MacLaurin à l'ordre $(2n+1)$ nous donne :

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^{2n+1}}{(2n+1)!}f^{(2n+1)}(0) + \frac{x^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi)$$

expression dans laquelle $x$ et $\xi$ appartiennent à l'intervalle $(-1, +1)$.

Intégrons terme à terme cette expression, on obtient :

$$I = \int_{-1}^{+1} f(x)\,\mathrm{d}x = 2\left\{f(0) + \frac{1}{3!}f''(0) + \cdots + \frac{1}{(2m+1)!}f^{2m}(0) + \cdots + \frac{1}{(2n+3)!}f^{(2n+2)}(\xi)\right\}.$$

On remarque que toutes les dérivées d'ordre impair disparaissent lors de l'intégration. L'exploitation directe de la relation (7.11) donne :

$$J = f(0)\sum_{k=0}^{n} H_k + \frac{f'(0)}{1!}\sum_{k=0}^{n} H_k x_k + \frac{f''(0)}{2!}\sum_{k=0}^{n} H_k x_k^2 + \cdots + \frac{f^{(2m)}(0)}{m!}\sum_{k=0}^{n} H_k x_k^m + \cdots + \frac{f^{(2n+1)}(0)}{(2n+1)!}\sum_{k=0}^{n} H_k x_k^{2n+1} + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\sum_{k=0}^{n} H_k x_k^{2n+2}.$$

* `http://www.edpsciences.com/guilpin/`

Si $f(x)$ est un polynôme de degré $(2n+1)$, la dérivée de $f(x)$ d'ordre $(2n+2)$ est nulle, et l'on a l'égalité :

$$I = J.$$

Dans ce cas il n'y a pas d'erreur mathématique due à une approximation ; on obtient en conséquence le système suivant :

$$\begin{aligned}
&\sum_{k=0}^{n} H_k = 2 \\
&\sum_{k=0}^{n} H_k x_k = 0 \\
&\sum_{k=0}^{n} H_k x_k^2 = \frac{2}{3} \\
&\cdots\cdots\cdots\cdots \\
&\sum_{k=0}^{n} H_k x_k^{2m-1} = 0 \\
&\sum_{k=0}^{n} H_k x_k^{2m} = \frac{2}{2m+1} \\
&\cdots\cdots\cdots\cdots \\
&\sum_{k=0}^{n} H_k x_k^{2n+1} = 0,
\end{aligned} \tag{7.19}$$

expressions dans lesquelles les sommations s'effectuent pour $k$ variant de 0 à $n$.

Il est intéressant de constater que le système de $(2n+2)$ équations ainsi défini est constitué de $(2n+2)$ équations non linéaires dans la mesure où l'on considère que les $H_k$ et les $x_k$ sont les inconnues. En revanche, si l'on connaît les $x_k$, alors, $(n+1)$ équations quelconques prises dans le système précédent constituent un système linéaire permettant de calculer les $H_k$.

Toujours est-il que l'erreur prend la forme suivante :

$$I - J = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \left\{ \frac{2}{2n+3} - \sum_{k=0}^{n} H_k x_k^{2n+2} \right\}, \tag{7.20}$$

en prenant soin de remarquer que :

$$\sum_{k=0}^{n} H_k x_k^{2n+2} \neq \frac{2}{2n+3}.$$

Bien entendu, comme toutes les fois en pareil cas, on recherche une majoration la plus raisonnable possible de la fonction $f^{(2n+2)}(\xi)$ dans l'intervalle $(-1, +1)$, en définitive, l'erreur s'exprime de la manière suivante :

$$E = \frac{M_{2n+2}}{(2n+2)!)} \left\{ \frac{2}{2n+3} - \sum_{k=0}^{n} H_k x_k^{2n+2} \right\} \tag{7.21}$$

où $M_{2n+2} = \sup \left| f^{(2n+2)}(x) \right|$ pour $x$ appartenant à $(-1, +1)$.

Dans la mesure où l'on peut obtenir une majoration $M_{2n+2}$, $E$ donne l'erreur mathématique liée à la troncature de la série des polynômes de Legendre. Viendront s'ajouter lors des calculs effectifs les inévitables erreurs d'arrondi propagées par l'exécution des opérations.

Bien qu'elle soit très intéressante, cette expression n'est pas d'un emploi très commode dans la mesure où figure une dérivée d'ordre $(2n + 2)$. Le calcul devient très souvent arborescent et rapidement inextricable, quand à la majoration, elle risque de devenir catastrophique dans la mesure où elle est très surestimée.

***Table des valeurs numériques de*** $\eta = \left\{ \frac{2}{2n+3} - \sum_{k=0}^{n} H_k x_k^{2n+2} \right\}$ – Les valeurs du tableau 7.1, sont directement tirées du du fichier Legendre.txt sur le Web(*) où figurent les racines et les poids associés des premiers polynômes de Legendre.

**Tableau 7.1.**

| Degré du polynôme | Erreur mathématique |
|---|---|
| 2 | 0,177 778 |
| 3 | $0{,}457\,143\,10^{-1}$ |
| 4 | $0{,}116\,100\,10^{-1}$ |
| 5 | $0{,}293\,181\,10^{-2}$ |
| 6 | $0{,}738\,079\,10^{-3}$ |
| 7 | $0{,}185\,466\,10^{-3}$ |
| 8 | $0{,}465\,483\,10^{-4}$ |
| 9 | $0{,}116\,731\,10^{-4}$ |
| 10 | $0{,}292\,559\,10^{-5}$ |
| 11 | $0{,}732\,912\,10^{-6}$ |
| 12 | $0{,}183\,0547\,10^{-6}$ |
| 13 | $0{,}459\,546\,10^{-7}$ |

## 2.7. Un exemple numérique, comparaison avec d'autres méthodes

Nous avons choisi un exemple dont on connaît directement le résultat par quadratures :

$$I = \int_0^{+1} x \log(1 + x)\, \mathrm{d}x = \frac{1}{4} \cdot$$

Voici les résultats obtenus par la méthode de Gauss, la méthode des trapèzes et la méthode de Simpson.

### *a – Méthode de Gauss*

5 points : $I = 0{,}249\,999\,996$

6 points : $I = 0{,}249\,999\,999\,909$

* http://www.edpsciences.com/guilpin/

***b – Méthode des trapèzes et méthode de Simpson*** – En 6 points la méthode de Gauss est un peu plus précise que la méthode de Simpson en 100 points... tandis qu'il faut 40 points à la méthode de Simpson pour obtenir une précision analogue à celle de la méthode de Gauss en 5 points.

**Tableau 7.2.**

| Méthode des trapèzes | Méthode de Simpson | Nbre de points |
|---|---|---|
| 0,250 248 | 0,250 000 086 | 20 |
| 0,250 062 | 0,250 000 005 4 | 40 |
| 0,250 027 | 0,250 000 001 07 | 60 |
| 0,250 009 9 | 0,250 000 000 138 | 100 |

***c – Calcul de l'erreur concernant l'exemple*** – Par chance, le calcul de la dérivée $n^{\text{e}}$ de $x\log(1+x)$ est relativement aisé, et l'on peut montrer que :

$$[x\log(1+x)]^{(n)} = (-1)^n \frac{(n-2)!(x+n)}{(1+x)^n}\,.$$

Effectuons le calcul d'erreur pour 6 points :

$$E = \frac{M_{12}}{12!}\left\{\frac{2}{13} - \sum_{k=0}^{5} H_k x_k^{12}\right\} = 7{,}381\,10^{-4}\frac{M_{12}}{12!}\,. \qquad \text{Ici,}\quad M_{12} = \frac{10!(x+12)}{2^{12}(1+x)^{12}}$$

car nous avons dû effectuer un changement de variable $x = \dfrac{1}{2} + \dfrac{1}{2}t$ pour nous ramener à l'intervalle canonique $(-1,+1)$. Alors nous pouvons écrire : $\mathrm{d}x/\,\mathrm{d}t = 1/2$ et $\left\{\dfrac{\mathrm{d}x}{\mathrm{d}t}\right\}^{12} = \dfrac{1}{2^{12}}$.

Rappelons que, avec l'écriture choisie, $x$ appartient toujours à $(0,1)$, et que $\dfrac{b-a}{2} = \dfrac{1}{2}$.

$M_{12}$ est une fonction monotone décroissante dans l'intervalle $(0,1)$, elle est maximum pour $x = 0$. Donc $\dfrac{b-a}{2}M_{12}$ est plus petit ou égal à $\dfrac{12!}{11\times 2\times 2^{12}}$.

On tire une valeur de $E$ :

$$E = \frac{7{,}381\,10^{-4}}{11\times 8\,192} = 0{,}8\,10^{-8}.$$

On s'aperçoit que l'erreur $E$ est environ 100 fois plus grande que l'erreur observée directement ; cela est dû à la majoration de la fonction $M_{12}$. En $x = 0$, la dérivée est 3 780 fois plus petite qu'au point $x = 1$. On en conclut que l'erreur est comprise entre $2\,10^{-12}$ et $0{,}8\,10^{-8}$. L'évaluation directe de l'erreur se situe bien dans ce créneau puisqu'elle est de l'ordre de $10^{-10}$ et cet exemple montre bien combien la majoration d'une dérivée d'ordre n peut être pénalisante pour un calcul d'erreur.

***Remarque*** : Pour utiliser aisément la méthode de Gauss, il suffit de fabriquer un sous-programme dans lequel on rentre une fois pour toutes les racines et les poids du polynôme

**Tableau 7.3.** *Racines et poids associés des polynômes de Legendre.*

| Degré | Racines $x$ | Poids $H_k$ |
|---|---|---|
| 13 | 0,0 | 0,232 551 553 230 874 |
| | ±0,230 458 315 955 13 | 0,226 283 180 262 897 |
| | ±0,448 492 751 036 45 | 0,207 816 047 536 88 |
| | ±0,642 349 339 440 33 | 0,178 145 980 761 94 |
| | ±0,801 578 090 733 3 | 0,138 873 510 219 8 |
| | ±0,917 598 399 223 | 0,092 121 499 837 7 |
| | ±0,984 183 054 718 5 | 0,040 484 004 765 |

de degré 12 ou 13 (ensuite, des problèmes de précision apparaissent avec des représentations sur 16 chiffres significatifs et des racines complexes apparaissent...) ; il suffit d'ailleurs de rentrer 6 racines et 6 poids associés à cause des symétries, les racines sont deux à deux opposées.

On donne sur le Web(*) un fichier texte appelé `legendre.txt` où l'on trouvera les racines et les poids des treize premiers polynômes de Legendre.

## 2.8. Éléments de bibliographie

A. ANGOT (1965) *Compléments de Mathématiques*, Éditions de la Revue d'Optique.

M. CROUZEIX et A.L. MIGNOT (1984) *Analyse numérique des équations différentielles*, Éditions Masson.

F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.

H.N. MHASKAR (1996) *Introduction to the theory of weighted polynomial approximation*, World Scientific.

H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

* `http://www.edpsciences.com/guilpin/`

# 8 Les polynômes de Tchebycheff. Application à la méthode de Gauss-Tchebycheff

## 1. Les polynômes de Tchebycheff (1821–1894)

### 1.1. Première définition

La relation de définition des polynômes de Tchebycheff est donnée par l'expression suivante :

$$T_n^*(x) = \cos(n \arccos x) \tag{8.1}$$

ce qui implique que :

$$T_0^*(x) = 1 \qquad \text{et} \qquad T_1^*(x) = x.$$

Ils obéissent à une relation de récurrence que nous allons établir. Pour cela, il suffit de poser $y = \arccos x$, on obtient :

$$T_n^*(x) = \cos ny.$$

Écrivons $T_{n+1}^*(x)$ et $T_{n-1}^*(x)$ et effectuons-en la somme :

$$\begin{aligned} T_{n+1}^*(x) &= \cos(n+1)y = \cos ny \cos y - \sin ny \sin y, \\ T_{n-1}^*(x) &= \cos(n-1)y = \cos ny \cos y + \sin ny \sin y, \\ T_{n+1}^*(x) + T_{n-1}^*(x) &= 2 \cos ny \cos y = 2xT_n^*(x) \end{aligned}$$

D'où la relation de récurrence cherchée entre trois polynômes consécutifs :

$$T_{n+1}^*(x) - 2xT_n^*(x) + T_{n-1}^*(x) = 0. \tag{8.2}$$

À présent nous allons montrer que ces polynômes sont orthogonaux. En partant de la relation :

$$\int_0^\pi \cos nt \cos mt \, \mathrm{d}t = \pi/2\delta_{mn}$$

$\delta_{mn}$ étant le symbole de Kronecker (1823–1891), sauf pour $n = m = 0$ l'intégrale vaut $\pi$. Effectuons le changement de variable $t = \arccos x$, soit $x = \cos t$, ce qui nous permet d'écrire :

$$\mathrm{d}t = -\sin x \, \mathrm{d}x$$

et

$$\mathrm{d}x = -\frac{\mathrm{d}t}{\sqrt{1-x^2}}$$

puis

$$\int_0^{\pi} \cos nt \cos mt \, \mathrm{d}t = \int_{-1}^{+1} \frac{T_n^*(x) T_m^*(x) \, \mathrm{d}x}{\sqrt{1-x^2}} = 0 \qquad \text{si } m \neq n$$

et

$$\int_{-1}^{+1} \frac{T_n^*(x) T_n^*(x) \, \mathrm{d}x}{\sqrt{1-x^2}} = \int_0^{\pi} \cos^2 nt \, \mathrm{d}t = \frac{\pi}{2} \qquad \text{si } n \text{ est différent de zéro}$$

$$= \pi \qquad \text{si } n = 0.$$

On voit que les polynômes de Tchebycheff sont orthogonaux sur l'intervalle fondamental $(-1, +1)$ relativement à la fonction poids $\frac{1}{\sqrt{1-x^2}}$ qui est une fonction définie positive sur cet intervalle.

### 1.2. Seconde définition

Cette suite de polynômes orthogonaux peut être avantageusement modifiée pour le propos qui nous préoccupe. On peut définir la suite de telle sorte que le coefficient de degré le plus élevé pour chaque polynôme soit égal à 1. Nous allons donc définir une suite polynômes à coefficient principal réduit. Si nous examinons la relation (8.2), nous nous apercevons que le coefficient de degré le plus élevé est multiplié par 2 lorsque l'on passe du polynôme de degré $k$ à celui de degré $(k+1)$. Il nous suffit de poser comme relation de définition une nouvelle expression :

$$T_n(x) = \frac{1}{2^{n-1}} \cos(n \arccos x). \tag{8.3}$$

La relation de récurrence est alors légèrement modifiée et devient :

$$T_{n+1}(x) - 4xT_n(x) + 4T_{n-1}(x) = 0. \tag{8.4}$$

Il va de même de la norme qui s'exprime :

$$\int_{-1}^{+1} \frac{T_n^2 \, \mathrm{d}x}{\sqrt{1-x^2}} = \frac{\pi}{2^{2n-1}} \qquad \text{si } n \text{ est différent de zéro}$$

$$= \pi \qquad \text{si } n = 0. \tag{8.5}$$

## 2. Une propriété essentielle des polynômes de Tchebycheff à coefficient principal réduit

### 2.1. Théorème

De tous les polynômes de degré $n$ à coefficient principal réduit, c'est le polynôme de Tchebycheff $T_n(x)$ qui approche le mieux l'axe des $x$ sur l'intervalle $(-1, +1)$ au sens du sup du module.

### 2.2. Démonstration

Désignons par $p_n(x)$ un polynôme répondant à la question et considérons alors la fonction :

$$f(x) = p_n(x) - T_n(x).$$

Puisque $p_n(x)$ et $T_n(x)$ sont à coefficient principal réduit, $f(x)$ est un polynôme au plus de degré $(n-1)$. Par ailleurs, $T_n$ possède $n$ extremums alternativement positifs et négatifs sur l'intervalle $(-1,+1)$, tous égaux en module à 1 et dont les abscisses sont données par la relation :

$$x_k = \cos\left(\frac{k\pi}{n}\right) \qquad \text{avec } k = 0, 1, 2, \ldots, (n-1).$$

Quand $T_n(x_k)$ est positif, $f(x_k)$ est négatif et quand $T_n(x_k)$ est négatif, $f(x_k)$ est positif, il s'ensuit que $f(x)$ est alors un polynôme $n$ fois positif et négatif alternativement sur l'intervalle $(-1,+1)$. Il change donc $(n-1)$ fois de signe et par conséquent possède $n$ racines. Donc, $f(x)$ doit être un polynôme de degré $n$, ce qui est en contradiction avec le fait que $f(x)$ est au plus un polynôme de degré $(n-1)$. On en conclut que le polynôme $f(x)$ est identiquement nul et que $p_n(x) = T_n(x)$ ce qui démontre le théorème.

## 3. Les racines des polynômes de Tchebycheff $T_{n+1}(x)$

On sait que le polynôme de degré $(n+1)$ possède $(n+1)$ racines distinctes et réelles sur l'intervalle $(-1,+1)$. On les obtient en écrivant $T_{n+1} = 0$, soit :

$$\cos[(n+1)\arccos x] = 0,$$

ce qui donne :

$$x_k = \cos\frac{\pi(2k+1)}{2(n+1)}. \tag{8.6}$$

## 4. Calcul des poids $H_k$ correspondant aux racines $x_k$ du polynôme $T_{n+1}(x)$

Pour cela, il suffit d'exploiter directement la relation (B.5) de l'annexe B. D'abord, on calcule le coefficient $K$ qui est égal à $I_n$ puisque nous traitons de polynômes à coefficient principal réduit. Donc :

$$K = \frac{\pi}{2^{2n-1}}$$

À présent il nous faut calculer $T'_{n+1}(x_k)$ et $T_{n+1}(x_k)$. On trouve les expressions suivantes :

$$T'_{n+1}(x_k) = \frac{n+1}{2^n}\left(\frac{1}{\sqrt{1-x^2}}\sin[(n+1)\arccos x_k]\right)$$

d'où

$$T'_{n+1}(x_k) = \frac{(-1)^k}{2^n}\frac{n+1}{\sin\dfrac{\pi(2k+1)}{2(n+1)}},$$

puis il vient

$$T_n(x_k) = \frac{1}{2^n}\cos(n \arccos x_k) = \frac{1}{2^{n-1}}\cos n\frac{\pi(2k+1)}{2(n+1)} \cdot$$

Cette dernière expression, après quelques transformations trigonométriques, se réduit à :

$$T_n(x_k) = \frac{(-1)^k}{2^{n-1}}\sin\frac{\pi(2k+1)}{2(n+1)} \cdot$$

De là on déduit la valeur des $H_k$ :

$$H_k = \frac{K}{T'_{n+1}(x_k)T_n(x_k)} = \frac{\pi}{n+1} \cdot \tag{8.7}$$

Cette expression montre que tous les $H_k$ sont égaux pour un polynôme donné, ce qui, notons-le au passage, conduira à des calculs particulièrement simples.

## 5. Méthode d'intégration de Gauss-Tchebycheff

Il est usuel d'associer au nom de Gauss le nom du savant qui a laissé son nom à une suite particulière de polynômes orthogonaux. C'est pour respecter cette tradition que nous désignons par les deux noms accolés la méthode qui nous permet de calculer numériquement les intégrales du type :

$$I = \int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x \tag{8.8}$$

que l'on approche donc par l'expression classique :

$$J = \sum_{k=0}^{n} H_k f(x_k)$$

laquelle se simplifie notablement :

$$J = H\sum_{k=0}^{n} f(x_k) = \frac{\pi}{n+1}\sum_{k=0}^{n} f\left(\cos\frac{\pi(2k+1)}{2(n+1)}\right) \cdot \tag{8.9}$$

avec $k = 0, 1, 2, \ldots, n$. Il est bien entendu que la fonction $f(x)$ est régulière sur l'intervalle $(-1, +1)$. Par ailleurs, nous ne croyons pas utile de devoir donner la liste des racines des polynômes de Tchebycheff et des poids $H_k$ qui leur sont associés, puisque les expressions mathématiques qui permettent leur calcul sont très simples.

## 6. Calcul de l'intégrale $\mathbf{I = \int_{-a}^{+a} \frac{f(x)}{\sqrt{(x-a)(b-x)}}\,dx}$

Le changement de variable :

$$x = \frac{b+a}{2} + \frac{b-a}{2}y$$

nous permet de nous ramener au problème précédent car l'intégrale s'écrit :

$$I = \int_{-1}^{+1} \frac{f\left(\dfrac{b+a}{2} + \dfrac{b-a}{2}y\right)}{\sqrt{1-y^2}}\,\mathrm{d}y.$$

Par conséquent, on calcule $I$ au moyen de l'approximation :

$$J = \frac{\pi}{n+2}\sum_{k=0}^{n} f(y_k) \qquad \text{avec } y_k = \frac{b+a}{2} + \frac{b-a}{2}x_k,$$

où, rappelons-le, les $x_k$ sont, dans cette expression, les racines du polynôme de Tchebycheff de degré $(n+1)$. Le lecteur pourra légitimement s'étonner du fait que l'on utilise les polynômes de degré $(n+1)$ et non des polynômes de degré $n$, la raison en est simple, c'est pour rester cohérent avec les expressions établies au chapitre précédent.

## 7. Calcul de l'erreur commise lors de l'approximation

Nous allons reprendre le calcul réalisé lors de l'étude de la méthode de Gauss-Legendre, seuls quelques points de détails vont différer.

Soit à calculer :

$$I = \int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}}\,\mathrm{d}x\cdot$$

Supposons que $f(x)$ soit continûment différentiable sur l'intervalle canonique $(-1,+1)$. Le développement de Taylor-MacLaurin à l'ordre $(2n+2)$ nous donne :

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^{2n+1}}{(2n+1)!}f^{(2n+1)}(0) + \frac{x^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi),$$

expression dans laquelle $x$ et $\xi$ appartiennent à l'intervalle $(-1,+1)$.

Calculons $I$ en remplaçant $f(x)$ par son développement, il faudra calculer des expressions du type :

$$Q_n = \int_{-1}^{+1} \frac{x^n}{\sqrt{1-x^2}}\,\mathrm{d}x\cdot$$

Pour cela, posons $x = \cos z$, soit $\mathrm{d}x = -\sin z\,dz$ ; on obtient alors :

$$Q_n = \int_{0}^{\pi} \cos^n x\,\mathrm{d}x.$$

On intègre par parties cette expression en posant :

$$\begin{aligned} & v = \cos^{n-1} x && \text{et} \quad \mathrm{d}u = \cos x\,\mathrm{d}x \\ \text{ce qui donne :} \quad & \mathrm{d}v = -(n-1)\cos^{n-2} x \sin x\,\mathrm{d}x && \text{et} \quad u = \sin x. \end{aligned}$$

On peut écrire :

$$Q_n = (n-1)\int_0^{\pi} \cos^{n-2} x(1-\cos^2 x)\,\mathrm{d}x = -(n-1)Q_{n+(n-1)}\int_0^{\pi} \cos^{n-2} x\,\mathrm{d}x$$

d'où la relation de récurrence :

$$nQ_n = (n-1)Q_{n-2} \quad \text{avec} \quad Q_0 = \int_0^{\pi} \mathrm{d}x = \pi \quad \text{et} \quad Q_1 = \int_0^{\pi} \cos x\,\mathrm{d}x = 0.$$

De là, on tire que :

$$Q_{2n+1} = 0,$$
$$Q_{2n} = \frac{(2n-1)!!}{(2n)!!}\pi = \frac{(2n-1)!!}{2^n n!}\pi\,.$$

Revenons au calcul de I :

$$I = \int_{-1}^{+1} \frac{\mathrm{d}x}{\sqrt{1-x^2}}\Big[f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^{2n+1}}{(2n+1)!}f^{(2n+1)}(0) + \frac{x^{2n+2}}{(2n+2)!}$$
$$\times f^{(2n+2)}(\xi)\Big] = \pi\left[f(0) + \cdots + \frac{f^{2q}(0)}{(2q)!}\frac{(2q-1)!!}{2^q q!} + \cdots + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\frac{(2n+1)!!}{2^{n+1}(n+1)!}\right].$$

L'exploitation directe de la formule d'intégration de Gauss-Tchebycheff donne :

$$J = \frac{\pi}{n+1}\left[(n+1)f(0) + \cdots + \frac{f^{(q)}(0)}{q!}\sum_{k=0}^{n} x_k^q + \cdots + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\sum_{k=0}^{n} x_k^{2n+2}\right].$$

L'erreur $E$ s'écrit :

$$E = |I-J| = \pi\left[f(0) - f(0) + \cdots + \frac{f^{(2q)}(0)}{(2q)!}\left(\frac{(2q-1)!!}{2^q q!} - \frac{1}{n+1}\sum_{k=0}^{n} x_k^q\right) + \cdots\right.$$
$$\left. + \frac{f^{(2q+1)}(0)}{(n+1)(2q+1)!}\sum_{k=0}^{n} x_k^{2q+1} + \cdots + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\left(\frac{(2n+1)!!}{2^{n+1}(n+1)!} - \frac{1}{n+1}\sum_{k=0}^{n} x_k^{2n+2}\right)\right]$$

Si $f(x)$ est un polynôme de degré $(2n+1)$, la dérivée de $f(x)$ d'ordre $(2n+2)$ est nulle, et l'on a l'égalité :

$$I = J.$$

Dans ce cas il n'y a pas d'erreur mathématique due à une approximation ; on obtient en conséquence le système suivant :

$$\frac{1}{n+1}\sum_{k=0}^{n} x_k^{2q} = \frac{(2q-1)!!}{2^q q!}$$
$$\frac{1}{n+1}\sum_{k=0}^{n} x_k^{2q+1} = 0.$$

Bien entendu, comme toutes les fois en pareil cas, on recherche une majoration la plus raisonnable possible de la fonction $f^{(2n+2)}(\xi)$ dans l'intervalle $(-1,+1)$ ; en définitive, l'erreur s'exprime de la manière suivante :

$$E(x) = \frac{M_{2n+2}}{(2n+2)!}\left\{\frac{(2n+1)!!}{2^{n+1}(n+1)!} - \frac{1}{n+1}\sum_{k=0}^{n} x_k^{2n+2}\right\}, \tag{8.10}$$

où $M_{2n+2} = \sup\left|f^{(2n+2)}(x)\right|$ pour $x$ appartenant à $(-1,+1)$.

Dans la mesure où l'on peut obtenir une majoration raisonnable $M_{2n+2}$, $E(x)$ donne l'erreur mathématique liée à la troncature de la série des polynômes de Tchebycheff. Viendront s'ajouter lors des calculs effectifs les inévitables erreurs d'arrondi propagées par l'exécution des opérations. Toujours est-il que pour $n = 12$, on trouve :

$$E_{24} = 1{,}47\,10^{-40} M_{24}.$$

## 8. Fonctions génératrices des polynômes de Tchebycheff

Nous admettrons les résultats suivants sans démonstration :

$$\frac{1-tx}{1-2tx+t^2} = T_0(x) + tT_1(x) + t^2T_2(x) + \cdots + t^nT_n(x) + \cdots$$

et

$$\exp(xt)\cos(t\sqrt{1-x^2}) = T_0(x) + \frac{t}{1}T_1(x) + \frac{t^2}{2}T_2(x) + \cdots + \frac{t^n}{n}T_n(x) + \cdots$$

## 9. Un exemple d'intégration

À titre d'exemple, calculons l'intégrale :

$$K = \frac{1}{\pi}\int_{-1}^{+1} \frac{\cos t}{\sqrt{1-t^2}}\,\mathrm{d}t = J_0(1).$$

Cet exemple n'est pas tout à fait gratuit ; en effet, la méthode d'intégration de Gauss-Tchebycheff constitue la meilleure façon de calculer les fonctions de Bessel de première espèce. Nous reviendrons sur ce point au cours de l'annexe F.

Pour l'instant, $J_0(1)$ est la valeur de la fonction de Bessel de première espèce d'ordre zéro pour la valeur 1 de l'argument. Nous avons trouvé dans la littérature la valeur suivante :

$$J_0(1) = 0{,}765\,197\,686\,557\,966\,4$$

avec une précision absolue de l'ordre de $10^{-15}$.

Le tableau 8.1, page suivante, donne les résultats obtenus.

Si l'on poursuit le calcul pour des valeurs plus élevées de $n$, seul le dernier chiffre significatif change, soit le 16$^{e}$ chiffre... ce qui constitue un résultat remarquable.

**Tableau 8.1.**

| Nombre de points | $J(1)$ |
|---|---|
| 3 | 0,765 239 563 |
| 4 | 0,765 197 498 111 |
| 5 | 0,765 197 687 084 089 |
| 6 | 0,765 197 686 556 966 |
| 7 | 0,765 197 686 557 967 7 |
| 8 | 0,765 197 686 557 966 2 |
| 9 | 0,765 197 686 557 966 4 |

## Note importante

Il faut faire attention à ne pas confondre la méthode d'intégration de Gauss-Tchebycheff avec la méthode d'intégration de Tchebycheff ; cette dernière consiste à évaluer l'intégrale :

$$I = \int_{-1}^{+1} f(x)\,\mathrm{d}x$$

au moyen de l'approximation

$$J = \sum_{k=0}^{n} H_k f(x_k)$$

où les $H_k$ ont été choisis *a priori*.

Comme précédemment, on cherchera à fixer les $x_k$ de telle sorte que $I = J$ pour un polynôme de degré le plus élevé possible. Cette modification des conditions, apparemment anodine, change complètement le fond du problème tant et si bien que non seulement la solution dépend alors des $H_k$ mais encore que les $x_k$ peuvent ne plus être réels. Même lorsque tous les $H_k$ sont égaux, les $x_k$ cessent d'être tous réels à partir de $n = 8$, excepté toutefois pour $n = 9$. Quoi qu'il en soit, la méthode est moins précise que celle de Gauss.

# 9 Les polynômes de Laguerre. Méthode d'intégration de Gauss-Laguerre

Les polynômes de Laguerre (1834–1886) sont orthogonaux relativement à la fonction poids $\exp(-x)$ sur l'intervalle $(0, \infty)$. On peut aborder l'étude de ces polynômes d'une façon tout à fait analogue à celle que nous avons adoptée pour les polynômes de Legendre, cependant, nous avons préféré utiliser une formule semblable à celle de Rodriguès pour établir plus rapidement leurs principales propriétés.

Par définition, les polynômes de Laguerre sont donnés par les expressions :

$$L_n(x) = \exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\}. \tag{9.1}$$

## 1. Relation de récurrence entre trois polynômes consécutifs

À partir de la relation de définition, nous allons établir une relation de récurrence entre trois polynômes consécutifs, ce qui nous permettra par la suite de calculer facilement les coefficients de l'un quelconque des polynômes de Laguerre. Il suffit d'écrire la relation de définition pour l'indice $(n+1)$, ce qui donne :

$$L_{n+1}(x) = \exp(x)\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}}\left\{\exp(-x)x^{n+1}\right\}.$$

Soit :

$$L_{n+1}(x) = \exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{-\exp(-x)x^{n+1} + (n+1)\exp(-x)x^n\right\}.$$

On transforme cette dernière expression à l'aide de la formule de Leibnitz :

$$\begin{aligned}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n+1}\right\} &= \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n x\right\} \\ &= x\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\} + n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{\exp(-x)x^n\right\}.\end{aligned}$$

L'expression donnant $L_{n+1}(x)$ devient :

$$L_{n+1}(x) = (n+1)L_n(x) - xL_n(x) - n\exp(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{\exp(-x)x^n\right\}.$$

On calcule directement le dernier terme du second membre à partir de la relation (9.1) :

$$L_{n+1}(x) = \exp(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{n\exp(-x)x^{n-1} - \exp(-x)x^n\right\}.$$

soit encore,

$$L_n(x) = \exp(x)\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{n\exp(-x)x^{n-1} - \exp(-x)x^n\right\}.$$

Il vient donc :

$$L_{n+1}(x) = (n+1)L_n(x) - xL_n(x) + nL_n(x) - n^2L_{n-1}(x)$$

De là nous tirons la relation utile pour déterminer les différents polynômes connaissant les deux premiers :

$$L_{n+1}(x) + (x - 2n - 1)L_n(x) + n^2L_{n-1}(x) = 0. \tag{9.2}$$

## 2. Relation de récurrence faisant intervenir la dérivée

À présent, nous allons établir une autre relation très intéressante pour le calcul des coefficients $H_k$. Il s'agit de pouvoir exprimer simplement la dérivée $L'_n(x)$. Cela nous sera bien utile pour calculer les $H_k$ de la formule de Gauss généralisée. Dans ce but, dérivons la relation de définition (9.1), ce qui permet d'écrire :

$$L'_n(x) = n\exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n-1}\right\}.$$

soit :

$$L'_n(x) = \frac{n}{x}x\exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n-1}\right\}.$$

Toujours à l'aide de la relation de Leibnitz, transformons cette dernière expression :

$$x\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n-1}\right\} = \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\} - n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left\{\exp(-x)x^{n-1}\right\}.$$

On en déduit directement que :

$$xL'_n(x) = nL_n(x) - nL_{n-1}(x). \tag{9.3}$$

## 3. Les premiers polynômes de Laguerre

La relation de définition permet de calculer les deux premiers polynômes, soit :

$$L_0(x) = 1,$$
$$L_1(x) = -x + 1$$

La relation de récurrence (9.2) montre que le polynôme $L_{n+1}(x)$ a un degré supérieur de 1 unité à celui du polynôme de degré $n$. Cela démontre au passage (en utilisant un raisonnement

par récurrence) que nous avons bien affaire à des polynômes, car $L_0(x)$ et $L_1(x)$ sont bien des polynômes donc $L_n(x)$ est aussi un polynôme et ainsi de suite. Voici les premiers polynômes de Laguerre :

$$\begin{aligned}
L_0(x) &= 1\\
L_1(x) &= -x+1\\
L_2(x) &= x^2-4x+2\\
L_3(x) &= -x^3+9x^2-18x+6\\
L_4(x) &= x^4-16x^3+72x^2-96x+24\\
L_5(x) &= -x^5+25x^4-200x^3+600x^2-600x+120.
\end{aligned}$$

## 4. Calcul des coefficients des n premiers polynômes de Laguerre

Il n'est pas très difficile de réaliser un programme destiné à calculer tous les coefficients des premiers polynômes de Laguerre. Ils seront entassés dans un tableau unique, par souci d'économie, selon les puissances décroissantes.

On trouvera sur le Web(*) le programme `laguerre.c` réalisant cette tâche.

## 5. Orthogonalité des polynômes de Laguerre

Les polynômes de Laguerre sont orthogonaux relativement à la fonction poids $\omega(x)=\exp(-x)$ sur l'intervalle fondamental $(0,\infty)$. Pour le montrer il suffit de calculer :

$$I_{nk}=\int_0^\infty \exp(-x)L_n(x)L_k(x)\,\mathrm{d}x=\int_0^\infty \exp(-x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left\{\exp(-x)x^k\right\}\,\mathrm{d}x.$$

Pour fixer les idées et sans nuire à la généralité, supposons que nous ayons $n<k$. Effectuons une intégration par parties en posant :

$$\begin{aligned}
u &= \exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\},\\
\mathrm{d}u &= n\exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n-1}\right\},
\end{aligned}$$

et

$$\begin{aligned}
\mathrm{d}v &= \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left\{\exp(-x)x^k\right\},\\
v &= \frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}}\left\{\exp(-x)x^k\right\}.
\end{aligned}$$

* `http://www.edpsciences.com/guilpin/`

Nous obtenons :

$$I_{nk} = \left[\exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^n\right\}\frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}}\left\{\exp(-x)x^k\right\}\right]_0^\infty - \int_0^\infty \exp(x)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\exp(-x)x^{n-1}\right\}\frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}}\left\{\exp(-x)x^k\right\}\,\mathrm{d}x.$$

Le premier terme est nul car il contient $x\exp(-x)$ en facteur ; en poursuivant l'intégration par parties, on arrive à l'expression suivante :

$$I_{nk} = \varepsilon n! \int_0^\infty \frac{\mathrm{d}^{k-n}}{\mathrm{d}x^{k-n}}\left\{\exp(-x)x^k\right\}\,\mathrm{d}x$$

où $\varepsilon$ représente le signe de $I_{nk}$. Le calcul de cette intégrale donne :

$$I_{nk} = \varepsilon n!\frac{\mathrm{d}^{k-n-1}}{\mathrm{d}x^{k-n-1}}\left\{\exp(-x)x^k\right\}_0^\infty,$$

car ici encore on fait apparaître le terme $x\exp(-x)$ en facteur en fin de calcul, et ce terme est nul pour $x = 0$ et $x$ infini. Donc les polynômes de Laguerre sont orthogonaux par rapport à la fonction poids $\exp(-x)$ sur l'intervalle fondamental $(0, \infty)$.

À présent, il nous reste à calculer la norme de l'intégrale, c'est-à-dire $I_{nn}$, car nous aurons besoin de cette valeur pour le calcul des $H_k$.

Réglons tout de suite le problème du signe : il est nécessairement positif puisque la forme est définie positive (carré scalaire). Donc reste à calculer :

$$I_{nn} = n! \int_0^\infty \left\{\exp(-x)x^n\right\}\,\mathrm{d}x.$$

En procédant par parties, on voit que :

$$\int_0^\infty \left\{\exp(-x)x^n\right\}\,\mathrm{d}x = n!,$$

il s'ensuit que :

$$I_{nn} = (n!)^2. \tag{9.4}$$

En conclusion, les polynômes de Laguerre sont orthogonaux relativement à la fonction poids $\omega(x) = \exp(-x)$ et à l'intervalle $(0, \infty)$, nous en déduisons donc que les racines de chacun des polynômes sont réelles, simples et comprises dans l'intervalle $(0, \infty)$.

## 6. Calcul des racines des premiers polynômes de Laguerre

On calcule au moyen de la méthode de Bairstow les racines des polynômes de Laguerre. Les valeurs numériques du polynôme de degré 12 sont données dans le tableau 9.3, page 149.

## 7. Calcul des poids $H_k$ correspondant aux racines $x_k$

Reprenons la relation (B.5) de l'annexe B dans laquelle nous explicitons le coefficient $K$. Nous avons :

$$K = -(n!)^2$$

car le rapport des coefficients de degré le plus élevé de deux polynômes consécutifs est égal à $-1$. Donc,

$$H_k = \frac{-(n!)^2}{L'_{n+1}(x_k)L_n(x_k)}$$

expression dans laquelle, rappelons-le, $x_k$ est une racine du polynôme de degré $(n+1)$. Si nous utilisons la relation de récurrence (9.3), on obtient :

$$xL'_{n+1}(x_k) = -(n+1)^2 L_n(x_k),$$

ce qui nous permet d'obtenir deux expressions plus simples :

$$H_k = \frac{(n!)^2 x_k}{(n+1)^2 L_n^2(x_k)} = \frac{(n+1)!^2}{x_k L'_{n+1}(x_k)} \cdot \tag{9.5}$$

## 8. Calcul numérique des poids $H_k$ associés aux racines

Les deux expressions que nous venons d'établir, très faciles à programmer, nous fournissent les valeurs numériques des $H_k$ dont on donne un tableau 9.3, page 149.

Sur le Web(*), on trouvera le programme `r_laguer.c` qui permet de calculer les racines et les poids des polynômes de Laguerre.

## 9. Calcul des intégrales du type $I = \int_0^\infty \exp(-x) f(x) dx$

Comme d'habitude on suppose que $f(x)$ est une fonction régulière sur $(0, \infty)$ et l'on approche l'intégrale au moyen de l'expression :

$$J = \sum_{k=0}^{n} H_k f(x_k).$$

### 9.1. Exemple 1

On se propose de calculer la constante d'Euler $\gamma = 0{,}577\,215\,664\,90$ au moyen de l'expression suivante :

$$\gamma = \int_0^\infty \left( \frac{1}{1-\exp(-x)} - \frac{1}{x} \right) \exp(-x)\, dx$$

Les résultats trouvés sont indiqués dans le tableau 9.1, page suivante.

* `http://www.edpsciences.com/guilpin/`

**Tableau 9.1.**

| Nombre de points | Constante d'Euler |
|---|---|
| 5 | 0,577 215 409 |
| 6 | 0,577 215 316 |
| 7 | 0,577 215 638 |
| 8 | 0,577 215 683 |
| 9 | 0,577 215 671 |
| 10 | 0,577 215 664 927 |
| 11 | 0,577 215 664 244 |

## 9.2. Exemple 2

Il s'agit de calculer la fonction factorielle, et nous avons choisi :

$$10! = 3\,628\,800.$$

Compte tenu de l'expression intégrale de la fonction factorielle, nous savons :

$$I = \int_0^\infty \exp(-x)x^{10}\,\mathrm{d}x = 10!$$

Puisque la fonction à intégrer est de degré 10, nous devons obtenir un résultat exact en 6 points (en négligeant les erreurs d'arrondi propagées par la machine. C'est ce que l'on se propose de vérifier. Les résultats obtenus sont donnés dans le tableau 9.2.

**Tableau 9.2.**

| Nombre de points | 10! |
|---|---|
| 4 | 3 033 216 |
| 5 | 3 614 400 |
| 6 | 3 628 799,999 999 97 |

## 9.3. Remarque importante

On note une fois de plus que la méthode donne d'excellents résultats. Toutefois il convient de se méfier des intégrales du type :

$$I = \int_0^\infty \exp(-x)\cos mx\,\mathrm{d}x = \frac{1}{m^2+1}$$

qu'il faut examiner soigneusement. En effet, si $m$ est « assez grand », $\cos mx$ peut être assez mal représenté par un polynôme de Laguerre de degré peu élevé. D'autres cas bien sûr peuvent présenter des dangers analogues liés au nombre de points où la fonction s'annule.

## 10. Calcul de l'erreur commise lors de l'approximation

Nous allons reprendre le calcul réalisé lors de l'étude de la méthode de Gauss-Legendre, seuls quelques points de détail différent.

Soit à calculer :

$$I = \int_0^\infty \exp(-x) f(x)\, \mathrm{d}x.$$

Supposons que $f(x)$ soit continûment différentiable sur l'intervalle fondamental $(0, \infty)$. Le développement de Taylor-MacLaurin à l'ordre $(2n+1)$ nous donne :

$$f(x) = f(0) + \frac{x}{1!} f'(0) + \frac{x^2}{2!} f''(0) + \cdots + \frac{x^{2n+1}}{(2n+1)!} f^{(2n+1)}(0) + \frac{x^{2n+2}}{(2n+2)!} f^{(2n+2)}(\xi),$$

expression dans laquelle $x$ et $\xi$ appartiennent à l'intervalle $(0, \infty)$.

Calculons $I$ en remplaçant $f(x)$ par son développement, il faudra utiliser des expressions du type :

$$Q_{2n+2} = \int_0^\infty \exp(-x) x^{2n+2}\, \mathrm{d}x = (2n+2)!$$

L'exploitation directe de la formule d'intégration de Gauss-Laguerre donne :

$$J = f(0) \sum_{k=0}^{n} H_k + f'(0) \sum_{k=0}^{n} H_k x_k + \frac{f''(0)}{2!} \sum_{k=0}^{n} H_k x_k^2 + \cdots + \frac{f^{(2m)}(0)}{(2m)!} \sum_{k=0}^{n} H_k x_k^m$$
$$+ \cdots + \frac{f^{(2n+1)}(0)}{(2n+1)!} \sum_{k=0}^{n} H_k x_k^{2n+1} + \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \sum_{k=0}^{n} H_k x_k^{2n+2}.$$

L'erreur $E$ s'écrit :

$$E = f(0) \left(1 - \sum_{k=0}^{n} H_k\right) + f'(0) \left(1 - \sum_{k=0}^{n} H_k x_k\right) + f^{(q)}(0) \left(1 - \frac{1}{q!} \sum_{k=0}^{n} H_k x_k^q\right)$$
$$+ \cdots + f^{(2n+2)}(\xi) \left(1 - \frac{1}{(2n+2)!} \sum_{k=0}^{n} H_k x_k^{2n+2}\right).$$

Si $f(x)$ est un polynôme de degré $(2n+1)$, la dérivée de $f(x)$ d'ordre $(2n+2)$ est nulle, et l'on a l'égalité :

$$I = J.$$

Dans ce cas il n'y a pas d'erreur mathématique due à une approximation ; on obtient en conséquence le système suivant :

$$\begin{aligned}
\sum_{k=0}^{n} H_k &= 1 \\
\sum_{k=0}^{n} H_k x_k &= 1! \\
\sum_{k=0}^{n} H_k x_k^2 &= 2! \\
&\cdots\cdots\cdots\cdots \\
\sum_{k=0}^{n} H_k x_k^{2m-1} &= (2m-1)! \\
\sum_{k=0}^{n} H_k x_k^{2m} &= (2m)! \\
&\cdots\cdots\cdots\cdots \\
\sum_{k=0}^{n} H_k x_k^{2n+1} &= (2n+1)!
\end{aligned}$$

Il est intéressant de constater que le système de $(2n+2)$ équations ainsi défini est constitué de $(2n+2)$ équations non linéaires dans la mesure où l'on considère que les $H_k$ et les $x_k$ sont les inconnues. En revanche, si l'on connaît les $x_k$, alors, $(n+1)$ équations quelconques prises dans le système précédent constituent un système linéaire permettant de calculer les $H_k$.

Toujours est-il que l'erreur prend la forme suivante :

$$E = |I - J| = f^{(2n+2)}(x) \left\{ 1 - \frac{1}{(2n+2)!} \sum_{k=0}^{n} H_k x_k^{2n+2} \right\}, \tag{9.6}$$

en prenant soin de noter que :

$$\sum_{k=0}^{n} H_k x_k^{2n+2} \neq (2n+2)!$$

Bien entendu, comme toutes les fois en pareil cas, on recherche une majoration la plus raisonnable possible de la fonction $f^{(2n+2)}(\xi)$ dans l'intervalle $(-1, +1)$, en définitive, l'erreur s'exprime de la manière suivante :

$$E(x) = M_{2n+2} \left\{ 1 - \frac{1}{(2n+2)!} \sum_{k=0}^{n} H_k x_k^{2n+2} \right\}, \tag{9.7}$$

où $M_{2n+2} = \sup \left| f^{(2n+2)}(x) \right|$ pour $x$ appartenant à $(0, \infty)$.

Dans la mesure où l'on peut obtenir une majoration raisonnable $M_{2n+2}$, $E(x)$ donne l'erreur mathématique liée à la troncature de la série des polynômes de Laguerre. Viendront s'ajouter lors des calculs effectifs les inévitables erreurs d'arrondi propagées par l'exécution des opérations. Pour $n = 12$, on trouve :

$$E_{24} = 3,698\, 10^{-7} M_{24}.$$

Le tableau 9.3 donne les racines et les poids du polynôme de degré 12. Sur sur le Web (*) on trouve le fichier texte appelé `laguerre.txt` qui donne les poids et les racines des 12 premiers polynômes.

**Tableau 9.3.** *Racines et poids associés des polynômes de Laguerre.*

| Degré | Racines $x$ | Poids $H_k$ |
|---|---|---|
| 12 | 0,611 757 484 515 13 | 0,377 759 275 873 127 |
| | 0,115 722 117 358 02 | 0,264 731 371 055 443 |
| | 2,833 751 337 743 36 | 0,090 449 222 211 682 |
| | 1,512 610 269 776 45 | 0,244 082 011 319 895 |
| | 6,844 525 453 118 77 | 0,002 663 973 541 842 |
| | 4,599 227 639 418 09 | 0,020 102 381 154 650 |
| | 13,006 054 993 319 3 | 0,000 008 365 055 857 |
| | 9,621 316 842 445 915 | 0,000 203 231 592 668 |
| | 22,151 090 379 373 | 0,000 000 001 342 391 |
| | 17,116 855 187 466 | 0,000 000 166 849 388 |
| | 37,099 121 044 461 | 0,000 000 000 000 001 |
| | 8,487 967 251 005 | 0,000 000 000 003 062 |

## 11. Fonction génératrice des polynômes de Laguerre

Nous admettrons le résultat suivant sans démonstration :

$$\frac{\exp\left(-\dfrac{xt}{1-t}\right)}{1-t} = L_0(x) + tL_1(x) + t^2L_2(x) + \cdots + t^nL_n(x) + \cdots$$

## 12. Calcul numérique de la transformée de Laplace

$F(p)$ la transformée de Laplace de la fonction $h(t)$ est une équation fonctionnelle qui s'écrit :

$$F(p) = \int_0^\infty h(t)\exp(-pt)\,\mathrm{d}t,$$

expression dans laquelle $t$ est une variable réelle (généralement le temps) et $p$ une variable complexe. Sans entrer dans les détails de la théorie, les calculs sont réalisables si l'intégrale est convergente. Sous réserve de cette convergence, on peut être amené à calculer numériquement $F(p)$. En effectuant le changement de variable $y = pt$, on obtient l'expression :

$$F(p) = \frac{1}{p}\int_0^\infty h(y/p)\exp(-y)\,\mathrm{d}y$$

* `http://www.edpsciences.com/guilpin/`

qui se calcule évidemment par la méthode de Gauss-Laguerre. D'un point de vue pratique, les calculs ne sont pas toujours aussi simples, notamment si la fonction $h(y/p)$ a à peu près l'allure d'un pic entre zéro et $x_{\min}$ ($x_{\min}$ est la plus petite racine de $L_n(x)$ polynôme de Laguerre de degré $n$ servant à l'intégration numérique). Bien entendu, plus le degré $n$ est élevé, plus $x_{\min}$ tend vers zéro et meilleure est la précision du calcul. Malheureusement, en calcul usuel, il est difficile de dépasser le degré 12 ou 13. Nous n'allons pas renoncer pour autant à utiliser cette technique. Il suffit de couper judicieusement l'intervalle $(0, \infty)$ en deux de la façon suivante :

$$F(p) = \frac{1}{p}\int_0^a h(y/p)\exp(-y)\,\mathrm{d}y + \frac{1}{p}\int_a^\infty h(y/p)\exp(-y)\,\mathrm{d}y$$

$a$ étant déterminé par des considérations sur la fonction $h(y/p)$. La première intégrale se calcule par la méthode de Gauss-Legendre, la méthode de Simpson etc., tandis que la seconde intégrale se calcule par la méthode de Gauss-Laguerre en posant $z = y - a$, soit :

$$\frac{1}{p}\int_a^\infty h(y/p)\exp(-y)\,\mathrm{d}y = \frac{1}{p}\int_0^\infty h[(z+a)/p]\exp(-z)dz.$$

L'intérêt est évident dans la mesure où l'on a une très bonne estimation de la queue de la fonction.

## 13. Appendice : Les polynômes de Laguerre généralisés

Les polynômes de Laguerre sont susceptibles d'être étendus en écrivant une formule de définition plus générale :

$$L_n^\alpha(x) = \exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[\exp(-x)x^{n+\alpha}] \tag{9.8}$$

avec toutefois une restriction concernant le paramètre $\alpha$, il faut que sa partie réelle soit plus grande que $-1$.

### 13.1. Relation de récurrence entre trois polynômes consécutifs

Partant de la relation de définition (9.8), écrite pour le polynôme $L_{n+1}^\alpha(x)$, on obtient :

$$L_{n+1}^\alpha(x) = \exp(x)\frac{x^{-\alpha}}{(n+1)!}\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}}\left[\exp(-x)x^{n+\alpha+1}\right]$$

expression qui peut s'écrire sous la forme :

$$(n+1)L_{n+1}^\alpha(x) = \exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\exp(-x)x^{n+\alpha}(n+\alpha+1-x)\right]$$

ce qui nous donne :

$$(n+1)L_{n+1}^\alpha(x) = (n+\alpha+1)L_n^\alpha(x).$$

En utilisant la relation de Leibnitz, on peut transformer la dernière expression :

$$\exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[\exp(-x)x^{n+\alpha}x] = \exp(x)\frac{x^{-\alpha+1}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[\exp(-x)x^{n+\alpha}] + \exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}[\exp(-x)x^{n+\alpha}]$$

Comme par ailleurs on peut écrire :

$$nL_n^\alpha(x) = n\exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[(n+\alpha)\exp(-x)x^{n+\alpha-1} - \exp(-x)x^{n+\alpha}\right]$$

puis,

$$-n\exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}[\exp(-x)x^{n+\alpha}] = nL_n^\alpha(x) - (n+\alpha)\exp(-x)L_{n-1}^\alpha(x)\,;$$

on trouve en définitive la relation de récurrence cherchée :

$$(n+1)L_{n+1}^\alpha(x) - (2n+1+\alpha-x)nL_n^\alpha(x) + (n+\alpha)L_{n-1}^\alpha(x) = 0. \tag{9.9}$$

## 13.2. Relation de récurrence faisant intervenir les dérivées

Dérivons la relation de définition (9.8) :

$$[L_n^\alpha(x)]' = \frac{1}{n!}[\exp(x)x^{-\alpha} - \alpha x^{-\alpha-1}\exp(x)]\frac{\mathrm{d}^n}{\mathrm{d}x^n}[\exp(-x)x^{n-\alpha}] + \exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[(n+\alpha) \times \exp(-x)x^{n+\alpha-1} - \exp(-x)x^{n+\alpha}]$$

On obtient alors :

$$[L_n^\alpha(x)]' = -\frac{\alpha}{x}L_n^\alpha(x) + (n+\alpha)\exp(x)\frac{x^{-\alpha}}{n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\exp(-x)x^{n+\alpha-1}\right].$$

Grâce à la formule de Leibnitz, nous allons transformer la dernière expression du membre de droite, ce qui donne :

$$(n+\alpha)\exp(x)\frac{x^{-\alpha}}{n!}\frac{x}{x}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\exp(-x)x^{n+\alpha-1}\right] = (n+\alpha)\exp(x)\frac{x^{-\alpha}}{n!x}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\exp(-x)x^{n+\alpha}\right] - n\frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\left[\exp(-x)x^{n+\alpha-1}\right] = \frac{n+\alpha}{x}L_n^\alpha(x) - \frac{n+\alpha}{x}L_{n-1}^\alpha(x).$$

De là nous tirons une relation de récurrence faisant apparaître la dérivée d'un polynôme dont nous aurons besoin pour calculer les $H_k$ :

$$x[L_n^\alpha(x)]' = n[L_n^\alpha(x)] - (n+\alpha)[L_{n-1}^\alpha(x)]. \tag{9.10}$$

## 13.3. Orthogonalité des polynômes de Laguerre généralisés

Il s'agit de calculer l'expression :

$$J = \int_0^\infty \exp(-x)x^\alpha[L_n^\alpha(x)][L_m^\alpha(x)]\,\mathrm{d}x$$
$$= \int_0^\infty \exp(-x)x^{-\alpha}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left[\exp(-x)x^{n+\alpha}\right]\frac{\mathrm{d}^m}{\mathrm{d}x^m}\left[\exp(-x)x^{m+\alpha}\right]\,\mathrm{d}x.$$

En intégrant par parties d'une façon analogue à celle que nous avons envisagée lors de l'étude des polynômes de Laguerre, on obtient $J = 0$ pour $n \neq m$.

Reste à calculer la norme :

$$I = \int_0^\infty \exp(-x)x^{-\alpha}[L_n^\alpha(x)]^2 \, \mathrm{d}x = n! \int_0^\infty \exp(-x)x^{-a+n} \, \mathrm{d}x,$$

ce qui donne :

$$I = n!\Gamma(\alpha + n + 1)$$

expression dans laquelle $\Gamma(u)$ est la fonction factorielle :

$$\Gamma(u) = \int_0^\infty \exp(-t)t^{u-1} \, \mathrm{d}t.$$

## 13.4. Calcul des racines de $L_n^\alpha$ et des poids associés $H_k$

On ne peut pas dresser une table des racines et des poids associés puisque ces valeurs dépendent du paramètre $\alpha$. Dans chaque cas particulier que l'on sera amené à considérer, il faudra construire la chaîne de calculs déjà réalisée pour les cas précédemment étudiés.

## 13.5. Technique de calcul des intégrales du type $I = \int_0^\infty \exp(-x)x^{-\alpha}f(x)dx$

Bien entendu, $f(x)$ est une fonction régulière sur $(0, \infty)$, et l'approximation s'effectue comme toujours au moyen de la relation :

$$I = \sum_{k=0}^{n} H_k f(x_k),$$

expression dans laquelle les $x_k$ sont les zéros du polynôme de Laguerre généralisé $L_{n+1}^\alpha(x)$ de degré $(n + 1)$.

## 13.6. Une autre relation de récurrence

On peut avantageusement utiliser une relation de récurrence portant sur les $L_{n+1}^\alpha(x)$ pour calculer les poids $H_k$. En effet, on établit que :

$$[L_n^\alpha(x)]' = -\frac{\alpha}{x}L_n^\alpha(x) + \frac{n+\alpha}{x}L_n^{a-1}(x).$$

Si $x_k$ est une racine de $L_n^\alpha(x)$, on simplifie l'expression précédente :

$$[L_n^\alpha(x)]' = \frac{n+\alpha}{x_k}L_n^{a-1}(x_k).$$

Nous obtenons ainsi une forme aisée à calculer pour les $H_k$ où nous n'avons plus à exprimer formellement la dérivée.

# 10 Les polynômes d'Hermite. La méthode d'intégration de Gauss-Hermite

Ce sont des polynômes orthogonaux relativement à la fonction poids gaussienne et sur l'intervalle $(-\infty, +\infty)$. Nous allons en effectuer une étude en tout point semblable à celle des polynômes de Laguerre. Pour des raisons qui apparaîtront un peu plus loin, le polynôme d'Hermite (1822–1901) de degré $n$ est défini par la relation :

$$K_n(x) = (-1)^n \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-x^2/2\right). \tag{10.1}$$

## 1. Relation de récurrence entre trois polynômes consécutifs

Pour obtenir la relation désirée, il nous suffit de développer l'expression de $K_{n+1}(x)$, soit :

$$K_{n+1}(x) = (-1)^{n+1} \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left[x \exp\left(-x^2/2\right)\right]$$

à laquelle on applique la formule de Leibnitz :

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n} x f(x) = x \frac{\mathrm{d}^n}{\mathrm{d}x^n} f(x) + \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}} f(x)$$

ce qui nous permet d'écrire :

$$(-1)^{n+1} K_{n+1}(x) = \exp(x^2/2) \left( -x \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-x^2/2\right) - n \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}} \exp\left(-x^2/2\right) \right).$$

On en déduit immédiatement :

$$(-1)^{n+1} K_{n+1}(x) = -(-1)^n x K_n(x) - (-1)^{n-1} n K_{n-1}(x)$$

soit encore :

$$K_{n+1}(x) = x K_n(x) - n K_{n-1}(x). \tag{10.2}$$

## 2. Relation de récurrence entre polynômes et dérivées

Dans le même esprit que précédemment, nous allons établir une relation de récurrence simple entre les polynômes et les dérivées à seule fin d'obtenir une expression de $H_k$ aisément calculable. Pour ce faire nous dérivons la relation de définition et nous obtenons :

$$K'_n(x) = (-1)^n \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-x^2/2\right) + (-1)^n \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \left[-x \exp\left(-x^2/2\right)\right],$$

ce qui donne grâce à la formule de Leibnitz :

$$K'_n(x) = nK_{n-1}(x). \tag{10.3}$$

## 3. Les premiers polynômes d'Hermite

Puisque nous bénéficions de la relation de récurrence (10.2), il suffit en réalité de connaître les deux premiers polynômes d'Hermite pour déterminer tous les autres. Nous obtenons :

$$\begin{aligned} K_0(x) &= 1 \\ K_1(x) &= x \\ K_2(x) &= x^2 - 1 \\ K_3(x) &= x^3 - 3x \\ K_4(x) &= x^4 - 6x^2 + 3 \end{aligned}$$

et ainsi de suite.

## 4. Calcul des coefficients des premiers polynômes d'Hermite

On trouvera sur le Web(*) le programme `hermite.c` destiné à calculer les coefficients des polynômes d'Hermite dans le but plus précis de calculer leurs racines.

## 5. Orthogonalité des polynômes d'Hermite

Les polynômes d'Hermite sont orthogonaux relativement à la fonction poids $\exp\left(-x^2/2\right)$ et à l'intervalle $(-\infty, +\infty)$. C'est ce que nous allons montrer en calculant l'intégrale $I_{nk}$. Sans nuire à la généralité, on peut supposer que $k > n$ :

$$I_{nk} = \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) K_n(x) K_k(x)\, \mathrm{d}x,$$

soit encore :

$$I_{nk} = (-1)^{n+k} \int_{-\infty}^{+\infty} \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-x^2/2\right) \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\left(-x^2/2\right)\, \mathrm{d}x.$$

* `http://www.edpsciences.com/guilpin/`

Effectuons une intégration par parties en posant :

$$\mathrm{d}v = \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\left(-x^2/2\right)\ \mathrm{d}x$$

soit

$$v = \frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}} \exp\left(-x^2/2\right)$$

et

$$u = \exp(x^2/2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-x^2/2\right)$$

soit, après transformation donnée par la formule de Leibnitz :

$$du = -n \exp(x^2/2) \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}} \exp\left(-x^2/2\right)$$

Dans le cas où $n \neq k$, le produit $u \cdot v$ est nul sur l'intervalle $(-\infty, +\infty)$ car il contient $\exp\left(-x^2/2\right)$ en facteur. On peut alors écrire :

$$I_{nk} = I_{n-1k-1}$$

et de proche en proche on obtient :

$$I_{nk} = n! \int_{-\infty}^{+\infty} \frac{\mathrm{d}^{k-n}}{\mathrm{d}x^{k-n}} \exp\left(-x^2/2\right)\ \mathrm{d}x$$

Or

$$\int_{-\infty}^{+\infty} \frac{\mathrm{d}^{k-n}}{\mathrm{d}x^{k-n}} \exp\left(-x^2/2\right)\ \mathrm{d}x = 0 \qquad \text{quand } k \neq n.$$

Donc les polynômes d'Hermite sont orthogonaux. On en déduit immédiatement que les zéros d'un polynôme d'Hermite de degré $n$ sont distincts, réels et répartis sur l'intervalle $(-\infty, +\infty)$.

Il nous reste à présent à calculer la norme d'un polynôme, il suffit de faire $n = k$ dans l'expression de $I_{nk}$ :

$$I_{nn} = I_n = n! \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right)\ \mathrm{d}x = n!\sqrt{2\pi}.$$

## 6. Calcul des racines des premiers polynômes d'Hermite

On trouvera un tableau 10.2, page 162, donnant les racines du polynôme d'Hermite de degré 12 que nous avons calculées en double précision.

Sur le Web(*), on donne le programme `r_hermit.c` qui calcule les racines et les poids des polynômes d'Hermite.

* `http://www.edpsciences.com/guilpin/`

## 7. Calcul des poids $H_k$ correspondant aux racines $x_k$

Rappelons tout d'abord que ce calcul est effectué dans l'hypothèse où les racines $x_k$ sont les $(n+1)$ zéros du polynôme de degré $(n+1)$. La relation (B.5) de l'annexe B s'écrit alors :

$$H_k = \frac{\mathcal{K}}{K'_{n+1}(x_k)K_n(x_k)}$$

expression dans laquelle on aura soin de ne pas confondre la constante $\mathcal{K}$ avec un quelconque polynôme d'Hermite. On trouve que la constante $\mathcal{K}$ vaut $n!\sqrt{2\pi}$.

Grâce à la relation (10.3), on obtient diverses expressions de $H_k$ aisées à calculer, soit :

$$H_k = \frac{n!\sqrt{2\pi}}{K'_{n+1}(x_k)K_n(x_k)} = \frac{(n+1)!\sqrt{2\pi}}{[K'_{n+1}(x_k)]^2} = \frac{n!\sqrt{2\pi}}{(n+1)[K_n(x_k)]^2}$$

En exploitant une de ces relations, on calcule numériquement sans difficultés les poids $H_k$ associés aux racines $x_k$. On trouvera dans le tableau 10.2, page 162 les valeurs numériques de ces poids en face des racines correspondantes pour le polynôme de degré 12.

## 8. Technique de calcul des intégrales du type $I = \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) f(x)dx$

où $f(x)$ est une fonction régulière sur l'intervalle $(-\infty, +\infty)$. L'approximation se réalise au moyen de la somme :

$$J = \sum_{k=0}^{n} H_k f(x_k).$$

expression dans laquelle les $x_k$ sont les racines du polynôme de degré $(n+1)$.

### Exemple

On se propose de calculer une intégrale dont on connaît la valeur numérique, soit :

$$I = \sqrt{\frac{2\pi}{e}} \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) \cos(x)\, \mathrm{d}x = 1.$$

Les résultats que nous avons obtenus sont donnés dans le tableau 10.1, page ci-contre.

***Remarque*** : On note qu'il n'y a pas de difficultés pour calculer les intégrales du type :

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{\sigma^2}\right) f(x)\, \mathrm{d}x$$

en effet, le changement de variable :

$$y = \frac{x\sqrt{2}}{\sigma}$$

nous permet de revenir au cas canonique que nous venons de traiter.

**Tableau 10.1.**

| Nombre de points | Résultat |
|---|---|
| 4 | 0,999 222 |
| 5 | 1,000 043 |
| 6 | 0,999 998 143 |
| 7 | 1,000 000 175 |
| 8 | 1,000 000 097 |
| 9 | 1,000 000 099 |
| 12 | 1,000 000 099 |

## 9. Autres notations très utiles

Il arrive que des auteurs adoptent une définition légèrement différente pour les polynômes d'Hermite, soit :

$$K_n^*(x,a) = (-1)^n \exp(ax^2)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\exp(-ax^2),$$

expression dans laquelle ***a* est un nombre positif** (il s'agit d'une formule analogue à celle de Rodriguès concernant les polynômes de Legendre).

Dans ce cas, la relation de récurrence entre trois polynômes consécutifs s'écrit :

$$K_{n+1}^*(x,a) = 2axK_n^*(x,a) - 2anK_{n-1}^*(x,a).$$

De même, la relation faisant intervenir les dérivées prend la forme :

$$K_n^{*\prime}(x,a) = 2anK_{n-1}^*(x,a),$$

et le carré scalaire de l'intégrale vaut alors :

$$I_{nn} = I_n = n!\frac{\sqrt{2}}{2}(2a)^n.$$

En remarquant que :

$$K_n(x) = \frac{1}{(2\sqrt{a})^n}K_n^*\left(\frac{x}{2\sqrt{a}}, a\right),$$

on obtient les modifications à apporter aux expressions des racines et des poids des polynômes correspondants, à savoir :

$$x_k^* = x_k\frac{1}{\sqrt{2a}}$$
$$H_k^* = H_k\frac{1}{\sqrt{2a}}\cdot$$

Les fonctions $D_n^*(x,a) = \exp(-ax^2/2)K_n^*(x)$, encore appelées fonctions du cylindre parabolique ou fonctions de Weber-Hermite, constituent une base orthogonale de l'espace $L^2$ qui est l'espace des fonctions de carré sommable sur $(-\infty,+\infty)$ ( $\int_{-\infty}^{\infty} f(x)^2\,\mathrm{d}x$).

La décomposition d'une fonction $f(x)$ appartenant à $L^2$ sur $(-\infty, +\infty)$ s'effectue en calculant les coefficients du développement de Fourier correspondant, soit :

$$f(x) = \sum_{k=0}^{\infty} c_k D_k^*(x, a).$$

On multiplie les deux membres de cette dernière expression par $D_m^*(x)$ puis on procède à l'intégration sur l'intervalle $(-\infty, \infty)$, on obtient :

$$\int_{-\infty}^{+\infty} f(x) D_m^*(x, a)\, \mathrm{d}x = \int_{-\infty}^{+\infty} \sum_{k=0}^{\infty} c_k D_k^*(x) D_m^*(x, a)\, \mathrm{d}x = \sum_{k=0}^{\infty} c_k \int_{-\infty}^{+\infty} D_k^*(x, a) D_m^*(x, a)\, \mathrm{d}x$$

$$= \sum_{k=0}^{\infty} c_k \int_{-\infty}^{+\infty} K_k^*(x) K_m^*(x) \exp(-ax^2)\, \mathrm{d}x$$

et compte tenu de l'orthogonalité, on trouve :

$$\int_{-\infty}^{+\infty} f(x) D_m^*(x, a)\, \mathrm{d}x = c_m I_m = c_m m! \frac{\sqrt{2}}{2} (2a)^m$$

et

$$c_m = \frac{1}{m! \dfrac{\sqrt{2}}{2} (2a)^m} \int_{-\infty}^{+\infty} f(x) D_m^*(x, a)\, \mathrm{d}x$$

ainsi que

$$c_0 = \sqrt{2} \int_{-\infty}^{+\infty} f(x) \exp(-ax^2)\, \mathrm{d}x.$$

On trouvera la démonstration de la convergence des intégrales $c_m$ ainsi que celle de la série associée à la fonction $f(x)$ dans les ouvrages de Arsac et de Bass cités en bibliographie.

## Un cas particulier intéressant

C'est le cas où $a = 2\pi$ qui permet d'obtenir des résultats simples lorsque l'on est amené à travailler dans l'espace réciproque de Fourier (*cf.* chapitre 16).

Nous écrivons :

$$K_n^*(x, 2\pi) = (-1)^n \left( n! \frac{\sqrt{2}}{2} (4\pi)^n \right)^{1/2} \exp\left(2\pi x^2\right) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp\left(-2\pi x^2\right),$$

expression qui rend les polynômes orthonormés.

En rappelant que la transformation de Fourier directe s'écrit :

$$F(u) = \int_{-\infty}^{+\infty} f(x) \exp(2\pi j x u)\, \mathrm{d}x,$$

soit :

$$f(x) \xrightarrow{TF} F(u)$$

avec une propriété importante :

$$x^p f(x) \xrightarrow{TF} \frac{1}{(2\pi j)^p} F^{(p)}(u).$$

Nous allons montrer que :

$$K_n^*(x, 2\pi)\exp(-\pi x^2) \xrightarrow{TF} j^n K_n^*(u, 2\pi)\exp(-\pi u^2).$$

Si $f(x) = x^p \exp(-\pi x^2)$ on obtient pour sa transformée de Fourier (*cf.* chapitre 16) :

$$x^p \exp(-\pi x^2) \xrightarrow{TF} \frac{1}{(2\pi j)^p} \frac{\mathrm{d}^p}{du^p} \exp(-\pi u^2).$$

Or, en faisant usage de la relation générale de définition des polynômes d'Hermite, on peut écrire :

$$\frac{1}{(2\pi j)^p} \frac{\mathrm{d}^p}{du^p} \exp(-\pi u^2) = P_p(u)\exp(-\pi u^2)$$

où $P_p(u)$ est un polynôme de degré $p$. Comme tout polynôme d'Hermite peut se développer de la façon suivante :

$$K_n^*(x, 2\pi) = \sum_{p=0}^{n} a_p x^p,$$

on en déduit que

$$K_n^*(x, 2\pi)\exp(-\pi x^2) \xrightarrow{TF} Q_n(u)\exp(-\pi u^2).$$

Il s'agit maintenant de déterminer le polynôme $Q_n(u)$. La démonstration repose sur les propriétés des produits scalaires dans $L^2$, et en particulier, de la relation suivante :

$$\langle f(x), f(-x)\rangle = \langle F(u), F(u)\rangle$$

dont on trouvera une démonstration dans le chapitre 16 concernant les transformées de Fourier. Comme la relation de définition permet d'écrire :

$$K_n^*(-x, 2\pi) = (-1)^n K_n^*(x, 2\pi),$$

on en déduit que :

$$\begin{aligned}\left\langle K_n^*(-x, 2\pi)\exp(-\pi x^2), K_m^*(x, 2\pi)\exp(-\pi x^2)\right\rangle \\ = (-1)^n \delta_{nm} = \left\langle Q_n(x, 2\pi)\exp(-\pi x^2), Q_m(x, 2\pi)\exp(-\pi x^2)\right\rangle.\end{aligned}$$

À présent, changeons $Q_m(x, 2\pi)$ en $(j)^m R_m(x, 2\pi)$, il vient :

$$\left\langle R_n(x, 2\pi)\exp(-\pi x^2), R_m(x, 2\pi)\exp(-\pi x^2)\right\rangle = \delta_{nm},$$

on en déduit que :

$$R_m(x, 2\pi) = K_m^*(x, 2\pi),$$

et que, par conséquent, on a :

$$K_m^*(x, 2\pi)\exp(-\pi x^2) \xrightarrow{TF} \pm j^n K_m^*(u, 2\pi)\exp(-\pi u^2),$$

l'ambiguïté du signe se trouve levée en examinant le terme de plus haute puissance (coefficient $a_p$), mais on a déjà établi la relation suivante :

$$a_p x^p \exp(-\pi x^2) \xrightarrow{TF} a_p \frac{1}{(2\pi j)^p}\frac{\mathrm{d}^p}{du^p}\exp(-\pi u^2),$$

ce qui permet de conclure que le signe à conserver est le signe plus.

## 10. Calcul de l'erreur commise lors de l'approximation

Nous allons reprendre le calcul réalisé lors de l'étude de la méthode de Gauss-Legendre, seuls quelques points de détails vont différer.

Soit à calculer :

$$I = \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) f(x)\,\mathrm{d}x$$

Supposons que $f(x)$ soit continûment différentiable sur l'intervalle fondamental $(-\infty, +\infty)$. Le développement de Taylor-MacLaurinà l'ordre $(2n+2)$ nous donne :

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^{2n+1}}{(2n+1)!}f^{(2n+1)}(0) + \frac{x^{2n+2}}{(2n+2)!}f^{(2n+2)}(\xi),$$

expression dans laquelle $x$ et $\xi$ appartiennent à l'intervalle $(-\infty, +\infty)$. Calculons $I$ en remplaçant $f(x)$ par son développement, il faudra calculer des expressions du type :

$$Q_{2n+2} = \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) x^{2n+2}\,\mathrm{d}x = \int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) x^{2n+1} x\,\mathrm{d}x.$$

On intègre par parties cette expression en posant :

$$v = x^{2n+1}x \qquad \text{et} \quad \mathrm{d}u = \exp\left(-x^2/2\right) x\,\mathrm{d}x$$

ce qui donne :

$$\mathrm{d}v = (2n+1)x^{2n}\,\mathrm{d}x \qquad \text{et} \quad u = -\exp\left(-x^2/2\right).$$

On peut écrire :

$$Q_{2n+2} = \left[-\exp\left(-x^2/2\right) x^{2n+1}\right]_{-\infty}^{+\infty} + (2n+1)\int_{-\infty}^{+\infty} \exp\left(-x^2/2\right) x^{2n}\,\mathrm{d}x,$$

comme

$$\int_{-\infty}^{+\infty} \exp\left(-x^2/2\right)\, \mathrm{d}x = \sqrt{2\pi},$$

on trouve que :

$$Q_{2n+2} = (2n+1)!!\sqrt{2\pi}.$$

Par ailleurs $Q_{2n+1}$ étant une fonction impaire, $Q_{2n+1} = 0$. Revenons au calcul de $I$ :

$$I = \sqrt{2\pi}\left[f(0) + \cdots + \frac{f^{2q}(0)}{(2q)!}(2q-1)!! + \cdots + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}(2n+1!!)\right]$$

L'exploitation directe de la formule d'intégration de Gauss-Hermite donne :

$$J = f(0)\sum_{k=0}^{n} H_k + \cdots + \frac{f^{(q)}(0)}{q!}\sum_{k=0}^{n} H_k x_k^q + \cdots + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\sum_{k=0}^{n} H_k x_k^{2n+2}.$$

L'erreur $E$ s'écrit :

$$\begin{aligned} E = |I-J| = {} & f(0)\left(\sqrt{2\pi} - \sum_{k=0}^{n} H_k\right) + f'(0)\sum_{k=0}^{n} H_k x_k + \frac{f''}{2!}\left(\sqrt{2\pi} - \sum_{k=0}^{n} H_k x_k^2\right) \\ & + \cdots + \frac{f^{(2q+1)}(0)}{(n+1)(2q+1)!}\sum_{k=0}^{n} x_k^{2q+1} + \cdots + \frac{f^{(2q)}(0)}{(2q)!}\left((2q-1)!!\sqrt{2\pi} - \sum_{k=0}^{n} H_k x_k^{2q}\right) \\ & + \cdots + \frac{f(2n+2)(\xi)}{(2n+2)!}\left(\sqrt{2\pi}(2n+1)!! - \sum_{k=0}^{n} H_k x_k^{2n+2}\right). \end{aligned}$$

Si $f(x)$ est un polynôme de degré $(2n+1)$, la dérivée de $f(x)$ d'ordre $(2n+2)$ est nulle, et l'on a l'égalité :

$$I = J.$$

Dans ce cas il n'y a pas d'erreur mathématique due à une approximation ; on obtient en conséquence le système suivant :

$$\begin{aligned} \sum_{k=0}^{n} H_k &= \sqrt{2\pi} \\ \sum_{k=0}^{n} H_k x_k^{2q+1} &= 0 \\ \sum_{k=0}^{n} H_k x_k^{2p} &= (2p-1)!!\sqrt{2\pi} \end{aligned}$$

donc

$$E = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\left(\sqrt{2\pi}(2n+1)!! - \sum_{k=0}^{n} H_k x_k^{2n+2}\right).$$

Bien entendu, comme toutes les fois en pareil cas, on recherche une majoration la plus raisonnable possible de la fonction $f^{(2n+2)}(\xi)$ dans l'intervalle $(-1,+1)$, en définitive, l'erreur s'exprime de la manière suivante :

$$E(x) = \frac{M_{2n+2}}{(2n+2)!}\left|\left(\sqrt{2\pi}(2n+1)!! - \sum_{k=0}^{n} H_k x_k^{2n+2}\right)\right| \tag{10.4}$$

où $M_{2n+2} = \sup\left|f^{(2n+2)}(x)\right|$ pour $x$ appartenant à $(-\infty,+\infty)$.

Dans la mesure où l'on peut obtenir une majoration raisonnable $M_{2n+2}$, $E(x)$ donne l'erreur mathématique liée à la troncature de la série des polynômes de Laguerre. Viendront s'ajouter lors des calculs effectifs les inévitables erreurs d'arrondi propagées par l'exécution des opérations. Pour $n = 12$, on trouve :

$$E_{24} = 1{,}935\,10^{-15} M_{24}.$$

Les poids et les racines du polynôme de degré 12 figurent sur le tableau 10.2, tandis que le fichier texte appelé `hermite.txt` que l'on trouvera sur le Web(*) donne les résultats concernant les douze premiers polynômes.

**Tableau 10.2.** *Racines et poids associés des premiers polynômes d'Hermite.*

| Degré | $x$ | $H$ |
|---|---|---|
| 12 | ±0,444 403 001 944 139 | 0,806 292 983 509 187 |
| | ±1,340 375 197 151 62 | 0,368 391 758 069 477 |
| | ±2,259 464 451 000 80 | 0,072 984 713 184 739 |
| | ±3,223 709 828 770 10 | 0,005 523 056 331 147 |
| | ±4,271 825 847 932 28 | 0,000 121 250 244 966 |
| | ±5,500 901 704 467 74 | 0,000 000 375 975 985 |

## 11. Fonction génératrice des polynômes d'Hermite

Nous admettrons les résultats suivants sans démonstration :

$$\exp\left(xt - \frac{t^2}{2}\right) = K_0(x) + \frac{t}{1}K_1(x) + \frac{t^2}{2}K_2(x) + \cdots + \frac{t^n}{n}K_n(x) + \cdots$$

ou encore si l'on fait usage des fonctions du cylindre parabolique :

$$\exp\left(-\frac{x^2}{4} + xt - \frac{t^2}{2}\right) = D_0(x) + \frac{t}{1}D_1(x) + \frac{t^2}{2}D_2(x) + \cdots + \frac{t^n}{n}D_n(x) + \cdots$$

Dans le cas où l'on utilise la définition :

$$K_n^*(x,a) = (-1)^n \exp(ax^2)\frac{\mathrm{d}^n}{\mathrm{d}x^n}\exp(-ax^2),$$

* `http://www.edpsciences.com/guilpin/`

on obtient :

$$\exp\left(2axt - \frac{at^2}{2}\right) = K_0^*(x,a) + \frac{t}{1}K_1^*(x,a) + \frac{t^2}{2}K_2^*(x,a) + \cdots + \frac{t^n}{n}K_n^*(x,a) + \cdots$$

et en particulier si $a = 2\pi$ :

$$\exp\left(4\pi xt - \frac{2\pi t^2}{2}\right) = K_0^*(x,2\pi) + \frac{t}{1}K_1^*(x,2\pi) + \frac{t^2}{2}K_2^*(x,2\pi) + \cdots + \frac{t^n}{n}K_n^*(x,2\pi) + \cdots$$

## 12. Éléments de bibliographie


A. ANGOT (1972) *Compléments de Mathématiques*, Éditions Masson.

J. ARSAC (1962) *Transformation de Fourier et théorie des distributions*, Éditions Dunod.

J. BASS (1971) *Cours de mathématiques*, Tome 3, Éditions Masson.

M. CROUZEIX et A.L. MIGNOT (1984) *Analyse numérique des équations différentielles*, Éditions Masson.

H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.

S. THANGAVELU (1993) *Lectures on Hermite and Laguerre expansions*, Princeton University Press.

# 11 Calcul de quelques intégrales relevant des études précédentes au moyen d'un changement de variable

Eu égard à la grande simplicité et à la grande précision de la méthode de Gauss généralisée, on est fortement tenté de passer en revue le plus grand nombre de polynômes orthogonaux dans le but évident de calculer le plus grand nombre d'intégrales présentant des singularités.

Tout d'abord il nous faudra définir les polynômes orthogonaux relativement à la fonction poids $\omega(x)$ sur l'intervalle fondamental fini ou infini $(a, b)$, puis calculer leurs zéros $x_k$ ainsi que les $H_k$ associés. Comme trois polynômes consécutifs de la suite orthogonale obéissent à une relation de récurrence (*cf.* annexe B), on cherche à obtenir les coefficients de ladite relation de récurrence ; en règle générale, les deux premiers polynômes sont faciles à calculer.

En désignant par $W_n(x)$ le polynôme de degré $n$ de la suite orthogonale envisagée, on peut écrire une relation de la forme suivante :

$$W_{n+1}(x) - [A_n(x)x + B_n(x)]W_n(x) + C_n(x)W_{n-1}(x) = 0.$$

Ensuite, il faut être en mesure de calculer la norme d'un polynôme :

$$I = \int_a^b \omega(x) W_n^2(x)\, \mathrm{d}x$$

car cette quantité intervient dans le calcul des $H_k$. Bien que le champ d'investigation soit en principe infini, il n'y a que peu d'issues conduisant à des expressions qui puissent être aisément traitées analytiquement et qui autorisent le calcul effectif des $x_k$ et des $H_k$. D'une manière tout à fait générale, les polynômes susceptibles de déboucher sur des résultats pratiques relèvent de l'étude des polynômes hypergéométriques encore appelés polynômes de Jacobi. Nous allons passer en revue un certain nombre de cas typiques dont l'étude se ramène aux cas précédemment abordés.

## 1. Intégrale de la forme : $\mathbf{I} = \int_0^1 \frac{\mathbf{f(x)}}{\sqrt{\mathbf{1-x}}}\mathbf{dx}$

Il s'agit donc d'étudier les polynômes $V_n(x)$ orthogonaux relativement à la fonction poids $\omega(x) = \frac{1}{\sqrt{1-x}}$) sur l'intervalle fondamental $(0, 1)$. Soit :

$$\int_0^1 \frac{V_n(x)V_k(x)}{\sqrt{1-x}}\, \mathrm{d}x = 0 \qquad \text{avec } m \neq k.$$

Le changement de variable $y = \sqrt{1-x}$, soit $x = 1-y^2$, nous permet de transformer l'intégrale de la façon suivante :

$$\int_0^1 V_n(1-y^2)V_k(1-y^2)\,\mathrm{d}y = 0.$$

Remarquons que l'on peut étendre l'intégrale à l'intervalle $(-1,+1)$ puisque l'élément différentiel est une fonction paire, d'où :

$$\int_{-1}^{+1} V_n(1-y^2)V_k(1-y^2)\,\mathrm{d}y = 0.$$

La suite des polynômes $V_n(x)$, lesquels sont des polynômes pairs, n'est rien d'autre que la suite des polynômes de Legendre d'indice pair. En effet, il suffit de se reporter à la formule de Rodriguès pour vérifier que les polynômes de Legendre d'indice pair sont pairs et que les polynômes de Legendre d'indice impairs sont impairs. Il s'ensuit que l'on peut écrire que :

$$V_{n+1}(1-x^2) = P_{2n+2}(x) = V_n(y),$$

et que les racines $y_k$ sont données par l'expression :

$$y_k = 1 - x_k^2.$$

Il convient de rappeler ici que les $x_k$ sont les zéros du polynôme de Legendre de degré $(2n+2)$. Comme les racines des polynômes de Legendre sont opposées, on ne conserve que les $x_k$ positifs pour calculer les $y_k$.

Il nous reste à voir comment les $H_k$ associés aux $x_k$ doivent être calculés. Pour cela reprenons l'expression générale :

$$H_k^* = \frac{1}{V'_{n+1}(y_k)} \int_0^1 \frac{V_{n+1}(y)}{\sqrt{1-y}} \frac{1}{y-y_k}\,\mathrm{d}y$$

dans laquelle on effectue le changement de variable :

$$y = 1 - x^2, \qquad \text{soit} \quad \mathrm{d}y = -2x\,\mathrm{d}x.$$

En remarquant au préalable que :

$$\frac{\mathrm{d}}{\mathrm{d}y}V_{n+1}(y) = \frac{\mathrm{d}}{\mathrm{d}x}P_{2n+2}(x)\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{1}{2x}P'_{2n+2}(x) = V'_{n+1}(y)$$

l'expression de $H_k^*$ devient :

$$H_k^* = \frac{4x_k}{P'_{2n+2}(x_k)} \int_0^1 \frac{V_{n+1}(y)}{\sqrt{1-y}} \cdot \frac{1}{y-y_k}\,\mathrm{d}y$$

Cette dernière expression se transforme en utilisant l'identité :

$$\frac{1}{x^2-x_k^2} = \frac{1}{2x_k}\left[\frac{1}{x-x_k} - \frac{1}{x+x_k}\right].$$

On obtient alors

$$H_k^* = \frac{2}{P'_{2n+2}(x_k)} \left[ \int_0^1 \frac{P_{2n+2}(x)}{x - x_k} \, \mathrm{d}x \int_0^1 \frac{P_{2n+2}(x)}{x + x_k} \, \mathrm{d}x \right].$$

Soit en définitive :

$$H_k^* = \frac{2}{P'_{2n+2}(x_k)} \int_{-1}^{+1} \frac{P_{2n+2}(x)}{x - x_k} \, \mathrm{d}x = 2H_k$$

où, n'hésitons pas à le répéter, les $H_k$ sont les poids correspondant aux $x_k$ du polynôme de Legendre de degré $(2n + 2)$. En résumé nous avons :

$$y_k = 1 - x_k^2$$
$$H_k^* = 2H_k.$$

## 1.1. Calcul numérique des $y_k$ et des $H_k^*$ des polynômes $V_n(x)$

Il n'est pas très difficile de calculer ces grandeurs, et le tableau 11.1 donne les résultats du calcul pour le polynôme de degré 11. Sur le Web(*), on trouvera les résultats concernant les premiers polynômes dans le fichier texte `vn_x.txt`.

**Tableau 11.1.** *Zéros et poids correspondants des polynômes $V_n(x)$.*

| Degré | $y_k$ | $H_k$ |
|---|---|---|
| 11 | 0,995 136 433 756 837 | 0,278 503 745 711 265 |
| | 0,956 794 043 016 991 | 0,273 082 996 692 025 |
| | 0,883 079 894 390 853 | 0,262 347 009 574 158 |
| | 0,779 705 097 347 766 | 0,246 504 753 620 871 |
| | 0,654 678 756 166 416 | 0,225 864 592 161 418 |
| | 0,517 687 441 268 599 | 0,200 828 288 893 594 |
| | 0,379 344 680 233 946 | 0,171 883 212 409 773 |
| | 0,250 368 580 202 265 | 0,139 592 936 925 331 |
| | 0,140 751 142 618 251 | 0,104 586 670 228 415 |
| | 0,058 982 630 398 875 | 0,067 549 803 115 903 |
| | 0,011 378 277 344 275 | 0,029 255 990 740 224 |

## 1.2. Technique d'intégration

On utilise toujours la même relation pour effectuer une approximation de la valeur de $I$, soit :

$$J = \sum_{k=0}^{n} H_k^* f(y_k).$$

* `http://www.edpsciences.com/guilpin/`

On peut également calculer les intégrales du type :

$$I = \int_a^b \frac{f(x)}{\sqrt{b-x}}\,\mathrm{d}x$$

qui se ramène au cas précédent au moyen du changement de variable :

$$y = \frac{x-a}{b-a}$$

ce qui nous permet d'écrire :

$$I = \sqrt{b-a}\int_0^1 \frac{f[a+(b-a)y]}{\sqrt{1-y}}\,\mathrm{d}y.$$

Il suffit à présent de poser $y_k^* = a+(b-a)y_k$ pour obtenir la forme désirée, $J$ est alors approchée par l'expression :

$$J = \sqrt{b-a}\sum_{k=0}^{n} H_k^* f(y_k^*).$$

## 1.3. Un exemple d'intégration

On se propose de calculer :

$$I = \int_0^1 \frac{x}{\sqrt{1-x}}\,\mathrm{d}x = 4/3 = 1{,}333\,333\,333\,333\,333.$$

Les résultats que nous avons obtenus sont donnés dans le tableau 11.2.

**Tableau 11.2.**

| Nombre de points | Résultat |
|---|---|
| 2 | 1,333 333 333 333 333 |
| 3 | 1,333 333 333 333 333 |
| 4 | 1,333 333 333 333 333 |
| 5 | 1,333 333 333 333 331 |
| 6 | 1,333 333 333 032 |

On vérifie que, puisque $f(x)$ est un polynôme du premier degré, les résultats doivent être rigoureux à partir de $n = 2$. Cependant, nous avons tenu à donner les autres résultats pour montrer que la précision se dégrade avec le nombre croissant de points car apparaissent les inévitables erreurs de troncature qui affectent la précision des racines et des poids associés ainsi que la précision de la somme.

## 2. Intégrale de la forme $\mathbf{I} = \int_0^1 \mathbf{f(x)}\sqrt{\mathbf{1-x}}\mathbf{dx}$

Il va de soi que nous allons nous intéresser aux polynômes orthogonaux $U_n(x)$ relativement à la fonction de base $\omega(x) = \sqrt{1-x}$ et l'intervalle $(0,1)$, soit :

$$\int_0^1 U_n(x)U_k(x)\sqrt{1-x}\,\mathrm{d}x = 0 \qquad \text{avec } m \neq k.$$

Le changement de variable $y = \sqrt{1-x}$, soit $x = 1-y^2$, nous permet d'écrire :

$$\int_0^1 y^2 U_n(1-y^2)U_k(1-y^2)\,\mathrm{d}x = 0 \qquad \text{avec } m \neq k.$$

Comme l'élément différentiel est une fonction paire, on en conclut que :

$$\int_{-1}^{+1} y^2 U_n(1-y^2)U_k(1-y^2)\,\mathrm{d}x = 0 \qquad \text{avec } m \neq k.$$

Par conséquent les polynômes $yU_n(1-y^2)$ sont des polynômes impairs orthogonaux sur l'intervalle $(0,1)$ relativement à la fonction de base $\omega(x) = \sqrt{1-x}$. Il s'agit là des polynômes de Legendre de degré impair qui sont de surcroît impairs. Donc :

$$xU_{n+1}(1-x^2) = P_{2n+2}(x)$$

de là on déduit que :

$$U_n(x) = \frac{1}{\sqrt{1-x}} P_{n+1}(\sqrt{1-x}).$$

On calculera les racines $y_k$ de $U_{2n+1}(y)$ à partir des racines du polynôme de degré $(2n+1)$.

Remarquons que les polynômes de Legendre de degré impair possèdent toujours une racine nulle et $2n$ racines réelles opposées, donc pour obtenir les $y_k$ il suffit de poser :

$$y_k = 1 - x_k^2.$$

La racine nulle de $P_{2n+3}(x)$ correspond à la racine nulle évidente du polynôme $yU_{n+1}(1-y^2)$.

À présent, calculons les $H_k^*$ correspondant aux $y_k$ :

$$H_k^* = \frac{1}{U'_{n+1}(y_k)} \int_0^1 \frac{U_{n+1}(y)\sqrt{1-y}}{y-y_k}\,\mathrm{d}y.$$

Le changement de variable $x = \sqrt{1-y}$ transforme cette expression en :

$$H_k^* = \frac{2}{U'_{n+1}(1-x_k^2)} \int_0^1 \frac{U_{n+1}(1-x^2)x^2}{x-x_k}\,\mathrm{d}y.$$

En notant que :

$$U'_{n+1}(y) = \frac{-2x_k^2}{P'_{2n+3}(x_k)}$$

et que

$$\frac{x}{x^2 - x_k^2} = \frac{1}{2}\left[\frac{1}{x - x_k} + \frac{1}{x + x_k}\right].$$

on aboutit finalement à une expression simple de $H_k^*$ :

$$H_k^* = \frac{2x_k^2}{P'_{2n+3}(x_k)} \int_{-1}^{+1} \frac{P_{2n+3}(x)}{x - x_k}\, \mathrm{d}x = 2H_k x_k^2.$$

Rappelons que les $x_k$ et les $H_k$ sont les zéros et les poids associés du polynôme de Legendre de degré $(2n + 3)$. En définitive, nous avons :

$$y_k = 1 - x_k^2$$
$$H_k^* = 2H_k \cdot x_k^2$$

## 2.1. Calcul numérique des $y_k$ et des $H_k^*$ des polynômes $U_n(y)$

On a reporté sur le tableau 11.3 les résultats obtenus pour le polynôme de degré 11, résultats directement déduits de l'étude des polynômes de Legendre. Sur le Web(*), on trouvera les résultats concernant les premiers polynômes dans le fichier texte `un_x.txt`.

**Tableau 11.3.** *Les zéros et les poids correspondants des polynômes $U_n(x)$.*

| Degré | $y_k$ | $H_k$ |
|---|---|---|
| 11 | 0,982 242 618 777 886 | 0,004 704 357 862 335 |
| | 0,930 232 342 038 332 | 0,017 986 900 669 840 |
| | 0,847 665 099 712 485 | 0,037 489 339 976 674 |
| | 0,740 408 244 072 688 | 0,059 704 359 522 125 |
| | 0,616 083 601 856 242 | 0,080 539 587 896 191 |
| | 0,483 525 845 139 551 | 0,095 977 183 512 452 |
| | 0,352 154 660 959 083 | 0,102 724 186 114 703 |
| | 0,231 305 302 178 075 | 0,098 750 243 709 796 |
| | 0,010 433 970 271 955 | 0,026 543 841 354 487 |
| | 0,054 161 141 423 958 | 0,058 619 320 141 303 |
| | 0,129 564 951 355 262 | 0,083 627 346 047 326 |

* `http://www.edpsciences.com/guilpin/`

## 2.2. Technique d'intégration

On approche l'intégrale par la même formule, à savoir :

$$J = \sum_{k=0}^{n} H_k^* f(y_k).$$

tandis que l'intégrale :

$$I = \int_a^b f(x)\sqrt{b-x}\,\mathrm{d}x$$

s'approche par l'expression :

$$J = (b-a)^{3/2} \sum_{k=0}^{n} H_k^* f(y_k).$$

où les valeurs $y_k^*$ sont données par la relation :

$$y_k^* = a + (b-a)y_k.$$

Ce dernier résultat est obtenu d'une manière tout à fait identique à celui établi dans le dernier paragraphe à cette différence près que l'on effectue le changement de variable $y = a + (b-a)x$ dans l'intégrale donnant $H_k^*$.

## 2.3. Un exemple d'intégration

Nous nous proposons de calculer :

$$I = \int_0^1 x3\sqrt{1-x}\,\mathrm{d}x = \frac{\Gamma(4)\Gamma(3/2)}{\Gamma(11/2)} = 0{,}101\,587\,301\,587\,301\,5,$$

où $\Gamma(x)$ est la fonction factorielle.

On donne dans le tableau 11.4 les valeurs trouvées.

**Tableau 11.4.**

| Nombre de points | Résultat |
|---|---|
| 2 | 0,101 587 301 587 301 5 |
| 3 | 0,101 587 301 587 301 5 |
| 4 | 0,101 587 301 587 301 3 |
| 5 | 0,101 587 301 587 301 2 |
| 6 | 0,101 587 301 587 301 0 |

Bien entendu, comme $f(x)$ est un polynôme de degré 3, nous devons trouver un résultat exact à partir de deux points. Ici encore, nous avons voulu montrer l'influence des troncatures et des arrondis sur les calculs.

## 3. Intégrales de la forme $I = \int_{-1}^{+1} \sqrt{1-x^2}\,dx$

Sans entrer dans les détails, disons que l'étude des polynômes $R_n(x)$ orthogonaux relativement à la fonction poids $\omega(x) = \sqrt{1-x^2}$ et à l'intervalle $(-1, +1)$ relève directement de l'étude des polynômes de Tchebycheff. Les zéros de $R_{n+1}(x)$ sont donnés par :

$$x_k = \cos\left(\frac{k+1}{n+2}\pi\right)$$

tandis que les $H_k^*$ correspondants sont donnés par :

$$H_k^* = \frac{\pi}{n+2}\sin^2\left(\frac{k+1}{n+2}\pi\right).$$

### 3.1. Calcul numérique des $x_k$ et des $H_k^*$ des polynômes $R_n(x)$

Dans le tableau 11.5 on a reporté les valeurs numériques calculées pour le polynôme de degré 11. Sur le Web(*), on trouvera les résultats concernant les premiers polynômes dans le fichier texte `rn_x.txt`.

**Tableau 11.5.** *Les zéros et les poids correspondants des polynômes $R_n(x)$.*

| Degré | $y_k$ | $H_k$ |
|---|---|---|
| 11 | ±0,965 925 826 289 068 | 0,017 537 233 634 936 |
| | ±0,866 025 403 784 439 | 0,065 449 846 949 787 |
| | ±0,707 106 781 186 548 | 0,130 899 693 899 575 |
| | ±0,500 000 000 000 000 | 0,196 349 540 849 362 |
| | ±0,258 819 045 102 521 | 0,244 262 154 164 213 |
| | 0,0 | 0,261 799 387 799 149 |

### 3.2. Technique d'intégration

L'intégrale $I$ est bien entendu approchée par l'expression devenue classique :

$$J = \sum_{k=0}^{n} H_k^* f(x_k).$$

Il est intéressant de noter que les intégrales de la forme :

$$I = \int_a^b f(x)\sqrt{(x-a)(b-x)}\,dx$$

* `http://www.edpsciences.com/guilpin/`

se ramènent au cas précédemment étudié à condition de faire usage du changement de variable suivant :

$$x_k^* = \frac{b+a}{2} + \frac{b-a}{2} x_k.$$

Cette dernière intégrale est alors approchée par :

$$J = \sum_{k=0}^{n} H_k^* f(x_k^*).$$

### 3.3. Un exemple numérique

Soit à calculer :

$$I = \int_{-1}^{+1} x\sqrt{1-x^2}\, \mathrm{d}x = 0.$$

Sur le tableau 11.6, nous avons porté les valeurs approchées. On vérifie que les erreurs sont de l'ordre de $10^{-16}$, et qu'elles ne sont dues qu'à l'exécution des calculs.

**Tableau 11.6.**

| Nombre de points | Résultats |
|---|---|
| 3 | $0{,}166\,10^{-15}$ |
| 4 | $0{,}971\,10^{-16}$ |
| 5 | $0{,}277\,10^{-16}$ |
| 6 | $0{,}555\,10^{-16}$ |
| 7 | $0{,}598\,10^{-16}$ |

## 4. Intégrales de la forme $I = \int_0^{+1} f(x)\sqrt{\frac{x}{1-x}}\,dx$

Ici encore l'étude des polynômes $S_n(x)$ orthogonaux relativement à la fonction poids $\omega(x) = \sqrt{\frac{x}{1-x}}$ et à l'intervalle $(-1, +1)$ relève directement de l'étude des polynômes de Tchebycheff. Les zéros de $S_n(x)$ s'expriment au moyen de la relation :

$$x_k = \cos^2\left(\frac{2k+1}{2n+3}\frac{\pi}{2}\right) \qquad \text{avec } k = 0, 1, 2, \ldots, n.$$

et les $H_k^*$ par la relation :

$$H_k^* = \frac{2\pi}{2n+3} x_k.$$

## 4.1. Calcul numérique des $x_k$ et des $H_k^*$ des polynômes $S_n(x)$

On trouvera sur le tableau 11.7 les valeurs numériques correspondant au polynôme de degré 11. Sur le Web (*), on trouvera les résultats concernant les premiers polynômes dans le fichier texte `sn_x.txt`.

**Tableau 11.7.** *Les zéros et les poids associés des polynômes* $S_n(x)$.

| Degré | $y_k$ | $H_k$ |
|---|---|---|
| | 0,995 342 973 018 165 | 0,271 909 754 072 704 |
| | 0,958 605 650 752 726 | 0,261 873 780 008 211 |
| | 0,887 855 645 352 210 | 0,242 546 154 164 063 |
| | 0,788 340 161 057 434 | 0,215 360 318 131 115 |
| | 0,667 439 806 085 493 | 0,182 332 521 001 007 |
| 11 | 0,534 121 206 682 336 | 0,145 912 283 394 760 |
| | 0,398 271 993 473 683 | 0,108 800 727 724 129 |
| | 0,269 967 481 134 424 | 0,073 750 248 299 135 |
| | 0,158 723 428 390 673 | 0,043 360 378 833 454 |
| | 0,072 790 297 726 756 | 0,019 884 996 920 956 |
| | 0,018 541 356 326 101 | 0,005 065 164 245 362 |

## 4.2. Technique d'intégration

L'intégrale $I$ est toujours approchée par l'expression :

$$J = \sum_{k=0}^{n} H_k^* f(x_k)$$

Il est intéressant de noter que les intégrales de la forme :

$$I = \int_a^b f(x)\sqrt{\frac{(x-a)}{(b-x)}}\,\mathrm{d}x$$

se ramènent au cas précédemment étudié en faisant usage du changement de variable : $x = a + (b-a)y$ ce qui permet d'approcher l'intégrale au moyen de l'expression :

$$J = (b-a)\sum_{k=0}^{n} H_k^* f\left[a + (b-a)x_k\right].$$

soit encore :

$$J = \frac{2\pi(b-a)}{(2n+3)}\sum_{k=0}^{n} x_k f\left[a + (b-a)x_k\right].$$

* `http://www.edpsciences.com/guilpin/`

### 4.3. Un exemple numérique

Soit à calculer :

$$I = \int_{-1}^{+1} \sqrt{\frac{1+x}{1-x}}\, \mathrm{d}x = \pi$$

Il s'agit d'intégrer un polynôme de degré zéro. Les résultats trouvés sont donnés dans le tableau 11.8. Ici encore les erreurs ne sont dues qu'à l'exécution des calculs, et l'on note que les deux derniers chiffres significatifs sont faux.

**Tableau 11.8.**

| Nombre de points | Résultats |
|---|---|
| 3 | 3,141 592 653 589 792 |
| 4 | 3,141 592 653 589 791 |
| 5 | 3,141 592 653 589 791 |
| 6 | 3,141 592 653 589 791 |

## 5. Éléments de bibliographie

M. CROUZEIX et A.L. MIGNOT (1984) *Analyse numérique des équations différentielles*, Éditions Masson.

H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.

# 12 Les polynômes de Bernoulli, formule d'Euler-MacLaurin. Méthode de Romberg et autres techniques d'intégration

## 1. Formule d'Euler-MacLaurin

Les polynômes et les nombres de Bernoulli (Jacques Bernoulli 1654–1705) jouent un rôle important en analyse mathématique. Ces éléments sont à la base d'une technique d'intégration numérique obtenue à partir de la formule d'Euler-MacLaurin, laquelle sert ensuite de tremplin à la méthode de Romberg. Au passage nous verrons une application des nombres de Bernoulli au calcul de la constante d'Euler.

Ces polyômes de Bernoulli possèdent une autre propriété concernant les séries de Fourier : on peut considérer les fonctions périodiques de période 1 représentées sur une période par les polynômes de Bernoulli de degré $q$ et noté $B_q(x)$. Le coefficient de rang $n$ de ce développement ($a_n$ ou $b_n$ selon la parité) est en $1/n^q$. Cette propriété relie aussi de près ces polynômes à la fonction dzeta de Riemann (1826–1866), et plus spécifiquement les nombres de Bernoulli. En outre, cela signifie que les fonctions périodiques qui sont définies sur une période par les polynômes de Bernoulli, non seulement sont continues, **mais ont toutes leurs dérivées continues partout.** Notamment, aux extrémités de l'intervalle $(0, 1)$, les fonctions sont continues à dérivées continues, bien entendu, nous ne nous intéressons qu'aux dérivées non nulles.

### 1.1. Fonction génératrice des polynômes de Bernoulli

Les polynômes de Bernoulli sont les coefficients du développement de l'expression suivante :

$$\frac{u\exp(ux)}{\exp(u)-1} = \sum_{n=0}^{\infty} \frac{u^n}{n!} B_n(x). \tag{12.1}$$

Par dérivation de la formule (12.1), on montre que :

$$B'_k(x) = kB_{k-1}(x), \tag{12.2}$$

et par identification directe, on voit que :

$$B_0(x) = 1. \tag{12.3}$$

En utilisant l'identité :

$$\frac{(-u)\exp(-xu)}{\exp(-u)-1} = \frac{u\exp[(1-x)u]}{\exp(u)-1}$$

puis la relation de définition (12.1), on déduit alors l'expression suivante :

$$B_k(1-x) = (-1)^k B_k(x) \tag{12.4}$$

ce qui permet d'écrire :

$$B_k(1) = (-1)^k B_k(0).$$

En intégrant la relation (12.1) par rapport à $x$, on établit que :

$$\int_0^1 B_k(x)\,\mathrm{d}x = 0 \qquad \text{pour } k > 0.$$

On obtient alors :

$$\begin{aligned} B_0(x) &= 1, \\ B_1(x) &= x - \frac{1}{2}, \\ B_2(x) &= x^2 - x + \frac{1}{6}, \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{x}{2}. \end{aligned}$$

## 1.2. Les nombres de Bernoulli

Par définition, les nombres de Bernoulli sont donnés par l'expression :

$$B_k = B_k(0).$$

En faisant $x = 0$ dans la formule (12.1) pour obtenir les nombres de Bernoulli, on voit que la fonction :

$$\frac{u}{\exp(u) - 1} + \frac{u}{2} = \frac{u}{2}\coth\frac{u}{2}$$

est une fonction paire. On déduit alors que :

$$B_1 = -\frac{1}{2}$$

et

$$B_{2m+1} = 0 \qquad \text{avec } m \geq 1.$$

Il s'ensuit que les nombres de Bernoulli de rang impair — hormis le premier — sont nuls et que seuls diffèrent de zéro les nombres de Bernoulli de rang pair. Voici donc les premiers

nombres de Bernoulli non nuls :

$$\begin{aligned} B_0 &= 1 & B_1 &= -\frac{1}{2} \\ B_2 &= \frac{1}{6} & B_4 &= -\frac{1}{30} \\ B_6 &= \frac{1}{42} & B_8 &= -\frac{1}{30} \\ B_{10} &= \frac{5}{66} & B_{12} &= -\frac{691}{2\,730} \\ B_{14} &= \frac{7}{6} & B_{16} &= -\frac{3\,617}{510} \\ B_{18} &= \frac{43\,867}{798} & B_{20} &= -\frac{174\,611}{330} \\ B_{22} &= \frac{854\,513}{138} & B_{24} &= -\frac{236\,364\,091}{2\,730}. \end{aligned}$$

## 1.3. Expression des polynômes de Bernoulli en fonction des nombres de Bernoulli

Pour $x = 0$, l'expression (12.1) nous donne :

$$\frac{u}{\exp(u)-1} = 1 + B_1 u + B_2 \frac{u^2}{2!} + B_4 \frac{u^4}{4!} + \cdots + B_{2n} \frac{u^{2n}}{(2n)!} + \cdots$$

il s'ensuit que :

$$\begin{aligned} \frac{u\exp(ux)}{\exp(u)-1} = \frac{u}{\exp(u)-1}\exp(ux) &= \left(1 + \frac{ux}{1!} + \frac{u^2x^2}{2!} + \frac{u^3x^3}{3!} + \cdots + \frac{u^kx^k}{k!} + \cdots\right) \\ &\quad \times \left(1 + B_1 u + B_2 \frac{u^2}{2!} + B_4 \frac{u^4}{4!} + \cdots + B_{2n}\frac{u^{2n}}{(2n)!} + \cdots\right) \\ &= \sum_{n=0}^{\infty} \frac{u^n}{n!} B_n(x) = 1 + B_1(x)\frac{u}{1!} + B_2(x)\frac{u^2}{2!} + B_k(x)\frac{u^k}{k!} + \cdots \end{aligned}$$

L'identification des termes en $u$ donne :

$$\frac{B_k(x)}{k!} = \frac{x^k}{k!} + \frac{x^{k-1}}{(k-1)!}B_1 + \frac{x^{k-2}}{(k-2)!}\frac{B_2}{2!} + \frac{x^{k-4}}{(k-4)!}\frac{B_4}{4!} + \frac{x^{k-6}}{(k-6)!}\frac{B_6}{6!}$$
$$\begin{cases} \cdots + \dfrac{xB_{k-1}}{(k-1)!} & \text{si } k \text{ est impair,} \\ \cdots + \dfrac{B_k}{k!} & \text{si } k \text{ est pair.} \end{cases}$$

De là nous tirons l'expression générale :

$$B_k(x) = x^k + x^{k-1}B_1 + x^{k-2}C_k^2 B_2 + x^{k-4}C_k^4 B_4 + \cdots$$
$$\begin{cases} \cdots + xB_{k+1} & \text{si } k \text{ est impair,} \\ \cdots + B_k & \text{si } k \text{ est pair.} \end{cases}$$

## 1.4. La formule d'Euler-Maclaurin dite formule de la somme

Il s'agit d'une méthode d'intégration qui utilise les nombres de Bernoulli ; pour l'obtenir, nous allons intégrer successivement par parties l'expression suivante :

$$F(x+h) - F(x) = \int_x^{x+h} F'(t)\,\mathrm{d}t = \int_0^h F'(x+h-t)\,\mathrm{d}t.$$

On aboutit alors à la formule usuelle :

$$F(x+h) - F(x) = hF'(x) + \frac{h^2}{2!}F''(x) + \cdots + \frac{h^{2p}}{(2p)!}F^{(2p)}(x) + \cdots + \int_0^h F^{(2p+1)}(x+h-t)\frac{t^{2p}}{(2p)!}\,\mathrm{d}t. \quad (12.5)$$

Maintenant, nous allons appliquer la relation (12.5) aux dérivées successives de $F(x)$, en nous arrêtant toutefois au même ordre à chaque fois. Nous obtenons les formules suivantes :

$$h[F'(x+h) - F'(x)] = h^2F''(x) + \cdots + \frac{h^{2p}}{(2p-1)!}F^{(2p)}(x) + h\int_0^h F^{(2p+1)}(x+h-t)\frac{t^{2p-1}}{(2p-1)!}\,\mathrm{d}t, \quad (12.6)$$

$$h^2[F''(x+h) - F''(x)] = h^3F'''(x) + \cdots + \frac{h^{2p}}{(2p-2)!}F^{(2p)}(x) + h^2\int_0^h F^{(2p+1)}(x+h-t)\frac{t^{2p-2}}{(2p-2)!}\,\mathrm{d}t, \quad (12.7)$$

$$h^k\left[F^{(k)}(x+h) - F^{(k)}(x)\right] = h^{k+1}F^{(k+1)}(x) + \cdots + \frac{h^{2p}}{(2p-k)!}F^{(2p)}(x) + h^k\int_0^h F^{(2p+1)}(x+h-t)\frac{t^{2p-k}}{(2p-k)!}\,\mathrm{d}t. \quad (12.8)$$

On poursuit le calcul jusqu'à l'ordre $(2p-1)$, soit :

$$h^{2p-1}\left[F^{(2p-1)}(x+h) - F^{(2p-1)}(x)\right] = h^{2p}F^{(2p)}(x) + h^{2p-1}\int_0^h F^{(2p+1)}(x+h-t)t\,\mathrm{d}t. \quad (12.9)$$

À présent, nous allons multiplier l'égalité (12.5) par $B_0$, puis l'égalité (12.6) par $B_1/1!$ et ainsi de suite, c'est-à-dire l'égalité (12.8) par $B_k/k!$. Additionnons toutes ces relations, puis, en se

souvenant que $B_{2m+1} = 0$, on obtient le coefficient de $F^{(k)}(x)$ que l'on désignera par $c$ :

$$c = \frac{1}{k!}\left(1 + \frac{B_1 k}{1!} + \frac{B_2 k(k-1)}{2!} + \frac{B_3 k(k-1)(k-2)}{3!} + \cdots + \frac{B_j k(k-1)\cdots(k-j-1)}{j!} + \cdots\right)$$

expression que l'on peut écrire sous la forme :

$$c = \frac{1}{k!}\left(1 + C_k^1 B_1 + C_k^2 B_2 + \cdots + C_k^{2j} B_{2j} + \cdots\right)$$

où $C_k^{2j}$ est le coefficient du binôme. En utilisant la relation (12.4) du paragraphe 1.1, nous remarquons que :

$$c = B_k(1) - B_k \quad (k \text{ est pair}).$$

On en déduit alors que $c = 0$. En définitive, il reste :

$$\begin{aligned} F(x+h) - F(x) = {} & -B_1 h\,[F'(x+h) + F'(x)] + B_2 \frac{h^2}{2!}\,[F''(x+h) - F''(x)] \\ & + \cdots + B_{2p-2}\frac{h^{2p-2}}{(2p-2)!}\left(F^{(2p-2)}(x+h) - F^{(2p-2)}(x)\right) \\ & + \cdots + \int_0^h F^{(2p+1)}(x+h-t)\left[\frac{t^{2p}}{(2p)!} + B_1 h\frac{t^{2p-1}}{(2p-1)!}\right. \\ & \left. + \cdots + \frac{B_{2k}h^{2k}t^{2p-2k}}{(2p-2k)!(2k)!} + \cdots + \frac{B_{2p-2}h^{2p-2}}{(2p-2)!2!}t^2\right]\mathrm{d}t. \end{aligned} \tag{12.10}$$

La dernière parenthèse s'exprime simplement en fonction du polynôme de Bernoulli de degré $2p$ ; pour cela on remarque que :

$$\frac{t^{2p}}{(2p)!} + \cdots + \frac{B_{2p-2}h^{2p-2}}{(2p-2)!2!}t^2 = \frac{h^{2p}}{(2p)!}B_{2p}\left(\frac{t}{h}\right) - \frac{h^{2p}}{(2p)!}B_{2p} = \frac{h^{2p}}{(2p)!}\left[B_{2p}\left(\frac{t}{h}\right) - B_{2p}\right].$$

Si maintenant on pose $F'(x) = f(x)$, nous obtenons une expression canonique qui prend alors la forme suivante :

$$\begin{aligned} \int_x^{x+h} f(t)\,\mathrm{d}t = {} & -B_1 h\,[f(x+h) + f(x)] + \frac{B_2 h^2}{2!}\,[f'(x+h) + f'(x)] \\ & + \cdots + \frac{h^{2p}}{(2p)!}\int_0^h f^{(2p)}(x+h-t)\left[B_{2p}\left(\frac{t}{h}\right) - B_{2p}\right]\mathrm{d}t. \end{aligned} \tag{12.11}$$

***Évaluation du reste*** – Le reste est donné par l'expression :

$$E = \left|\frac{h^{2p}}{(2p!)}\int_0^h f^{(2p)}(x+h-t)\left[B_{2p}\left(\frac{t}{h}\right) - B_{2p}\right]\mathrm{d}t\right|, \tag{12.12}$$

soit :

$$E = \left| \frac{h^{2p}}{(2p!)} \int_0^h f^{(2p)}(x+h-t) B_{2p}\, \mathrm{d}t \right| + \left| \frac{h^{2p}}{(2p)!} \int_0^h f^{(2p)}(x+h-t) B_{2p}\left(\frac{t}{h}\right)\, \mathrm{d}t \right|.$$

Comme $B_{2p}$ et $B_{2p}\left(\frac{t}{h}\right)$ conservent chacun un signe constant, on peut appliquer à chacun des deux termes de cette relation la formule de la moyenne en trouvant un nombre $\theta$ compris entre 0 et 1 tel que :

$$E = \frac{h^{2p+1}}{(2p)!} \left| f^{(2p)}(x+\theta h) B_{2p} \right| + \left| \frac{h^{2p}}{(2p)!} f^{(2p)}(x+\theta h) \int_0^h B_{2p}\left(\frac{t}{h}\right)\, \mathrm{d}t \right|$$

$$\text{mais : } \int_0^h B_{2p}(t/h)\, \mathrm{d}t = 0,$$

on obtient en définitive :

$$E = \frac{h^{2p+1}}{(2p)!} \left| f^{(2p)}(x+\theta h) \right| |B_{2p}|.$$

## 1.5. Application au calcul des intégrales simples

Considérons un intervalle fini $(a,b)$ divisé en $n$ sous-intervalles de longueur identique $h$, autrement dit, on définit des points $x_k$ en progression arithmétique sur l'intervalle $(a,b)$ tels que :

$$x_0 = a, \quad x_1 = a+h, \quad x_2 = a+2h, \quad \ldots, \quad x_n = a+nh = b.$$

Appliquons la relation (12.11) à chaque sous-intervalle puis effectuons la somme de toutes ces relations. En posant pour plus de clarté $y_k = f(a+kh)$ avec $k = 1, 2, \ldots, n$, nous obtenons :

$$\int_a^b f(x)\, \mathrm{d}x = h\left(\frac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n\right) + \frac{B_2 h^2}{2!}[f'(b) - f'(a)]$$
$$+ \frac{B_4 h^4}{4!}[f'''(b) - f'''(a)] + \cdots + \frac{B_{2p-2} h^{2p-2}}{(2p-2)!}\left[f^{(2p-3)}(b) - f^{(2p-3)}(a)\right]$$

l'erreur étant majorée par

$$E = \frac{|B_{2p}| h^{2p+1}}{(2p)!} \sum_{i=1}^{n} f^{(2p)}(x_i).$$

***Remarque*** : On reconnaît dans la première parenthèse du membre de droite l'expression de la formule des trapèzes que du reste on obtient rigoureusement en faisant $p = 1$ dans la dernière relation.

***Application au calcul de la constante d'Euler*** – Par définition, la constante d'Euler (1707–1783) est donnée par l'expression :

$$\gamma = \lim_{m\to y}\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\cdots+\frac{1}{m}-\log_e(m)\right)$$

Cette série ne converge pas très rapidement, aussi lui préfère-t-on un développement asymptotique encore appelé série semi-convergente que l'on obtient de la façon suivante : dans l'expression précédente, on fait $f(x)=\log_e(x)$, $h=1$ et $x_1=m$ (entier positif quelconque). On obtient alors sans problème :

$$\sum_{k=1}^{m-1}\frac{1}{k}=\log_e(m)+\gamma-\frac{1}{2m}-\frac{1}{12m^2}-\frac{1}{120m^4}-\frac{1}{252m^6}+\cdots+\frac{(-1)^nB_{2k}}{2km^{2k}}+\cdots$$

Retenons les treize premiers nombres de Bernoulli (non nuls), et calculons la somme pour $m=13$ : alors, on obtient $\gamma$ avec quinze chiffres significatifs exacts en utilisant la représentation des nombres sur huit octets (16 chiffres significatifs).

## 1.6. Application au modèle des chaleurs spécifiques selon Debye (1884–1966)

En 1912, Debye a proposé un modèle de la chaleur spécifique des solides qui aboutit au résultat suivant (le modèle ainsi que l'utilisation des nombres de Bernoulli sont présentés dans l'ouvrage de Rocard cité en bibliographie) :

$$C_v=3R\left\{\frac{12}{x^3}\int_0^x\frac{y^3}{\exp(y)-1}\,\mathrm{d}y-\frac{3x}{\exp(x)-1}\right\}$$

où $x=U/T$ ($U$ est une constante dépendant uniquement de la nature du corps considéré, elle s'appelle température caractéristique). On se propose d'établir une table de la chaleur spécifique de l'argent pour lequel $U=215$ K. On écrit les relations suivantes :

$$\frac{x}{\exp(x)-1}=\sum_{n=0}^{\infty}\frac{x^n}{n!}B_n\qquad\text{avec}\quad B_{2k+1}=0,$$

$$\frac{y^3}{\exp(y)-1}=\sum_{n=0}^{\infty}\frac{y^{n+2}}{n!}B_n=y^2-\frac{y^3}{2}+\sum_{n=1}^{\infty}\frac{y^{2n}}{(2n)!}B_{2n}$$

puis on calcule l'intégrale :

$$\int_0^x\frac{y^3}{\exp(y)-1}\,\mathrm{d}y=\left\{\sum_{n=0}^{\infty}\frac{y^{n+3}}{(n+3)n!}B_n\right\}=\frac{x^3}{3}-\frac{x^4}{8}+\sum_{k=1}^{\infty}\frac{x^{2k+3}}{(2k+3)(2k)!}B_{2k}$$

on déduit que :

$$\frac{C_v}{3R}=4-\frac{3}{2}x+12\sum_{k=1}^{\infty}\frac{x^{2k}}{(2k+3)(2k)!}B_{2k}-3+\frac{3}{2}x-3\sum_{k=1}^{\infty}\frac{x^{2k}}{(2k)!}B_{2k}$$

d'où

$$C_v=R\left(1+3\sum_{k=1}^{\infty}\frac{x^{2k}}{(2k)!}B_{2k}\frac{1-2k}{2k+3}\right).$$

Cette dernière expression a fait l'objet du programme `argent.c` qui est donné sur le Web(*).

* `http://www.edpsciences.com/guilpin/`

## 1.7. La méthode de Romberg

Quand on calcule une intégrale $I = \int_a^b f(x)\,\mathrm{d}x$ par la méthode des trapèzes avec un pas $h_n = \frac{b-a}{n}$, on obtient la relation suivante :

$$I_n = \frac{h_n}{2}\left(f(a) + f(b) + \sum_{k=1}^{n-1} f(a + kh_n)\right)$$

avec comme expression de l'erreur mathématique de troncature :

$$E_n = -\frac{(b-a)^3}{n^2} f''(x) \qquad \text{avec } a < x < b.$$

Si $n$ tend vers l'infini, la suite $I_n$, générée par la méthode des trapèzes, converge vers la valeur de l'intégrale. On peut alors songer à accélérer la procédure de convergence au moyen d'un des algorithmes présentés au chapitre 2, mais on peut aussi utiliser le fait que l'erreur est donnée par la formule d'Euler-MacLaurin en soulignant toutefois que la fonction $f(x)$ doit être $(2p+2)$ fois dérivable sur l'intervalle $(a, b)$. Revenons à cette expression de l'erreur :

$$E = \frac{h_n^{2p+2}}{(2p+2)!} \int_0^{h_n} f^{2p+2}(x + h_n - t)\left[B_{2p}\left(\frac{t}{h_n}\right) - B_{2p+2}\right]\,\mathrm{d}t$$

et nous notons que cette erreur est un polynôme en $h_n^2$. L'idée de la méthode de Romberg repose sur le calcul de $I_n$ pour des valeurs différentes de $h_n$, et de construire le polynôme d'interpolation de la variable $h_n^2$ passant par les $I_n$ puis de calculer la valeur de ce polynôme d'interpolation pour $h_n = 0$. Il faut bien remarquer que la valeur obtenue de $I_n$ pour en $h_n$ tendant vers zéro est une extrapolation du polynôme d'interpolation.

Pour réaliser l'extrapolation, on aura recours à la procédure de Richardson généralisée — généralisation qui semble avoir été conçue à cet effet. Comme l'erreur est un polynôme en $h_n^2$, on prendra :

$$x_n = h_n$$
$$(x) = x$$

Malheureusement, si nous appliquons le procédé de Richardson à la suite des $I_n$, la condition assurant la convergence ne sera pas vérifiée car :

$$\lim \frac{h_n}{h_{n+1}} = 1 \quad \text{quand } n \text{ tend vers l'infini.}$$

Pour pallier cet inconvénient, et assurer la condition de convergence, on choisit la suite $S_j$ formée par les valeurs $I_{2n}$. Autrement dit, on calculera les intégrales $I_n$ par la méthode des trapèzes avec un pas donné par l'expression :

$$h_{2n} = \frac{b-a}{2^n} \qquad \text{avec } h_0 = b - a.$$

On vérifie alors que l'on obtient :

$$\lim \frac{x_n}{x_{n+1}} = \frac{h_0}{2^n}\frac{2^{n+1}}{h_0} = 2.$$

Reprenant l'expression de la méthode de Richardson généralisée et en posant :

$$F(0) = 0 \quad \text{et} \quad F(x_n) = \frac{h_0}{2^n},$$

nous pouvons écrire :

$$T_n^{(k+1)} = \frac{(h_0/2^n)T_{n+1}^{(k)} - (h_0/2^{n+k+1})T_n^{(k)}}{(h_0/2^n)^2 - (h_0/2^{n+k+1})^2},$$

ce qui s'écrit encore :

$$T_n^{(k+1)} = \frac{4^{k+1}T_{n+1}^{(k)} - T_n^{(k)}}{4^{k+1} - 1}.$$

En résumé, on utilise les relations suivantes :

$$h_0 = \frac{b-a}{n},$$
$$h_n = \frac{h}{2^n},$$
$$S_n^{(0)} = h_n \left( \frac{f(a) + f(b)}{2} + \sum_{k=1}^{m} f(a + kh_n) \right), \qquad \text{avec } m = 2^n - 1.$$

## 1.8. Un exemple numérique

On se propose de calculer l'intégrale suivante :

$$I = \int_0^1 \frac{\log_e(1+x)}{1+x}\, \mathrm{d}x = \frac{1}{2}\left[\log_e(2)\right]^2 = 0{,}240\,226\,507$$

et de former la suite des $I_k$ à partir de $h_0 = \frac{1}{2}$. On obtient alors le tableau 12.1.

**Tableau 12.1.**

| | | | | |
|---|---|---|---|---|
| Suite initiale | | | | |
| 0,22 8 | | | | |
| | 0,240 066 6 | | | |
| 0,235 5 | | 0,240 216 00 | | |
| | 0,240 206 6 | | 0,240 226 03 | |
| 0,239 03 | | 0,240 225 87 | | 0,240 226 653 |
| | 0,240 224 6 | | 0,240 226 651 | |
| 0,239 926 | | 0,240 226 64 | | |
| | 0,240 226 5 | | | |
| 0,240 151 | | | | |

L'erreur est $E = 0{,}14\,10^{-6}$. L'epsilon-algorithme appliqué à la même suite donne le résultat suivant : $I = 0{,}240\,225\,49$ avec une erreur $E = 0{,}1\,10^{-5}$.

Si la méthode de Romberg donne un résultat avec un ordre de grandeur meilleur que le résultat donné par la méthode de l'epsilon-algorithme, on obtient encore de meilleurs résultats avec beaucoup moins de calculs en utilisant la méthode de Gauss-Legendre dont voici les résultats en 4, 6 et 12 points.

$$\begin{aligned} 4 \text{ points} \quad & I = 0{,}240\,228 \\ 6 \text{ points} \quad & I = 0{,}240\,226\,509 \\ 12 \text{ points} \quad & I = 0{,}240\,226\,507. \end{aligned}$$

Ces derniers calculs ont été réalisés avec une machine fonctionnant avec 10 chiffres significatifs.

## 2. Autres méthodes d'intégration

Si les méthodes fondées sur la technique de Gauss sont, d'une façon générale, les plus performantes, il existe un certain nombre d'algorithmes très simples qui méritent d'être présentés, ne serait-ce que pour la réflexion qu'ils suscitent.

Il est bien entendu que l'on s'intéresse ici à l'intégration numérique des fonctions dont l'intégrale sur un intervalle fini $(a, b)$ a un sens, d'ailleurs cette fonction peut être continue par morceaux avec des points de discontinuité de première espèce. Cela signifie que l'on effectue le calcul sur les différents morceaux selon une des méthodes que nous allons exposer, et que nous en ferons ensuite la somme.

Une intégrale peut avoir un sens bien que la fonction à intégrer présente une apparence de discontinuité de seconde espèce : nous avons rencontré un exemple à propos de la méthode de Gauss généralisée : les singularités étaient alors contenues dans la fonction-poids aux bornes de l'intervalle fondamental.

Quoi qu'il en soit, nous nous intéressons au calcul de :

$$I = \int_a^b f(x)\,\mathrm{d}x$$

où $a$ et $b$ sont des nombres finis et $f(x)$ une fonction continue possédant des dérivées jusqu'à l'ordre nécessaire pour le besoin des calculs d'erreur.

### 2.1. La méthode des rectangles

L'intervalle sur lequel s'effectue l'intégration est divisé en $N$ sous-intervalles de longueur égale $h = (b-a)/N$ ce qui nous définit une suite d'abscisses $x_i$ en progression arithmétique. La valeur de l'intégrale entre deux points consécutifs $x_i$ et $x_{i+1}$ notée $s_i$ est approchée par $s'_i$ :

$$s'_i = hf(x_i)\,;$$

la sommation des $s'_i$ fournit une approximation de l'intégrale $I$. Désignons par $I'$ l'approximation obtenue :

$$I' = \sum_{i=0}^{N-1} s'_i = \sum_{i=0}^{N-1} f(x_i).$$

***Estimation de l'erreur mathématique*** – L'erreur due à l'approximation mathématique s'écrit :

$$E = |I - I'|.$$

Désignons par $F(u)$ la primitive de la fonction $f(x)$. Nous pouvons écrire :

$$s_i = F(x_{i+1}) - F(x_i);$$

puis :

$$E_i = s_i - s'_i = F(x_{i+1}) - F(x_i) - hf(x_i) = F(x_{i+h}) - F(x_i) - hf(x_i).$$

Cette expression va nous permettre d'obtenir une évaluation de l'erreur en effectuant un développement limité de $F(x_{i+h})$ à l'ordre deux avec l'expression du reste de Lagrange, soit :

$$E_i = F(x_i) + hf(x_i) + \frac{h^2}{2} f'(\eta_i) - F(x_i) - hf(x_i)$$

où $\eta_i$ est un nombre compris entre $x_i$ et $x_{i+1}$. À partir de là, on obtient une estimation de $E_i$, soit :

$$E_i = \frac{h^2}{2} f'(\eta_i)$$

suivie de celle de E :

$$E = \frac{h^2}{2} \sum_{i=0}^{N-1} f'(\eta_i).$$

Cette expression prend une forme plus simple en majorant la somme, soit :

$$E = \frac{h^2}{2} N M_1 = \frac{(b-a)^2}{2N} M_1,$$

où $M_1$ est le sup de la valeur absolue de la dérivée $f'(x)$ sur l'intervalle $(a, b)$.

## 2.2. La méthode des trapèzes

En conservant les notations du précédent paragraphe, on désigne par $s_i$ la valeur de l'intégrale calculée entre deux points consécutifs $x_i$ et $x_{i+1}$ ; elle est approchée par $s'_i$ :

$$s'_i = \frac{h}{2} [f(x_{i+1}) + f(x_i)]$$

qui représente la surface du trapèze déterminé par les points $x_i$, $x_{i+1}$, $f(x_i)$ et $f(x_{i+1})$. En désignant par $I'$ l'approximation de l'intégrale $I$ obtenue par sommation des $s'_i$, nous pouvons écrire :

$$I' = \sum_{i=0}^{N-1} s'_i = \frac{h}{2} \left\{ f(x_0) + f(x_N) + 2 \sum_{i=1}^{N-1} f(x_i) \right\}.$$

***Estimation de l'erreur mathématique*** – Cette technique va nous permettre de gagner un ordre de grandeur en précision par rapport à la méthode des rectangles. Écrivons la nouvelle valeur de l'erreur que nous appelons $E_i$ :

$$\begin{aligned} E_i = s_i - s'_i &= F(x_{i+1}) - F(x_i) - \frac{h}{2}\left[f(x_{i+1}) + f(x_i)\right] \\ &= F(x_{i+h}) - F(x_i) - \frac{h}{2}\left[f(x_{i+h}) + f(x_i)\right]. \end{aligned}$$

Effectuons un développement limité de $F(x_{i+h})$ et de $f(x_{i+h})$ respectivement à l'ordre trois et deux avec les expressions associées du reste de Lagrange, soit :

$$E_i = F(x_i) + hf(x_i) + \frac{h^2}{2!}f'(h_i) + \frac{h^3}{3!}f''(\eta_i) - F(x_i) - \frac{h}{2}f(x_i) - \frac{h}{2}f(x_i) - \frac{h^2}{4}f'(\xi_i)$$

où $\eta_i$ et $\xi_i$ sont deux nombres compris entre $x_i$ et $x_{i+1}$. À partir de là, on obtient une estimation de $E_i$ :

$$E_i = -\frac{h^3}{12}f''(\theta_i)$$

où $\theta_i$ est un nombre compris entre $x_i$ et $x_{i+1}$. Il s'ensuit l'expression de $E$ :

$$E = \frac{h^3}{12}\sum_{i=0}^{N-1} f''(\theta_i).$$

Cette expression prend une forme plus simple en majorant la somme, soit :

$$E = \frac{h^3}{12}NM_2 = \frac{(b-a)^3}{12N^2}M_2,$$

où $M_2$ est le sup de la valeur absolue de la dérivée deuxième $f''(x)$ sur l'intervalle $(a, b)$.

## 3. La méthode de Simpson (1811–1870)

La méthode de Simpson repose sur l'utilisation de la formule d'intégration dite des trois niveaux. Cela signifie que l'on réalise une interpolation parabolique au lieu de l'interpolation linéaire utilisée dans la méthode des trapèzes, mais le gain est plus conséquent car la formule des trois niveaux étant exacte pour un polynôme de degré trois, l'interpolation réalisée est celle d'une parabole cubique. Le calcul de l'erreur nous montrera qu'il en est bien ainsi, en constatant que son expression est proportionnelle à $1/N^4$.

Toutefois, une légère différence existe dans le découpage de l'intervalle $(a, b)$ qui est maintenant réalisé à l'aide d'un nombre pair de points, soit $2N$ ce nombre, ainsi $h$ est donné par l'expression $h = (b - a)/2N$. Dans ces conditions, la formule d'intégration qui porte sur trois points consécutifs s'écrit :

$$s'_i = \frac{h}{3}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right],$$

puis :

$$I' = \sum_{i=0}^{N-1} s'_i = \frac{h}{3}\left\{f(x_0 + f(x_N) + 2\sum_{i=1}^{N-1} f(x_{2i}) + 4\sum_{i=0}^{N-1} f(x_{2i+1})\right\}.$$

***Estimation de l'erreur mathématique*** – En adoptant le procédé exploité pour les méthodes précédentes, nous écrivons la nouvelle valeur de $E_i$ :

$$E_i = s_i - s_i' = F(x_{2i+2}) - F(x_{2i}) - \frac{h}{3}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right],$$
$$= F(x_{2i} + 2h) - F(x_{2i}) - \frac{h}{3}\left[f(x_{2i} + 2h) + f(x_{2i}) + 4f(x_{2i+1})\right].$$

Effectuons un développement limité de $F(x_{2i} + h)$, $f(x_{2i} + 2h)$ et de $f(x_{2i} + h)$ respectivement à l'ordre cinq pour le premier et quatre pour les deux autres ; on obtient :

$$E_i = 2hf(x_i) + 2h^2 f'(x_i) + \frac{4h^3}{3}f''(x_i) + \frac{2h^4}{3}f'''(x_i) + \frac{4h^5}{15}f^{IV}(\xi_i) - \frac{h}{3}\Big(f(x_i) + 2hf'(x_i)$$
$$+ 2h^2 f''(x_i) + \frac{4h^3}{3}f'''(x_i) + \frac{2h^4}{3}f^{IV}(\varphi_i)\Big) - \frac{h}{3}\Big(f(x_i) + 4f(x_i) + 4hf'(x_i) + 2h^2 f''(x_i)$$
$$+ \frac{2h^3}{3}f'''(x_i) + \frac{h^4}{6}f^{IV}(\theta_i)\Big)$$

où $\xi_i$, $\varphi_i$ et $\theta_i$ sont deux nombres compris entre $x_{2i}$ et $x_{2i+2}$. À partir de là, on obtient une estimation de $E_i$ :

$$E_i = \frac{h^5}{90}f^{IV}(\rho_i)$$

où $\rho_i$ est un nombre compris entre $x_i$ et $x_{i+1}$. Il s'ensuit l'expression de $E$ :

$$E = \frac{h^5}{90}\sum_{i=0}^{N-1} f^{IV}(\rho_i).$$

Il est aisé d'obtenir une forme plus simple en majorant la valeur de la dérivée sur l'intervalle $(a, b)$, on obtient le résultat suivant :

$$E = \frac{h^5}{90}NM_4 = \frac{(b-a)^5}{2\,880N^4}M_4, \quad \text{(il y a } 2N \text{ points)},$$

où $M_4$ est le sup de la valeur absolue de la dérivée d'ordre quatre sur l'intervalle $(a, b)$.

***Remarque*** **:** Autant il est aisé de tirer des conclusions quand au rôle joué par $N$ dans chacune des méthodes, autant il est impossible de savoir quoi que ce soit concernant les majorations des dérivées d'ordre de plus en plus élevé. Les quelques exercices que l'on peut mener facilement le montrent fort bien : dans bien des cas, la majoration d'une dérivée d'ordre m conduit à des résultats décevants qui contrebalancent le gain obtenu par la puissance de $N$ (grand dénominateur).

Prenons un exemple simple :

$$I = \int_0^{10} \exp(-x^2)\,\mathrm{d}x.$$

Les dérivations successives de $\exp(-x^2)$ génèrent des polynômes qui ont une grosse parenté avec les polynômes d'Hermite (*cf.* la formule de Rodriguès générant les polynômes d'Hermite). Tous calculs faits, on trouve les résultats suivants : $M_1 = 0{,}43$, $M_2 = 0{,}893$ et $M_4 = 10{,}95$, ainsi, sur l'intervalle $(0, 10)$, et l'on voit que

$$\frac{M_4}{M_1} = 25{,}5.$$

## 3.1. Méthode de Newton-Cotes (1682–1716)

Soit une fonction $f(x)$ continue sur $(-1,+1)$ laquelle possède des dérivées continues sur le même intervalle jusqu'à l'ordre $(n+1)$ au moins. On connaît un échantillon de la fonction $f(x)$ constitué de $(n+1)$ points $y_k = f(x_k)$ correspondant à des abscisses en progression arithmétique $x_k = -1 + \frac{2k}{n}$

On se propose de calculer l'intégrale $I = \int_{-1}^{+1} f(x)\,\mathrm{d}x$ au moyen d'une approximation $J$ qui consiste à remplacer $f(x)$ par le polynôme de Lagrange donné par la formule de Lagrange, soit :

$$J = \sum_{k=0}^{n} f(x_k) \int_{-1}^{+1} \frac{\prod_{j\neq k}^{n}(x - x_j)}{\prod_{j\neq k}^{n}(x_k - x_j)}\,\mathrm{d}x = \sum_{k=0}^{n} f(x_k) H_k.$$

Le calcul des $H_k$ s'effectue au moyen du changement de variable $x_k = -1 + \frac{2u}{n}$, ce qui donne :

$$H_k = \frac{(-1)^{n-k}2}{nk!(n-k)!} \int_0^n u(u-1)\dots(u-k+1)(u-k-1)\dots(u-n)\,\mathrm{d}u,$$

$$H_k = \frac{(-1)^{n-k}2}{nk!(n-k)!} \int_0^n P_k(u)\,\mathrm{d}u.$$

Nous allons transformer $P_k(u)$ en l'écrivant sous la forme suivante :

$$P_k(u) = \sum_{j=0}^{n} a_j^{(k)} u^j,$$

l'exposant $(k)$ rappelant qu'il s'agit des coefficients du $k^{\mathrm{e}}$ polynôme. Cette expression peut faire l'objet d'une intégration terme à terme à condition de connaître les $a_j^{(k)}$. Cela est aisé à l'aide des relations de Newton puisque l'on connaît les racines de $P_k(u)$ lesquelles sont les entiers positifs successifs à l'exception de $k$. On calcule au préalable les $S^m$ qui sont les sommes des racines à la puissance $m$.

L'intégration de $P_k(u)$ entre 0 et $n$ est immédiate. Si l'intervalle d'intégration est $(a,b)$, le changement de variable permet d'obtenir :

$$\int_{-1}^{+1} f(x)\,\mathrm{d}x = \frac{b-a}{2} \sum_{k=0}^{n} H_k f(a+kh) \qquad \text{avec } h = \frac{b-a}{n}\cdot$$

L'expression de l'erreur $E = |I - J|$ est estimée à partir de l'expression de l'erreur commise en remplaçant $f(x)$ par le polynôme de Lagrange $P_L(x)$, soit :

$$E = \left| \int_{-1}^{+1} \{f(x) - P_L(x)\}\,\mathrm{d}x \right|,$$

$$E = \left| \frac{M_{n+1}}{(n+1)!} \frac{2^{n+2}}{n^{n+2}} \int_0^n \prod_{j=0}^{n}(u-j)\,\mathrm{d}u \right|,$$

où $M_{n+1}$ est le sup du module de $f^{(n+1)}(x)$ sur l'intervalle $(a, b)$. Rappelons que le calcul de $M_{n+1}$ est rarement simple. L'intégration du polynôme sous le signe somme est effectuée de la même manière que celle de $H_k$.

Sur le Web(*), on trouvera le programme `cotes_1.c` qui réalise cette technique d'intégration suivi d'un autre programme `cotes_2.c` qui calcule la quantité :

$$Q(n) = \frac{2^{n+2}}{n^{n+2}} \frac{1}{(n+1)!} \left| \int_0^n \prod_{j=0}^{n} (u-j)\, \mathrm{d}u \right|.$$

Il faut prendre quelques précautions pour calculer cette dernière intégrale. Si le principe que nous venons d'utiliser pour calculer les $H_k$ demeure, la valeur absolue pose un problème que l'on surmonte de la façon suivante : comme la fonction à intégrer s'annule pour les entiers successifs, dans chaque intervalle constitué par deux entiers successifs, la fonction est soit positive soit négative (elle est alternativement négative puis positive etc.). Pour obtenir une majoration de l'erreur, il suffit d'intégrer entre deux entiers consécutifs, de prendre la valeur absolue du résultat et d'en effectuer la somme depuis zéro jusqu'à $n$. Au préalable, on aura écrit le polynôme de degré $n+1$ sous la forme :

$$P_n(u) = \sum_{j=0}^{n+1} a_j^{(k)} u^j,$$

à l'aide des relations de Newton. Les premières valeurs calculées sont données dans le tableau 12.2.

**Tableau 12.2.**

| $n$ | $Q(n)$ |
|---|---|
| 3 | $0{,}823\,10^{-1}$ |
| 4 | $0{,}968\,10^{-2}$ |
| 5 | $0{,}973\,10^{-3}$ |
| 6 | $0{,}888\,10^{-4}$ |
| 7 | $0{,}717\,10^{-5}$ |
| 8 | $0{,}529\,10^{-6}$ |
| 9 | $0{,}354\,10^{-7}$ |
| 10 | $0{,}219\,10^{-8}$ |

Il est intéressant de s'arrêter sur le calcul d'erreur concernant l'exemple de l'intégrale de Gauss :

$$I = \frac{1}{\sqrt{2\pi}} \int_0^2 \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

* `http://www.edpsciences.com/guilpin/`

La dérivée d'ordre $n+1$ de l'expression sous le signe somme, notée $g(x)$ est donnée par les polynômes d'Hermite (*cf.* chapitre 10) :

$$g(x) = \frac{d^{n+1}}{\mathrm{d}x^{n+1}} \exp\left(-\frac{x^2}{2}\right) = (-1)^{n+1} K_{n+1}(x) \exp\left(-\frac{x^2}{2}\right).$$

Cherchons le sup de la valeur absolue de cette expression dans l'intervalle $(0,2)$. Désignons par $M_{n+1}$ cette valeur. Elle est obtenue pour l'une des valeurs qui annulent la dérivée de $g(x)$, soit :

$$g'(x) = \frac{\mathrm{d}^{n+2}}{\mathrm{d}x^{n+2}} \exp\left(-\frac{x^2}{2}\right) = (-1)^{n} K_{n+2}(x) \exp\left(-\frac{x^2}{2}\right).$$

Ce sont donc les zéros du polynôme d'Hermite de degré $n+2$ qui sont les extremums de $g(x)$, on les note $a_k$. Il est facile de calculer les valeurs de $g(a_k)$ correspondant aux racines $a_k$ contenues dans l'intervalle $(a,b)$ ici $(0,2)$, et de choisir la plus grande valeur absolue. Cette opération va être rendue aisée par l'usage des coefficients $H_k$ calculés lors de l'intégration de Gauss-Hermite ; en effet nous avons :

$$H_k^{(n+2)} = \frac{(n+1)!\sqrt{2}}{(n+2)\left[K_{n+1}(a_k)\right]^2},$$

de là, on tire :

$$K_{n+1}(a_k) \exp\left(-\frac{a_k^2}{2}\right) = \sqrt{\frac{(n+1)!\sqrt{2}}{(n+2)H_k^{(n+2)}}} \exp\left(-\frac{a_k^2}{2}\right).$$

Pour fixer les idées, choisissons $n = 8$, alors nous trouvons :

$$J = 0{,}477\,250\,155$$

La consultation de la table donnée chapitre 10 (concernant le polynôme de degré 10) nous montre que seules deux racines nous intéressent dans l'intervalle $(0,2)$, elles donnent des extremums du même ordre de grandeur et l'on trouve :

$$M_{n+1} = 217,$$

d'où

$$E(n=8) < \frac{217}{\sqrt{2\pi}} 0{,}529\,10^{-6} = 4{,}6\,10^{-5},$$

et

$$J = 0{,}477\,25 \pm 0{,}000\,46.$$

Ce résultat peut apparaître médiocre, en fait il est dû à la majoration excessive de la dérivée d'ordre $(n+1)$. En général, il est très difficile de calculer formellement la dérivée d'ordre $n$ d'une fonction, et rares sont les exercices qui le permettent, aussi ne s'attardera-t-on qu'exceptionnellement sur leur majoration.

Cette raison explique pourquoi il n'est pas possible de comparer directement, dans le cas général, les résultats fournis par la méthode des trapèzes et par la méthode de Simpson puisque la première méthode donne une précision en fonction de la dérivée deuxième et la seconde en fonction de la dérivée quatrième. La comparaison n'a de sens que dans l'étude de cas particuliers au cours desquels on sait expliciter les dérivées et les majorer convenablement...

Quoi qu'il en soit, ici, pour $n = 8$, les six premiers chiffres significatifs sont exacts...

## 4. Éléments de bibliographie

A. ANGOT (1965) *Compléments de Mathématiques*, Éditions de la Revue d'Optique.
N. BAKHVALOV (1976) *Méthodes Numériques*, Éditions MIR.
H. BESTOUGEFF CH. GUILPIN M. JACQUES (1975) *La Technique Informatique*, Tomes I et II, Éditions Masson.
M. CROUZEIX et A.L. MIGNOT (1984) *Analyse numérique des équations différentielles*, Éditions Masson.
B. DÉMIDOVITCH et L. MARON (1979) *Éléments de Calcul Numériques*, Éditions MIR.
G. HACQUES (1971) *Mathématiques pour l'Informatique*, Éditions Armand Colin.
P. HENRICI (1964) *Elements of Numerical Analysis*, Édition Wiley.
F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Édition Mc Graw-Hill.
R. KRESS (1998) *Numerical analysis*, Springer.
D. MCCRACKEN et W. DORN (1964) *Numerical Methods and Fortran Programming*, Éditions Wiley.
A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Éditions Dunod.
J.R. RICE (1969) *Approximation des Fonctions*, Éditions Dunod.
Y. ROCARD (1967) *Thermodynamique*, Éditions Masson.
G. VALIRON (1955) *La Théorie des Fonctions*, Éditions Masson.

# 13 Intégration des équations différentielles dans le champ réel

Ces types de problème relèvent du vaste cadre de l'analyse fonctionnelle. S'il est rare de trouver une primitive à une fonction donnée, il est encore plus rare de pouvoir exhiber une fonction $f$ qui soit solution d'une équation différentielle donnée. Pour s'en convaincre, il suffit de considérer une des toutes premières équations différentielles rencontrées dans la physique : l'équation du pendule simple dans le plan. Désignons par $\theta$ l'angle que fait le pendule avec la verticale et écrivons l'équation différentielle à laquelle obéit $\theta$ au cours du temps $t$ :

$$\frac{\mathrm{d}^2\theta(t)}{\mathrm{d}t^2} + \omega^2 \sin[\theta(t)] = 0.$$

Cette équation différentielle ne possède pas de solution formelle connue c'est-à-dire de solution exprimable au moyen des transcendances usuelles. D'autant plus que la nature des conditions aux limites peut singulièrement compliquer le problème à savoir :

**a.** Conditions de Cauchy (1789–1857). On connaît à l'instant $t = t_0$ la valeur de la fonction inconnue et de toutes les dérivées jusqu'à l'ordre le plus élevé figurant dans l'équation.
**b.** Conditions aux limites (ou de tir). On connaît les valeurs de la fonction inconnue à des instants différents.

Ici nous avons choisi le temps comme variable puisque l'exemple retenu est le pendule simple, mais rien n'est changé en faisant usage d'autres variables.

Ici, nous ne cherchons pas à résoudre le problème différentiel dans le cas le plus général implicite que l'on écrit :

$$F\left(\frac{\mathrm{d}^n y}{\mathrm{d}t^n}, \frac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}}, \ldots, \frac{\mathrm{d}y}{\mathrm{d}t}, y\right) = 0,$$

mais nous nous contenterons d'étudier le cas explicite suivant :

$$\frac{\mathrm{d}^n y}{\mathrm{d}t^n} = f\left(\frac{\mathrm{d}^{n-1} y}{\mathrm{d}t^{n-1}}, \ldots, \frac{\mathrm{d}y}{\mathrm{d}t}, y\right).$$

Celui-ci se ramène à l'étude d'un système différentiel de $n$ équations du premier ordre, en posant successivement :

$$\frac{dy}{dt} = u,$$
$$\frac{du}{dt} = v,$$
$$\cdots\cdots$$
$$\frac{dw}{dt} = z,$$

ce qui donne $(n-1)$ équations lesquelles jointes à $\frac{dz}{dt} = f(y, u, v, \ldots, z)$ constituent un système de $n$ équations différentielles du premier ordre.

Dans un premier temps nous étudierons les techniques usuelles de résolution des équations différentielles du premier ordre dans le cadre du problème de Cauchy, puis nous procéderons aux généralisations qui permettront l'intégration des équations différentielles dont l'ordre est supérieur à un.

Enfin, nous donnerons quelques indications à propos du problème aux limites qui ne concerne, cela va de soi, que les équations d'ordre supérieur à un.

## 1. Les équations différentielles du premier ordre

C'est donc le problème de Cauchy qui seul nous concerne ici. Sous la forme explicite, nous aurons à intégrer l'équation générale suivante :

$$\frac{dy}{dx} = f(x, y),$$

avec la condition que la solution doit prendre la valeur $y_0$ pour la valeur $x_0$ de la variable.

## 2. Les théorèmes d'Arzelà (1847–1912) et de Cauchy-Lipschitz (1832–1903)

Il convient de s'assurer de l'existence d'une solution et la réponse est apportée par le théorème d'Arzelà que l'on admettra sans démonstration :

Sous la seule condition que $f(x, y)$ soit continue dans le rectangle $x_0 \leq x \leq x_0 + a$, $|y - y_0| \leq b$, alors l'équation différentielle admet au moins une solution $\phi(x)$ qui prend pour $x = x_0$ la valeur $y_0$ et qui est définie pour $x_0 \leq x \leq x_0 + h$, $h$ étant le plus petit des deux nombres $a$ et $b/M$ où $M = \sup |f(x, y)|$ dans le rectangle considéré. Rien n'est changé si l'on remplace la condition $x_0 \leq x \leq x_0 + a$ par $x_0 - a \leq x \leq x_0$.

À présent, c'est l'unicité de la solution qui va retenir notre attention, elle est assurée par le théorème de Cauchy-Lipschitz :

Si de plus la fonction $f(x, y)$ est lipschitzienne (cela signifie qu'il n'y a pas de tangente verticale), soit :

$$|f(x, y_1) - f(x, y_2)| < A\,|y_1 - y_2| \qquad \text{avec } A \text{ fini},$$

alors la solution $\phi(x)$ qui prend la valeur $y_0$ pour la valeur $x_0$ de la variable est unique. En outre, dans le domaine précédemment défini, $\phi(x)$ est continue.

***Remarque*** : Si $f(x,y)$ est infinie en un point particulier, il faut examiner s'il s'agit d'un pôle simple ou d'un point singulier essentiel. Dans le premier cas, si la fonction $\dfrac{1}{f(x,y)}$ est régulière, elle conserve les bonnes propriétés puisque l'on pourra intégrer $x$ en fonction de $y$.

On trouvera dans l'ouvrage de G. Valiron cité en bibliographie une démonstration de ces théorèmes au moyen de la méthode itérative de Picard (1890). Comme elle constitue aussi une méthode numérique simple à mettre en œuvre, nous allons nous intéresser à cet aspect.

## 3. La méthode de Picard (1858–1941)

La technique est intéressante dans la mesure où elle substitue une équation intégrale à une équation différentielle, soit :

$$\int_{y_0}^{y(x)} \mathrm{d}z = y(x) - y_0 = \int_{x_0}^{x} f[t, y(t)]\,\mathrm{d}t.$$

Nous allons former une suite de fonctions de la manière suivante :

$$y_1(x) = y_0 + \int_{x_0}^{x} f(t, y_0)\,\mathrm{d}t,$$

$$y_2(x) = y_0 + \int_{x_0}^{x} f[t, y_1(t)]\,\mathrm{d}t,$$

$$\cdots\cdots\cdots$$

$$y_n(x) = y_0 + \int_{x_0}^{x} f[t, y_{n-1}(t)]\,\mathrm{d}t,$$

$$\cdots\cdots\cdots$$

On montre alors que la limite de $y_n(x)$ quand $n$ tend vers l'infini est la solution unique de l'équation différentielle dans le cadre des hypothèses énoncées dans les théorèmes d'Arzelà et de Cauchy-Lipschitz. De plus, il convient de signaler que la série de terme général $y_n(x) - y_{n-1}(x)$ est absolument et uniformément convergente pour $x$ appartenant à l'intervalle $(x_0, x_0 + h)$.

La conduite du calcul est fort simple :

**a.** On découpe l'intervalle d'intégration $(a,b)$, dans lequel la fonction $f(x,y)$ a les bonnes propriétés, en $n$ sous-intervalles égaux définis par les points $x_k$ en progression arithmétique :

$$h = \frac{b-a}{n} \quad \text{et} \quad x_k = a + kh.$$

**b.** Nous calculons la fonction $y_1(x) = y_0 + \int_{x_0}^{x} f(t, y_0)\,\mathrm{d}t$, c'est-à-dire l'ensemble des échantillons de $y_1(x)$ aux points $x_k$.

**c.** Ensuite, on calcule $y_2(x)$ à partir de $y_1(x)$ et ainsi de suite.

**d.** On arrête les calculs lorsque les valeurs de $y_n(b)$ et $y_{n+1}(b)$ sont égales à la précision que l'on s'est fixé à l'avance.

***Remarque 1*** : Cette belle méthode itérative présente l'inconvénient d'être lentement convergente.

***Remarque 2*** : Comme la méthode est absolument et uniformément convergente il n'est pas utile de conserver la fonction $y_n(x)$ intégralement avant de procéder au calcul de la fonction suivante $y_{n+1}(x)$. Il suffit donc d'écraser les valeurs au fur et à mesure des calculs. Non seulement cette remarque simplifie grandement la programmation mais encore elle accélère la convergence du procédé.

### Précision de la méthode

La suite de fonctions étant absolument et uniformément convergente, sur le plan strictement mathématique, il suffira d'effectuer les calculs à un ordre suffisamment élevé pour obtenir la précision souhaitée à l'avance. Par suite, on est en droit de penser que c'est la technique numérique retenue pour calculer les intégrales qui limite la précision des résultats (elle est fonction du pas $h$). Les problèmes sont malheureusement plus compliqués que cela.

On trouvera sur le Web (*) le programme `picard.c` qui réalise cette procédure.

## 4. Méthode de la série de Taylor

Nous allons développer la fonction $y(x_{k+1})$ au voisinage de $y(x_k)$ en série de Taylor en utilisant les notations $y_k$ et $y_{k+1}$ :

$$y_{k+1} = y_k + \frac{h}{1!}y'_k + \frac{h^2}{2!}y''_k + \frac{h^3}{3!}y'''_k + \cdots$$

$y'_k$ est donné par l'équation différentielle et vaut simplement $f(x_k, y_k)$, ensuite, par dérivation, on obtient :

$$y''_k = \left(\frac{\partial f}{\partial x}\cdot\frac{\mathrm{d}x}{\mathrm{d}x} + \frac{\partial f}{\partial y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x}\right)_{x_k} = \left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\cdot y'\right)_{x_k} = g(x_k, y_k),$$

et

$$y'''_k = \left(\frac{\partial^2 f}{\partial x^2} + 2y'\frac{\partial^2 f}{\partial x \partial y} + (y')^2\frac{\partial^2 f}{\partial x^2} + y''\frac{\partial f}{\partial y}\right)_{x_k} = l(x_k, y_k).$$

On peut aller aussi loin que ce que l'on désire.

La technique de calcul est simple : à partir de $x_0$ et $y_0$ on calculera $y_1$ au point $x_1$, puis $y_2$ au point $x_2$ et ainsi de suite jusqu'à $x_n = b$.

Lorsque l'on arrête le développement en série de Taylor au premier ordre, la méthode s'appelle **méthode d'Euler**.

On trouvera sur le Web (*) le programme `taylor.c` qui met en œuvre cette technique.

* `http://www.edpsciences.com/guilpin/`

# 5. Méthodes de Runge (1856–1927) et Kutta (1867–1944)

## 5.1. La méthode en trois points

Nous allons présenter cette technique de façon géométrique, puis nous verrons ce qu'elle signifie analytiquement et nous proposerons une généralisation à plusieurs points.

À partir du point $A(x_0, y_0)$, nous calculons au moyen de la méthode d'Euler les coordonnées du point $B\,[x_0 + h/2, y(x_0 + h/2)]$, soit :

$$y\left(x_0 + \frac{h}{2}\right) = y_0 + hf(x_0, y_0)/2.$$

On détermine le point $C\,[x_0 + h, y(x_0 + h)]$ en effectuant une interpolation parabolique, c'est-à-dire en supposant que les trois points $A$, $B$ et $C$ sont sur une parabole. Pour cela on utilise une propriété géométrique de la parabole selon laquelle la tangente en $B$ à la parabole est parallèle à la sécante $AC$. On peut alors écrire :

$$y_1 = y(x_0 + h) = y_0 + hf\,[x_0 + h/2, y_0 + hf(x_0, y_0)/2]\,.$$

L'ordonnée de $C$ n'est pas la valeur définitive que nous allons conserver comme approximation de l'intégration en $x_0 + h$. La formule des trois niveaux permet de trouver une meilleure approximation de $C$ en écrivant :

$$\int_{x_0}^{x_0+h} \mathrm{d}y = \int_{x_0}^{x_0+h} f[x, y(x)]\,\mathrm{d}x = y_1 - y_0 = \frac{h}{6}\,[f(A) + 4f(B) + f(C)]$$

expression dans laquelle $f(A)$ signifie $f(x_0, y_0)$

$$\begin{aligned} f(B) &\quad \text{signifie} \quad f\,[x_0 + h/2, y_0 + hf(x_0, y_0)/2] \\ f(C) &\quad \text{signifie} \quad f\,(x_0 + h, y_0 + hf\,(x_0 + h/2, y_0 + hf(x_0, y_0)/2))\,. \end{aligned}$$

La technique de calcul consiste à écrire :

$$\begin{aligned} U_0 &= f(x_0, y_0), \\ U_1 &= f\,(x_0 + h/2, y_0 + hU_0/2)\,, \\ U_2 &= f\,(x_0 + h, y_0 + hU_1)\,, \\ \text{puis} \quad y_1 &= y_0 + \frac{h}{6}\,[U_0 + 4U_1 + U_2]\,. \end{aligned}$$

Ensuite, à partir de $y_1$, on calcule $y_2$ et ainsi de suite jusqu'à atteindre la borne $b$ souhaitée.

On trouvera sur le Web(*) le programme `runge.c` qui réalise cette procédure.

Cette procédure montre que l'on itère trois fois la fonction $f(x, y)$. Nous pouvons retrouver ce même résultat par voie analytique. Cette façon de procéder est généralisable ce qui permet d'augmenter la précision des résultats. Sans s'attarder sur les développements, nous donnons les résultats du calcul en quatre points.

* `http://www.edpsciences.com/guilpin/`

### 5.2. Les méthodes en quatre points

Il existe plusieurs variantes de la méthode, en voici deux :

1. On calcule les valeurs suivantes :

$$\begin{aligned} U_0 &= hf(x_0, y_0), \\ U_1 &= hf(x_0 + h/3, y_0 + U_0/3), \\ U_2 &= hf(x_0 + 2h/3, y_0 - U_0/3 + U_1), \\ U_3 &= hf(x_0 + h, y_0 + U_0 - U_1 + U_2), \\ \text{et} \quad y(x_0 + h) &= y(x_0) + \frac{U_0}{8} + \frac{3U_1}{8} + \frac{3U_2}{8} + \frac{U_3}{8}. \end{aligned}$$

2. On peut aussi calculer cette autre suite :

$$\begin{aligned} U_0 &= hf(x_0, y_0), \\ U_1 &= hf(x_0 + h/2, y_0 + U_0/2), \\ U_2 &= hf(x_0 + h/2, y_0 + U_1/2), \\ U_3 &= hf(x_0 + h, y_0 + U_2), \\ \text{et} \quad y(x_0 + h) &= y(x_0) + \frac{U_0}{6} + \frac{U_1}{3} + \frac{U_2}{3} + \frac{U_3}{6}. \end{aligned}$$

Il faut savoir que les coefficients numériques sont calculés par identification de la formule présentée et du développement en série de Taylor le plus élevé possible.

### 5.3. Les méthodes en seize points (ordre 8)

Il faut résoudre un système de 285 équations pour satisfaire les conditions d'identification et en 1981, P.J. Prince et J.R. Dormand ont réalisé ces calculs et ont produit les coefficients numériques jusqu'à 16 points.

Souvent nous utilisons leurs résultats pour intégrer les équations différentielles, et ceci de la façon suivante : on intègre avec la méthode d'ordre 7 et l'on estime l'erreur à l'aide de la méthode d'ordre 8 qui sert de référence. Ceci permet de calculer aussi le pas d'intégration adapté au type de l'équation différentielle que l'on traite.

On retrouve ce type d'intégration dans quelques programmes fournis avec les corrections de problèmes (*cf.* annexes H et I).

## 6. Les méthodes d'Adams (1819–1892)

On se propose de calculer la valeur de la solution $y_j$ pour la valeur correspondante de l'abscisse $x_j$. Les abscisses ne sont pas nécessairement en progression arithmétique et l'on les note simplement dans l'ordre croissant $\{x_0 < x_1 < x_2 < \cdots < x_n \cdots < x_m\} = T$, $T$ étant l'intervalle sur lequel on désire effectuer les calculs et sur lequel la fonction $f[x, y(x)]$ a les bonnes propriétés énoncées dans le théorème de Cauchy-Lipschitz. On note $h_n = x_{n+1} - x_n$.

On suppose que l'on connaît les $n$ premières valeurs qui ont pu être obtenues au moyen d'une méthode de Runge et Kutta par exemple. On se propose de calculer la valeur de $y_{n+1}$ au point $x_{n+1}$ au moyen de la relation :

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f[x, y(x)]\, \mathrm{d}x,$$

expression dans laquelle nous allons approcher la fonction $f[x, y(x)]$ au moyen du polynôme d'interpolation de Lagrange de degré $\nu$, ce qui signifie que la détermination de ce polynôme nécessite la connaissance de $\nu + 1$ valeurs calculées précédemment.

## 6.1. Méthode de Adams-Bashforth

On désigne par $P_{\nu n}(x)$ le polynôme de Lagrange de degré $\nu$ qui interpole la fonction $f[x, y(x)]$ dans l'intervalle $(x_n, x_{n+1})$. On adopte la notation :

$$P_{\nu n}(x_{n-j}) = f[x_{n-j}, y(x_{n-j})] = f_{n-j} \qquad \text{avec } j = 0, 1, \ldots \nu.$$

La valeur approchée de $y_{n+1}$ est donnée par l'expression :

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} P_{\nu n}(x)\, \mathrm{d}x,$$

on en déduit la valeur de $f_{n+1}$ :

$$f_{n+1} = f\left(x_{n+1}, y_{n+1}\right).$$

Explicitons le polynôme de Lagrange :

$$P_{\nu n}(x) = \sum_{j=0}^{\nu} f_{n-j} \prod_{\substack{k=0 \\ k \neq j}}^{\nu} \frac{x - x_{n-k}}{x_{n-j} - x_{n-k}}.$$

On peut poser :

$$L_{\nu nj}(x) = \prod_{\substack{k=0 \\ k \neq j}}^{\nu} \frac{x - x_{n-k}}{x_{n-j} - x_{n-k}},$$

puis :

$$B_{\nu nj}(x) = \frac{1}{h_n} \int_{x_n}^{x_{n+1}} L_{\nu nj}(x)\, \mathrm{d}x = \frac{1}{h_n} \int_{x_n}^{x_{n+1}} \prod_{\substack{k=0 \\ k \neq j}}^{\nu} \frac{x - x_{n-k}}{x_{n-j} - x_{n-k}}\, \mathrm{d}x.$$

En définitive, la méthode d'Adams-Bashforth à $(\nu + 1)$ pas s'écrit :

$$y_{n+1} = y_n + \sum_{j=0}^{\nu} h_n B_{\nu nj} f_{n-j} \qquad \text{avec } n \geq \nu,$$

et

$$f_{n+1} = f\left(x_{n+1}, y_{n+1}\right).$$

***Remarque 1*** : Il faut une plate-forme de démarrage pour appliquer l'algorithme d'Adams-Bashforth, elle peut être obtenue par une méthode à un seul pas telle que celles que nous avons présentées précédemment.

***Remarque 2*** : Rien n'interdit de modifier le nombre de pas au cours des calculs.

***Remarque 3*** : La taille des pas $h_j$ n'est commandée que par des considérations sur la précision et la stabilité.

***Remarque 4*** : Comme la méthode existe pour $\nu = 0$ (il s'agit de la méthode d'Euler), il suffit de concevoir un programme qui fonctionne avec un pas variable de 0 à $\nu$, ainsi l'algorithme devient « auto-démarrant ».

## 6.2. Cas où les $x_j$ sont en progression arithmétique

On désigne alors simplement par $h$ le pas, et l'on remarque que le polynôme d'interpolation de Lagrange est également le polynôme ascendant de Newton (*cf.* chapitre 6) ; ainsi $L_{\nu nj}(x)$ prend une forme plus simple que l'on désigne par $l_{\nu j}(x)$ car, comme on va le voir, cette expression ne dépend pas de $n$ :

$$l_{\nu j}(x) = \prod_{\substack{k=0 \\ k \neq j}}^{\nu} \frac{x-k}{k-j},$$

de même $B_{\nu nj}$ ne dépend plus de $n$ et l'on désigne par $b_{\nu j}$ cette nouvelle expression qui s'écrit :

$$b_{\nu j} = \int_0^1 l_{\nu j}(u)\, \mathrm{d}u = \int_0^1 \prod_{\substack{k=0 \\ k \neq j}}^{\nu} \frac{u-k}{k-j}\, \mathrm{d}u.$$

Par ailleurs, nous avons vu que le polynôme de Newton ascendant (*cf.* chapitre 6) pouvait s'écrire, en posant $u = \dfrac{x - x_n}{h}$, sous la forme :

$$P_{\nu n}(u) = f_n + u\Delta^1 f_n + \frac{u(u+1)}{2!}\Delta^2 f_n + \cdots + \frac{u(u+1)\cdots(u+\nu-1)}{\nu!}\Delta^\nu f_n$$

qui s'exprime avec la notation plus concise :

$$\binom{u+j-1}{j} = \frac{u(u+1)\cdots(u+j-1)}{j!},$$

de la manière suivante :

$$P_{\nu n}(u) = \sum_{j=0}^{\nu} \Delta^j f_n \binom{u+j-1}{j}.$$

De là, on obtient :

$$y_{n+1} = y_n + h\sum_{j=0}^{\nu} \Delta^j f_n \int_0^1 \binom{u+j-1}{j}\, \mathrm{d}u = y_n + h\sum_{j=0}^{\nu} \gamma_j \Delta^j f_n$$

avec

$$\gamma_j = \int_0^1 \binom{u+j-1}{j} \, \mathrm{d}u.$$

Il y a moyen de calculer directement les $\gamma_j$ ainsi que les $b_{\nu j}$ qui en découlent, si bien qu'**il n'est plus utile de calculer les différences successives des $f_n$**. Les expressions à exploiter s'écrivent :

$$y_{n+1} = y_n + h \sum_{j=0}^{\nu} b_{\nu j} f_{n-j} \qquad \text{avec } n \geq \nu,$$
$$\text{et } f_{n+1} = f(x_{n+1}, y_{n+1}).$$

## 6.3. Calcul des coefficients $\gamma_j$ et $b_{\nu j}$

Nous allons montrer que les $\gamma_k$ sont les coefficients du développement en série entière de la fonction :

$$S(x) = \int_0^1 (1-x)^{-\alpha} \, \mathrm{d}\alpha = \sum_{k=0}^{\infty} \gamma_k x^k.$$

En effet, la dérivation successive sous le signe somme, puis l'égalisation à zéro de $x$ conduit au développement en série de MacLaurin de $S(x)$, et l'on trouve que :

$$\gamma_j = \int_0^1 \binom{u+j-1}{j} \, \mathrm{d}u.$$

D'un autre côté, l'intégrale est directement calculable :

$$S(x) = \int_0^1 (1-x)^{-\alpha} \, \mathrm{d}\alpha = \int_0^1 \exp[-\alpha \log_e(1-x)] \, \mathrm{d}\alpha = \frac{-1}{\log_e(1-x)} \left( \frac{x}{1-x} \right).$$

Nous pouvons donc écrire :

$$S(x) \log_e(1-x) \frac{1-x}{x} = -1.$$

Le développement en série entière de cette expression donne :

$$\left( \sum_{k=0}^{\infty} \gamma_k x^k \right) \left( -1 + \sum_{m=1}^{\infty} \frac{x^m}{m(m+1)} \right) = -1.$$

L'identification des termes de même puissance en $x$ conduit aux relations :

$$\gamma_0 = 1$$
$$\gamma_m = \sum_{k=0}^{m-1} \frac{\gamma_k}{(m-k)(m-k+1)} = \sum_{k=0}^{m-1} \gamma_k \left[ \frac{1}{(m-k)} - \frac{1}{(m-k+1)} \right],$$

soit encore :

$$\gamma_m = \sum_{k=0}^{m-1} \frac{\gamma_k}{(m-k)} - \sum_{k=0}^{m-1} \frac{\gamma_k}{(m-k+1)},$$

$$\sum_{k=0}^{m-1} \frac{\gamma_k}{(m-k)} = \sum_{k=0}^{m} \frac{\gamma_k}{(m-k+1)}.$$

Cette dernière relation est donc vraie quel que soit $m$, et l'égalité demeure en remplaçant $m$ par $m-1$ ou $(m+1)$ autant de fois que cela reste compatible avec le fait que $m$ est supérieur ou égal à zéro. En particulier, elle est vraie pour $m = 1$ dans le premier membre et $m$ quelconque dans le second, ainsi on trouve :

$$\gamma_0 = 1 = \sum_{k=0}^{m} \frac{\gamma_k}{(m-k+1)} \qquad \text{quel que soit } m \text{ positif.}$$

À partir de cette relation on calcule la suite $\gamma_1$, $\gamma_2$, $\gamma_3$ et ainsi de suite.

Le calcul des $b_{\nu j}$ s'effectue en fonction de $b_{\nu-1,j}$ et de $\gamma_\nu$ pour $0 \leq j < \nu$. D'abord, il est aisé de calculer $b_{\nu\nu}$ à partir de la relation de définition :

$$b_{\nu\nu} = \int_0^1 l_{\nu\nu}(u)\,\mathrm{d}u = \int_0^1 \prod_{k=0}^{\nu-1} \frac{u-k}{k-\nu}\,\mathrm{d}u = (-1)^\nu \int_0^1 \prod_{k=0}^{\nu-1} \frac{u-k}{\nu!}\,\mathrm{d}u = (-1)^\nu \gamma_\nu.$$

À présent, exprimons $l_{\nu j}(x)$ en fonction de $\binom{x+\nu-1}{\nu}$. On peut écrire, pour $j < \nu$ :

$$l_{\nu j}(x) = (-1)^j C_\nu^j \binom{x+\nu-1}{\nu} \frac{x+\nu}{x+j}.$$

De la même façon, on calcule $l_{\nu-1,j}(x)$ :

$$l_{\nu-1,j}(x) = (-1)^j C_\nu^j \binom{x+\nu-1}{\nu} \frac{\nu-j}{x+j}.$$

Maintenant, effectuons la différence $l_{\nu j}(x) - l_{\nu-1,j}(x)$ :

$$l_{\nu j}(x) - l_{\nu-1,j}(x) = (-1)^j C_\nu^j \binom{x+\nu-1}{\nu} \left( \frac{x+\nu}{x+j} - \frac{\nu-j}{x+j} \right) = (-1)^j C_\nu^j \binom{x+\nu-1}{\nu}.$$

Intégrons terme à terme sur $(0,1)$, on obtient :

$$b_{\nu j} = b_{\nu-1,j} + (-1)^j C_\nu^j \gamma\nu \qquad \text{avec } 0 \leq j < \nu.$$

En définitive, nous savons calculer les $b_{\nu j}$ sans difficulté à partir des $\gamma_\nu$. Sur le Web(*), on trouvera le programme `bashfort.c` qui utilise cette technique de calcul.

* `http://www.edpsciences.com/guilpin/`

## 6.4. Méthode de Adams-Moulton (1872–1952)

Au fond, il s'agit d'une variante de la méthode qui vient d'être exposée. La différence provient du fait que le polynôme d'interpolation passe par $f_{n+1} = f(x_{n+1}, y_{n+1})$ valeur qui ne sera connue que lorsque la valeur de $y_{n+1}$ sera connue. On comprend alors que la nouvelle technique soit une méthode implicite car il nous faudra résoudre une équation implicite pour obtenir la valeur de $y_{n+1}$. Le gain de la méthode repose sur le fait que le polynôme d'interpolation gagne un degré.

Nous posons donc :

$$\begin{aligned} Q_{\nu+1,n}(x_{n-j}) &= f[x_{n-j}, y(x_{n-j})] \qquad \text{avec } j = 0, 1, \ldots \nu, \\ \text{plus } Q_{\nu+1,n}(x_{n+1}) &= f[x_{n+1}, y(x_{n+1})] = f_{n+1}, \end{aligned}$$

cette dernière valeur est inconnue.

Le polynôme de Lagrange qui interpole dans le $n^{\text{e}}$ intervalle est alors de degré $(\nu+1)$. Il prend la forme :

$$Q_{\nu+1,n}(x) = \sum_{j=0}^{\nu+1} f_{n+1-j} \prod_{\substack{k=0 \\ k\neq j}}^{\nu+1} \frac{x - x_{n+1-k}}{x_{n+1-j} - x_{n+1-k}} = \sum_{j=-1}^{\nu} f_{n-j} \prod_{\substack{k=-1 \\ k\neq j}}^{\nu} \frac{x - x_{n-k}}{x_{n-j} - x_{n-k}},$$

soit encore :

$$Q_{\nu+1,n}(x) = \sum_{j=-1}^{\nu} f_{n-j} L_{\nu+1,n+1,j+1}(x).$$

Comme précédemment, posons :

$$D_{\nu nj} = \frac{1}{h_n} \int_{x_n}^{x_{n+1}} L_{\nu+1,n+1,j+1}(x)\, \mathrm{d}x.$$

La technique de résolution consiste à séparer ce qui est connu de ce qui doit être calculé par résolution d'une équation implicite :

$$y_{n+1} - h_n D_{\nu n,-1} f(x_{n+1}, y_{n+1}) = y_n + \sum_{j=0}^{\nu} h_n D_{\nu nj} f_{n-j} \qquad \text{avec } n \geq \nu.$$

(Le membre de droite est une constante calculable directement, et il reste à résoudre l'équation implicite figurant dans le premier membre égale à ladite constante.)

## 6.5. Cas où les $x_j$ sont en progression arithmétique

L'expression du polynôme de Newton devient :

$$\begin{aligned} Q_{\nu+1,n}(x) &= f_{n+1} + \frac{u}{1!}\Delta^1 f_{n+1} + \frac{u(u+1)}{2!}\Delta^2 f_{n+1} + \cdots + \frac{u(u+1)\cdots(u+n)}{(n+1)!}\Delta^{n+1} f_{n+1} \\ &= \sum_{k=0}^{\nu+1} \binom{u+k-1}{k} \Delta^k f_{n+1} \qquad \text{avec } u = \frac{x - x_{n+1}}{h}. \end{aligned}$$

Ici encore, nous pouvons nous dispenser du calcul effectif des différences successives, seules les valeurs $f_k$ pour$k = n+1, \ldots, n-\nu$ devront être connues.

La quantité $D_{\nu nj}$ ne dépend plus de $n$ et l'on désigne par $d_{\nu j}$ cette nouvelle expression qui s'écrit (attention au changement de variable qui définit $u$) :

$$d_{\nu j} = \int_{-1}^{0} l_{\nu+1,j+1}(u)\,\mathrm{d}u = \int_{-1}^{0} \prod_{\substack{k=0 \\ k\neq j}}^{\nu+1} \frac{u-k}{k-j}\,\mathrm{d}u.$$

De là, on extrait :

$$y_{n+1} = y_n + h\sum_{k=0}^{\nu+1} \Delta^k f_{n+1} \int_{-1}^{0} \binom{u+k-1}{k}\,\mathrm{d}u = y_n + h\sum_{j=0}^{\nu+1} \delta_j \Delta^j f_n$$

avec

$$\delta_j = \int_{-1}^{0} \binom{u+j-1}{j}\,\mathrm{d}u = \int_{-1}^{0} \frac{u(u+1)\cdots(u+j-1)}{j!}\,\mathrm{d}u.$$

L'expression générale à itérer devient plus simple :

$$y_{n+1} - h_n d_{\nu j-1} f\,(x_{n+1}, y_{n+1}) = y_n + \sum_{j=0}^{\nu} h_n d_{\nu j} f_{n-j}.$$

## 6.6. Calcul des coefficients $\delta_j$ et $d_{\nu j}$

Nous allons relier les coefficients $\delta_j$ aux coefficients $\gamma_k$. Écrivons $\delta_j$ et effectuons le changement de variable $u - 1 = x$ :

$$\delta_j = \int_{-1}^{0} \binom{u+j-1}{j}\,\mathrm{d}u = \int_{0}^{1} \frac{(x-1)x(x+1)\cdots(x+j-2)}{j!}\,\mathrm{d}x,$$

puis décomposons le dernier membre sous le signe somme :

$$\begin{aligned}
\frac{(x-1)x(x+1)\cdots(x+j-2)}{j!} &= \frac{x(x+1)\cdots(x+j-2)}{(j-1)!}\left(\frac{x-1}{j}\right) \\
&= \frac{x(x+1)\cdots(x+j-2)}{(j-1)!}\left(\frac{x-1}{j} - 1 + 1\right) \\
&= \frac{x(x+1)\cdots(x+j-2)}{(j-1)!}\left(\frac{x+j-1}{j} - 1\right) \\
&= -\frac{x(x+1)\cdots(x+j-2)}{(j-1)!} + \frac{x(x+1)\cdots(x+j-1)}{j!},
\end{aligned}$$

à présent il suffit d'intégrer de zéro à l'infini et l'on obtient :

$$\delta_j = \gamma_j - \gamma_{j-1} \qquad \text{avec } \delta_0 = \gamma_0 = 1.$$

Nous pouvons également calculer directement la suite des $\delta_j$ en réutilisant la relation établie à propos des $\gamma_j$, à savoir :

$$\sum_{k=0}^{m-1} \frac{\gamma_k}{(m-k)} - \sum_{k=0}^{m} \frac{\gamma_k}{(m-k+1)} = 0.$$

Le développement puis la factorisation des coefficients qui ont le même dénominateur permet d'écrire :

$$0 = \frac{\gamma_0}{m+1} + \frac{\gamma_0 - \gamma_1}{m} + \cdots + \frac{\gamma_{m-j} - \gamma_{m-j+1}}{j} + \cdots + (\gamma_{m-1} - \gamma_m),$$

ce qui donne :

$$0 = \sum_{k=0}^{m} \frac{\delta_k}{(m-k+1)} \quad \text{quel que soit } m \text{ positif.}$$

Il nous reste à exprimer les $d_{\nu j}$ en fonction des $\delta_j$. Nous avons d'abord :

$$d_{\nu\nu} = \int_{-1}^{0} l_{\nu+1,\nu+1}(u)\,\mathrm{d}u = \int_{-1}^{0} \prod_{k=0}^{\nu} \frac{u-k}{k-\nu}\,\mathrm{d}u = (-1)^{\nu+1} \int_{-1}^{0} \prod_{k=0}^{\nu} \frac{u-k}{(\nu+1)!}\,\mathrm{d}u = (-1)^{\nu+1}\delta_{\nu+1}.$$

Comme précédemment, nous écrivons $l_{\nu+1 j+1}(x)$ en fonction de $\binom{x+\nu}{\nu+1}$. On peut écrire, pour $j < \nu$ :

$$l_{\nu+1,j+1}(x) = (-1)^{j+1} C_{\nu+1}^{j+1} \binom{x+\nu}{\nu+1} \frac{x+\nu+1}{x+j+1}.$$

Maintenant, effectuons la différence $l_{\nu+1,j+1}(x) - l_{\nu j+1}(x)$ :

$$\begin{aligned} l_{\nu+1,j+1}(x) - l_{\nu j+1}(x) &= (-1)^{j+1} C_{\nu+1}^{j+1} \binom{x+\nu}{\nu+1} \left( \frac{x+\nu+1}{x+j+1} - \frac{\nu-j}{x+j+1} \right) \\ &= (-1)^{j} (-1)^{j+1} C_{\nu+1}^{j+1} \binom{x+\nu}{\nu+1}. \end{aligned}$$

L'intégration de deux membres sur l'intervalle $(-1, 0)$ permet d'obtenir :

$$d_{\nu j} = d_{\nu-1,j} + (-1)^{j+1} C_{\nu+1}^{j+1} \delta_{\nu+1} \qquad \text{avec } 0 \le j < \nu.$$

Ici encore, on sait facilement calculer les $d_{\nu j}$ en fonction des $\delta_{\nu+1}$.

On trouvera sur le Web(*) le programme `moulton.c` qui réalise l'intégration numérique des équations différentielles du premier ordre au moyen de cette procédure.

## 7. La méthode des différentiations rétrogrades

Il existe une catégorie de problèmes différentiels qui sont particulièrement rebelles aux méthodes d'analyse usuelles, c'est-à-dire les méthodes de Runge et Kutta et les méthodes d'Adams. Bien que les conditions de Cauchy-Lipschitz soient vérifiées, il peut arriver que la structure de l'équation différentielle soit tout à fait mauvais, c'est-à-dire que la discrétisation explicite fait apparaître des instabilités. Ce type de problème porte un nom : il s'agit de **problèmes raides**, on n'obtient pas d'approximations numériques convenables à l'aide des méthodes usuelles.

* `http://www.edpsciences.com/guilpin/`

Lorsqu'on rencontre un problème raide, ces difficultés peuvent être éventuellement contournées en utilisant des méthodes adaptées, et il convient de tester les méthodes de Runge et Kutta implicites ou la méthode des différentiations rétrogrades. Nous présentons maintenant cette dernière méthode.

L'idée consiste à interpoler la suite des $y_n$ par le polynôme de Lagrange $\Pi_{\nu n}(x)$ de degré $\nu$ $(n \geq \nu + 1)$. Le polynôme sert alors à trouver une approximation de $y_{n+1}$ pour la valeur $x = x_{n+1}$, ainsi nous écrivons :

$$\Pi_{\nu n}(x_{n+1-k}) = y_{n+1-k} \qquad \text{avec } k = 0, 1, \ldots, \nu.$$

Pour obtenir la valeur de $y_{n+1}$, on écrit que :

$$\left[\frac{\mathrm{d}}{\mathrm{d}x}\Pi_{\nu n}(x)\right]_{x=x_{n+1}} = f\left(x_{n+1}, y_{n+1}\right).$$

Nous allons expliciter le polynôme de Lagrange dans le cas particulier utile où le pas est constant et vaut $h$ :

$$\Pi_{\nu n}(x) = \sum_{j=0}^{\nu} y_{n+1-j} \prod_{\substack{k=0 \\ k\neq j}}^{\nu} \frac{x - x_{n+1-k}}{x_{n+1-j} - x_{n+1-k}}.$$

On désigne alors par $\pi_{\nu n}(u)$ la nouvelle expression qui prend la forme suivante :

$$\pi_{\nu n}(u) = \sum_{j=0}^{\nu} y_{n+1-j} \prod_{\substack{k=0 \\ k\neq j}}^{\nu} \frac{-u+k}{k-j} \quad \text{en posant : } u = \frac{x_{n+1} - x}{h}.$$

On obtient alors :

$$\pi'_{\nu n}(0) = -hf\left(x_{n+1}, y_{n+1}\right),$$

$$\text{et } \pi_{\nu n}(\xi_k) = y_{n+1-k} \qquad \text{avec } k = 0, 1, \ldots, \nu \qquad \text{où } \xi_k = \frac{x_{n+1} - x_{n+1} - k}{h}.$$

Nous allons dériver $\pi_{\nu n}(u)$ par rapport à $u$, après avoir donné son expression en fonction de :

$$q_{\nu n}(u) = \prod_{m=1}^{\nu} \left(1 - \frac{u}{m}\right)$$

ainsi que de sa dérivée $q'_{\nu n}(u)$. Développons $\pi_{\nu n}(u)$ de la manière suivante :

$$\pi_{\nu n}(u) = y_{n+1} \prod_{k=1}^{\nu} \frac{-u+k}{k} + \sum_{j=1}^{\nu} y_{n+1-j} \prod_{\substack{k=0 \\ k\neq j}}^{\nu} \frac{-u+k}{k-j}.$$

On note que, puisque $j \neq 0$ :

$$\prod_{\substack{k=0 \\ k\neq j}}^{\nu} \frac{-u+k}{k-j} = u \prod_{\substack{k=1 \\ k\neq j}}^{\nu} \frac{-u+k}{k-j} = u \frac{q_{\nu n}(u)}{1 - \dfrac{u}{j}} \Bigg/ \prod_{\substack{k=1 \\ k\neq j}}^{\nu} \left(1 - \frac{j}{k}\right),$$

et que

$$q'_{\nu n}(u) = -q_{\nu n}(u) \cdot \sum_{m=1}^{\nu} \frac{1/m}{\left(1 - \dfrac{u}{m}\right)},$$

ce qui donne :

$$q'_{\nu n}(0) = -\sum_{j=1}^{\nu} \frac{1}{m} \qquad \text{car } q_{\nu n}(0) = 1.$$

Donc :

$$\pi_{\nu n}(u) = y_{n+1} q_{\nu n}(u) + \sum_{j=1}^{\nu} y_{n+1-j} u \frac{q_{\nu n}(u)}{1 - \frac{u}{j}} \Bigg/ \prod_{\substack{k=1 \\ k \neq j}}^{\nu} \left(1 - \frac{j}{k}\right).$$

Dérivons par rapport à $u$ et faisons $u = 0$ :

$$\pi'_{\nu n}(0) = y_{n+1} q'_{\nu n}(0) + \sum_{j=1}^{\nu} \frac{y_{n+1-j}}{j \prod\limits_{\substack{k=1 \\ k \neq j}}^{\nu} \left(1 - \frac{j}{k}\right)} = -hf\left(x_{n+1}, y_{n+1}\right).$$

En réécrivant la dernière égalité, on obtient :

$$y_{n+1} + \frac{hf\left(x_{n+1}, y_{n+1}\right)}{q'_{\nu n}(0)} = -\sum_{j=1}^{\nu} \frac{y_{n+1-j}}{j q'_{\nu n}(0) \prod\limits_{\substack{k=1 \\ k \neq j}}^{\nu} \left(1 - \frac{j}{k}\right)},$$

soit encore :

$$y_{n+1} - \frac{hf\left(x_{n+1}, y_{n+1}\right)}{\sum\limits_{j=1}^{\nu} \frac{1}{m}} = \sum_{j=1}^{\nu} \frac{y_{n+1-j}}{j \sum\limits_{j=1}^{\nu} \frac{1}{m} \prod\limits_{\substack{k=1 \\ k \neq j}}^{\nu} \left(1 - \frac{j}{k}\right)}.$$

Formellement, on peut écrire une expression semblable à celles établies à propos des méthodes d'Adams :

$$y_{n+1} - h\alpha_\nu f(x_{n+1}, y_{n+1}) = \sum_{j=1}^{\nu} a_{\nu j} y_{n+1-j},$$

avec :

$$\alpha_\nu = \frac{1}{\sum\limits_{j=1}^{\nu} \frac{1}{m}}$$

et

$$a_{\nu j} = \frac{\alpha_\nu}{j \prod\limits_{\substack{k=1 \\ k \neq j}}^{\nu} \left(1 - \frac{j}{k}\right)}.$$

Sur le Web (*), on trouvera le programme `retrogra.c` qui réalise cet algorithme.

* `http://www.edpsciences.com/guilpin/`

# 8. Les équations différentielles du deuxième ordre

L'étude des équations différentielles du deuxième ordre relève de l'étude des systèmes d'équations différentielles du premier ordre, et elles en constituent un cas particulier. En effet, écrivons l'équation résolue par rapport à la dérivée seconde qui soit la plus générale possible dans ces conditions :

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = g(x, y, y')$$

il est alors possible, comme nous l'avons vu précédemment d'écrire :

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= z \\ \frac{\mathrm{d}z}{\mathrm{d}x} &= g(x, y, z). \end{aligned}$$

Il ne coûte rien d'étudier le système général suivant :

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} &= f(x, y, z) \\ \frac{\mathrm{d}z}{\mathrm{d}x} &= g(x, y, z). \end{aligned}$$

## 8.1. Cas du problème de Cauchy

En $x_0$, on connaît les valeurs $y_0$ et $z_0$, alors le théorème de Cauchy-Lipschitz se généralise de la manière suivante :

Si les fonctions $f(x, y, z)$ et $g(x, y, z)$ sont continues dans le parallélépipède $x_0 \leq x \leq x_0 + a$, $|y - y_0| \leq b$ et $|z - z_0| \leq c$, et si les fonctions sont lipschitziennes c'est-à-dire que :

$$\begin{aligned} |f(x, y, z) - f(x, y_1, z_1)| &< A\,|y - y_1| + B\,|z - z_1|\,, \\ |g(x, y, z) - g(x, y_1, z_1)| &< A\,|y - y_1| + B\,|z - z_1|\,, \end{aligned}$$

alors le système différentiel du premier ordre (ou l'équation différentielle du deuxième ordre) admet **une solution unique** $y = \phi(x)$ et $z = \psi(x)$ laquelle prend pour $x = x_0$ les valeurs $\phi(x_0) = y_0$ et $\psi(x_0) = z_0$ et qui est définie pour $x_0 \leq x \leq x_0 + h$, $h$ étant le plus petit des trois nombres $a$, $b/M$ et $c/M$ où $M$ étant la borne supérieure de $|f|$ et $|g|$ dans le parallélépipède considéré.

***a – Méthode de Picard –*** Nous formons la suite de fonctions de la manière suivante :

$$\begin{aligned} y_1(x) &= y_0 + \int_{x_0}^{x} f(t, y_0, z_0)\,\mathrm{d}t \\ z_1(x) &= z_0 + \int_{x_0}^{x} g(t, y_0, z_0)\,\mathrm{d}t \end{aligned}$$

puis

$$\begin{aligned} y_2(x) &= y_0 + \int_{x_0}^{x} f(t, y_1, z_1)\,\mathrm{d}t \\ z_2(x) &= z_0 + \int_{x_0}^{x} g(t, y_1, z_1)\,\mathrm{d}t \\ &\cdots\cdots \\ y_{n+1}(x) &= y_0 + \int_{x_0}^{x} f(t, y_n, z_n)\,\mathrm{d}t \\ z_{n+1}(x) &= z_0 + \int_{x_0}^{x} g(t, y_n, z_n)\,\mathrm{d}t. \\ &\cdots\cdots \end{aligned}$$

***b – Méthode des séries de Taylor*** – Il suffit d'écrire les expressions suivantes :

$$\begin{aligned} y_{k+1} &= y_k + \frac{h}{1!}y'_k + \frac{h^2}{2!}y''_k + \frac{h^3}{3!}y'''_k + \cdots \\ z_{k+1} &= z_k + \frac{h}{1!}z'_k + \frac{h^2}{2!}z''_k + \frac{h^3}{3!}z'''_k + \cdots \\ \text{avec}\quad y'_k &= f\left(x_k, y_k, z_k\right) \\ z'_k &= g\left(x_k, y_k, z_k\right), \end{aligned}$$

ensuite, par dérivation, on obtient :

$$y''_k = \left(\frac{\partial f}{\partial x}\cdot\frac{\mathrm{d}x}{\mathrm{d}x} + \frac{\partial f}{\partial y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x} + \frac{\partial f}{\partial z}\cdot\frac{\mathrm{d}z}{\mathrm{d}x}\right)_{x_k} = \left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\cdot y' + \frac{\partial f}{\partial z}\cdot z'\right)_{x_k} = k(x_k, y_k, z_k),$$

$$z''_k = \left(\frac{\partial g}{\partial x}\cdot\frac{\mathrm{d}x}{\mathrm{d}x} + \frac{\partial g}{\partial y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x} + \frac{\partial g}{\partial z}\cdot\frac{\mathrm{d}z}{\mathrm{d}x}\right)_{x_k} = \left(\frac{\partial g}{\partial x} + \frac{\partial g}{\partial y}\cdot y' + \frac{\partial g}{\partial z}\cdot z'\right)_{x_k} = l(x_k, y_k, z_k),$$

et ainsi de suite.

***c – Méthode de Runge et Kutta en trois points*** – La technique de calcul consiste à écrire :

$$\begin{aligned} U_0 &= f\left(x_0, y_0, z_0\right), \\ V_0 &= f\left(x_0 + h/2, y_0 + hU_0/2, z_0\right), \\ W_0 &= f\left(x_0 + h, y_0 + hV_0, z_0\right), \\ R_0 &= g\left(x_0, y_0, z_0\right), \\ S_0 &= g\left(x_0 + h/2, y_0, z_0 + hR_0/2\right), \\ T_0 &= g\left(x_0 + h, y_0, z_0 + hS_0\right), \\ \text{puis}\quad y_1 &= y_0 + \frac{h}{6}\left[U_0 + 4V_0 + W_0\right], \\ z_1 &= z_0 + \frac{h}{6}\left[R_0 + 4S_0 + T_0\right]. \end{aligned}$$

Ensuite, à partir de $y_1$ et $z_1$, on calcule $y_2$ et $z_2$ jusqu'à atteindre la borne $b$ souhaitée.

### d – Méthode de Runge et Kutta en quatre points

1. On calcule les valeurs suivantes :

$$
\begin{aligned}
U_0 &= hf\,(x_0, y_0, z_0)\,,\\
V_0 &= hg\,(x_0, y_0, z_0)\,,\\
U_1 &= hf\,(x_0 + h/3, y_0 + U_0/3, z_0)\,,\\
V_1 &= hg\,(x_0 + h/3, y_0, z_0 + V_0/3)\,,\\
U_2 &= hf\,(x_0 + 2h/3, y_0 - U_0/3 + U_1, z_0)\,,\\
V_2 &= hg\,(x_0 + 2h/3, y_0, z_0 - V_0/3 + V_1)\,,\\
U_3 &= hf\,(x_0 + h, y_0 + U_0 - U_1 + U_2, z_0)\,,\\
V_3 &= hg\,(x_0 + h, y_0, z_0 + V_0 - V_1 + V_2)\,,
\end{aligned}
$$

$$
\text{et}\quad y(x_0+h) = y(x_0) + \frac{U_0}{8} + \frac{3U_1}{8} + \frac{3U_2}{8} + \frac{U_3}{8},
$$

$$
z(x_0+h) = z(x_0) + \frac{V_0}{8} + \frac{3V_1}{8} + \frac{3V_2}{8} + \frac{V_3}{8}.
$$

2. On peut aussi calculer cette autre suite :

$$
\begin{aligned}
U_0 &= hf\,(x_0, y_0, z_0)\,,\\
V_0 &= hg\,(x_0, y_0, z_0)\,,\\
U_1 &= hf\,(x_0 + h/2, y_0 + U_0/2, z_0)\,,\\
V_1 &= hg\,(x_0 + h/2, y_0, z_0 + V_0/2)\,,\\
U_2 &= hf\,(x_0 + h/2, y_0 + U_1/2, z_0)\,,\\
V_2 &= hg\,(x_0 + h/2, y_0, z_0 - V_1/2)\,,\\
U_3 &= hf\,(x_0 + h, y_0 + U_2, z_0)\,,\\
V_3 &= hg\,(x_0 + h, y_0, z_0 + V_2)\,,
\end{aligned}
$$

$$
\text{et}\quad y(x_0+h) = y(x_0) + \frac{U_0}{6} + \frac{U_1}{3} + \frac{U_2}{3} + \frac{U_3}{6},
$$

$$
z(x_0+h) = z(x_0) + \frac{V_0}{6} + \frac{V_1}{3} + \frac{V_2}{3} + \frac{V_3}{6}.
$$

### e – Méthode d'Adams-Bashforth dans le cas d'un pas constant

$$
y_{n+1} = y_n + h\sum_{j=0}^{\nu} b_{\nu j} f_{n-j} \qquad \text{avec } n \geq \nu,
$$

et

$$
f_{n+1} = f(x_{n+1}, y_{n+1}, z_{n+1}),
$$

puis

$$z_{n+1} = z_n + h\sum_{j=0}^{\nu} b_{\nu j} g_{n-j} \qquad \text{avec } n \geq \nu,$$

et

$$g_{n+1} = g(x_{n+1}, y_{n+1}, z_{n+1}),$$

Sur le Web(*) figure le programme `pendul_0.c` qui met en œuvre cette technique.

### *f – Méthode d'Adams-Moulton dans le cas d'un pas constant*

$$y_{n+1} - hd_{\nu j-1} f(x_{n+1}, y_{n+1}, z_{n+1}) = y_n + \sum_{j=0}^{\nu} hd_{\nu j} f_{n-j} \qquad \text{avec } n \geq \nu,$$

$$z_{n+1} - hd_{\nu j-1} g(x_{n+1}, y_{n+1}, z_{n+1}) = z_n + \sum_{j=0}^{\nu} hd_{\nu j} g_{n-j} \qquad \text{avec } n \geq \nu.$$

***Remarque*** : Ici, il y a à résoudre un système non linéaire de deux équations à deux inconnues dont les racines sont $y_{n+1}$ et $z_{n+1}$. Toutefois, si le prix du calcul est trop élevé, on pourra se contenter de l'algorithme suivant :

$$y_{n+1} - hd_{\nu j-1} f(x_{n+1}, y_{n+1}, z_n) = y_n + \sum_{j=0}^{\nu} hd_{\nu j} f_{n-j} \qquad \text{avec } n \geq \nu,$$

$$z_{n+1} - hd_{\nu j-1} g(x_{n+1}, y_{n+1}, z_{n+1}) = z_n + \sum_{j=0}^{\nu} hd_{\nu j} g_{n-j} \qquad \text{avec } n \geq \nu,$$

dans lequel on n'a plus qu'à rechercher la racine locale d'une première équation implicite qui donne $y_{n+1}$ puis la racine d'une seconde équation implicite qui permet de déterminer $z_{n+1}$.

### *g – Méthode des différences rétrogrades*

$$y_{n+1} - h\alpha_\nu f(x_{n+1}, y_{n+1}, z_{n+1}) = \sum_{j=1}^{\nu} a_{\nu j} y_{n+1-j},$$

$$z_{n+1} - h\alpha_\nu g(x_{n+1}, y_{n+1}, z_{n+1}) = \sum_{j=1}^{\nu} a_{\nu j} z_{n+1-j}.$$

Ici encore on retrouve le problème rencontré lors de l'utilisation de la méthode d'Adams-Moulton, on le traite de la même façon.

---

* `http://www.edpsciences.com/guilpin/`

### 8.2. Le problème aux limites

On peut rechercher une solution $y = \phi(x)$ et $z = \psi(x)$ telle que, par exemple, elle admette $y_0 = \phi(x_0)$ et $y_1 = \phi(x_1)$ ou encore $y_0 = \phi(x_0)$ et $z_1 = \psi(x_1)$, avec dans les deux cas $x_0 \neq x_1$.

Il n'y a plus de théorème d'unicité analogue à celui de Cauchy-Lipschitz. Le problème est plus délicat, et pour s'en convaincre, il suffit de considérer l'équation du pendule. On se donne $x_0$ et $y_0$ à $t_0$. Il est bien clair que l'on ne peut pas imposer une seconde condition n'importe comment : elle doit être compatible avec les liaisons du système étudié. Lorsque ce problème de conditions aux limites se présente, on peut l'aborder de la façon suivante :

On fixe une condition, par exemple $y_0 = \phi(x_0)$ et l'on choisit arbitrairement l'autre condition à la même limite pour $x_0$, soit $z_0$. La technique consiste à ajuster $z_0$ de telle sorte qu'à la limite $x_1$ la seconde condition imposée soit remplie. L'ajustement de $z_0$ s'effectue par tâtonnements en essayant de trouver deux valeurs $z_a$ et $z_b$ (qui encadrent la valeur cherchée $z_0$) à partir desquelles on recherche $z_0$ par dichotomie, d'où son nom : **la méthode de tir**.

## 9. Équations différentielles d'ordre supérieur à deux

Ce qui a été dit pour le deuxième ordre peut être généralisé sans grande difficulté pour les ordres supérieurs et les algorithmes se transposent de la même manière que celle qui a été présentée pour le deuxième ordre.

## 10. Éléments de bibliographie

M. AINSWORTH (1995) *Theory and numerics of ordinary and partial equations*, Clarendon Press.

M. BACH (1998) *Analysis, numerics and applications of differential and integral equations*, Longman.

J. BARANGER (1991) *Analyse numérique*, Éditions Hermann.

M. CROUZEIX et A.L. MIGNOT (1984) *Analyse numérique des équations différentielles*, Éditions Masson.

J.P. DEMAILLY (1991) *Analyse numérique et équations différentielles*, Presses Universitaires de Grenoble.

J. FAVARD (1962) *Cours d'Analyse de l'École Polytechnique*, Tome III : Théorie des équations, fascicule 1 : Équations différentielles, Éditions Gauthier-Villars.

P. HENRICI (1964) *Elements of Numerical Analysis*, Wiley.

F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.

D. MC CRACKEN et W. DORN (1964) *Numerical Methods and Fortran Programming*, Wiley.

H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

Op.J. PRINCE et J.R. DORMAND (1981) High order embedded Runge-Kutta formulæ, *Journal of Computation and Applied Mathematics*, volume 7, n° 1.

# 14 Intégration des équations aux dérivées partielles

Nous nous proposons d'étudier quelques méthodes d'intégration d'équations aux dérivées partielles du premier et du deuxième ordre dans le champ réel. Ce problème généralise l'intégration des équations différentielles ordinaires lesquelles ne dépendent que d'une seule variable ; à partir du moment où nous sommes en présence de plusieurs variables, dans les équations, figurent, presque de façon inéluctable, des dérivées partielles.

Cette remarque induit une conséquence immédiate : les méthodes que nous allons développer peuvent également s'appliquer à la résolution des équations différentielles ordinaires.

Hormis quelques techniques très pointues attachées à des problèmes d'espèce et qui sortent du cadre de cet enseignement, on peut dire que, grosso modo, les méthodes utilisées ont toutes en dénominateur commun la réalisation d'un maillage se superposant au domaine $D$ sur lequel on désire effectuer l'intégration numérique. C'est aux nœuds de ce maillage que l'on calcule les valeurs de la fonction faisant l'objet de l'équation aux dérivées partielles. On voit bien qu'il s'agit de la généralisation de la notion de pas, celui-ci pouvant être constant ou variable selon les besoins.

Le premier problème qui se pose concerne les domaines ayant un point à l'infini, tandis que le second concerne les techniques à pas variables. Il ne faut voir là que des problèmes liés à la géométrie du domaine ainsi qu'à la nature de l'équation aux dérivées partielles. Ce sont les raisons qui expliquent qu'il existe deux types de méthodes numériques pour intégrer les équations aux dérivées partielles :

**a.** les méthodes aux différences finies,
**b.** les méthodes aux éléments finis.

Fondamentalement, ce sont les mêmes techniques qui utilisent les mêmes approximations des opérateurs différentiels, et ce qui les différencie repose sur la technique de maillage du domaine considéré. Les méthodes aux différences finies exploitent un maillage à pas constants (quel que soit le type de coordonnées utilisées) et proposent deux types de résolution : l'une est explicite, c'est-à-dire que les inconnues aux nœuds du maillage sont données explicitement par les équations, l'autre est implicite, alors les inconnues constituent un système linéaire qu'il convient d'inverser. Dans ce dernier cas, on a affaire à une matrice « creuse », c'est-à-dire ne comportant que des zéros hormis la diagonale et son voisinage, il existe alors des méthodes d'inversion spécifiques de ce type de matrices (*cf.* le chapitre 5 consacré au calcul matriciel).

Les méthodes aux éléments finis proposent un maillage adapté au type de problème considéré : là où la fonction inconnue varie beaucoup, le maillage est serré, là où la fonction inconnue varie

peu le maillage est très grand. Cela conduit à une sorte d'optimisation du temps de calcul et de la quantité de calcul, mais, en revanche, conduit à des études préalables longues et coûteuses qui restent en général du domaine du spécialiste. C'est précisément la méthode retenue dans l'étude de la résistance des matériaux ; on note à ce propos que l'on a affaire à des domaines de même type.

## 1. Considérations sur les équations aux dérivées partielles d'ordre au plus égal à deux

Dans un repère cartésien orthonormé tridimensionnel de coordonnées $(x, y, z)$, l'équation aux dérivées partielles du deuxième ordre la plus générale à laquelle obéit la fonction $\Phi$ s'écrit de la façon suivante :

$$f\left(\Phi, x, y, z, \frac{\partial\Phi}{\partial x}, \frac{\partial\Phi}{\partial y}, \frac{\partial\Phi}{\partial z}, \frac{\partial^2\Phi}{\partial x^2}, \frac{\partial^2\Phi}{\partial y^2}, \frac{\partial^2\Phi}{\partial z^2}, \frac{\partial^2\Phi}{\partial x\partial y}, \frac{\partial^2\Phi}{\partial x\partial z}, \frac{\partial^2\Phi}{\partial y\partial z}\right) = 0.$$

Ici encore, les cas offerts par la physique sont généralement des formes explicites, et c'est plutôt l'étude des formes linéaires qui va retenir notre attention. Nous limitons toutefois notre étude aux **équations linéaires** à deux variables indépendantes $x$ et $y$ (ou $t$), ainsi l'équation devient :

$$A(x,y)\frac{\partial^2\Phi}{\partial x^2} + B(x,y)\frac{\partial^2\Phi}{\partial x\partial y} + C(x,y)\frac{\partial^2\Phi}{\partial y^2} + D = 0,$$

où $D$ est une fonction pouvant dépendre linéairement de $\Phi$, $\frac{\partial\Phi}{\partial x}$ et $\frac{\partial\Phi}{\partial y}$ uniquement.

L'analyse de ce cas particulier permet de distinguer trois types d'équations aux dérivées partielles linéaires du deuxième ordre selon les racines de l'équation caractéristique (théorie de Monge (1746–1818) - Ampère (1775–1836)) :

1. $B^2(x,y) - 4A(x,y)C(x,y) < 0$ : c'est le type elliptique, il s'agit de l'équation de Poisson (1781–1840) qui admet comme cas particulier l'équation de Laplace,
2. $B^2(x,y) - 4A(x,y)C(x,y) = 0$ : c'est le type parabolique, le prototype est fourni par l'équation de la propagation de la chaleur ou encore de la diffusion,
3. $B^2(x,y) - 4A(x,y)C(x,y) > 0$ : c'est le type hyperbolique dont l'exemple est l'équation de propagation des ondes (cordes vibrantes par exemple).

La recherche d'une solution numérique à l'intérieur d'un domaine $D$ passe par la connaissance de certains renseignements concernant la fonction $\Phi$, il s'agit des conditions aux limites (éventuellement à un instant $t$) : ce sont les conditions sur la frontière $F$ du domaine $D$. Parmi les différentes manières d'imposer les conditions aux limites, nous évoquerons les deux plus importantes :

**a. Les conditions de Dirichlet** (1805–1859). En tout point de la frontière $F$, nous connaissons la valeur de la fonction $\Phi$, c'est aussi le problème de Cauchy.

**b. Les conditions de von Neumann** (1832–1925). En tout point de la frontière $F$ du domaine, nous connaissons la valeur de la dérivée normale de la fonction $\Phi$. Généralement ce type de conditions conduit à utiliser une méthode d'intégration implicite.

***Remarque 1*** : Les conditions aux limites peuvent dépendre du temps. Par exemple, il est possible d'étudier la propagation d'une perturbation thermique sinusoïdale dans un milieu matériel.

***Remarque 2*** : L'obtention de solutions générales aux différents types de problèmes présentés dépend de la nature de l'équation aux dérivées partielles, mais c'est surtout la forme géométrique de la frontière (qui possède ou non des propriétés de symétrie) qui permet de savoir si l'on doit s'orienter vers la recherche de solutions formelles ou vers la recherche de solutions numériques.

## 2. Les opérateurs de différence

Au domaine $D$ on superpose donc un maillage le plus adapté possible à la forme géométrique de $D$. Quelle que soit la méthode utilisée, il faut pouvoir exprimer les opérateurs différentiels en fonction des différences fournies par le choix discret des points (nœuds du maillage). Ici, notre propos se limite aux opérateurs **différence première et différence seconde**, cependant, le cas échéant, la généralisation peut se faire sans difficulté, mais il est rare de rencontrer des dérivées partielles au-delà de deux.

Dans un repère cartésien orthonormé à deux dimensions $(x, y)$, on considère une fonction « bien élevée » $\Phi(x, y)$ qui possède les bonnes propriétés de régularité (continuité, dérivabilité au moins à l'ordre trois, etc.). Soient trois points tous distincts $A(x_a, y_a)$, $B(x_b, y_b)$ et $C(x_c, y_c)$ pour lesquels nous avons : $x_a < x_b < x_c$, et $y_a < y_b < y_c$. La dérivée partielle en $B$ par rapport à $x$ est approchée par l'opérateur différence première ainsi défini :

$$\left.\frac{\partial \Phi(x, y_b)}{\partial x}\right|_{x_b} = \left.\frac{\delta \Phi(x, y_b)}{\delta x}\right|_{x_b} \approx \frac{\Phi(x_b, y_b) - \Phi(x_a, y_b)}{x_b - x_a} \approx \frac{\Phi(x_c, y_b) - \Phi(x_b, y_b)}{x_c - x_b}\,.$$

La dérivée partielle en $B$ par rapport à $y$ est donnée par l'expression :

$$\left.\frac{\partial \Phi(x_b, y_b)}{\partial x}\right|_{y_b} = \left.\frac{\delta \Phi(x_b, y)}{\delta y}\right|_{y_b} \approx \frac{\Phi(x_b, y_b) - \Phi(x_b, y_a)}{y_b - y_a} \approx \frac{\Phi(x_b, y_c) - \Phi(x_b, y_b)}{y_c - y_b}\,.$$

La dérivée partielle d'ordre deux par rapport à $x$, calculée en $B$, est approchée par l'opérateur différence seconde ainsi défini :

$$\begin{aligned}\left.\frac{\partial^2 \Phi(x, y_b)}{\partial x^2}\right|_{x_b} &= \frac{\delta}{\delta x}\left(\left.\frac{\delta \Phi(x, y_b)}{\delta x}\right)\right|_{x_b} \\ &\approx \left(\frac{\Phi(x_c, y_b) - \Phi(x_b, y_b)}{x_c - x_b} - \frac{\Phi(x_b, y_b) - \Phi(x_a, y_b)}{x_b - x_a}\right)\frac{1}{x_c - x_b},\end{aligned}$$

la dérivée partielle d'ordre deux par rapport à $y$ s'écrit :

$$\begin{aligned}\left.\frac{\partial^2 \Phi(x_b, y)}{\partial y^2}\right|_{y_b} &= \frac{\delta}{\delta y}\left(\left.\frac{\delta \Phi(x_b, y)}{\delta y}\right)\right|_{y_b} \\ &\approx \left(\frac{\Phi(x_b, y_c) - \Phi(x_b, y_b)}{y_c - y_b} - \frac{\Phi(x_b, y_b) - \Phi(x_b, y_a)}{y_b - y_a)}\right)\frac{1}{y_c - y_b}\,.\end{aligned}$$

Il n'est pas très difficile de voir que, compte tenu des bonnes propriétés de la fonction $\Phi$, les opérateurs de différence tendent vers les opérateurs différentiels lorsque $x_b - x_a$ tend vers zéro (ainsi que $x_c - x_b$) ou encore lorsque $y_b - y_a$ tend vers zéro (ainsi que $y_c - y_b$).

Dans le cas particulier où le maillage est régulier (points en progression arithmétique), on obtient des expressions plus simples en désignant par $\delta x$ et $\delta y$ les raisons des progressions arithmétiques :

$$\begin{aligned}
\frac{\partial\Phi(x,y_b)}{\partial x}\Big|_{x_b} &\approx \frac{\Phi(x_c,y_b)-\Phi(x_b,y_b)}{\delta x},\\
\frac{\partial\Phi(x_b),y_b)}{\partial x}\Big|_{y_b} &\approx \frac{\Phi(x_b,y_c)-\Phi(x_b,y_b)}{\delta y},\\
\frac{\partial^2\Phi(x,y_b)}{\partial x^2}\Big|_{x_b} &\approx \frac{\Phi(x_c,y_b)-2\Phi(x_b,y_b)+\Phi(x_a,y_b)}{\delta x^2},\\
\frac{\partial^2\Phi(x_b,y)}{\partial y^2}\Big|_{y_b} &\approx \frac{\Phi(x_b,y_c)-2\Phi(x_b,y_b)+\Phi(x_b,y_a)}{\delta y^2}.
\end{aligned}$$

## Valeur de l'opérateur de différence sur les bords du domaine

Il n'est pas toujours simple de faire coïncider les nœuds du maillage avec les bords du domaine quand bien même aurions-nous choisi un système de coordonnées correctement adapté au problème ; alors nous sommes amené à calculer la valeur des opérateurs de différence sur les points qui sont à l'intersection du réseau formant le maillage et la ligne constituant le bord du domaine. Voyons ce qui se passe à deux dimensions (*cf.* Fig. 14.1).

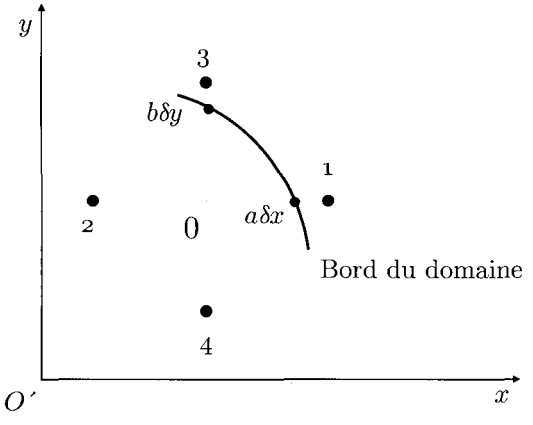


**Figure 14.1.** *Valeur de l'opérateur de différence sur les bords du domaine.*

Nous pouvons écrire :

$$\begin{aligned}
\frac{\delta\Phi}{\delta x}\Big|_{01} &= \frac{\Phi_1-\Phi_0}{a\delta x}, \qquad \frac{\delta\Phi}{\delta x}\Big|_{20} = \frac{\Phi_0-\Phi_2}{\delta x},\\
\frac{\delta\Phi}{\delta y}\Big|_{03} &= \frac{\Phi_3-\Phi_0}{b\delta y}, \qquad \frac{\delta\Phi}{\delta x}\Big|_{40} = \frac{\Phi_0-\Phi_4}{\delta y}.
\end{aligned}$$

On en déduit que :

$$\begin{aligned}
\frac{\delta^2\Phi}{\delta x^2}\Big|_0 &= \frac{2}{\delta x(1+a)}\left(\frac{\Phi_1-\Phi_0}{a\delta x}+\frac{\Phi_2-\Phi_0}{\delta x}\right),\\
\frac{\delta^2\Phi}{\delta y^2}\Big|_0 &= \frac{2}{\delta y(1+b)}\left(\frac{\Phi_3-\Phi_0}{b\delta y}+\frac{\Phi_4-\Phi_0}{\delta y}\right),
\end{aligned}$$

en généralisant sur la troisième dimension :

$$\left.\frac{\delta^2\Phi}{\delta z^2}\right|_0 = \frac{2}{\delta z(1+c)}\left(\frac{\Phi_5-\Phi_0}{c\delta z}+\frac{\Phi_6-\Phi_0}{\delta z}\right).$$

## 3. L'opérateur laplacien

Compte tenu de son importance, il mérite qu'on s'arrête quelque peu sur lui et notamment que l'on donne son expression en fonction des opérateurs de différence dans le repère cartésien ainsi que dans les systèmes curvilignes orthogonaux classiques.

### 3.1. Repère cartésien orthonormé

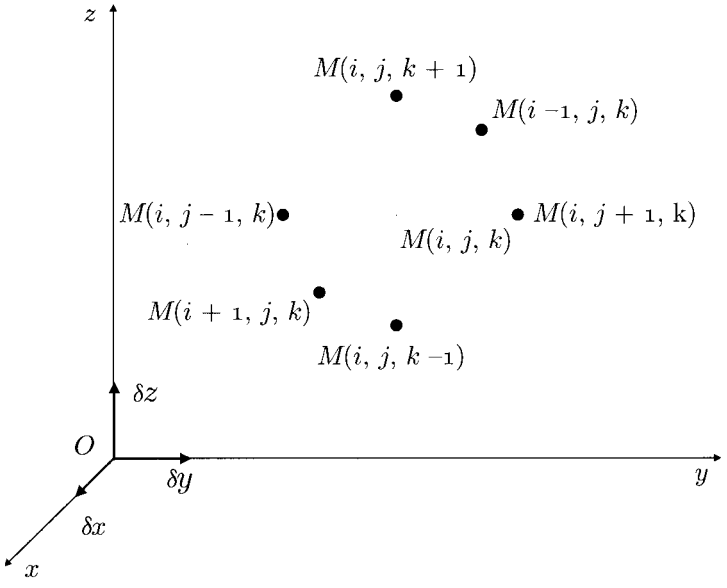


**Figure 14.2.** *Repère cartésien orthonormé.*

Désignons par $\delta x$, $\delta y$ et $\delta z$ les dimensions de la maille élémentaire, l'orientation du maillage à trois dimensions étant définie conformément à la figure 14.2. Il vient immédiatement que :

$$\begin{aligned}\Delta\Phi(x,y,z)_{i,j,k} = {} & \frac{1}{\delta x^2}\left[\Phi(i+1,j,k)+\Phi(i-1,j,k)-2\Phi(i,j,k)\right]\\ & +\frac{1}{\delta y^2}\left[\Phi(i,j+1,k)+\Phi(i,j-1,k)-2\Phi(i,j,k)\right]\\ & +\frac{1}{\delta z^2}\left[\Phi(i,j,k+1)+\Phi(i,j,k-1)-2\Phi(i,j,k)\right].\end{aligned}$$

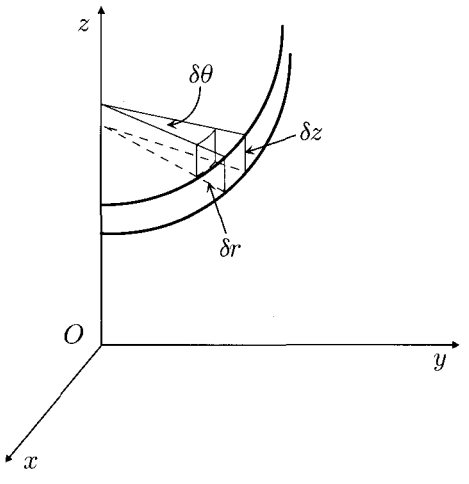


**Figure 14.3.** *Coordonnées cylindriques.*

### 3.2. Coordonnées cylindriques (r, $\theta$, z)

Désignons par $\delta r$, $r\delta\theta$ et $\delta z$ les dimensions de la maille élémentaire (*cf.* Fig. 14.3).

$$\Delta\Phi_{r_i,\theta_j,z_k} = \frac{1}{r_i}\left[\frac{2r_i+\delta r}{2\delta r^2}\left[\Phi(i+1,j,k)-\Phi(i,j,k)\right]+\frac{2r_i-\delta r}{2\delta r^2}\left[\Phi(i-1,j,k)-\Phi(i,j,k)\right]\right.$$
$$+\frac{1}{r_i\delta\theta^2}\left[\Phi(i,j+1,k)+\Phi(i,j-1,k)-2\Phi(i,j,k)\right]$$
$$\left.+\frac{r_i}{\delta z^2}\left[\Phi(i,j,k+1)+\Phi(i,j,k-1)-\Phi(i,j,k)\right]\right].$$

### 3.3. Coordonnées sphériques (r, $\theta$, $\phi$)

Désignons par $\delta r$, $r\delta\theta$ et $r\delta\phi$ les dimensions de la maille élémentaire (*cf.* Fig. 14.4, page ci-contre).

$$\Delta\Phi_{r_i,\theta_j,\phi_k} = \frac{1}{r_i^2}\left[\frac{(2r_i+\delta r/2)^2\left[\Phi(i+1,j,k)-\Phi(i,j,k)\right]+(2r_i-\delta r/2)^2\left[\Phi(i-1,j,k)-\Phi(i,j,k)\right]}{\delta r^2}\right.$$
$$+\frac{\Phi(i,j,k+1)+\Phi(i,j,k-1)-2\Phi(i,j,k)}{\sin^2(\theta_j)\delta\phi^2}+\frac{\sin(\theta_j+\delta\theta/2)\left[\Phi(i,j+1,k)-\Phi(i,j,k)\right]}{\sin(\theta_j)\delta\theta^2}$$
$$\left.+\frac{\sin(\theta_j-\delta\theta/2)\left[\Phi(i,j-1,k)-\Phi(i,j,k)\right]}{\sin(\theta_j)\delta\theta^2}\right].$$

## 4. Résolution des équations de type elliptique

On se propose de calculer les valeurs d'une fonction $\Phi$ obéissant à une équation de Poisson :

$$\Delta\Phi(\vec{\mathrm{r}}) = \rho(\vec{\mathrm{r}}),$$

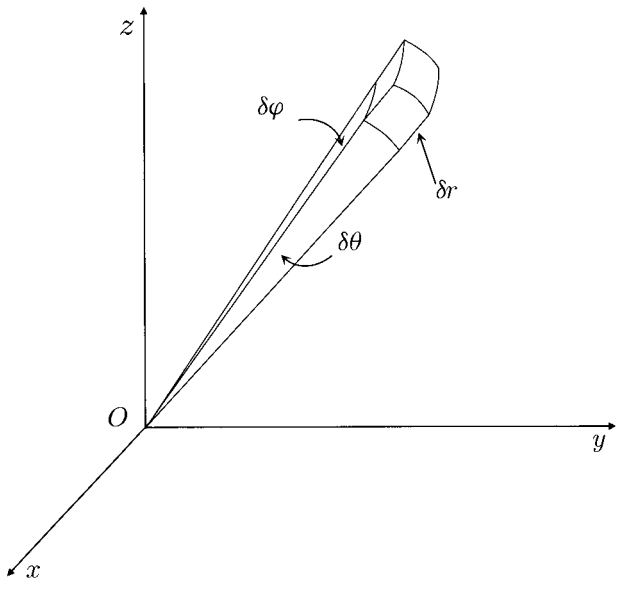


**Figure 14.4.** *Coordonnées sphériques.*

avec $\vec{r} = \overrightarrow{OM}$, $M$ appartenant au domaine $D$, connaissant les valeurs que $\Phi$ prend sur la frontière $F$.

Dans de nombreux cas $\rho(\vec{r}) = 0$, et nous devons résoudre une équation de Laplace :

$$\Delta\Phi(\vec{r}) = 0.$$

Les fonctions qui obéissent à cette équation portent un nom, on les appelle **fonctions harmoniques**. La recherche de ces solutions dans le cas des symétries usuelles s'effectue sous la forme de produits de Laplace et elle aboutit à un bon nombre de **fonctions dites spéciales**, autrement dit celles qui jouent un rôle important en analyse : sinus, cosinus, fonctions de Legendre associées, fonctions de Weber-Hermite, de Bessel, de Neumann, de Mathieu etc.

Dans le cas de l'équation de Poisson, nous connaissons une solution particulière :

$$\Phi(r) = -\frac{1}{4\pi}\int_D \frac{\rho(\vec{r})}{r}\,\mathrm{d}\tau,$$

$\mathrm{d}\tau$ étant l'élément de volume. Il suffit d'ajouter cette solution aux solutions de l'équation de Laplace (équation sans second membre) pour obtenir la solution complète (équation linéaire).

Malheureusement, lorsque le problème étudié n'admet pas de symétries simples, les solutions formelles deviennent exceptionnelles et pratiquement dans tous les cas il faut avoir recours au calcul numérique.

## 4.1. Conduite du calcul permettant la résolution numérique de l'équation de Poisson

Par raison de simplicité, nous étudierons le problème dans un repère cartésien orthonormé et nous choisirons $\delta x = \delta y = \delta z = h$.

Maintenant, exprimons $\Phi(i,j,k)$ au point $M(i,j,k)$ en fonction des valeurs prises aux points plus proches voisins de $M(i,j,k)$ :

$$\Phi(i,j,k) = \frac{1}{6}\big[\Phi(i+1,j,k) + \Phi(i-1,j,k) + \Phi(i,j+1,k) + \Phi(i,j-1,k) + \Phi(i,j,k+1) + \Phi(i,j,k-1) - h^2\rho(i,j,k)\big].$$

## 4.2. Méthode de Jacobi (1804–1851)

C'est une méthode itérative qui consiste à affecter aux nœuds du maillage une valeur arbitraire de l'ordre de grandeur de celles qui se situent sur la frontière. La dernière relation établie permet alors de calculer tous les points adjacents à la frontière, lesquels sont les seuls à être modifiés. La réitération de l'exploitation de la formule permet de proche en proche de modifier tous les points du maillage. On montre que la méthode est convergente. Au $n^{\text{e}}$ tour d'itération, on écrit les relations qui font appel aux calculs effectués au $(n-1)^{\text{e}}$ :

$$\Phi^{(n)}(i,j,k) = \frac{1}{6}\big[\Phi^{(n-1)}(i+1,j,k) + \Phi^{(n-1)}(i-1,j,k) + \Phi^{(n-1)}(i,j+1,k) + \Phi^{(n-1)}(i,j-1,k) + \Phi^{(n-1)}(i,j,k+1) + \Phi^{(n-1)}(i,j,k-1) - h^2\rho(i,j,k)\big].$$

D'un point de vue pratique, il faut deux tableaux pour mettre en œuvre cet algorithme. De toute façon, il n'est pas très utile de programmer cette procédure car la méthode de Gauss-Seidel en est une importante amélioration.

## 4.3. Méthode de Gauss-Seidel (1821—1896)

On peut montrer que la technique itérative de Jacobi est convergente et absolument convergente. Gauss et Seidel ont profité de cette propriété pour améliorer la méthode de Jacobi : au cours du calcul, on peut faire immédiatement usage des valeurs qui viennent d'être calculées et nous n'avons plus besoin de sauvegarder les différents tours d'itération. Nous n'avons plus besoin que d'un seul tableau dans lequel les valeurs sont « écrasées » successivement. Cette méthode exige environ deux fois moins de calculs que la méthode de Jacobi car elle converge beaucoup plus rapidement.

## 4.4. Application

On se propose d'étudier les équipotentielles d'une lampe triode que nous présentons assez simplifiée. Une lampe triode est constituée d'une cathode $K$, d'une anode $A$ et d'une grille $G_1$, $G_2$, $G_3$, $G_4$, $G_5$, $G_6$. La figure 14.5, page ci-contre montre la coupe de cette lampe simplifiée.

L'anode $A$ est un cylindre métallique de rayon 1 cm porté au potentiel 100 volts par rapport à la cathode qui se trouve au potentiel zéro. La cathode est un simple fil métallique de rayon négligeable placé sur l'axe du cylindre de l'anode. La grille est constituée de six fils parallèles à la cathode disposés régulièrement sur le cylindre de rayon $r = 0{,}2$ cm, ces six fils étant maintenus au potentiel $-5$ volts par rapport à la cathode. On se propose de déterminer la carte du potentiel dans un plan perpendiculaire à l'axe de la triode.

Pour ce faire, nous ferons usage des coordonnées polaires à cause de la symétrie cylindrique. Comme le problème admet une symétrie d'ordre six, nous n'avons donc besoin pour effectuer les calculs que de faire varier l'angle $\theta$ dans l'intervalle $(0, \pi/3)$ et nous compléterons ensuite par symétrie.

Nous choisissons le maillage suivant : $\delta\theta = \pi/36$ et $\delta r = 0{,}1$ cm (*cf.* Fig. 14.6, page ci-contre).

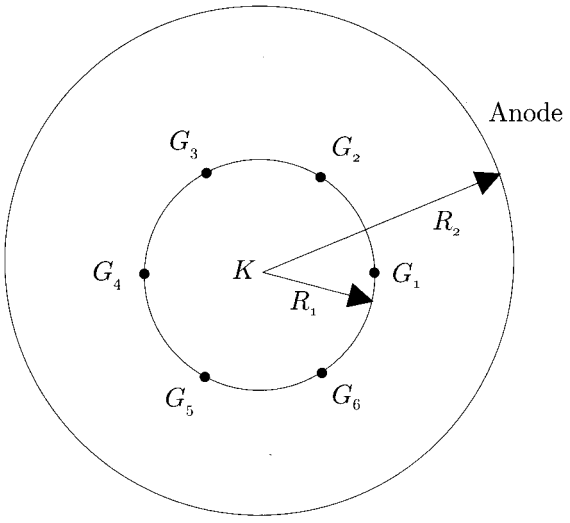


**Figure 14.5.** *Représentation schématique d'une lampe triode.*

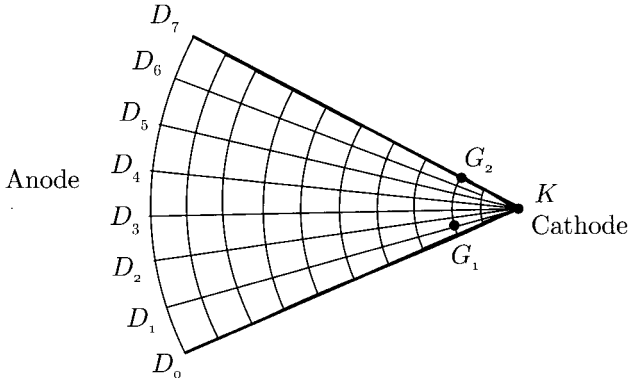


**Figure 14.6.** *Maillage d'une partie du domaine de la triode.*

## 4.5. Technique opératoire

On met des zéros partout aux nœuds du maillage hormis aux points $A$, $G_1$ et $K$. Nous calculons le potentiel sur les droites $D_1$, $D_2$, $D_3$ et $D_4$. Nous complétons par symétrie car sur la droite $D_5$ le potentiel est le même que sur la droite $D_3$, et celui de la droite $D_0$ est le même que celui de la droite $D_2$. Nous avons donc besoin d'un tableau contenant 88 valeurs et nous arrêtons les calculs quand deux valeurs calculées au même point au centre du maillage, lors de deux itérations consécutives, sont égales à une certaine précision près.

On trouvera sur le Web(*) le programme `triode.c` qui réalise le calcul des potentiels par la méthode de Gauss-Seidel.

* `http://www.edpsciences.com/guilpin/`

# 5. Résolution des équations de type parabolique (méthode explicite)

Nous recherchons une fonction $\Phi$ qui prend certaines valeurs sur le bord d'un domaine et qui obéit à l'équation suivante :

$$\Delta\Phi = \frac{1}{D}\frac{\partial\Phi}{\partial t},$$

expression dans laquelle $D$ est une constante appelée coefficient de diffusion thermique. Ici la fonction dépend des trois variables d'espace $x$, $y$, $z$ et de la variable de temps $t$. Pour étudier l'intégration de cette équation, nous ne conserverons qu'une seule coordonnée d'espace, $x$ par exemple, sans pour autant nuire à la généralité. L'équation se réduit :

$$\frac{\partial^2\Phi(x,t)}{\partial x^2} = \frac{1}{D}\frac{\partial\Phi}{\partial t}.$$

## 5.1. Maillage pour une coordonnée d'espace et la coordonnée de temps

Comme précédemment, nous choisissons un maillage régulier, et la maille élémentaire a la taille $\delta x \delta t$. Nous pouvons écrire :

$$\left.\frac{\partial^2\Phi(x,t)}{\partial x^2}\right|_{x,t} \approx \frac{\Phi(x+\delta x,t)+\Phi(x-\delta x,t)-2\Phi(x,t)}{\delta x^2},$$

$$\left.\frac{\partial\Phi(x,t)}{\partial t}\right|_{x,t} \approx \frac{\Phi(x,t+\delta t)-\Phi(x,t)}{\delta t}.$$

Au moyen des différences finies, l'expression de l'équation devient :

$$\Phi(x+\delta x,t)+\Phi(x-\delta x,t)-2\Phi(x,t) = \frac{\delta x^2}{D\delta t}\left[\Phi(x,t+\delta t)-\Phi(x,t)\right].$$

## 5.2. Conduite du calcul pour obtenir la résolution approchée de l'équation de diffusion

Les valeurs en chaque point du maillage au temps $t+\delta t$ ne dépendent que des valeurs calculées à l'instant $t$. La solution du problème est déterminée par une suite d'opérations élémentaires effectuées à partir des conditions initiales, et il est facile de s'apercevoir que cette méthode n'est pas itérative (les valeurs changent au cours du temps) :

$$\Phi(x,t+\delta t) = \Phi(x,t) + \frac{D\delta t}{\delta x^2}\left[\Phi(x+\delta x,t)+\Phi(x-\delta x,t)-2\Phi(x,t)\right].$$

Au temps $t=0$, nous connaissons la valeur de $\Phi$ en tout point du maillage. Au temps $\delta t$, nous calculons au moyen de la relation ci-dessus les valeurs en chaque point du maillage, en tenant compte des conditions imposées au système étudié (conditions aux limites). Puis nous effectuons les calculs au temps $2\delta t$ par le même procédé, et ainsi de suite au temps $3\delta t, \ldots, n\delta t, \ldots$ Nous obtenons alors l'évolution du système au cours du temps.

## 5.3. Problème de stabilité

Nous ne pouvons pas choisir n'importe comment le quadrillage de l'espace $(x,t)$ et le calcul proposé est stable lorsque la condition suivante est remplie :

$$\delta t \leq \frac{\delta x^2}{\alpha D},$$

$\alpha$ valant 2, 4 ou 6 selon qu'il y a une, deux ou trois coordonnées d'espace.

Le non-respect de la condition de stabilité est sanctionné immédiatement : on obtient très rapidement n'importe quoi : on aboutit à un dépassement de capacité (overflow) qui produit un arrêt de la machine.

### 5.4. Application : Refroidissement d'une plaque homogène

Une plaque homogène d'épaisseur $l = 10$ cm est portée à la température uniforme $\theta_0 = 90\,°\text{C}$. Au temps $t = 0$, elle est plongée dans un réservoir d'eau à la température $\theta_1 = 10\,°\text{C}$ suffisamment grand et bien agité pour que l'on puisse admettre que l'énergie cédée par la plaque ne change pas de façon notable la température du réservoir. On se propose de déterminer la température à l'intérieur de la plaque en fonction de la distance $x$ et du temps $t$ (on notera la symétrie du problème). La plaque est en cuivre dont le coefficient de diffusion est $D = 1{,}12\ \text{cm}^2\text{s}^{-1}$. Au bout de combien de temps pouvons-nous admettre que la température est à nouveau uniforme dans la plaque, c'est-à-dire lorsque la température la plus élevée ne dépasse que de 0,1 °C la température du thermostat ?

Il est tout à fait intéressant de comparer les résultats du calcul numérique avec les résultats du calcul formel. C'est ce genre de problème que Fourier a traité et pour lequel il a trouvé la solution sous forme d'un développement en série de sinus et de cosinus qui, aujourd'hui, porte son nom (Théorie analytique de la chaleur, 1822).

On trouvera sur le Web (*) le programme `chaleur.c` qui traite ce problème.

## 6. Résolution des équations de type hyperbolique (méthode explicite)

Nous recherchons une fonction $\Phi$ qui prend certaines valeurs sur le bord d'un domaine et qui obéit à l'équation suivante :

$$\Delta\Phi = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2},$$

expression dans laquelle $c$ est une constante appelée célérité de l'onde libre dans le milieu considéré (vitesse de phase).

La fonction dépend des trois variables d'espace $x$, $y$, $z$ et de la variable de temps $t$. Comme précédemment, pour étudier l'intégration de cette équation, nous ne conservons que la seule coordonnée d'espace $x$. L'équation se réduit alors à l'expression :

$$\frac{\partial^2\Phi(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2}\,.$$

### 6.1. Maillage pour une coordonnée d'espace et la coordonnée de temps

Comme précédemment, nous choisissons un maillage régulier, et la maille élémentaire a la taille $\delta x \delta t$. Nous pouvons écrire :

$$\left.\frac{\partial^2\Phi(x,t)}{\partial x^2}\right|_{x,t} \approx \frac{\Phi(x+\delta x,t)+\Phi(x-\delta x,t)-2\Phi(x,t)}{\delta x^2},$$

$$\left.\frac{\partial^2\Phi(x,t)}{\partial t^2}\right|_{x,t} \approx \frac{\Phi(x,t+\delta t)+\Phi(x,t-\delta t)-2\Phi(x,t)}{\delta t^2}\,.$$

Au moyen des différences finies, l'expression de l'équation devient :

$$\Phi(x+\delta x,t)+\Phi(x-\delta x,t)-2\Phi(x,t) = \frac{\delta x^2}{c^2\delta t^2}\left[\Phi(x,t+\delta t)+\Phi(x,t-\delta t)-2\Phi(x,t)\right].$$

* `http://www.edpsciences.com/guilpin/`

## 6.2. Conduite du calcul pour obtenir la résolution numérique de l'équation de propagation

Les valeurs en chaque point du maillage au temps $t+\delta t$ ne dépendent que des valeurs calculées à l'instant $t$ et à l'instant $t-\delta t$. La solution du problème est déterminée par une suite d'opérations élémentaires effectuées à partir des conditions initiales, et l'on voit ici encore que cette méthode n'est pas itérative :

$$\Phi(x,t+\delta t) = r^2\Phi(x+\delta x,t) + r^2\Phi(x-\delta x,t) + 2(1-r^2)\Phi(x,t) - \Phi(x,t-\delta t),$$

expression dans laquelle on a posé :

$$r^2 = \frac{\delta x^2}{c^2\delta t^2}.$$

Au temps $t \leq 0$, nous connaissons la valeur de $\Phi$ en tout point du maillage. Au temps $\delta t$, nous calculons au moyen de la relation ci-dessus les valeurs en chaque point du maillage, en tenant compte des conditions imposées au système étudié. Puis nous effectuons les calculs au temps $2\delta t$ au moyen de ce qui s'est passé au temps $t = 0$ et au temps $t = \delta t$, et ainsi de suite au temps $3\delta t, \ldots, n\delta t, \ldots$ Nous obtenons par ce moyen l'évolution du système au cours du temps.

## 6.3. Problème de stabilité

Ici encore, nous ne pouvons pas choisir n'importe comment le quadrillage de l'espace $(x,t)$ et l'algorithme proposé est stable lorsque la condition suivante est remplie :

$$\delta t^2 \leq \frac{\delta x^2}{\alpha c^2},$$

$\alpha$ valant 2, 4 ou 6 selon qu'il y a une, deux ou trois coordonnées d'espace.

Ici encore, le non-respect de la condition de stabilité provoque l'arrêt quasi immédiat de la machine.

***Application : Profil d'une corde vibrante pincée*** – Une corde vibrante $ABC$ est fixée à ces deux extrémités $B$ et $C$, et, à l'instant initial $t = 0$, cette corde est disposée conformément à la figure 14.7.

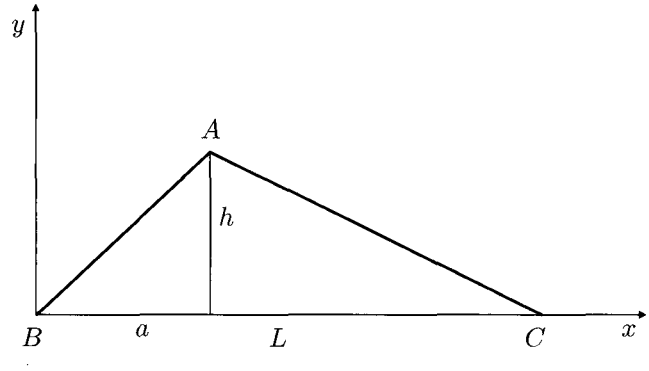


**Figure 14.7.** *Profil d'une corde pincée.*

Sachant que $L = 100$ cm, $h = 1$ cm, $c = 300\ \mathrm{m\,s^{-1}}$ et $a = 15$ cm, on se propose de déterminer les profils successifs de la corde sur une demi-période. On note que l'on doit retrouver un profil

symétrique au profil initial. On observera aussi la propagation des ébranlements de chaque côté du point $A$.

Sur le Web [*], on donne le programme `corde.c` qui réalise la simulation d'une corde pincée. Il est intéressant de comparer les résultats avec ceux fournis par la solution analytique.

## 7. Éléments de bibliographie


M. AINSWORTH (1995) *Theory and numerics of ordinary and partial equations*, Clarendon Press.

J. AMES (1972) *Nonlinear partial differential equations in engineering*, Volumes 1 and 2, Academic Press.

J. FAVARD (1962) *Cours d'Analyse de l'École Polytechnique*, Tome III : Théorie des équations, fascicule 1 : Équations différentielles, Gauthier-Villars.

J. GIRERD et W. KARPLUS (1968) *Traitement des équations différentielles sur calculateurs arithmétiques*, Gauthier-Villars.

P. HENRICI (1964) *Elements of Numerical Analysis*, Wiley.

F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.

D. MC CRACKEN et W. DORN (1964) *Numerical Methods and Fortran Programming*, Wiley.

H. MINEUR (1966) *Techniques de Calcul Numérique*, Éditions Dunod.

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

P.A. RAVIART et J.M. THOMAS (1988) *Introduction à l'analyse numérique des équations aux dérivées partielles*, Dunod.

A. TVEITO (1998) *Introduction to partial differential equations : a computational approach*, Springer.


* `http://www.edpsciences.com/guilpin/`

# 15 | Les séries de Fourier

Les séries de Fourier jouent un rôle important en calcul numérique, et on les retrouve lors des problèmes d'approximation des fonctions, de lissage, d'interpolation et aussi dans les problèmes de filtrage numérique. En toute rigueur le calcul pratique des séries de Fourier n'est pas à dissocier du calcul des transformées de Fourier, et nous aborderons l'étude du très puissant algorithme de Cooley-Tukey. On se bornera donc ici à étudier les fonctions périodiques de période $2\pi$ puisqu'un changement linéaire de variable permet toujours de se ramener à ce cas. Cependant on pourrait tout aussi bien s'intéresser aux fonctions périodiques de période 2, c'est-à-dire dans l'intervalle $(-1, +1)$, dans la mesure où l'on souhaiterait étudier en même temps les développements en série de Tchebycheff.

Au cours de ce chapitre, on se contentera de rappeler un certain nombre de théorèmes fondamentaux dont on omettra la démonstration.

## 1. Petit aperçu historique

L'origine des séries de Fourier remonte à 1822 quand est paru l'ouvrage de Jean-Baptiste-Joseph Fourier (1768–1830) sous le titre « Théorie Analytique de la Chaleur ». C'est en résolvant « le problème du mur » en régime transitoire que Fourier a fourni une expression formelle constituée d'une somme trigonométrique infinie. Il est intéressant de rappeler ici les grandes lignes de ce problème ainsi que sa solution. On recherche donc la solution de l'équation de la chaleur dans un milieu homogène et isotrope constitué par un mur indéfini d'épaisseur $h$. On désigne par $D$ le coefficient de diffusion thermique dans le milieu, par $\theta$ la température, par $x$ la seule variable spatiale et par $t$ la variable temporelle. Il convient alors de résoudre l'équation :

$$\frac{\partial^2 \theta}{\partial x^2} = \frac{1}{D} \frac{\partial \theta}{\partial t} \cdot \tag{15.1}$$

Posons $\theta(x,t) = X(x) \cdot T(t)$ afin de rechercher les solutions sous la forme de produits de Laplace. Nous obtenons :

$$DX''(x) \cdot T(t) = X(x) \cdot T'(t)$$

d'où l'on tire :

$$D \frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)} \cdot \tag{15.2}$$

Comme le premier membre de (15.2) est constitué d'une fonction ne dépendant que de $x$ et comme le second membre est constitué d'une fonction ne dépendant que de $t$, leur intersection commune ne peut être égale qu'à une constante que l'on désigne par $-k^2$. Comme on va le voir le signe moins résulte du fait physique que la température ne peut pas croître indéfiniment de façon spontanée. On obtient alors les équations :

$$X''(x) + k^2 X(x) = 0, \tag{15.3}$$

$$T'(t) + k^2 T(t) = 0. \tag{15.4}$$

Maintenant, pour déterminer complètement la solution il est nécessaire de faire intervenir les conditions initiales et aux limites.

Au temps $t < 0$, $T = T_i$ pour $0 \leq x \leq h$ et $T = T_0$ partout hors du mur.

Au temps $t = 0$, on a $T = T_0$ sur les parois du mur c'est-à-dire pour $x = 0$ et $x = h$. Les solutions des équations (15.3) et (15.4) s'écrivent :

$$T(t) = A' \exp(-kDt) \qquad X(x) = B' \cos(kx) + C' \sin(kx),$$

$A'$, $B'$ et $C'$ étant des constantes d'intégration. Une solution de (15.1) s'écrit :

$$\theta(x,t) = [A \cos(kx) + B \sin(kx)] \exp(-k^2 Dt), \tag{15.5}$$

expression dans laquelle $A$ et $B$ sont des constantes, ce sont des combinaisons linéaires des autres constantes $A'$, $B'$ et $C'$. Il reste à appliquer les conditions aux limites et les conditions initiales :

$$\theta(0, t \geq 0) = \theta(0,0) = T_0 = A,$$
$$\theta(h, t > 0) = \theta(h,0) = T_0 = A \cos(kh) + B \sin(kh).$$

Pour simplifier les équations, on peut réaliser un changement d'origine des températures et poser :

$$\Theta = \theta - T_0$$

$T_0$ est la nouvelle référence, et l'on obtient :

$$A = 0$$
$$0 = B \sin(kh).$$

La solution non triviale impose :

$$kh = n\pi \qquad \text{avec } n = 0, 1, 2, 3 \ldots,$$
$$\text{donc } k = n\frac{\pi}{h} \qquad \text{avec } n = 0, 1, 2, 3, \ldots,$$

et l'expression (15.5) s'écrit alors :

$$\theta(x,t) - T_0 = B \sin\left(\frac{n\pi}{h} x\right) \exp\left[-\left(\frac{n\pi}{h}\right)^2 Dt\right].$$

Il ne reste plus qu'à déterminer le coefficient $B$, et, en toute rigueur, $B$ dépend de $n$. Compte tenu de la linéarité de l'équation (15.1), la solution la plus générale doit donc s'écrire :

$$\theta(x,t) - T_0 = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi}{h}x\right) \exp\left(-Dn^2\frac{\pi^2}{h}t\right).$$

On remarque qu'à l'instant $t = 0$ : $\theta(x,0) - T_0 = T_i - T_0 = T_i'$ pour $0 < x < h$ donc :

$$T_i' = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi}{h}x\right) \cdot \tag{15.6}$$

Multiplions les deux membres de (15.6) par $\sin\left(\frac{n\pi}{h}x\right)$ et intégrons de 0 à $h$. On obtient :

$$\int_0^h T_i' \sin\left(\frac{n\pi}{h}x\right)\,\mathrm{d}x = B_n \int_0^h \sin^2\left(\frac{n\pi}{h}x\right)\,\mathrm{d}x$$

à cause de l'orthogonalité des fonctions sinus (voir le prochain paragraphe). Le membre de gauche est différent de zéro quand $n$ est impair soit $n = 2p + 1$, quant au membre de droite, il vaut $h/2$. On déduit donc la solution générale :

$$\theta(x,t) - T_0 = 4\frac{T_i'}{\pi} \sum_{\pi=1}^{\infty} \frac{1}{2p+1} \sin\left[(2p+1)\frac{\pi}{h}x\right] \exp\left[-D(2p+1)^2\frac{\pi^2}{h^2}t\right]. \tag{15.7}$$

D'une façon plus générale, si au temps $t < 0$, la condition dans le mur s'écrit $F(x)$ pour $x$ compris entre 0 et $h$, alors la condition (15.6) prend la forme :

$$F(x) = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi}{h}x\right) \cdot \tag{15.8}$$

La somme est le développement de $F(x)$ en série de sinus, les propriétés de parité ou de symétrie ayant annulé les coefficients des termes en cosinus.

En définitive, la solution spatiale est donnée par le développement de la condition initiale dans le mur (ici, c'est la fonction « fenêtre » qui est une constante dans un intervalle fini et qui est nulle ailleurs).

## 2. Orthogonalité des fonctions sinus et cosinus sur une période

Il s'agit de calculer les expressions suivantes :

$$K = \int_0^{2\pi} \sin(nx)\sin(mx)\,\mathrm{d}x$$

$$L = \int_0^{2\pi} \sin(nx)\cos(mx)\,\mathrm{d}x$$

$$M = \int_0^{2\pi} \cos(nx)\cos(mx)\,\mathrm{d}x.$$

Pour cela on utilise les formules usuelles de trigonométrie à savoir :

$$\begin{aligned}
\sin(a+b) &= \sin(a)\cos(b) + \sin(b)\cos(a)\\
\sin(a-b) &= \sin(a)\cos(b) - \sin(b)\cos(a)\\
\cos(a+b) &= \cos(a)\cos(b) - \sin(a)\sin(b)\\
\cos(a-b) &= \cos(a)\cos(b) + \sin(a)\sin(b)
\end{aligned}$$

qui nous permettent d'obtenir :

$$\begin{aligned}
\sin(a)\cos(b) &= [\sin(a+b) + \sin(a-b)]/2\\
\cos(a)\cos(b) &= [\cos(a+b) + \cos(a-b)]/2\\
\sin(a)\sin(b) &= [\cos(a-b) - \cos(a+b)]/2.
\end{aligned}$$

L'intégration des expressions $K$, $L$ et $M$ au moyen des relations précédentes fournit les égalités suivantes :

$$\begin{aligned}
K &= \pi\delta_{mn}\\
L &= 0\\
M &= \pi\delta_{mn}.
\end{aligned}$$

## 3. Série de Fourier associée à une fonction périodique

On s'intéresse à une fonction $f(x)$ périodique de période $2\pi$, à laquelle on associe la série trigonométrique :

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx),$$

susceptible de représenter la fonction $f(x)$ dans l'intervalle $(0, 2\pi)$. Supposons que nous ayons l'égalité :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx), \tag{15.9}$$

nous préciserons ultérieurement les conditions pour que la fonction et le développement trigonométrique proposé soient égaux mais, pour l'instant cherchons à calculer les coefficients $a_n$ et $b_n$. Pour cela, multiplions les deux membres de la relation (15.9) par $\cos(kx)$ et intégrons sur la période :

$$\int_0^{2\pi} f(x)\cos(kx)\,\mathrm{d}x = \int_0^{2\pi} \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(nx) + b_n \sin(nx)]\cos(kx)\,\mathrm{d}x,$$

et en tenant compte des relations d'orthogonalité, on obtient :

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(kx)\,\mathrm{d}x\,; \tag{15.10}$$

et, comme

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, \mathrm{d}x,$$

ceci explique l'origine du coefficient $a_0/2$ : tous les $a_k$ sont calculés au moyen de la même formule.

De la même façon, multiplions les deux membres de la relation (15.9) par $\sin(kx)$, puis intégrons sur une période ; en tenant compte des relations d'orthogonalité, on obtient :

$$b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(kx)\, \mathrm{d}x. \tag{15.11}$$

## 4. Conditions d'égalité de f(x) et de la série de Fourier associée

On admettra sans démonstration les théorèmes suivants :

### 4.1. Théorème I

Si $f(x)$ est une fonction périodique de période $2\pi$, continue et pourvue d'une dérivée première continue $f'(x)$, sauf éventuellement en un nombre fini de points sur une période et qui sont alors des points de discontinuité de première espèce pour $f(x)$ et $f'(x)$, la série trigonométrique (15.7) est convergente pour tout $x$ et a pour somme $f(x)$ dans le cas de la continuité et $[f(x+0) + f(x-0)]/2$ dans le cas de la discontinuité (convergence simple).

### 4.2. Théorème II

Le théorème I est conservé si $f(x)$ est à variations bornées sur un intervalle de période $2\pi$.

***Définition d'une fonction à variations bornées*** – Soit $f(x)$ une fonction définie sur un intervalle $(a, b)$. On divise $(a, b)$ en $n$ sous-intervalles partiels au moyen de $(n-1)$ points tels que :

$$a < x_1 < x_2 < x_3 < \cdots < x_{n-1} < b,$$

soit $S$ la somme définie par :

$$S_n = |f(x_1) - f(a)| + |f(x_2) - f(x_1)| + \cdots + |f(b) - f(x_{n-1})|.$$

Si, quelle que soit la décomposition de $(a, b)$, la somme $S_n$ reste inférieure à un nombre $M$ fini, on dit que la fonction $f(x)$ est à variations bornées sur $(a, b)$.

### 4.3. Théorème III

Appelé théorème de Jordan (1838–1922). Si $f(x)$ est périodique de période $2\pi$, et à variations bornées sur tout intervalle fini, alors sa série de Fourier est convergente et **uniformément** convergente dans tout intervalle où $f(x)$ est continue.

# 5. Quelques propriétés remarquables

## 5.1. Sur la vitesse de décroissance des coefficients $a_n$ et $b_n$

Si la série trigonométrique est convergente, alors $a_n$ et $b_n$ tendent vers zéro au moins comme $1/n$, cependant, la convergence de la série peut être très lente.

Si la série trigonométrique est convergente et uniformément convergente, $a_n$ et $b_n$ tendent vers zéro comme $1/n^2$, $f(x)$ est alors continue partout sur une période.

Réciproquement, si $f(x)$ n'est pas continue, l'un au moins des coefficients tend vers zéro comme $1/n$.

## 5.2. À propos du théorème de Jordan

Si les conditions du théorème de Jordan sont vérifiées, alors la série de Fourier est intégrable terme à terme. À propos de ce théorème, on peut se poser la question de savoir si toutes les fonctions continues sont développables en série de Fourier. La réponse est non. En effet, toutes les fonctions continues ne sont pas à variations bornées et l'on peut trouver des exemples de fonctions continues dont la série trigonométrique associée diverge en un point parce que la fonction n'est pas à variations bornées.

## 5.3. Sur la parité des fonctions f(x)

Si $f(x)$ est une fonction paire, la série trigonométrique associée n'admet qu'un développement en cosinus (fonction paire), les coefficients des termes en sinus sont nuls.

Si $f(x)$ est une fonction impaire, le développement ne comprend que des termes en sinus, et tous les $a_n$ sont nuls.

## 5.4. Dérivabilité d'une série de Fourier

Si $f(x)$ et $f'(x)$ possèdent des discontinuités de première espèce en nombre fini sur l'intervalle $(0, 2\pi)$, et sont continues ailleurs, alors $f(x)$ est développable en série de Fourier, mais la série converge lentement ; l'un des coefficients $a_n$ ou $b_n$ tend vers zéro avec $1/n$ et la série dérivée ne converge pas. En revanche, si $f'(x)$ et $f''(x)$ existent et sont continues sauf éventuellement en un nombre fini de points dans l'intervalle $(0, 2\pi)$ où il existe pour ces deux fonctions des discontinuités de première espèce, alors, on peut calculer directement un développement de $f'(x)$ en série de Fourier.

***Remarque*** : Deux séries de Fourier dont les sommes coïncident sur un intervalle intérieur à une période, ne sont pas nécessairement identiques en dehors de cet intervalle.

## 5.5. Quelques développements traditionnels

***a*** – Soit la fonction paire de période $2\pi$ définie ainsi :

$$f(x) = \pi - x \qquad \text{si} \qquad 0 \leq x < \pi$$
$$f(x) = x - \pi \qquad \text{si} \qquad -\pi \leq x < 0.$$

Comme $f(x)$ est une fonction continue et paire, on en déduit immédiatement que les coefficients $b_n$ sont nuls. Il reste à calculer les $a_n$ :

$$a_n = \frac{2}{\pi} \int_0^{2\pi} (\pi - x) \cos(nx)\, \mathrm{d}x.$$

Les termes correspondant à une valeur de $n$ pair sont nuls, il s'ensuit que :

$$a_{2p+1} = \frac{4}{\pi(2p+1)^2},$$

avec $p = 0, 1, 2, \ldots$ On trouve $a_0 = \pi$.

***b*** – Soit la fonction $f(x) = x - 1/2$ si $0 < x < 1$.
On obtient le développement suivant :

$$f(x) = -\frac{1}{\pi}\left(\frac{\sin(2\pi x)}{1} + \frac{\sin(4\pi x)}{2} + \cdots + \frac{\sin(2n\pi x)}{n} + \cdots\right).$$

***c*** – Soit la fonction qui vaut 1 sur $(0, \pi)$ et $-1$ sur $(-\pi, 0)$.
Cette fonction est impaire, donc les $a_n$ sont nuls. On trouve :

$$b_{2p+1} = \frac{4}{\pi(2p+1)},$$

d'où

$$f(x) = \frac{4}{\pi}\left(\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \cdots + \frac{\sin[2(p+1)x]}{(2p+1)} + \cdots\right).$$

La dérivée de ce développement n'est pas un développement convergent, cependant on pourra l'utiliser quand même pour calculer $f'(x)$... (*cf.* § 8.2 de ce chapitre).

### 5.6. Note sur le calcul effectif des coefficients de Fourier

Il n'est pas nécessaire de calculer les coefficients de Fourier au moyen des intégrales, et il est plus efficace sur le plan de la précision des calculs ainsi que du temps d'exécution d'utiliser la transformée de Fourier numérique que l'on abordera au chapitre 16. Donc, il n'est pas utile de fournir un programme qui calculerait directement les intégrales.

## 6. Approximation des fonctions par une série de Fourier tronquée

La première application des séries de Fourier concerne l'approximation d'une fonction $f(x)$. Le développement alors est limité à l'ordre $n$. La démonstration repose sur le théorème d'approximation de Weierstrass selon lequel toute fonction continue sur l'intervalle $(a, b)$ peut être approchée par un polynôme $P_n(x)$ tel que :

$$|f(x) - P_n(x)| < \varepsilon$$

quel que soit $x$ appartenant à $(a, b)$. Dans cette expression $\varepsilon$ est un nombre positif aussi petit que l'on veut. On trouvera une démonstration complète dans l'ouvrage de J. Bass, *Cours de Mathématiques*, Tome I, p. 312, Masson, Paris 1961.

## 6.1. Théorème

Si l'on veut approcher au sens des moindres carrés une fonction $f(x)$ définie sur $(a, b)$ par une série trigonométrique et que cette série est tronquée à l'ordre $n$, alors les coefficients du développement trigonométrique sont égaux aux coefficients du développement en série de Fourier.

Bien entendu ce théorème est valable sous réserve que les conditions de validité de développement en série de Fourier soient vérifiées. Autrement dit, le développement en série de Fourier tronqué à l'ordre $n$ est celui qui, parmi tous les développements trigonométriques possibles tronqués à l'ordre $n$, assure la meilleure approximation, au sens des moindres carrés, d'une fonction développable en série trigonométrique.

Nous allons démontrer cette proposition dans le cas d'une fonction échantillonnée.

***Approximation d'une fonction échantillonnée*** – Considérons une fonction $f(x)$ périodique de période $2\pi$ dont on connaît uniquement $2N+1$ valeurs $y_k$, pour $2N+1$ valeurs $x_k$ de la variable. Les $2N+1$ valeurs de la variable sont réparties en progression arithmétique sur toute la période, soit :

$$-\pi, \frac{N-1}{N}\pi, \ldots, -\frac{\pi}{N}, 0, \frac{\pi}{N}, \ldots, \frac{N-1}{N}\pi.$$

Comme $f(\pi) = f(-\pi)$, seules $2N$ valeurs $f(x_l)$ sont indépendantes. On a posé :

$$x_l = \frac{\pi}{N} l \qquad \text{avec } l = -N+1, -N+2, \ldots, -1, 0, +1, \ldots, N.$$

Nous allons donc approcher $f(x)$ par une expression trigonométrique dont la forme la plus générale est :

$$f(x) = A_0 + \sum_{k=1}^{n} [A_k \cos(kx) + B_k \sin(kx)] \qquad \text{avec } n \leq N,$$

de telle sorte qu'elle vérifie le critère des moindres carrés. Autrement dit, on désire rendre l'expression :

$$E^2 = \sum_{l=-N+1}^{N} \left\{ f(x_l) - A_0 - \sum_{k=1}^{n} [A_k \cos(kx) + B_k \sin(kx)] \right\}^2 \tag{15.12}$$

la plus petite possible. On y parvient en écrivant que :

$$\frac{\partial E^2}{\partial A_j} = 0 \qquad \text{et} \qquad \frac{\partial E^2}{\partial B_j} = 0.$$

Pour obtenir des expressions commodes d'emploi, il faut utiliser un certain nombre de relations trigonométriques qui s'établissent en se servant de la formule de Moivre (1667–1754) et de l'expression de la somme d'une progression géométrique. Voici ces relations dans lesquelles $j$ et

$k$ sont des entiers appartenant à $(0, N)$ :

$$\sum_{l=-N+1}^{N} \sin(jx_l)\cos(kx_l) = 0$$

$$\sum_{l=-N+1}^{N} \sin(jx_l)\sin(kx_l) = 0 \quad \text{si } j \neq k$$

$$\sum_{l=-N+1}^{N} \cos(jx_l)\cos(kx_l) = 0 \quad \text{si } j \neq k$$

$$\sum_{l=-N+1}^{N} \sin(kx_l) = \sum_{l=-N+1}^{N} \cos(kx_l) = N \quad \text{si } k \neq 0, N$$

$$\sum_{l=-N+1}^{N} \cos(Nx_l) = 2N$$

$$\sum_{l=-N+1}^{N} \cos(mx_l) = \begin{cases} 0 & \text{si } m \neq 2\nu N \\ 2N & \text{si } m = 2\nu N \end{cases} \qquad \text{où } \nu \text{ est un entier quelconque.}$$

$$\sum_{l=-N+1}^{N} \sin(mx_l) = 0.$$

Compte tenu de ces relations, on calcule les dérivées de l'expression (15.12) par rapport aux coefficients $A_k$ et $B_k$. On obtient alors les relations suivantes :

$$A_0 = \frac{1}{2N} \sum_{l=-N+1}^{N} f(x_l) \qquad A_N = \frac{1}{2N} \sum_{l=-N+1}^{N} f(x_l)\cos(Nx_l)$$

et si $k \neq 0, N$

$$A_k = \frac{1}{N} \sum_{l=-N+1}^{N} f(x_l)\cos(kx_l) \qquad B_k = \frac{1}{N} \sum_{l=-N+1}^{N} f(x_l)\sin(kx_l).$$

On reconnaît sans grande difficulté les coefficients du développement en série de Fourier calculés par la méthode des rectangles.

***Remarque très importante*** – Lorsque l'on effectue le calcul numérique des coefficients de Fourier de fonctions périodiques, on est tenté d'utiliser des méthodes offrant en apparence plus de précision, telles, par exemple, la méthode des trapèzes ou la méthode de Simpson. À bien y réfléchir, ces méthodes privilégient certains points en leur affectant des poids différents, mais on ne doit pas perdre de vue que le calcul de toutes les intégrales peut être réalisé sur n'importe quel intervalle pourvu qu'il ait la longueur d'une période. Conclusion, nous devons utiliser une méthode qui affecte à chaque point le même poids : il n'y en a qu'une seule, c'est la méthode des rectangles.

On retrouvera rigoureusement le même procédé lors du calcul des transformées de Fourier par les algorithmes appropriés.

***Un exemple numérique*** – On a fabriqué une dent de scie tout à fait analogue à la première fonction développée, et il est intéressant de noter le comportement de l'approximation au voisinage de la discontinuité de la dérivée première. On a tronqué la série trigonométrique à l'ordre 10.

On trouvera sur le Web[*] le programme `echantil.c` réalisant la décomposition en série de Fourier de fonctions périodiques connues au moyen d'un échantillonnage déterminé selon des abscisses en progression arithmétique.

## 7. Cas où la fonction est discontinue à l'origine

Dans ce cas, la notion de périodicité au sens strict ne s'applique plus, cependant, cela ne change rien à l'analyse précédemment réalisée ; en effet deux solutions peuvent être apportées à ce problème :

1. On choisit comme nouvelle origine un point où la fonction est continue ; ainsi il n'y a plus de discontinuité à l'origine et ce faux problème a disparu.
2. On sait très bien obtenir le développement en série de Fourier des fonctions en dent de scie présentant une discontinuité à l'origine. Par conséquent, il suffit de retrancher convenablement une fonction en dent de scie qui annule la discontinuité. En faisant usage de la propriété de linéarité des intégrales, on déduit que le développement en série de Fourier de deux fonctions développables est la somme de chacun des deux développements ; ce qui résout notre problème.

## 8. Le phénomène de Gibbs (1839–1903) et l'epsilon-algorithme

Il s'agit d'étudier le comportement du développement en série de Fourier de la fonction $f(x)$ au voisinage d'un point de discontinuité. Nous allons nous intéresser à un cas très classique constitué par la fonction « rectangle » ou fonction « fente » :

$$g(x) = \begin{cases} -1 & \text{si} \quad -\pi < x < 0 \\ +1 & \text{si} \quad 0 < x < \pi \end{cases}$$

dont nous avons déjà établi le développement en série de Fourier, soit :

$$f(x) = \frac{4}{\pi}\left(\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \cdots + \frac{\sin[2(p+1)x]}{(2p+1)} + \cdots\right).$$

Bien que $g(x)$ tende vers $\pm 1$, le développement associé $f(x)$ qui est continu tend vers zéro pour la valeur $x = 0$. À présent nous allons étudier le comportement de $f(x)$ au voisinage de zéro. Pour cela, considérons le développement tronqué à l'ordre $n$ de $f(x)$ que nous désignons par $S_n(x)$ :

$$S_n(x) = \frac{4}{\pi}\left(\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \cdots + \frac{\sin[2(n-1)x]}{(2n-1)}\right).$$

On peut encore écrire :

$$\frac{\pi}{4}S_n(x) = \int_0^x \{\cos(t) + \cos(3t) + \cdots + \cos[(2n-1)t]\}\ \mathrm{d}t,$$

* `http://www.edpsciences.com/guilpin/`

et puisque

$$\cos(t) + \cos(3t) + \cdots + \cos[(2n-1)t] = \frac{\sin(2nt)}{2\sin(t)},$$

on obtient :

$$\frac{\pi}{4} S_n(x) = \int_0^x \frac{\sin(2nt)}{2\sin(t)}\, \mathrm{d}t.$$

Les extremums de $S_n(x)$ sont donnés par $\sin(2nx) = 0$. Soit encore :

$$x_k = \frac{k\pi}{2n} \quad \text{avec } k = 1, 2, 3, \ldots n.$$

La courbe représentant $S_n(x)$ va effectuer une série d'oscillations pour un point $x$ quelconque choisi dans le voisinage de l'ordonnée $y = 1$. On voit sans trop de difficulté que le premier maximum est le plus grand ; pour le calculer on peut écrire :

$$\frac{2}{\pi} \int_0^{\pi/2n} \frac{\sin(2nt)}{2\sin(t)}\, \mathrm{d}t = \frac{2}{\pi} \int_0^{\pi} \frac{\sin(u)}{2n\sin(u/2n)}\, \mathrm{d}u$$

en ayant posé $u = 2nx$.

Pour les grandes valeurs de $n$ et les petites valeurs de $x$, on peut effectuer un développement limité de $\sin(u/2n)$ et remplacer cette expression par $u/2n$. Ainsi quand $n$ tend vers l'infini et $x$ vers zéro, on obtient :

$$\frac{2}{\pi} \int_0^{\pi} \frac{\sin(u)}{u}\, \mathrm{d}u = \frac{2}{\pi} Si(\pi),$$

où $Si(x)$ est la fonction sinus intégral que l'on peut calculer au moyen du développement en série entière suivant :

$$Si(x) = \frac{x}{1!1} - \frac{x^3}{3!3} + \frac{x^5}{5!5} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!(2n+1)} + \cdots$$

La fonction représentée par la série trigonométrique fait un saut brusque dépassant de 17 % la valeur de $g(x)$. Plus on prend de termes, plus cette anomalie se trouve repoussée sur l'axe des $y$.

## 8.1. Usage de l'epsilon-algorithme et disparition du phénomène de Gibbs

Nous allons appliquer l'epsilon-algorithme à la suite des sommes partielles $S_n(x)$. Nous avons considéré les termes depuis $S_{71}$ jusqu'à $S_{85}$ pour la valeur $x = 0{,}1$ de l'argument. Les résultats trouvés sont présentés dans le tableau 15.1, page suivante.

L'epsilon-algorithme appliqué à ces quinze valeurs donne le résultat suivant :

0,999 998 031

On voit que l'epsilon-algorithme fait disparaître le fâcheux phénomène de Gibbs et les incessantes oscillations au voisinage d'un point de discontinuité.

**Tableau 15.1.**

| |
|---|
| 0,990 709 731 |
| 0,999 733 562 |
| 1,008 519 556 |
| 1,016 728 834 |
| 1,024 054 265 |
| 1,035 047 814 |
| 1,038 349 421 |
| 1,040 045 437 |
| 1,038 581 660 |
| 1,035 560 117 |
| 1,025 708 447 |
| 1,019 325 920 |

Sur le Web[*], on donne le programme `epsil2.c` qui montre ce genre de calcul.

***Remarque*** : Le phénomène de Gibbs n'est pas propre au développement d'une fonction présentant une discontinuité de première espèce sur une base de sinus et de cosinus. Le problème est inhérent à chaque point de discontinuité quelle que soit la base de **fonctions continues** utilisée et le développement continu ne peut en aucun cas représenter convenablement une discontinuité. Il n'en est plus de même si la base des fonctions (orthogonales si possible) est constituée de fonctions discontinues telles que les fonctions de Hadamard, Walsh, Paley et de Haar. Les points de discontinuité ne sont plus la source d'oscillations, et les fonctions échantillonnées sont parfaitement rendues.

## 8.2. Retour sur les fonctions f(x) présentant des discontinuités

Nous avons dit que la fonction $f(x)$ admettait un développement en série de Fourier mais qu'au moins une série de coefficients ($a_n$ ou $b_n$) tendait vers zéro comme $1/n$. Si l'on dérive chacun des deux membres, on voit que le seconde membre est une série divergente puisque c'est une série de terme en $\cos(kx)$ ou $\sin(kx)$. L'epsilon-algorithme tire néanmoins parti de ces données en calculant les sommes partielles de la série divergente. Nous avons pris pour exemple la fonction qui vaut $-1$ pour $x$ appartenant à l'intervalle $(-\pi, 0)$ et $+1$ pour $x$ appartenant à l'intervalle $(0, \pi)$ :

$$f(x) = \frac{4}{\pi}\left(\frac{\sin(x)}{1} + \frac{\sin(3x)}{3} + \cdots + \frac{\sin[2(p+1)x]}{(2p+1)} + \cdots\right)$$

la dérivée s'écrit :

$$f'(x) = \frac{4}{\pi}\{\cos(x) + \cos(3x) + \cdots + \cos[2(p+1)x] + \cdots\}$$

le membre de droite est divergent alors que le membre de gauche est nul partout sauf aux points $x = -\pi$, $x = 0$ et $x = \pi$ où $f'(x)$ est un pic de Dirac. C'est bien ce que donne l'epsilon-algorithme,

* `http://www.edpsciences.com/guilpin/`

et l'on trouvera sur le Web (*) le programme `dirac.c` qui permet de calculer $f'(x)$ en exploitant le développement en série.

## 9. Représentation des séries de Fourier avec un terme de phase

Dans bien des applications, et notamment en électricité, il est souhaitable d'utiliser une autre forme pour le développement en série de Fourier (qui rend plus facile d'emploi la notion d'impédance par exemple). On peut alors préférer conserver uniquement ou les sinus ou les cosinus et faire apparaître à l'intérieur des expressions un terme de phase, soit :

$$f(x) = \frac{d_0}{2} + \sum_{n=1}^{\infty} d_n \cos(nx - \varphi_n) = \frac{d_0}{2} + \sum_{n=1}^{\infty} d_n \left[\cos(nx)\cos(\varphi_n) + \sin(nx)\sin(\varphi_n)\right].$$

En rapprochant cette expression du développement (15.9), nous obtenons par identification :

$$a_0 = d_0 \qquad a_n = d_n \cos(\varphi_n) \qquad b_n = d_n \sin(\varphi_n) \qquad d_n = (\text{signe de } a)\sqrt{a_n{}^2 + b_n{}^2}.$$

## 10. Écriture du développement sous forme complexe

Comme c'est à partir de cette étude que nous allons établir les transformées de Fourier, nous procéderons d'abord à l'extension de la période que l'on désigne par $T$ au lieu de $2\pi$. En posant $\omega = 2\pi/T$, la série s'écrit alors :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega x) + \sum_{n=1}^{\infty} b_n \sin(n\omega x), \tag{15.13}$$

et les coefficients deviennent :

$$a_n = \frac{2}{T}\int_0^T f(x)\cos(n\omega x)\,\mathrm{d}x\,;$$

$$b_n = \frac{2}{T}\int_0^T f(x)\sin(n\omega x)\,\mathrm{d}x.$$

À présent, écrivons les cosinus et les sinus en termes complexes :

$$\cos(n\omega x) = \frac{\exp(jn\omega x) + \exp(-jn\omega x)}{2},$$

$$\sin(n\omega x) = \frac{\exp(jn\omega x) - \exp(-jn\omega x)}{2j},$$

ce qui donne pour $f(x)$ l'expression :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \frac{\exp(jn\omega x) + \exp(-jn\omega x)}{2} + \sum_{n=1}^{\infty} b_n \frac{\exp(jn\omega x) - \exp(-jn\omega x)}{2j}.$$

* `http://www.edpsciences.com/guilpin/`

Soit encore :

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \frac{a_n - jb_n}{2} \exp(jn\omega x) + \sum_{n=1}^{\infty} \frac{a_n + jb_n}{2} \exp(-jn\omega x). \tag{15.14}$$

Les coefficients sont alors donnés par les expressions suivantes :

$$\frac{a_n - jb_n}{2} = \frac{1}{T} \int_0^T f(x)[\cos(n\omega x) - j \sin(n\omega x)] \, dx = \frac{1}{T} \int_0^T f(x) \exp(-jn\omega x) \, dx$$

$$\frac{a_n + jb_n}{2} = \frac{1}{T} \int_0^T f(x) \exp(jn\omega x)) \, dx.$$

Si l'on remarque que l'on obtient $(a_n - jb_n)/2$ à partir de $(a_n + jb_n)/2$ en changeant $n$ en $-n$, on trouve des formules plus faciles d'emploi en posant :

$$c_n = \frac{a_n - jb_n}{2} \quad \text{et} \quad c_{-n} = \frac{a_n + jb_n}{2} \quad \text{avec} \quad c_0 = a_0.$$

Alors, la série trigonométrique s'écrit :

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \exp(jn\omega x),$$

et les coefficients sont donnés par l'expression :

$$c_n = \frac{1}{T} \int_0^T f(x) \exp(-jn\omega x) \, dx.$$

Il faut bien noter que ces dernières expressions n'apportent rien de plus que les précédentes du point de vue du calcul, mais elles servent d'introduction commode aux transformées de Fourier.

## 11. Approximation des fonctions au sens de Tchebycheff

Les problèmes d'approximation au sens de Tchebycheff ne sont pas à dissocier de l'étude des séries de Fourier. Soit une fonction $f(x)$ continue à dérivées continues sur un intervalle $(a, b)$. Nous recherchons une approximation $g(x)$ de la fonction $f(x)$ dans cet intervalle de telle sorte que $\sup |f(x) - g(x)|$ soit le plus petit possible. C'est ce qui est convenu d'appeler l'approximation au sens de Tchebycheff.

Examinons le cas où $g(x)$ est un polynôme $P_n(x)$ de degré $n$. Pour l'amour de la simplicité et en faisant usage d'une transformation linéaire classique, ramenons l'intervalle $(a, b)$ à l'intervalle canonique $(-1, +1)$. Rappelons rapidement quelques résultats fondamentaux relatifs à l'interpolation par des polynômes. Le polynôme qui passe par les $(n + 1)$ points d'abscisse $\alpha_i$ est le polynôme de Lagrange $L_n(x)$ de degré $n$. L'erreur commise $E_n(x)$ est donnée par l'expression :

$$E_n(x) = f(x) - L_n(x) = \prod_{i=1}^{n+1} (x - \alpha_i) \frac{f^{(n+1)}(\eta)}{(n+1)!}$$

où $\eta$ appartient à l'intervalle $(-1, +1)$.

L'erreur sera rendue minimum en utilisant la propriété essentielle des polynômes de Tchebycheff, ce qui nous imposera le choix des abscisses, à savoir :

$$\alpha_i = \cos\left(\frac{2i+1}{n+1}\frac{\pi}{2}\right) \quad \text{avec} \quad i = 0, 1, 2, 3, \ldots, n.$$

Ces valeurs reportées dans l'expression du polynôme permettent d'obtenir une formule fournissant l'erreur minimum au sens de Tchebycheff. Pour des raisons pratiques, il est possible d'opérer de la façon suivante : toute puissance de $x$, notée $x^k$, peut s'exprimer au moyen d'une combinaison linéaire de polynômes de Tchebycheff dont les degrés sont égaux et inférieurs à $k$. Donc, formellement :

$$L_n(x) = \sum_{k=0}^{n} p_k T_k(x).$$

Il nous reste alors à calculer les coefficients $p_k$. Pour cela, nous allons écrire l'erreur minimum $E_n(x)$ sous la forme d'un polynôme de Tchebycheff :

$$E_n(x) = f(x) - L_n(x) = \frac{T_n(x)}{2^n(n+1)!} f^{(n+1)}(\eta).$$

Posons alors $x = \cos(Q)$. Eu égard à la définition des polynômes de Tchebycheff, on peut écrire :

$$f[\cos(Q)] = \sum_{k=0}^{n} p_k \cos(kQ) + E_n \cos(Q) = \Psi(Q),$$

en se souvenant que pour les valeurs $\alpha_i$, nous avons :

$$f(\alpha_i) = \sum_{k=0}^{n} p_k T_k(\alpha_i)$$

et pour les valeurs

$$Q_i = \frac{2i+1}{n+1}\frac{\pi}{2},$$

nous avons :

$$\Psi(Q_i) = \sum_{k=0}^{n} p_k \cos(kQ_i).$$

Les coefficients $p_k$ sont donc les coefficients du développement en série de Fourier, tronqué à l'ordre $n$, de la fonction paire $f[\cos(Q)]$. On peut écrire :

$$p_0 = \frac{1}{1+n}\sum_{k=0}^{n} \Psi(Q_i) \qquad p_k = \frac{2}{1+n}\sum_{k=0}^{n} \Psi(Q_i)\cos(kQ_i),$$

ou encore :

$$p_0 = \frac{1}{1+n}\sum_{i=0}^{n} f(\alpha_i) \qquad p_k = \frac{2}{1+n}\sum_{i=0}^{n} f(\alpha_i) T_i(k\alpha_i).$$

## 12. Application des séries de Fourier au filtrage numérique

D'un point de vue très simple, le filtrage numérique est une opération linéaire que l'on réalise dans l'espace réciproque de Fourier dans le but de modifier le spectre de fréquences ou de nombre d'ondes (filtre passe-bas, filtre passe-haut et filtre passe-bande pour faire référence aux applications radioélectriques). Ensuite on retourne à l'espace direct. Nous reviendrons plus en détail sur ces espaces, car pour ce qui concerne les séries de Fourier cette distinction, pourtant fondamentale, est peu perceptible dans la mesure où les fonctions $\sin(n\omega x)$ et $\cos(n\omega x)$ dépendent directement des arguments de l'espace direct et de l'espace réciproque, c'est-à-dire $x$ et $\omega$.

Pour fixer les idées nous allons étudier un filtre sans contrainte. On peut dire qu'un filtre linéaire est tout simplement un ensemble fini de coefficients $A_k$ qui dépendent uniquement de l'opération que l'on souhaite réaliser mais certainement pas de la fonction $f(t)$ à filtrer, coefficients qui viennent pondérer les valeurs des échantillons :

$$\Phi(t) = \sum_{k=-M}^{N} A_k f(t + k\delta t), \tag{15.15}$$

où $\Phi(t)$ représente la fonction filtrée à l'instant $t$, $\delta t$ représente l'intervalle de temps entre deux échantillonnages, $M$ et $N$ sont deux entiers positifs.

L'étude des filtres numériques linéaires peut passer par la notion de fonction de transfert. On étudie donc l'action du filtre sur un signal sinusoïdal que nous écrirons :

$$f(t) = \exp(j\omega t).$$

Par définition, la fonction de transfert en $\omega$ du filtre (15.15) est donnée par l'expression :

$$T(\omega) = \frac{\Phi(t)}{\exp(j\omega t)},$$

on obtient donc :

$$T(\omega) = \sum_{k=-M}^{N} A_k \exp(jk\omega\delta t). \tag{15.16}$$

Très souvent on se donne une fonction de transfert $T(\omega)$, et il s'agit alors d'exprimer les coefficients $A_k$ correspondant à un filtre dont la fonction de transfert $T_{-MN}(\omega)$ soit la plus voisine de $T(\omega)$ au sens des moindres carrés. Autrement dit, il convient de minimiser, dans l'espace réciproque de Fourier, l'intégrale calculée sur une période $2\pi/\delta t$ :

$$I = \int_{-\pi/\delta_t}^{+\pi/\delta t} |T_{-MN}(\omega) - T(\omega)|^2 \, d\omega.$$

Il s'ensuit que les $A_k$ sont tels que :

$$\frac{\partial I}{\partial A_k} = 0,$$

et, en reportant la relation (15.16) dans la dernière intégrale, il est possible d'écrire :

$$\frac{\partial I}{\partial A_m} = 2 \int_{-\pi/\delta_t}^{+\pi/\delta t} \left[ \sum_{k=-M}^{N} A_k \exp(jk\omega\delta t) - T(\omega) \right] \exp(jm\omega\delta t) d\omega.$$

En faisant usage de la relation d'orthogonalité :

$$\int_{-\pi/\delta_t}^{+\pi/\delta t} \exp[j\omega(m+k)\delta t] d\omega = \frac{2\pi}{\delta t} \delta_{-m,k},$$

on en déduit que :

$$A_k = \frac{\delta t}{2\pi} \int_{-\pi/\delta_t}^{+\pi/\delta t} T(\omega) \exp(-jk\omega\delta t) d\omega.$$

Deux remarques s'imposent :

**a.** Les $A_k$ sont indépendants de $M$ et $N$.
**b.** Les $A_k$ sont les coefficients de Fourier du développement de la fonction de transfert $T(\omega)$.

## 12.1. Application à l'étude d'un filtre passe-bas

***a*** – Nous allons porter notre attention sur un des filtres les plus simples à réaliser : la fonction « fenêtre » dans l'espace réciproque de Fourier. Ce filtre se définit de la façon suivante :

$$\begin{aligned} 0 \leq |\omega| \leq \omega_1 \qquad & T(\omega) = 1, \\ \omega_1 \leq |\omega| \leq \pi/\delta t \qquad & T(\omega) = 0. \end{aligned}$$

Comme c'est manifestement un filtre symétrique, les termes en sinus sont nuls. Il reste à calculer les termes en cosinus :

$$A_k = \frac{1}{\pi k} \cos(k\omega_1 \delta t) \qquad \text{avec } A_0 = \frac{\omega_1 \delta t}{\pi}.$$

***b*** – Il est préférable d'utiliser un filtre constitué d'une fonction continue afin de réduire les oscillations données par l'approximation. Aussi, pourra-t-on choisir un filtre du type suivant :

$$\begin{aligned} 0 \leq |\omega| \leq \omega_1 \qquad & T(\omega) = 1 \\ \omega_1 \leq |\omega| \leq \omega_2 \qquad & T(\omega) = \left\{ 1 + \cos\left( \pi \frac{\omega - \omega_1}{\omega_2 - \omega_1} \right) \right\} / 2 \\ \omega_2 \leq |\omega| \leq \pi/\delta t \qquad & T(\omega) = 0. \end{aligned}$$

Tous calculs faits, on trouve les coefficients suivants :

$$A_k = \frac{\pi}{2k\left[\pi^2 - k^2 \delta t^2 (\omega_2 - \omega_1)^2\right]} \left[\sin(k\omega_2\delta t) + \sin(k\omega_1\delta t)\right] \qquad \text{avec } A_0 = \frac{(\omega_2 + \omega_1)\delta t}{2\pi}.$$

Un bon exercice consiste à représenter le filtre théorique ainsi que la fonction de transfert approchée pour $N = M = 20$ (*cf.* équation (15.15)), ensuite, d'appliquer ce filtre à une fonction présentant des irrégularités (courbe expérimentale entachée d'erreur).

### 12.2. Remarques sur les opérations mathématiques réalisées lors du filtrage

L'opération de filtrage que nous venons de réaliser traduit une convolution dans l'espace direct. Dans l'espace réciproque de Fourier nous avons réalisé le produit simple de deux fonctions qui sont la transformée de Fourier de la fonction $f(t)$ à filtrer et de la fonction « fenêtre » $T(\omega)$ qui est un filtre passe-bas. De retour dans l'espace direct, le produit simple de deux fonctions est transformé en produit de convolution. Nous reviendrons sur ces aspects des problèmes en abordant l'étude des transformées de Fourier. Cette dernière remarque prend tout son sens quand on passe à la programmation qui se réduit au calcul du produit de convolution de fonctions périodiques. Notons bien, ici, que la fonction de transfert est symétrique ce qui peut donner l'illusion de calculer une fonction de corrélation...

On trouvera sur le Web (*) le programme `filtre.c` qui réalise le filtrage de données numériques entachées d'erreur.

## 13. À propos du développement des fonctions non périodiques

Il est parfaitement légitime de développer en série de Fourier (au sens large) une fonction $f(x)$ non périodique dans la mesure où l'intervalle de définition $(a, b)$ est fini et que la fonction satisfait aux conditions légitimes du développement. Il existe une restriction importante dans la mesure où l'on ne peut pas se servir du développement en dehors de l'intervalle de définition de la fonction. Par ailleurs, il est convenable de dire que le développement en série de Fourier constitue un prolongement analytique de la fonction $f(x)$ en dehors de l'intervalle $(a, b)$.

En ce qui concerne les fonctions non périodiques définies sur un intervalle infini, elles font l'objet de traitements spéciaux appelés transformations de Fourier.

## 14. Calcul des séries de Fourier à coefficients approchés dans $L^2$

Dans certains cas, le calcul d'une série au moyen de ses coefficients est un problème mal posé dont on trouvera l'étude dans l'ouvrage de Tikhonov cité en bibliographie. Voici un exemple concernant une série convergente :

$$f_1(t) = \sum_{n=0}^{\infty} a_n \cos(nt).$$

Supposons que les coefficients $a_n$ soient entachés d'erreur et que l'on ait :

$$c_n = a_n + \frac{\varepsilon}{n} \quad \text{avec} \quad c_0 = a_0.$$

À la place de $f_1(t)$ on obtient la fonction

$$f_2(t) = \sum_{n=0}^{\infty} c_n \cos(nt).$$

Dans la métrique de $L^2$, les coefficients différent de :

$$\varepsilon_1 = \left\{ \sum_{n=0}^{\infty} (c_n - a_n)^2 \right\}^{1/2} = \varepsilon \left\{ \sum_{n=0}^{\infty} \frac{1}{n^2} \right\}^{1/2} = \varepsilon \frac{\pi}{\sqrt{6}}.$$

donc $\varepsilon_1$ est aussi petit que l'on veut.

* `http://www.edpsciences.com/guilpin/`

À présent, examinons la distance entre les fonctions $f_1(t)$ et $f_2(t)$. Elle est donnée par l'expression :

$$f_2(t) - f_1(t) = \varepsilon \sum_{n=0}^{\infty} \frac{1}{n} \cos(nt).$$

Cette quantité peut être faite aussi grande que l'on veut : la série diverge pour $t = 0$.

## 15. Éléments de bibliographie

A. ANGOT (1972) *Compléments de mathématique*, Éditions Masson.
J. BASS (1961) *Cours de mathématiques*, Éditions Masson.
R. LASSER (1996) *Introduction to Fourier series*, M. Dekker.
A. TIKHONOV et V. ARSÉNINE (1976) *Méthodes de résolution des problèmes mal posés au sens de Hadamard*, Éditions MIR, Moscou.

# 16 | Les transformées de Fourier

Depuis la vulgarisation des algorithmes rapides (1965), le rôle des transformées de Fourier dans les problèmes de traitement de fonctions échantillonnées s'est notablement accru. Ce n'est pas pour autant qu'il faille négliger l'apport considérable de la théorie qui a fourni notamment un catalogue de solutions formelles de problèmes. Cependant, le strict calcul numérique posé par un problème n'ayant pas de solution formelle connue était une entreprise à la fois très longue et très coûteuse. En calculant les intégrales au moyen de la technique des rectangles pour une fonction échantillonnées en 2 048 points il fallait, avec une calculatrice de la fin des années 60, environ 3 heures pour obtenir la transformée de Fourier. Le calcul des fonctions sinus et cosinus au moyen de relations de récurrence mentionnées par J. Arsac a ramené le même calcul à 3/4 d'heure tandis que l'utilisation de l'algorithme de Cooley-Tukey a permis d'effectuer la transformation en 45 secondes... Il ne faut pas oublier qu'ils ont eu des prédécesseurs méconnus qui ont œuvré tout à fait dans la même direction et il est probable que l'idée originale de cet algorithme soit due à Carl Runge (1856–1927) et à König. Sur ce sujet, on trouvera un petit aperçu historique dans l'ouvrage de E. Oran Brigham cité en bibliographie.

Les transformées de Fourier trouvent leurs principales applications en optique, en cristallographie, en analyse harmonique, en traitement du signal, en résolution des équations de convolution, en corrélation (théorème de Wiener (1894–1952) - Kintchine), en théorie des probabilités (fonctions caractéristiques), et d'une façon générale, dans tous les problèmes où il y a avantage à travailler dans l'espace réciproque.

## 1. Extension des séries de Fourier au cas où la période est infinie

L'idée consiste à isoler une partie de la fonction $f(t)$ sur un intervalle fini $(-T/2, +T/2)$, puis à procéder à son développement en série de Fourier, et enfin à faire tendre $T$ vers l'infini et à examiner le résultat. En posant :

$$\Omega = 2\pi/T,$$

on obtient le développement :

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(n\Omega t) + b_n \sin(n\Omega t)] \tag{16.1}$$

$$\text{avec } a_n = \frac{2}{T}\int_{-T/2}^{+T/2} f(x)\cos(n\Omega x)\,\mathrm{d}x$$

$$b_n = \frac{2}{T}\int_{-T/2}^{+T/2} f(x)\sin(n\Omega x)\,\mathrm{d}x.$$

À présent, posons $\omega = n\Omega$, cela nous permet d'obtenir une autre forme du développement :

$$f(t) = \frac{a_0}{2} + \frac{1}{\Omega}\sum_{\omega=\Omega}^{\infty} \Omega\left[a_n\cos(\omega t) + b_n\sin(\omega t)\right]. \tag{16.2}$$

Dans cette nouvelle représentation, on peut dire que $\Omega$ est l'accroissement de $\omega$ pris entre deux valeurs consécutives. Autrement dit, on peut considérer que $\Omega$ est l'accroissement $\delta\omega$ de $\omega$. Le développement (16.2) s'écrit alors :

$$f(t) = \frac{1}{T}\int_{-T/2}^{+T/2} f(x)\,\mathrm{d}x + \frac{1}{\pi}\sum_{\omega=\Omega}^{\infty}\delta\omega\int_{-T/2}^{+T/2} f(x)\left[\cos(\omega x)\cos(\omega t) + \sin(\omega x)\cos(\omega t)\right]\,\mathrm{d}x. \tag{16.3}$$

Soit encore :

$$f(t) = \frac{1}{T}\int_{-T/2}^{+T/2} f(x)\,\mathrm{d}x + \frac{1}{\pi}\sum_{\omega=\Omega}^{\infty}\delta\omega\int_{-T/2}^{+T/2} f(x)\cos[\omega(x-t)]\,\mathrm{d}x.$$

Si, pour $T$ tendant vers l'infini, l'on fait l'hypothèse que :

$$\lim\frac{1}{T}\int_{-T/2}^{+T/2} f(x)\,\mathrm{d}x = 0, \quad \text{que } \delta\omega \longrightarrow \mathrm{d}\omega \quad \text{et que } \sum_{\omega=\Omega}^{\infty} \longrightarrow \int_0^{\infty},$$

alors on peut écrire :

$$f(t) = \frac{1}{\pi}\int_0^{\infty}\mathrm{d}\omega\int_{-\infty}^{+\infty} f(x)\cos[\omega(x-t)]\,\mathrm{d}x. \tag{16.4}$$

C'est le développement en intégrale de Fourier de la fonction $f(t)$. On peut encore écrire :

$$f(t) = \frac{1}{\pi}\int_0^{\infty}\left\{\cos(\omega t)\int_{-\infty}^{+\infty} f(x)\cos(\omega x)\,\mathrm{d}x + \sin(\omega t)\int_{-\infty}^{+\infty} f(x)\sin(\omega x)\,\mathrm{d}x\right\}\mathrm{d}\omega.$$

Ces formes de développement correspondent aux développements en série de Fourier de fonctions offrant « **un spectre continu** » encore appelé « **spectre de bandes** ».

## 1.1. Intégrale de Fourier en termes complexes

Partons du développement en série de Fourier en termes complexes :

$$f(t) = \sum_{-\infty}^{+\infty} c_n \exp(jn\Omega t),$$

$$\text{avec } c_n = \frac{1}{T} \int_{-T/2}^{+T/2} f(x) \exp(-jn\Omega x)\, \mathrm{d}x.$$

En conservant les mêmes conventions que précédemment :

$$n\Omega = \omega \quad \text{et} \quad \delta\omega \longrightarrow \mathrm{d}\omega = \Omega,$$

on écrit :

$$f(t) = \frac{1}{\Omega} \sum_{-\infty}^{+\infty} c_n \exp(j\omega t),$$

d'où

$$f(t) = \frac{1}{\Omega} \sum_{\omega=-\infty}^{+\infty} \left( \frac{\Omega}{T} \int_{-T/2}^{+T/2} f(x) \exp(-j\omega x)\, \mathrm{d}x \right) \exp(j\omega t),$$

$$\text{soit encore} \quad f(t) = \frac{1}{2\pi} \sum_{\omega=-\infty}^{+\infty} \left( \delta\omega \int_{-T/2}^{+T/2} f(x) \exp(-j\omega x)\, \mathrm{d}x \right) \exp(j\omega t).$$

En effectuant une extension à l'infini, nous obtenons :

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{d}\omega \int_{-\infty}^{+\infty} f(x) \exp[j\omega(t-x)]\, \mathrm{d}x \tag{16.5}$$

On peut encore poser :

$$f(t) = \int_{-\infty}^{+\infty} G(\omega) \exp(j\omega t)\, \mathrm{d}\omega$$

$$G(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t) \exp(-j\omega t)\, \mathrm{d}t.$$

Il s'agit dans cette dernière forme d'une des expressions les plus employées appelées l'une transformée de Fourier directe, l'autre transformée de Fourier réciproque. Bien des tables de transformées de Fourier utilisent cette notation, cependant, on préfère utiliser des formes plus

symétriques en adoptant les expressions suivantes :

$$f(t) = \int_{-\infty}^{+\infty} G(x) \exp(2\pi j\omega t)\, \mathrm{d}\omega$$

$$G(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(-2\pi j\omega t)\, \mathrm{d}t \tag{16.6}$$

Les équations (16.6) sont des équations fonctionnelles.

### 1.2. Relation entre différentes définitions

Les tables d'intégrales de Fourier utilisent généralement une notation plus concise qui ne fait pas usage du facteur $2\pi$ :

$$\Psi(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(-j\omega t)\, \mathrm{d}t,$$

que l'on peut écrire encore :

$$\Psi(\omega) = \int_{-\infty}^{+\infty} f(t) \exp\left(-2\pi j\omega \frac{t}{2\pi}\right)\, \mathrm{d}t = G\left(\frac{t}{2\pi}\right).$$

***Remarque*** : Rien n'empêche de choisir des signes opposés dans les exponentielles pour définir les transformées directes et réciproques, soit :

$$f(t) = \int_{-\infty}^{+\infty} G(x) \exp(-2\pi j\omega t)\, \mathrm{d}\omega$$

$$G(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(2\pi j\omega t)\, \mathrm{d}t.$$

Il est évident qu'il faut se tenir à un choix pour effectuer les calculs.

## 2. Conditions d'existence des transformées de Fourier dans les espaces $L^1$ et $L^2$

### 2.1. Définition

L'espace $L^1$ est l'espace fonctionnel des fonctions intégrables en module, tandis que l'espace $L^2$ est l'espace fonctionnel des fonctions de carré sommable.

Si $f(x)$ appartient à $L^1$ alors l'intégrale $\displaystyle\int_{-\infty}^{+\infty} |f(x)|\, \mathrm{d}x$ existe.

Si $f(x)$ appartient à $L^2$ alors l'intégrale $\displaystyle\int_{-\infty}^{+\infty} f^2(x)\, \mathrm{d}x$ (ou $\displaystyle\int_{-\infty}^{+\infty} f(x)f^*(x)\, \mathrm{d}x$) existe.

Dans ce paragraphe, nous désirons rappeler les résultats théoriques essentiels. Il faut bien reconnaître que l'usage pratique des transformées de Fourier passe exceptionnellement par les considérations qui vont suivre, le physicien rencontrant en général «de bonnes fonctions ». Quoi qu'il en soit, un survol rapide de la théorie doit permettre d'éviter quelques pièges que l'on peut rencontrer dans la théorie des fonctions échantillonnées. Le problème fondamental repose sur les conditions d'existence de l'intégrale de Fourier qui est l'une des expressions (16.6).

Disons tout de suite que la notion d'intégrale de Riemann ne suffit plus pour l'étude de tels problèmes, et que l'on doit lui substituer la notion d'intégrale au sens de Lebesgue (1875–1941). Voici pourquoi : le point essentiel repose sur le fait que la limite d'une suite de fonctions intégrables n'est pas nécessairement une fonction intégrable ni au sens de la convergence simple ni au sens de la convergence uniforme. Le lecteur intéressé trouvera des développements et des exemples dans les ouvrages cités en bibliographie ; en particulier, des exemples de fonctions bornées qui ne sont pas intégrables au sens de Riemann.

### 2.2. Conclusions

**a.** L'espace $L^1$ des fonctions sommables au sens de Riemann n'est pas complet puisque la limite d'une suite de fonctions intégrables au sens de $L^1$ n'est pas nécessairement une fonction intégrable.

**b.** Il existe des fonctions parfaitement définies qui ne sont pas intégrables au sens de Riemann. C'est le cas de fonction discontinues ne comportant que des discontinuités de première espèce.

Les mathématiciens ont trouvé une définition plus générale de l'intégrale de Riemann afin que $L^1$ soit toujours un espace vectoriel, complet au sens de la norme $L^1$. Au fond, il s'agit d'un élargissement de la classe des fonctions intégrables. C'est à Henri Lebesgue (1875–1941) que revient le mérite d'avoir découvert la nouvelle intégrale en 1902. Non seulement l'espace $L^1$ devient un espace complet mais aussi $L^2$, ce qui lui confère un intérêt de tout premier plan lors de l'étude des séries de Fourier et des développements en séries de fonctions orthogonales. Nous ne parlerons pas davantage de cette intégrale qui passe par la notion d'intégration de mesures abstraites. Insistons toutefois sur un point essentiel : lorsque l'intégrale de Riemann est réellement calculable, elle donne le même résultat que l'intégrale de Lebesgue.

## 3. La transformée de Fourier dans l'espace $L^1$

On dit que $f(x)$ appartient à l'espace $L^1$ si, pour la fonction $f(x)$ à valeurs réelles ou complexes, on a :

$$\int_{-\infty}^{+\infty} |f(x)| \, \mathrm{d}x = f_0 \quad (f_0 \text{ fini}). \tag{16.7}$$

Ici, la norme est celle de la convergence en moyenne.

### 3.1. Théorème

La transformée de Fourier

$$F(t) = \int_{-\infty}^{+\infty} f(x) \exp(2\pi j x t) \, \mathrm{d}x$$

existe si $f(x)$ appartient à $L^1$.

En effet, on a les inégalités suivantes :

$$\left| \int_{-\infty}^{+\infty} f(x) \exp(2\pi jxt)\, \mathrm{d}x \right| < \int_{-\infty}^{+\infty} |f(x)|\, \mathrm{d}x = f_0.$$

Donc $F(t)$ existe pour tout $t$, de plus $F(t)$ est bornée uniformément :

$$|F(t)| < |f(x)|\,.$$

## 3.2. Propriétés essentielles dans $\mathbf{L^1}$

Rappelons que l'on note :

$$f(x) \quad \xrightarrow{TF} \quad F(t)$$

1. La $T.F.$ est linéaire.
2. La translation qui consiste à changer $x$ en $(x - y)$ donne :

$$f(x-y) \quad \xrightarrow{TF} \quad F(t)\exp(2\pi jty). \tag{16.8}$$

3. La symétrie donne :

$$f(x)\exp(-2\pi jzx) \quad \xrightarrow{TF} \quad F(t-z). \tag{16.9}$$

4. La dilatation d'abscisse, si $a \neq 0$, donne :

$$f(ax) \quad \xrightarrow{TF} \quad \frac{1}{|a|}F(t/a). \tag{16.10}$$

   Il faut bien noter que le cas particulier $a = 0$ ne présente pas beaucoup d'intérêt. En effet, si $a = 0$, $f(0)$ est une constante et la fonction constante n'appartient pas à $L^1$.
5. Si $f(x)$ appartient à $L^1$ et si la fonction $xf(x)$ appartient aussi à $L^1$, alors $F(t)$ admet partout une dérivée $F'(t)$ bornée qui est la transformée de Fourier de $2\pi jxf(x)$ :

$$2\pi jxf(x) \quad \xrightarrow{TF} \quad F'(t).$$

   Si $f(x)$ est dérivable $n$ fois, et si $f(x)$ ainsi que ses $n$ premières dérivées appartiennent à $L^1$, alors :

$$f^{(p)}(x) \quad \xrightarrow{TF} \quad (-2\pi jt)^p F(t)$$

   ou encore

$$x^p f(x) \quad \xrightarrow{TF} \quad \frac{1}{(2\pi jt)^p}F^{(p)}(t).$$

6. Continuité de la transformée. Si $f(x)$ appartient à $L^1$ sa transformée de Fourier $F(t)$ est uniformément continue pour toute valeur de $t$.

7. La réciprocité de la transformée de Fourier pose un problème délicat. En effet, l'intégrale (16.7) est prise au sens de Lebesgue, et sa valeur n'est pas modifiée quand on change les valeurs de $f$ sur un ensemble de mesure nulle.
8. Si $f$ et $F$ appartiennent ensemble à $L^1$, alors nous avons l'égalité de Parseval (1755–1836) :

$$\int_{-\infty}^{+\infty} |f(x)|^2 \, \mathrm{d}x = \int_{-\infty}^{+\infty} |F(t)|^2 \, \mathrm{d}t. \tag{16.11}$$

Il faut alors des propriétés de continuité pour déterminer $f(x)$ en tout point à partir de $F(t)$. Du point de vue du physicien, l'égalité de Parseval ne fait que traduire le principe de conservation de l'énergie.

### 3.3. Exemples de fonctions transformées de Fourier l'une dans l'autre

$$\begin{aligned} f(x) &\overset{TF}{\longleftrightarrow} F(t) \\ \exp(-\pi x^2) &\overset{TF}{\longleftrightarrow} \exp(-\pi t^2) \\ \exp(-2\pi|x|) &\overset{TF}{\longleftrightarrow} \frac{1}{\pi(1+t^2)} \end{aligned}$$

$$f(x) \begin{cases} = 1 \text{ si } x \text{ appartient à } (-1/2, +1/2) \\ \qquad\qquad \overset{TF}{\longleftrightarrow} \quad \dfrac{\sin(\pi t)}{\pi t} \\ = 0 \text{ ailleurs} \end{cases}$$

$$f(x) \begin{cases} = 1 \text{ si } x \text{ appartient à } (a, b) \\ \qquad\qquad \overset{TF}{\longleftrightarrow} \quad \dfrac{\sin[\pi t(b-a)]}{\pi t} \exp[j\pi t(b-a)] \\ = 0 \text{ ailleurs} \end{cases}$$

Dans ces deux derniers cas, $F(t)$ n'appartient pas à $L^1$, cependant la transformée de Fourier inverse existe si l'on prend la valeur principale au sens de Cauchy.

$$f(x) \begin{cases} = 1 - |x| \text{ si } x \text{ appartient à } (-1, +1) \\ \qquad\qquad \overset{TF}{\longleftrightarrow} \quad \left(\dfrac{\sin(\pi t)}{\pi t}\right)^2 \\ = 0 \text{ ailleurs .} \end{cases}$$

## 4. Les transformées de Fourier dans l'espace $L^2$

On dit que $f(x)$ à valeurs réelles ou complexes appartient à $L^2$ si l'on a :

$$\int_{-\infty}^{+\infty} f^2(x) \, \mathrm{d}x = |f|^2, \tag{16.12}$$

c'est-à-dire que l'intégrale existe.

### Propriétés fondamentales

Si $f(x)$ appartient à $L^2$, on peut démontrer que $F(t)$ appartient aussi à $L^2$ et que nous conservons l'égalité de Parseval :

$$\int_{-\infty}^{+\infty} f^2(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F^2(t)\,\mathrm{d}t.$$

Ces démonstrations reposent sur les propriétés des suites de Cauchy. Les autres propriétés rencontrées dans l'espace $L^1$ restent vraies dans l'espace $L^2$ à savoir : la linéarité, la translation, la dilatation d'abscisse.

## 5. Produit de convolution dans les espaces $L^1$ ou $L^2$

C'est un problème qui joue un rôle important en physique et en traitement du signal. Notamment, toutes les mesures expérimentales sont les résultats du produit de convolution du signal fondamental et de la fonction d'appareil. Par exemple, en spectroscopie, l'observation d'une raie est le produit de convolution du profil donné par la source lumineuse et de la fonction d'appareil qui a permis l'observation. Si l'on utilise un spectromètre à fentes, en première approximation, la fonction d'appareil est constituée de la « fonction fente » (triangle isocèle).

### 5.1. Définition

On appelle produit de convolution de deux fonctions $f(x)$ et $g(x)$ une fonction $h(x)$ définie par l'intégrale :

$$h(x) = \int_{-\infty}^{+\infty} f(y)g(x-y)\,\mathrm{d}y \tag{16.13}$$

qui est notée d'une façon plus concise : $h = f * g$.

***Remarque*** : Le produit de convolution est commutatif, c'est-à-dire que $h = f*g = g*f$. Il faut respecter certaines conditions pour que cette intégrale existe. L'existence est assurée notamment quand $f$ et $g$ appartiennent à $L^1$ ou $L^2$.

Supposons que les fonctions $f(x)$ et $G(t)$ appartiennent à $L^1$, de telle sorte que :

$$F(t) = \int_{-\infty}^{+\infty} f(x)\exp(j2\pi xt)\,\mathrm{d}x$$

et

$$g(x) = \int_{-\infty}^{+\infty} G(t)\exp(-j2\pi xt)\,\mathrm{d}t.$$

Alors, $g(x)$ est bornée et l'intégrale définissant le produit de convolution existe. Remplaçons $g(x)$ par sa transformée de Fourier dans l'intégrale (16.13) :

$$h(x) = \int_{-\infty}^{+\infty} f(y)g(y-x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} f(y)\left(\int_{-\infty}^{+\infty} G(t)\exp[-j2\pi t(x-y)]\,\mathrm{d}t\right)\,\mathrm{d}y.$$

En inversant l'ordre des intégrations, on obtient :

$$h(x) = \int_{-\infty}^{+\infty} G(t)\exp(-2j\pi tx)\left(\int_{-\infty}^{+\infty} f(y)\exp(2j\pi ty)\,\mathrm{d}y\right)\,\mathrm{d}t$$

et par suite :

$$h(x) = \int_{-\infty}^{+\infty} G(t)F(t)\exp(-2j\pi tx)\,\mathrm{d}t,$$

$h(x)$ étant une fonction bornée et la fonction $[G(t)\cdot F(t)]$ appartenant à $L^1$, il s'ensuit que $h(x)$ est la transformée de Fourier de $G(t)\cdot F(t)$.

La démonstration est encore plus simple si $f(x)$ et $G(t)$ appartiennent à $L^2$ ; en effet, on sait que les transformées de Fourier $g(x)$ et $F(t)$ appartiennent à $L^2$. Il s'ensuit que la transformée de Fourier transforme un produit de convolution en un produit simple de transformées de Fourier. On résumera les opérations ainsi :

$$\begin{aligned} f(x) &\xrightarrow{TF} F(t) \\ g(x) &\xrightarrow{TF} G(t) \\ f*g &\xrightarrow{TF} F\cdot G \\ f\cdot g &\xrightarrow{TF} F*G \end{aligned}$$

### 5.2. Remarque sur le sens de la transformée de Fourier

Du point de vue de la physique, la connaissance de la fonction $f(x)$ est strictement équivalente à la connaissance de la fonction $F(t)$ ; l'une et l'autre fonctions contiennent exactement **la même information**. Par ailleurs, l'égalité de Parseval n'est rien d'autre que l'expression du principe de conservation de l'énergie.

### 5.3. Sur la symétrie et la parité des fonctions

Si la fonction $f(x)$ possède certaines propriétés de symétrie il en va de même de sa transformée de Fourier $F(t)$. Le tableau 16.1, page suivante donne les principales correspondances.

### 5.4. Transformées de Fourier et produit scalaire

Le produit scalaire de deux fonctions $f(x)$ et $g(x)$ est défini par l'intégrale du produit des deux fonctions :

$$\langle f(x), g(x)\rangle = \int_{-\infty}^{+\infty} f(x)g(x)\,\mathrm{d}x \quad \left(= \int_{-\infty}^{+\infty} f(x)g^*(x)\,\mathrm{d}x \text{ si les fonctions sont complexes}\right).$$

**Tableau 16.1.**

| | | |
|---|---|---|
| $f(x)$ | $\overset{TF}{\longleftrightarrow}$ | $F(t)$ |
| réelle paire | $\overset{TF}{\longleftrightarrow}$ | réelle paire |
| réelle impaire | $\overset{TF}{\longleftrightarrow}$ | imaginaire impaire |
| imaginaire paire | $\overset{TF}{\longleftrightarrow}$ | imaginaire paire |
| imaginaire impaire | $\overset{TF}{\longleftrightarrow}$ | réelle impaire |
| complexe paire | $\overset{TF}{\longleftrightarrow}$ | complexe paire |
| complexe impaire | $\overset{TF}{\longleftrightarrow}$ | complexe impaire |
| réelle quelconque | $\overset{TF}{\longleftrightarrow}$ | partie réelle paire ; partie imaginaire impaire |
| imaginaire quelconque | $\overset{TF}{\longleftrightarrow}$ | partie imaginaire paire ; partie réelle impaire |
| partie réelle paire ; partie imaginaire impaire | $\overset{TF}{\longleftrightarrow}$ | réelle quelconque |
| partie réelle impaire ; partie imaginaire paire | $\overset{TF}{\longleftrightarrow}$ | imaginaire quelconque |

Si $f(x)$ et $g(x)$ appartiennent à $L^2$, ces fonctions admettent chacune une transformée de Fourier désignée respectivement par $F(u)$ et $G(u)$. Comme nous avons l'égalité de Parseval :

$$\int_{-\infty}^{+\infty} f(x)f^*(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(t)F^*(t)\,\mathrm{d}t,$$

nous allons en déduire quelques propriétés sur le produit scalaire des fonctions appartenant à $L^2$. La linéarité de l'opération d'intégration nous permet d'écrire :

$$\int_{-\infty}^{+\infty} [f(x)+g(x)]\,[f^*(x)+g^*(x)]\,\mathrm{d}x = \int_{-\infty}^{+\infty} [F(t)+G(t)]\,[F^*(t)+G^*(t)]\,\mathrm{d}t,$$

et

$$\int_{-\infty}^{+\infty} [f(x)+jg(x)]\,[f^*(x)-jg^*(x)]\,\mathrm{d}x = \int_{-\infty}^{+\infty} [F(t)+jG(t)]\,[F^*(t)-jG^*(t)]\,\mathrm{d}t.$$

La combinaison de ces deux relations jointe à l'égalité de Parseval nous donne :

$$\int_{-\infty}^{+\infty} f(x)g^*(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(t)G^*(t)\,\mathrm{d}t.$$

L'exploitation de cette dernière égalité en liaison avec la convolution donne également une expression intéressante concernant une forme de produit scalaire :

$$\langle f(x-u), g^*(x)\rangle = \int_{-\infty}^{+\infty} f(x-u)g^*(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(t)G^*(t)\exp(2\pi jtu)\,\mathrm{d}t,$$

d'après le théorème de translation sur l'axe des abscisses. Par ailleurs, nous avons les relations :

$$\int_{-\infty}^{+\infty} f(-x)\exp(2\pi jtx)\,\mathrm{d}x = F(-t)$$

$$\int_{-\infty}^{+\infty} f(-x+u)\exp(2\pi jtx)\,\mathrm{d}x = F(-t)\exp(2\pi jtu),$$

il s'ensuit :

$$\langle f(-x+u), g^*(x)\rangle = \int_{-\infty}^{+\infty} f(x-u)g^*(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(-t)G^*(t)\exp(2\pi jtu)\,\mathrm{d}t.$$

En rappelant que :

$$\int_{-\infty}^{+\infty} g^*(x)\exp(2\pi jtx)\,\mathrm{d}x = \left[\int_{-\infty}^{+\infty} g(x)\exp(-2\pi jtx)\,\mathrm{d}x\right]^* = G^*(-t),$$

et en remplaçant $g^*(x)$ par $g(x)$ on change $G^*(t)$ en $G(-t)$, on peut écrire :

$$\langle f(-x+u), g(x)\rangle = \int_{-\infty}^{+\infty} f(x-u)g(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(-t)G(-t)\exp(2\pi jtu)\,\mathrm{d}t.$$

Maintenant, remplaçons $t$ par $-t$ dans la dernière égalité, nous obtenons :

$$\langle f(u-x), g(x)\rangle = \int_{-\infty}^{+\infty} f(x-u)g(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} F(t)G(t)\exp(-2\pi jtu)\,\mathrm{d}t$$

soit encore :

$$\langle f(u-x), g(x)\rangle = \int_{-\infty}^{+\infty} F(t)G(t)\exp(-2\pi jtu)\,\mathrm{d}t.$$

Si l'on fait $u=0$, on obtient l'expression :

$$\langle f(-x), g(x)\rangle = \langle F(t), G(t)\rangle.$$

Comme $f(x)$, $g(x)$ d'une part, $F(t)$, $G(t)$ d'autre part jouent des rôles symétriques, compte tenu de la réciprocité des transformations de Fourier, on a les autres relations :

$$\begin{aligned}\langle f(x), g(-x)\rangle &= \langle F(t), G(t)\rangle,\\ \langle f(x), g(x)\rangle &= \langle F(-t), G(t)\rangle,\\ \langle f(x), g(x)\rangle &= \langle F(t), G(-t)\rangle.\end{aligned}$$

Certaines de ces relations sont utilisées dans les paragraphes évoquant l'approximation de fonctions de $L^2$ au moyen des fonctions du cylindre parabolique, fonctions rattachées aux polynômes d'Hermite (*cf.* Chapitre 10).

## 6. Sur le calcul numérique des transformées de Fourier

Il existe quelques fonctions dont on connaît la transformée de Fourier, lesquelles ont été réunies dans des tables. Malheureusement, les équations fonctionnelles (16.6) n'admettent pas toujours des expressions littérales de transcendances élémentaires. Un autre problème se pose lorsque l'on a affaire à des données expérimentales composées d'échantillons — le plus souvent distribués selon une progression arithmétique. En apparence il suffit de calculer numériquement les intégrales (16.6), mais l'expérience montre que les temps de calcul sont prohibitifs et qu'il faut trouver des algorithmes extrêmement élaborés pour parvenir à des temps de calcul raisonnables. Nous allons donc examiner le problème sous cet aspect en accordant notre attention aux fonctions échantillonnées.

## 7. Cas des fonctions échantillonnées

Puisqu'il s'agit essentiellement d'un point de vue pratique, il faut préciser que les fonctions $f(x)$ ne peuvent être connues que sur un intervalle fini $(-a/2, +a/2)$. Ceci est dû aux impératifs de nature physique et humaine : le temps d'observation est limité et la densité d'information recueillie pouvant être manipulée est finie. Par ailleurs, on supposera que la fonction $f(x)$ est donnée par un ensemble d'échantillons qui sont pris en progression arithmétique, soit $(2n+1)$ leur nombre et l'on notera $(x_k, y_k)$ les couples de points représentant la fonction. D'ores et déjà, on est en mesure de dire qu'il n'y aura aucune différence entre une transformée de Fourier calculée numériquement et la série de Fourier. En effet nous avons procédé à l'extension de l'intervalle de définition des fonctions périodiques pour parvenir à l'étude des transformées de Fourier, mais maintenant nous ne sommes en mesure que de traiter numériquement les fonctions définies sur un intervalle fini... Il va de soi que les algorithmes que nous allons décrire permettront de calculer les coefficients des développements en série de Fourier. Avant d'entreprendre à proprement parler l'étude des algorithmes fondamentaux, il nous faut rappeler les propriétés essentielles des fonctions échantillonnées et de leur transformée de Fourier.

**a.** Le pas d'échantillonnage est $h = a/(2n)$ et la période de la fonction $f(x)$ est $a$, alors, dans l'espace réciproque, la période de la transformée de Fourier $F(t)$ est $T = [-1/(2h), +1/(2h)]$ et son pas d'échantillonnage est $r = 1/a$.

**b.** Le calcul des échantillons de la transformée de Fourier $F(t)$ se fait selon une progression arithmétique de raison $r$. On calcule donc $2n$ points dans l'espace réciproque.

On démontre qu'aucune information n'est perdue en opérant de cette sorte. $2n$ points en progression arithmétique sont nécessaires et suffisants. Le fait de prendre davantage de points n'apportera rien en précision et cela ne fera qu'accroître le temps de calcul. En revanche, il existe un problème beaucoup plus délicat qui est celui de déterminer convenablement la période d'échantillonnage $h$. Nous reviendrons en détail sur ces problèmes.

## 8. Calcul par un algorithme ordinaire

Il s'agit ici de calculer les intégrales (16.6) au moyen de la méthode des rectangles ; la raison en est la suivante : comme les fonctions sont périodiques, il est indispensable d'affecter à chaque point d'échantillonnage le même poids pour obtenir le même résultat quel que soit le point choisi pour définir le début d'une période. On approche $F(t)$ au moyen de l'expression suivante :

$$F(t) = h \sum_{k=-n}^{n-1} f(kh) \exp(2\pi jkht). \tag{16.14}$$

Dans le cas où $f(x)$ est une fonction réelle, nous pouvons écrire :

$$F(t) = h \sum_{k=-n}^{n-1} f(kh) \cos(2\pi kht) + jh \sum_{k=-n}^{n-1} f(kh) \sin(2\pi kht).$$

Il est plus commode de poser :

$$\begin{aligned} & r_q = f(qh) + f(-qh) \\ & s_q = f(qh) - f(-qh) \qquad \text{avec } q = 0, 1, 2, \ldots, n-1 \\ \text{et } & r_n = f(-nh) \\ & s_n = -f(-nh). \end{aligned}$$

Cette façon de procéder revient à décomposer une fonction en une somme de deux fonctions, l'une paire et l'autre impaire ; en posant $F(t) = F_R(t) + jF_J(t)$, cela nous permet d'écrire :

$$\begin{aligned} F_R(t) &= h \sum_{k=0}^{n} r_k \cos(2\pi kht) \\ F_J(t) &= h \sum_{k=1}^{n} s_k \sin(2\pi kht). \end{aligned} \tag{16.15}$$

***Remarque*** : Nous avons déjà dit que les relations (16.15) étaient simples à calculer mais qu'elles exigeaient un temps de calcul considérable. Cependant, il y a moyen de diviser le temps de calcul par trois environ en effectuant la remarque suivante : L'argument de base des sinus et cosinus est : $\Theta_0 = 2\pi ht$, ce qui fait que les arguments des sinus et cosinus sont en progression arithmétique de raison $\Theta_0$. Comme l'appel des fonctions de bibliothèque SIN et COS demande un temps important, il est beaucoup plus astucieux de calculer ces fonctions au moyen des relations de récurrence du type :

$$\begin{aligned} \cos[(p+1)\Theta_0] &= 2\cos(\Theta_0)\cos(p\Theta_0) - \cos[(p-1)\Theta_0] \\ \sin[(p+1)\Theta_0] &= 2\cos(\Theta_0)\sin(p\Theta_0) - \sin[(p-1)\Theta_0]. \end{aligned} \tag{16.16}$$

Les relations (16.16) ne sont pas uniques, et l'on peut trouver des relations de récurrence tout à fait analogues. Quoi qu'il en soit, il suffit de calculer au départ seulement $A_0 = \cos(\Theta_0)$ et $B_0 = \sin(\Theta_0)$ puis d'utiliser les formules de récurrence :

$$\begin{aligned} A_{p+1} &= 2A_0A_p - A_{p-1} \\ B_{p+1} &= 2A_0B_p - B_{p-1}. \end{aligned}$$

Quelles que soient les formules de récurrence utilisées, elles mettent en œuvre toujours une soustraction, et l'on va observer une lente dégradation de la précision des valeurs calculées pour les sinus et cosinus. C'est pourquoi il est utile de les calibrer de temps en temps, c'est-à-dire d'utiliser les fonctions de bibliothèque tous les 1 024 points par exemple et de calculer les points intermédiaires au moyen des relations de récurrence.

La transformée de Fourier réciproque ne présente pas plus de difficulté à calculer, on écrira simplement :

$$f(x) = r \sum_{l=-n}^{n-1} F(lh) \exp(-2\pi j l h t). \tag{16.17}$$

D'une façon générale, dans les relations (16.15) et (16.17), les fonctions $f(x)$ et $F(t)$ sont des fonctions complexes de variables réelles respectivement $x$ et $t$.

## 9. L'algorithme de Cooley-Tukey (1915– )

Il s'agit d'un algorithme que l'on désigne souvent sous le nom de FFT (fast Fourier transform).

### 9.1. La transformée de Fourier discrète

Il suffit de reprendre les formules (16.15) et (16.17) en précisant que les points qui doivent être calculés sont ceux qui sont en progression arithmétique de raison $r$ ou de raison $h$ selon l'espace que l'on considère. Donc, si l'on fait $x = mr$, on obtient :

$$F(mr) = h \sum_{k=-n}^{n-1} f(kh) \exp(2\pi j k h m r);$$

mais comme $r = 1/(2nh)$ il est plus commode d'écrire :

$$F(mr) = h \sum_{k=-n}^{n-1} f(kh) \exp\left(2\pi j \frac{mk}{2n}\right). \tag{16.18}$$

Réciproquement, on a la transformée inverse :

$$f(kh) = r \sum_{q=-n}^{n-1} F(qh) \exp\left(-2\pi j \frac{kq}{2n}\right). \tag{16.19}$$

Par ailleurs, si la fonction $f(x)$ est périodique de période $a$, on aura les égalités suivantes :

$$F(mr) = h \sum_{k=-n+p_1}^{n-1+p_1} f(kh) \exp\left(2\pi j \frac{mk}{2n}\right) \tag{16.20}$$

$$f(kh) = r \sum_{q=-n+p_2}^{n-1+p_2} F(qh) \exp\left(-2\pi j \frac{kq}{2n}\right). \tag{16.21}$$

quels que soient les entiers $p_1$ et $p_2$. On s'aperçoit que le calcul des transformées de Fourier fait appel à des valeurs numériques qui sont strictement liées à la nature de l'échantillonnage.

Comme cela est sans intérêt, on préfère travailler sur une **transformée de Fourier normalisée**, c'est-à-dire que l'on va effectuer une transformation linéaire des abscisses de telle sorte que l'intervalle arbitraire $(-a/2, +a/2)$ soit transformé en un intervalle de longueur unité, soit l'intervalle $(-1/2, +1/2)$. Il faudra se souvenir de cette opération lorsque l'on désirera effectuer une interpolation dans les transformées de Fourier.

## 9.2. Normalisation de la transformée de Fourier

Pour obtenir la normalisation, il suffit donc de poser :

$$a = 2nh = 1, \quad N = 2n, \quad \text{et } p_1 = p_2 = 0.$$

Il s'ensuit que :

$$h = 1/N, \quad r = 1, \quad T = N = 1/h.$$

Rappelons que $T$ est la période de la transformée $F(t)$. Puisqu'on ne calcule qu'un nombre fini de points de la transformée, il est plus commode d'écrire en indice le numéro du point c'est-à-dire que l'on notera désormais : $F(mr) = F_m$ et $f(kh) = f_k$

Comme dans le cas du calcul des fonctions trigonométriques, nous poserons : $W_N = \exp(2\pi j/N)$. Par ce procédé, les transformées de Fourier discrètes normalisées s'écrivent simplement :

$$F_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j W_N^{jk}$$

$$f_m = \frac{1}{N} \sum_{k=0}^{N-1} F_k W_N^{-mk}. \tag{16.22}$$

Il faut bien reconnaître que, jusqu'à présent, nous n'avons pas amélioré la technique de calcul, et nous avons toujours besoin de $N$ opérations complexes pour mener à bien les calculs.

Maintenant, nous allons étudier un algorithme qui rend le temps de calcul non plus proportionnel à $N^2$ mais à $N \log_2 (N)$. Pour des raisons qui deviendront évidentes par la suite, on choisira un nombre $N$ de données qui est une puissance de deux, et l'on posera $N = 2^n$.

## 9.3. Calcul de la transformée de Fourier au moyen de la technique de partage

***Théorème*** – La transformée de Fourier d'une fonction quelconque connue en $N$ points est une combinaison linéaire d'une transformée de Fourier de deux fonctions issues de la première et ne comportant que $N/2$ points chacune.

Pour démontrer cette proposition, nous allons décomposer la fonction $f(x)$ en deux fonctions, l'une constituée des indices impairs $f_{2k+1}$ et l'autre des indices pairs $f_{2k}$. Il est bien entendu que l'on conserve l'ordre des échantillons dans la fonction $f(x)$.

On désigne par $u$ la fonction constituée par les échantillons d'indice pair et par $v$ la fonction constituée par les échantillons d'indice impair, chacune de ces fonctions comprenant $N/2$ points. Les transformées de Fourier respectives de $u$ et $v$ sont désignées par $U$ et $V$ ; on obtient leurs

expressions en utilisant les relations (16.22) :

$$U_k = \frac{2}{N} \sum_{j=0}^{N/2-1} u_j W_{N/2}^{jk} = \frac{2}{N} \sum_{j=0}^{N/2-1} u_j W_N^{2jk}$$

$$\text{et } V_k = \frac{2}{N} \sum_{j=0}^{N/2-1} v_j W_N^{2jk}.$$

Si à présent nous revenons à l'expression de $F$ dans laquelle nous séparons les parties constituées par les indices pairs et les indices impairs, nous pouvons écrire :

$$NF_k = \sum_{j=0}^{N/2-1} u_j W_N^{2jk} + \sum_{j=0}^{N/2-1} v_j W_N^{(2j+1)k}.$$

Il n'y a plus qu'à substituer les expressions de $U$ et $V$, ce qui donne :

$$2F_k = U_k + V_k W_N^k. \tag{16.23}$$

Cette expression ne permet de calculer que la moitié de la transformée de Fourier, et il faut une autre relation pour obtenir l'autre moitié. Pour cela, il suffit de remarquer que les fonctions $U$ et $V$ sont périodiques de période $N/2$, par conséquent nous pouvons dire que :

$$U_k = U_{k+N/2} \quad \text{et } V_k = V_{k+N/2}.$$

Comme par ailleurs nous avons :

$$W_N^{k+N/2} = -W_N^k = -\exp\left(2\pi j \frac{k}{N}\right),$$

nous obtenons la relation (16.24) :

$$2F_{k+N/2} = U_k - V_k W_N^k. \tag{16.24}$$

À présent on comprend l'intérêt d'avoir choisi pour $N$ une puissance de deux, puisqu'il va de soi que l'on va calculer chacune des transformées $U$ et $V$ par le même procédé.

## 9.4. Mise en œuvre de l'algorithme

Il y a plusieurs réalisations pratiques possibles pour utiliser cet algorithme.

a. Si l'on dispose d'un langage récursif, c'est le cas du Pascal, du C, du PL1... — et dans la mesure où le temps de calcul n'est pas prohibitif — il n'y aura pas de difficultés particulières pour réaliser un programme car le tri des données initiales se fera par appel récursif.

b. Si l'on ne dispose pas d'un langage récursif — c'était le cas du FORTRAN et du BASIC — il est indispensable de procéder à un tri préalable des données initiales afin de commencer les calculs de transformées par les fonctions contenant deux éléments, puis en poursuivant par les fonctions à quatre éléments, puis à huit, seize... Donc le problème qui se pose est celui de la disposition convenable des échantillons au départ, lesquels ont été pris dans l'ordre séquentiel. Autrement dit, il s'agit de savoir à quelle place (indice) il convient de mettre la donnée d'indice $k$. Il y a deux méthodes usuelles qui permettent de réaliser élégamment cette tâche à savoir le tri direct et le tri par inversion de bit.

***a – Le tri direct*** – Dans l'ordre séquentiel on considère la donnée d'indice $k$, cette donnée sera placée à l'indice $j$ ; nous allons donner la correspondance entre $k$ et $j$ en faisant l'hypothèse que le premier indice est zéro et le dernier $N-1$. Désignons par $q$ l'exposant de deux qui permet d'encadrer le nombre $k$ de telle sorte que :

$$2^q \leq k < 2^{q+1},$$

alors nous obtiendrons la valeur de $j$ au moyen de la relation de récurrence suivante :

$$j = R_N(k) = R_N(k - 2^q) + \frac{N}{2^{q+1}}$$

et l'on prendra au départ : $R_N(0) = 0$.

Cette méthode s'applique pour $k = 0, 1, 2, 3, \ldots, N-1$. La manière de l'exploiter sur un ensemble de huit données numérotées de un à huit est présentée dans le tableau 16.2.

**Tableau 16.2.**

| Indices initiaux | | | Indices finals |
|---|---|---|---|
| 0 | $q = 0$ | $R_8(0) = 0$ | 0 |
| 1 | $q = 0$ | $R_8(1) = R_8(0) + 8/2$ | 4 |
| 2 | $q = 1$ | $R_8(2) = R_8(0) + 8/4$ | 2 |
| 3 | $q = 1$ | $R_8(3) = R_8(1) + 8/4$ | 6 |
| 4 | $q = 2$ | $R_8(4) = R_8(0) + 8/8$ | 1 |
| 5 | $q = 2$ | $R_8(5) = R_8(1) + 8/8$ | 5 |
| 6 | $q = 2$ | $R_8(6) = R_8(2) + 8/8$ | 3 |
| 7 | $q = 2$ | $R_8(7) = R_8(3) + 8/8$ | 7 |

***b – Tri par inversion de bit*** – Pour des raisons de simplicité, on considère encore que les indices commencent à la valeur zéro et se terminent donc à la valeur $N-1$. Il faut donc $m = \log_2(N)$ bits pour exprimer tous les indices de la fonction échantillonnée dans le système binaire. Pour obtenir l'indice $j$, il suffit d'intervertir l'ordre des chiffres binaires représentant le nombre $k$ avec $m$ bits et l'on obtient ainsi la représentation binaire du nombre $j$.

## 9.5. Exemple

Nous avons huit données numérotées de zéro à sept et nous obtenons une représentation binaire sur trois bits. La façon d'opérer est schématisée sur le tableau 16.3, page suivante.

## 9.6. Remarques générales

1. Certaines données conservent le même indice.
2. Il est simple de passer d'une numérotation à l'autre selon que la première donnée à la valeur zéro ou la valeur un, et il suffira de retrancher ou d'ajouter une unité pour exploiter à notre convenance l'un des deux algorithmes que nous venons de présenter.

**Tableau 16.3.**

| Indices initiaux | | Indices finals | |
|---|---|---|---|
| Décimal | Binaire | Binaire | Décimal |
| 0 | 000 | 000 | 0 |
| 1 | 001 | 100 | 4 |
| 2 | 010 | 010 | 2 |
| 3 | 011 | 110 | 6 |
| 4 | 100 | 001 | 1 |
| 5 | 101 | 101 | 5 |
| 6 | 110 | 011 | 3 |
| 7 | 111 | 111 | 7 |

## 10. Programmes de calcul des transformées de Fourier

On trouvera sur le Web(*) les sous-programmes `dfft0.h` et `dfftinv0.h` pour calculer les transformées de Fourier. Nous avons retenu une méthode non récursive agrémentée d'un algorithme de tri direct des données. Ces programmes nécessitent quelques commentaires.

1. Lors du calcul de la transformée directe, le calcul successif des transformée au moyen des relations (16.23) et (16.24) impose la division des résultats par $N$. Cette opération a été réalisée au cours du tri et non pas après les calculs (les opérations sont linéaires).
2. Comme le montrent les relations (16.22), il ne faut pas diviser les données par $N$ lorsque l'on calcule la transformation réciproque.
3. Il est très facile de se retrouver en présence d'un fâcheux décalage de $\pi$ lors du calcul des transformées de Fourier. Généralement ceci est dû au fait que les données ont été mal numérotées au départ : **il faut que le premier échantillon se trouve à l'origine.** La plupart des programmes exigent cette façon de procéder. Nous ne l'avons pas retenue pour la raison suivante : lorsque l'on traite des données échantillonnées au cours du temps, la première donnée qui se présente correspond à l'abscisse normalisée $-1/2$ et non pas zéro. Dans ce cas il ne semble pas habile d'attendre d'être en possession de toutes les données pour effectuer le décalage d'abscisse avant d'entreprendre le calcul effectif de la transformée de Fourier. C'est pour cela que la transformée de Fourier directe que nous proposons tient compte de ce décalage. Par conséquent, si l'on utilise ce programme avec des données numérotées à partir de l'origine, on se trouvera également en présence d'un déphasage de $\pi$. Il convient alors de noter que **le module de la transformée de Fourier est correct**.
4. Quelle que soit la manière retenue, les résultats de la transformée de Fourier directe sont donnés à partir de l'origine. Autrement dit, dans l'espace réciproque, le premier point calculé est celui correspondant à l'origine, et l'on aura les points suivants jusqu'à la demi-période $1/(2h)$ ensuite viendront les points de la demi-période $-1/(2h)$ jusqu'à zéro non compris.
5. C'est pour les raisons expliqués au cours du paragraphe précédent que la transformée de Fourier réciproque effectue les calculs à partir de l'échantillon correspondant à l'origine.

* `http://www.edpsciences.com/guilpin/`

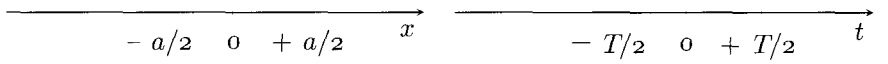


**Figure 16.1.** *Avec*

$h = a/N$ $\quad$ $r = 1/a,$
$h = 1/T$ $\quad$ $T = 1/h,$
$a = 1/r$ $\quad$ $N = T/r.$

6. Lors de la mise au point d'un programme de transformées de Fourier, il est raisonnable de réaliser simultanément la transformée directe et la transformée réciproque qui n'admettent que peu de différences. Ainsi, après avoir obtenu la transformée directe, on revient dans l'espace initial au moyen de la transformée réciproque... et l'on doit retrouver les données initiales (à partir de l'échantillon correspondant à l'origine). Cela constitue un bon test de mise au point. Par ailleurs, ce procédé permet d'apprécier de façon tangible les erreurs qui viennent altérer les résultats, car on observera de toute façon des divergences par rapport aux données initiales, divergences qu'il est souhaitable de trouver de quelques ordres de grandeur de l'erreur de troncature affectant les nombres traités en machine. Cependant, il convient de se souvenir que le calcul rapide des fonctions trigonométriques intervenant dans les calculs sont des sources d'incertitude qu'il ne faut pas négliger.

## 11. Un problème fondamental : quelle doit être la période d'échantillonnage de la fonction f(x) ?

Rappelons que la taille de la période d'échantillonnage $h$ de la fonction $f(x)$ va déterminer la taille du domaine de représentation de la transformée de Fourier $F(t)$ dans l'espace réciproque. Ajoutons que la taille du domaine de définition de $f(x)$ déterminera la taille de la fréquence d'échantillonnage r dans l'espace réciproque. Au fond, chaque fois que l'on traite de transformées de Fourier, il est indispensable d'avoir présent à l'esprit la figure 16.1.

La détermination de la période d'échantillonnage $h$, pour une expérience donnée, n'est pas un problème banal : il n'est pas question d'effectuer un échantillonnage de n'importe quelle façon quand bien même ne serions-nous pas intéressés par la partie de la transformée de Fourier au-delà d'une certaine valeur de l'abscisse $f_M$.

Il y a deux façons de procéder avec une fonction $f(t)$, laquelle a fait l'objet d'un enregistrement continu, afin d'obtenir une bonne période d'échantillonnage dans l'espace direct.

### 11.1. Détermination expérimentale de $h_0$

On effectue un échantillonnage à la période $h$ puis on calcule la transformée de Fourier $F(\omega)$. Pour que $F(\omega)$ appartienne à $L^1$ ou $L^2$, il est nécessaire qu'au moins elle soit nulle à l'infini, c'est-à-dire au bord du domaine $(-0,5, +0,5)$. (Elle peut évidemment devenir nulle avant le bord...)

Si tel n'est pas le cas, on modifie le pas d'échantillonnage $h$ pour s'assurer de la nullité de la fonction $F(\omega)$ au-delà de $|\omega_0|$. Bien entendu, on a tout intérêt à ce que l'annulation, dans l'espace normalisé, s'effectue au bord du domaine $(-0,5, +0,5)$ et non à l'intérieur pour la bonne raison qu'il est inutile de réaliser des calculs qui n'apprennent rien sur la TF. Dans ces

conditions, on déduit que l'étendue de l'espace réciproque est $(-\omega_0, +\omega_0)$ soit $2\omega_0$ et que la fréquence d'échantillonnage est :

$$r_0 = \frac{2\omega_0}{N},$$

il s'ensuit que la « bonne période » d'échantillonnage $h_0$ est donnée par l'expression :

$$h_0 = \frac{1}{2\omega_0}$$

c'est la période d'échantillonnage de Shannon (1916– ). Autrement dit, il faut échantillonner à une fréquence au moins deux fois plus grande que la plus grande fréquence apparaissant dans le signal à traiter pour ne rien perdre de l'information contenue dans le signal. Quand cette condition est remplie, Shannon a montré que l'on pouvait effectuer une **interpolation rigoureuse** $G(\omega)$ dans l'espace réciproque de la fonction $F(\omega)$ au moyen de la relation :

$$G(\omega) = \sum_{k=-\infty}^{+\infty} F(kr_0) \frac{\sin[2\pi T(\omega - kr_0)]}{2\pi T(\omega - kr_0))}.$$

***Démonstration*** – On isole une période $a$ de la fonction $f(t)$ de l'espace direct en la multipliant par une « fonction porte » $r(t)$ :

$$g(t) = f(t) \cdot r(t).$$

La transformée de Fourier de $g(t)$ notée $G(\omega)$ est continue, périodique et définie numériquement pour les valeurs $\omega = kr_0$ :

$$G(\omega) = F(\omega) * R(\omega) \quad \text{(produit de convolution)},$$

où $R(\omega)$ est la transformée de Fourier de la fonction porte, soit :

$$R(\omega) = \frac{\sin(2\pi T\omega)}{2\pi T\omega};$$

de là, nous tirons :

$$G(\omega) = \sum_{k=-\infty}^{+\infty} F(kr_0) \frac{\sin[2\pi T(\omega - kr_0)]}{2\pi T(\omega - kr_0)}.$$

## 11.2. On impose $h_0$

Dans ce cas, il est indispensable de procéder au filtrage du signal avant de l'échantillonner et il ne devra pas comporter de fréquences supérieures à :

$$\omega_0 = \frac{1}{2h_0},$$

sous peine de voir apparaître dans le spectre des « battements » avec la fréquence d'échantillonnage. C'est réellement ce problème qui est traité dans la réalisation d'un disque numérique. La fréquence d'échantillonnage est fixée aux environs de 45 kHz. Il s'ensuit que le signal ne doit pas comporter de fréquences supérieures à 22, 5 kHz ou tout au plus ces fréquences, si elles existent, auront une amplitude très atténuée — au moins 20 dB par octave.

Si tel n'est pas le cas, supposons alors qu'il existe, par exemple, dans le signal une fréquence de 30 kHz. Nous verrons, ou plutôt entendrons, une fréquence résultant du battement de cette fréquence avec la fréquence d'échantillonnage, c'est-à-dire 15 kHz. On comprend très bien que toutes les fréquences comprises entre 22,5 et 45 kHz vont battre avec la fréquence d'échantillonnage et recouvrir le spectre qui a été obtenu entre 0 et 22,5 kHz. C'est ce que certains auteurs appellent le «repliement du spectre ».

On comprend tout le danger qu'il y a à effectuer un échantillonnage sans connaître la plus grande fréquence composant le signal et qui soit non négligeable en amplitude.

## 12. La distribution de Dirac (1902–1984)

La distribution de Dirac joue un rôle intéressant dans l'étude des transformées de Fourier, et l'on consultera avec intérêt l'ouvrage d'Arsac sur les transformées de Fourier des distributions tempérées.

### 12.1. Définition

On considère la suite de fonctions :

$$f_n(x) \begin{cases} = 0 & \text{si } x \leq 0 \quad \text{et } x \geq \dfrac{1}{n} \\ = n & \text{si } 0 < x < \dfrac{1}{n}. \end{cases}$$

La distribution de Dirac, notée $\delta(x)$, est la limite de $f_n(x)$ lorsque $n$ tend vers l'infini. $\delta(x)$ est encore appelée impulsion unité car sa surface est égale à 1 (comme toutes les fonctions $f_n(x)$ par construction). L'intégrale de la fonction de Dirac est notée $Y(t)$ :

$$Y(t) = \int_{-\infty}^{t} \delta(x)\, \mathrm{d}x,$$

elle vaut 1 si $t \geq 0$ et 0 si $t < 0$. Elle s'appelle fonction de Heaviside (1850–1925) ou échelon unité.

### 12.2. Quelques propriétés

***a*** –

$$\int_{-\infty}^{+\infty} f(x)\delta(x)\, \mathrm{d}x = f(0)$$

***b*** – De même, si $a < 0$ et $b > 0$, on peut écrire :

$$\int_{a}^{b} f(x)\delta(x)\, \mathrm{d}x = f(0),$$

car $\delta(x)$ est nulle partout sauf en $x = 0$.

Réciproquement, la dérivée de la fonction de Heaviside est la fonction de Dirac.

$Y(t)$ est encore appelée la fonction échelon unité.

***c – Théorème*** – Dans l'algèbre de convolution, $\delta(x)$ joue le rôle d'élément unité :

$$T^* f = f.$$

Par exemple,

$$\int_{-\infty}^{+\infty} f(x-t)\delta(t-a)\,\mathrm{d}t = \int_{-\infty}^{+\infty} f(t)\delta_\alpha(x-t)\,\mathrm{d}t = f(x-a)$$

$\delta_\alpha(x)$ étant la fonction $\delta(x)$ qui a subi la translation $\alpha$.

***d – Transformée de Fourier de $\delta(x)$ et propriétés connexes***

1. La transformée de Fourier de la fonction $f(x) = 1$ est la fonction de Dirac.

$$\int_{-\infty}^{+\infty} 1 \exp(2\pi j\nu t)\,\mathrm{d}t = \delta(\nu)$$

2. Réciproquement :

$$\int_{-\infty}^{+\infty} \delta(\nu) \exp(-2\pi j\nu t)\,\mathrm{d}\nu = 1.$$

***Remarque*** : Considérons la fonction $\frac{1}{2}\delta(t)$ à qui l'on fait subir deux translations symétriques $\pm a$ :

$$\frac{1}{2}\delta(t-a) \quad \overset{TF}{\longleftrightarrow} \quad \frac{1}{2}\exp(-2\pi ja\nu)$$
$$\frac{1}{2}\delta(t+a) \quad \overset{TF}{\longleftrightarrow} \quad \frac{1}{2}\exp(2\pi ja\nu)$$

En additionnant membre à membre, on obtient :

$$\frac{1}{2}\left[\delta(t+a)+\delta(t-a)\right] \quad \overset{TF}{\longleftrightarrow} \quad \cos(2\pi a\nu)$$
$$\text{Si } f(t) \quad \overset{TF}{\longleftrightarrow} \quad F(\nu),$$

alors :

$$\int_{-\infty}^{+\infty} \delta(t) f(t) \exp(-2\pi j\nu t)\,\mathrm{d}t = F(0)$$

et $$\int_{-\infty}^{+\infty} \delta(t-a) f(t) \exp(-2\pi j\nu t)\,\mathrm{d}t = F(a)$$

### 12.3. La fonction « peigne » de Dirac encore appelée sha(t)

On peut montrer que la fonction (distribution) $\sum_{k=-\infty}^{\infty} \delta(t-k)$, avec $k$ entier est sa propre transformée de Fourier. On en déduit ensuite la formule de Poisson :

$$\sum_{k=-\infty}^{\infty} f(k) = \sum_{n=-\infty}^{\infty} F(n),$$

laquelle permet d'écrire :

$$\sum_{k=-\infty}^{\infty} \delta(t-k) = \sum_{n=-\infty}^{\infty} \exp(-2\pi jtn).$$

La fonction sha$(t)$ permet de représenter la fonction d'échantillonnage d'une fonction quelconque à la fréquence $F_e$ (ou à la période $T_e$) :

$$\text{sha}(t) = T_e \sum_{k=-\infty}^{\infty} \delta(t - kF_e).$$

Soit $f(t)$ la fonction que l'on désire échantillonner dont la forme échantillonnée s'appelle $\Phi(t)$ et dont la transformée de Fourier est $F(\nu)$ :

$$\Phi(t) = T_e \sum_{k=-\infty}^{\infty} f(kT_e)\delta(t-kT_e) = f(t)T_e \sum_{k=-\infty}^{\infty} \delta(t-kT_e).$$

La formule de Poisson permet d'écrire :

$$\Phi(t) = F(\nu) * \sum_{n=-\infty}^{+\infty} \delta(\nu - nF_e).$$

## 13. Transformées de Fourier multidimensionnelles

Il n'y a pas de difficultés à traiter des transformées de Fourier à plusieurs dimensions, car les grands principes que nous avons évoqués à propos des transformées de Fourier à une dimension demeurent. Nous allons nous intéresser aux transformées à deux dimensions dont l'intérêt évident concerne le traitement des images, quoique d'autres applications soient également bien utiles.

Dans un espace cartésien à deux dimensions, la transformée directe s'écrit :

$$f(u,v) = \int_{-\infty}^{+\infty}\int G(x,y) \exp\left[2\pi j(ux+vy)\right] \, \mathrm{d}x \, \mathrm{d}y$$

ainsi que la transformée réciproque sous réserve de son existence :

$$G(x,y) = \int_{-\infty}^{+\infty}\int f(u,v) \exp\left[-2\pi j(ux+vy)\right] \, \mathrm{d}u \, \mathrm{d}v.$$

Il est possible encore d'écrire :

$$f(u,v) = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} G(x,y) \exp(2\pi jux)\, \mathrm{d}x \right\} \exp(2\pi jvy)\, \mathrm{d}y$$

$$\text{et } G(x,y) = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} f(u,v) \exp(-2\pi jux) du \right\} \exp(-2\pi jvy)\, \mathrm{d}v.$$

Dans la pratique, le domaine $D$ d'intégration est fini et rectangulaire dans un espace cartésien. Il s'ensuit que l'on pourra utiliser l'algorithme de Cooley-Tukey en effectuant un maillage du domaine $D$ qui comporte $2^n$ points sur l'axe des $x$ et $2^m$ points sur l'axe des $y$.

Ainsi on calculera toutes les transformées de Fourier ligne à ligne suivies de toutes les transformées effectuées colonne à colonne. Cela nécessite un aménagement de la transformée de Fourier à une dimension en ce sens que les adresses des éléments ainsi que la table des cosinus ne sont à calculer qu'une seule fois pour chacune des deux dimensions.

Sur le Web(*), on trouvera le sous-programme `dfft2.h` qui réalise la transformée de Fourier directe à deux dimensions. La transformée réciproque `dfftinv2.c` est obtenue en modifiant ce programme de la façon suivante : on supprime la normalisation, puis on change convenablement les signes exactement de la même manière que ce qui a été fait à propos des transformées à une dimension.

## 14. Éléments de bibliographie

J. ARSAC (1961) *Transformation de Fourier et théorie des distributions*, Éditions Dunod.

J. BASS (1971) *Cours de mathématiques*, Tome III, Éditions Masson.

O. BRIGHAM (1988) *The fast Fourier transform and its applications*, Prentice Hall.

H. DELOUIS (1973) *Exploitation des spectres obtenus par transformation de Fourier, Thèse de Doctorat d'État*, Paris.

H. LEBESGUE (1903) *Leçons sur l'intégration et la recherche des fonctions primitives*, Gauthier-Villars, réédité aux Éditions Jacques Gabay.

L. SCHWARTZ (1965) *Méthodes mathématiques pour les sciences physiques*, Hermann.

* `http://www.edpsciences.com/guilpin/`

# 17 Initiation aux problèmes mal posés : équations intégrales, systèmes linéaires mal conditionnés et équations de convolution

La résolution numérique de certains problèmes mathématiques peut amener un type particulier d'ennuis lié à l'instabilité des solutions vis-à-vis de faibles variations des données initiales. En d'autres termes, des variations des données aussi petites que l'on veut peuvent amener des variations aussi grandes que l'on veut de la solution. C'est bien ce type de problème qui est évoqué par les météorologistes quand ils disent qu'un battement d'aile de papillon peut provoquer un cyclone n'importe où...

Grosso modo, la solution approchée de ce type de problème n'est pas unique dans les limites de précision que l'on s'est fixé et devient difficile à interpréter.

C'est Hadamard (1865–1963) qui le premier a mis en évidence ce type de problèmes, dits mal posés aujourd'hui, et l'on trouvera en bibliographie ses articles originaux. On peut ajouter que ces problèmes ont fait l'objet d'une récente actualité au cours du développement de la théorie du chaos.

Ce chapitre doit énormément à l'ouvrage de référence écrit par Tikhonov (1906–1993) et Arsénine aux Éditions de Moscou (*cf.* bibliographie). La lecture de ce livre demande une solide connaissance en mathématique, et nous avons seulement cherché dans ce chapitre à résumer la philosophie sous-jacente. Cependant, dans une première approche de ces problèmes, la lecture des exemples et l'usage des programmes sont amplement suffisants.

## 1. Un exemple de problème mal posé : le calcul des séries de Fourier à coefficients approchés dans $L^2$

Considérons une série de Fourier convergente que nous écrivons :

$$f(x) = \sum_{n=0}^{\infty} a_n \cos(nx)$$

et supposons que chaque coefficient $a_n$ soit entaché d'une erreur $\varepsilon/n$ (à l'exception de $a_0$ évidemment) ; nous écrivons donc :

$$c_n = a_n + \varepsilon/n \qquad \text{et} \qquad c_0 = a_0.$$

La fonction $f(x)$ est donc remplacée par la fonction $g(x)$ :

$$g(x) = \sum_{n=0}^{\infty} c_n \cos(nx).$$

Dans la métrique de $L^2$, les coefficients diffèrent de :

$$E_1 = \left\{\sum_{n=0}^{\infty}(c_n - a_n)^2\right\}^{1/2} = \varepsilon\left\{\sum_{n=1}^{\infty}\frac{1}{n^2}\right\}^{1/2} = \frac{\varepsilon\pi}{\sqrt{6}},$$

par conséquent, $E_1$ est aussi petit que l'on veut. D'autre part, si la distance entre les fonctions $f(x)$ et $g(x)$ est donnée par la norme de la convergence uniforme, nous avons :

$$\sup|g(x) - f(x)| = \sup\left|\varepsilon\sum_{n=1}^{\infty}\frac{\cos(nx)}{n}\right| = |\varepsilon|\sum_{n=1}^{\infty}\frac{1}{n},$$

cette quantité est aussi grande que l'on veut.

***Remarque*** : Si l'on choisit la norme de la convergence en moyenne quadratique, le problème devient bien posé. En effet :

$$\left\{\int_0^{\pi}[f(x) - g(x)]^2\,\mathrm{d}x\right\}^{1/2} = \left\{\frac{\pi}{2}\sum_{n=1}^{\infty}(c_n - a^n)^2\right\}^{1/2} = E^1\sqrt{\frac{\pi}{2}}\,.$$

## 2. L'équation intégrale de Fredholm (1866–1927) de première espèce

L'équation de Fredholm de première espèce constitue un problème mal posé. Considérons l'équation fonctionnelle suivante :

$$\int_a^b K(x,t)z(t)\,\mathrm{d}t = u(x) \qquad \text{pour } x \text{ appartenant à } (c,d),$$

expression dans laquelle $z(t)$ est une fonction inconnue de l'espace $F$ des fonctions continues sur $(a,b)$ et $u(x)$ une fonction connue de l'espace $U$. Mentionnons au passage que **l'équation de convolution** est un cas particulier de cette équation fonctionnelle, $K(x,t)$ devenant $K(x,-t)$.

Ajoutons une hypothèse supplémentaire sur le noyau $K(x,t)$ qui est connu : c'est une fonction continue en $x$ qui possède une dérivée partielle $\partial K(x,t)/\partial x$ également continue.

Si la solution n'est qu'une solution approchée, la mesure de l'écart du second membre s'effectue, par exemple, dans une métrique quadratique $\mu_u$ que nous écrivons ($U$ est l'espace des fonctions de carré sommable $L^2$) :

$$\mu_u(u_1,u_2) = \left\{\int_c^d[u_1(x) - u_2(x)]^2\,\mathrm{d}x\right\}^{1/2}.$$

L'écart de la solution $z(t)$ peut être mesuré, par exemple, dans la métrique $\mu_z$ de la convergence uniforme ($F$ est l'espace des fonctions $L^\infty$) :

$$\mu_z(z_1,z_2) = \sup\ |z_1(t) - z_2(t)| \qquad \text{pour } t \text{ appartenant à } (a,b).$$

À présent, montrons que le problème est un problème mal posé. Supposons que pour un second membre $u_1(x)$ nous connaissions une solution exacte $z_1(t)$ ; on peut alors écrire :

$$\int_a^b K(x,t)z_1(t)\,\mathrm{d}t = u_1(x).$$

Maintenant, supposons que l'on connaisse un second membre approché peu différent de $u_1(x)$ dans la métrique $L^2$, et l'on se propose de rechercher une solution voisine de $z_1(t)$.

Considérons la fonction $z_2(t) = z_1(t) + \Gamma\sin(\omega t)$, elle est solution de l'équation :

$$\Gamma\int_a^b K(x,t)\sin(\omega t)\,\mathrm{d}t + u_1(x) = u_2(x).$$

Cette expression permet de calculer la distance de $[u_1(x) - u_2(x)]$ :

$$\mu_u(u_1,u_2) = |\Gamma|\left\{\int_c^d\left[\int_a^b K(x,t)\sin(\omega t)\,\mathrm{d}t\right]^2\mathrm{d}x\right\}^{1/2},$$

cette expression peut être rendue aussi petite que l'on veut pourvu que $\omega$ soit suffisamment élevé. Calculons maintenant la distance des solutions correspondantes :

$$\mu_z(z_1,z_2) = \sup|z_1(t) - z_2(t)| = \sup|\Gamma\sin(\omega t)| = |\Gamma| \quad \text{pour } t \text{ appartenant à } (a,b).$$

et cet écart pourra être arbitrairement grand.

***Remarque*** : Rien n'est changé si la norme $\mu_z(z_1,z_2)$ est remplacée par l'autre norme $\mu_u(z_1,z_2)$ ; on calcule alors :

$$\mu_u(z_1,z_2) = \left\{\left[\int_a^b z_1(t) - z_2(t)\right]^2\mathrm{d}t\right\}^{1/2} = |\Gamma|\left[\int_a^b \sin^2(\omega t)\,\mathrm{d}t\right]^{1/2}$$
$$= |\Gamma|\left[\frac{b-a}{2} - \frac{1}{2\omega}\sin[w(b-a)]\cos[\omega(b-a)]\right]^{1/2}.$$

Cette dernière expression montre que $\Gamma$ et $\omega$ peuvent être choisis de telle façon que la distance $\mu_u(z_1,z_2)$ puisse être arbitrairement grande. Ici le problème n'a pas été modifié par le choix de la norme, mais ce n'est pas une loi générale.

La conclusion peut s'énoncer de la manière suivante : la recherche d'une solution de l'équation intégrale ne peut pas être déterminée par la condition :

$$\mu_u(A_z,u) \leq \delta$$

expression dans laquelle :

**a.** l'opérateur intégral est désigné par $A$, ce qui permet d'écrire formellement $Az = u$ à la place de l'équation intégrale.

**b.** $\delta$ est un infiniment petit positif.

# 3. Notion de problèmes bien et mal posés

Résoudre le problème linéaire $Az = u$, c'est trouver sa solution $z = \Re(u)$ à partir des données initiales $u(x)$. Il va de soi que l'on travaille dans des espaces métriques et que la distance est suggérée par le type de problème traité.

Supposons que la « solution » soit définie, et qu'à tout $u$ appartenant à $U$ corresponde une solution unique $z = \Re(u)$ appartenant à $F$.

Par définition, on dit que **le problème de la recherche d'une solution est stable vis-à-vis des données initiales** si $\forall \varepsilon > 0$, il existe $\delta(\varepsilon) > 0$ tel que :

$$\begin{aligned} \mu_u(u_1, u_2) \leq \delta(\varepsilon) \quad &\text{entraîne} \quad \mu_z(z_1, z_2) \leq \varepsilon \\ \text{avec} \quad z_1 = \Re(u_1) \quad &\text{et} \quad z_2 = \Re(u_2), \end{aligned}$$

$u_1$ et $u_2$ appartenant à $U$ et $z_1$ et $z_2$ appartenant à $F$.

Par définition, **le problème est bien posé sur les espaces $U$ et $F$** si l'on vérifie les conditions suivantes :

1. $\forall u$ appartenant à $U$, il existe une solution $z$ appartenant à $F$.
2. La solution est définie de façon unique.
3. Le problème est stable sur les espaces $U$ et $F$.

Dans le cas contraire, on dit que le **problème est mal posé.**

### Quelques remarques

1. Le rôle de la troisième condition est fondamental pour l'exploitation des méthodes numériques.
2. La recherche d'une solution approchée d'un problème mal posé est généralement non univoque. Supposons que l'équation $Az = u$ soit mal posée dans l'espace des fonctions $F$, même si $A$ est un opérateur absolument continu, $A^{-1}$ ne sera pas en général un opérateur continu sur $U$ et la solution ne sera pas stable. De plus on suppose que le second membre $u_{ap}$ est connu à $\delta$ près tel que $\mu_u(u_{ex}, u_{ap}) \leq \delta$, ($u_{ex}$ étant la solution exacte et $u_{ap}$ étant la solution approchée). Naturellement on recherche la solution approchée dans la classe $Q_\delta$ de $z$ pour laquelle la distance $\mu_u(Az, u_{ap}) \leq \delta$ ; malheureusement, la classe $Q_\delta$ est en général trop vaste et il faut introduire une **contrainte** c'est-à-dire un **principe de sélection** sur les solutions possibles. Pour ce faire, on exploite une information supplémentaire sur la solution qui peut être par exemple la régularité des solutions.

# 4. Méthode de régularisation

À présent nous envisageons le cas où $F$ la classe des solutions possibles n'est plus compacte, ainsi les variations du second membre de l'équation $Az = u$ sont susceptibles de sortir des frontières de $AF$ ; on dit alors qu'il s'agit de problèmes essentiellement mal posés, et A. Tikhonov a proposé une méthode de résolution fondée sur la notion d'opérateur régularisant.

## 4.1. Notion d'opérateur régularisant

L'opérateur$A^{-1}$ n'est plus continu sur $AF$ et l'ensemble des solutions de $F$ n'est plus compact. Même dans le cas où l'on est en présence d'un second membre approché $u_{ap}$ appartenant à $U$ tel que $\mu_u(u_{ex}, u_{ap}) \leq \delta$, on sait que la solution :

$$z_{ap} = A^{-1} u_{ap}$$

n'approche pas la solution $z_{ex}$. Pour pallier cet inconvénient, on définit un opérateur régularisant qui dépend de $\delta$, de telle sorte que, si $u_{ap}$ tend vers $u_{ex}$ c'est-à-dire que $\delta$ tend vers zéro cela doit entraîner que $z_{ap}$ tend vers $z_{ex}$ dans les métriques appropriées.

***Définition*** – On dit que l'opérateur $\mathbf{R}(u, \delta)$ est régularisant dans le voisinage de $u = u_{ex}$ pour l'équation $Az = u$, s'il possède les propriétés suivantes :

1. Il existe $\delta_1 > 0$ tel que $\mathbf{R}(u, d)$ soit défini quel que soit $\delta \in (0, \delta_1)$ et pour tout $u_\delta$ tel que $\mu_u(u_{ex}, u_\delta) \leq \delta$.
2. Quel que soit $\varepsilon > 0$, il existe $\delta_0(\varepsilon, u_{ex}) \leq \delta_1$ tel que l'inégalité :

$$\mu_u(u_{ex}, u_\delta) \leq \delta \leq \delta_0 \text{ entraîne l'inégalité : } \mu_z(z_\delta, z_{ex}) \leq \varepsilon \text{ avec } z_\delta = \mathbf{R}(u_\delta, \delta).$$

On note que nous ne faisons pas l'hypothèse d'un opérateur univoque, il peut exister un ensemble $\{\mathbf{R}(u_\delta, \delta)\}$ et $z_\delta$ est un de ces éléments.

À la place de $\delta$, on peut substituer un **paramètre de régularisation** noté $\alpha$ ($\alpha \geq 0$), qui permet d'apporter une définition plus commode.

**a.** $\mathbf{R}(u, \alpha)$ est défini quel que soit $\alpha \geq 0$ et pour tout $u_\delta \in U$ tel que :

$$\mu_u(u_{ex}, u_\delta) \leq \delta \leq \delta_1.$$

**b.** Il existe d'une part $\alpha$ fonction de $\delta$ tel que, quel que soit $\varepsilon > 0$, et il existe d'autre part $\delta(\varepsilon) \leq \delta_1$ tel que si $u_\delta \in U$ et $\mu_u(u_{ex}, u_\delta) \leq \delta(\varepsilon)$, cela entraîne :

$$\mu_z(z_\delta, z_{ex}) \leq \varepsilon \text{ avec } z_{ap} = \mathbf{R}[u_\delta, \alpha(\delta)].$$

On choisit comme solution approchée la solution $z_\delta$ obtenue au moyen de l'équation $z_\delta = \mathbf{R}[u_\delta, \alpha(\delta)]$, équation dans laquelle le paramètre régularisant est lié à l'erreur entachant les données initiales $u_\delta$.

Il nous reste donc à chercher des opérateurs régularisants et à définir le paramètre $\alpha$ connaissant l'erreur $\delta$ entachant le deuxième membre.

Nous n'allons pas rentrer dans les détails de l'obtention d'opérateurs régularisants dont nous donnerons quelques exemples en nous limitant aux idées générales. La construction d'opérateurs régularisants est fondée sur le **principe variationnel** qui minimise une fonctionnelle stabilisatrice $\Omega(z)$ non négative continue et définie sur un sous-ensemble $F_1$ dense partout sur $F$. En outre, cette fonctionnelle est telle que l'ensemble des éléments $z$ de $F_1$ pour lesquels $\Omega(z) \leq \delta$ quel que soit $ap > 0$ est un compact sur $F_1$. On cherche la solution approchée dans l'ensemble $M$ de $F$ tel que :

$$\mu_u(Az, u_\delta) \leq \delta,$$

comme la solution n'est pas continue en $\delta$, on ne peut pas retenir un élément arbitraire de $M$. En général, $M$ est très grand et c'est notre principe de sélection variationnel qui va nous permettre de choisir un ou plusieurs éléments de $M$ qui dépendent continûment de $\delta$.

Revenons à notre fonctionnelle stabilisatrice $\Omega(z)$ définie sur l'ensemble $F_1$ de $F$ et désignons par $F_{1\delta}$ l'ensemble des éléments de $M$ sur lequel $\Omega(z)$ est défini ; nous pouvons écrire : $F_{1\delta} = M \cap F_1$. L'exploitation du principe de sélection consiste à rechercher les éléments de $F_{1\delta}$ qui minimisent la fonctionnelle $\Omega(z)$.

Dans le cas général, il est possible de montrer que, plutôt que de chercher à minimiser la fonctionnelle stabilisatrice $\Omega(z)$ sur $F_{1\delta}$, il est équivalent de chercher à minimiser $\Omega(z)$ sur $F_1$ avec la contrainte :

$$\mu_u(Az, u_\delta) = \delta.$$

Ce problème se traite au moyen de la méthode des multiplicateurs de Lagrange (1736–1813). Autrement dit, nous minimisons la fonctionnelle suivante :

$$M_\alpha(z, u_\delta) = \mu_u^2(Az, u_\delta) + \alpha\Omega(z),$$

cependant, il faut qu'il existe un paramètre $\alpha$ et un élément $z_\alpha$ tels que non seulement $M_\alpha(z, u_\delta)$ atteigne sa borne inférieure, mais encore qui assurent la condition :

$$\mu_u(Az_\alpha, u_\delta) = \delta.$$

On peut introduire formellement la fonctionnelle $M_\alpha(z, u_\delta)$ et chercher l'élément $z_\alpha$ susceptible de constituer le minimum sur $F_1$ et alors, le paramètre $\alpha$ qu'il convient de déterminer se présente comme une fonction de $\delta$. D'une façon générale, $M_\alpha(z, u_\delta)$ s'appelle fonctionnelle lissante.

Il est tout à fait équivalent de considérer que $z_\alpha$ est donné par l'application d'un certain opérateur $\mathbf{R}[u_\delta, \alpha(\delta)]$ dépendant de tel que :

$$z_\delta = \mathbf{R}[u_\delta, \alpha(\delta)],$$

où $\alpha$ **se détermine au moyen du résidu** comme on le verra dans le paragraphe 4.4 qui suit.

Le choix de la fonctionnelle stabilisatrice est guidé par la nature du problème à traiter et cela sera mis en évidence sur les exemples que nous présenterons.

## 4.2. Théorème important (sans démonstration)

Si un sous-ensemble $\phi$ de $F$ admet une métrique $\mu_\phi(z_1, z_2)$ qui majore la métrique $\mu_F(z_1, z_2)$, il est possible de montrer que, sous réserve que la sphère $\mu_\phi(z, z_0) \leq \rho$ (de centre $z_0$) soit compacte dans $F$, il existe alors un élément $z_\alpha$ appartenant à $\phi$ qui minimise la fonctionnelle :

$$M_\alpha(z, u_\delta) = \mu_u^2(Az, u_\delta) + \alpha\Omega(z), \quad \text{avec } \Omega(z) = \mu_\phi^2(z, 0).$$

***a – Exemple 1 –*** Si $F$ est muni de la métrique :

$$\mu_F(z_1, z_2) = \sup |z_1(x) - z_2(x)|$$

pour $x$ appartenant à $(a, b)$, il est possible de choisir l'ensemble $\phi$ des fonctions continûment dérivables sur $(a, b)$ muni de la métrique :

$$\mu_\phi(z_1, z_2) = \sup \{|z_1(x) - z_2(x)| + |z_1'(x) - z_2'(x)|\}$$

pour $x$ appartenant à $(a, b)$.

***b* – *Exemple 2*** – Si $F$, espace des fonctions continues sur $(a, b)$, est muni de la métrique :

$$\mu_F(z_1, z_2) = \sup |z_1(x) - z_2(x)|$$

on peut choisir l'ensemble $\phi$ des fonctions de carré sommable qui admettent des dérivées jusqu'à l'ordre $p$ sur $(a, b)$ (espace de Sobolev (1908–1989) $W_2^p$) muni de la métrique :

$$\mu_\phi(z_1, z_2) = \left\{ \int_a^b \sum_{k=0}^{p} q_k(x) \left( \frac{d^k z(x)}{\mathrm{d}x^k} \right)^2 \mathrm{d}x \right\}^{1/2}$$

expression dans laquelle les $q_k(x)$ sont des fonctions continues connues $\geq 0$, avec toutefois $q_p(x) > 0$. On aura donc :

$$\Omega(z) = \int_a^b \sum_{k=0}^{p} q_k(x) \left( \frac{d^k z(x)}{\mathrm{d}x^k} \right)^2 \mathrm{d}x,$$

et comme la solution régularisée $z_\alpha$ minimise la fonctionnelle stabilisante $\Omega(z)$ avec la contrainte $\mu_u(Az_\alpha, u_\delta) = \delta$, alors, la solution sera approchée par les fonctions les plus lisses jusqu'à l'ordre $p$.

Les stabilisateurs s'appellent dans ce cas stabilisateurs d'ordre $p$, et si les fonctions $q_k(x)$ sont des coefficients constants $\geq 0$, on parle alors de stabilisateurs d'ordre $p$ à coefficients constants.

## 4.3. Recherche de la solution régularisée sous forme d'une série

On suppose ici que $A$ est un opérateur complètement continu de $F$ dans $U$, et que $U$ et $F$ sont des espaces hilbertiens. En outre, on suppose que le sous-ensemble $\phi$ de $F$ est aussi un espace hilbertien qui est muni de la métrique majorante $\|z\|$ pour laquelle l'ensemble des éléments $z$ de $\phi$, tels que $\|z\| \leq d$, est compact dans $F$. On peut alors prendre comme stabilisateur la fonctionnelle :

$$\Omega(z) = \|z\|^2.$$

Puisqu'il s'agit d'un problème variationnel, on écrit l'équation d'Euler associée à la fonctionnelle lissante $M_\alpha(z, u)$. On montre qu'elle se met sous la forme :

$$A^*Az + \alpha z = A^*u,$$

où $A^*A$ est un opérateur auto-adjoint.

Si l'on désigne par $\{f_k\}$ un système complet de l'opérateur $A^*A$ (vecteurs propres) et par $\{\lambda_k\}$ les valeurs propres associées, on peut mettre $A^*u$ sous la forme d'une série :

$$A^*u = \sum_{k=0}^{\infty} c_k f_k,$$

et $z$ sous la forme :

$$z = \sum_{k=0}^{\infty} b_k f_k,$$

expression dans laquelle les $\{b_k\}$ sont donnés par les expressions :

$$b_k = \frac{c_k}{\lambda_k + \alpha}.$$

### 4.4. Choix du paramètre de régularisation

Il n'est pas évident de connaître la fonction $\alpha(\delta)$ pour laquelle l'opérateur $\mathbf{R}[u_\delta, \alpha(\delta)]$ soit effectivement régularisant, cependant, on connaît une valeur de $\delta$ caractérisant l'imprécision générale et il nous faut trouver une valeur correspondante de $\alpha$ choisie parmi les valeurs possibles. On connaît la valeur $\delta$ d'après la relation :

$$\mu_u(u_{ex}, u_\delta) \leq \delta.$$

Alors, on peut estimer $\alpha$ d'après le résidu, c'est-à-dire au moyen de la condition :

$$\mu_u(Az_\alpha, u_\delta) = \delta.$$

D'un point de vue pratique, on va opérer de la façon suivante :

1. On estime $\delta$ l'erreur sur le second membre.
2. On va choisir une suite de valeur de $\alpha$ qui sont en progression géométrique, soit :

   $$\alpha_k = \alpha_0 q^k$$

   et pour chaque $\alpha_k$ on calcule la valeur de la fonction $z_{\alpha_k}$ qui minimise la fonctionnelle lissante $M_{\alpha_k}(z, u_\delta)$.
3. On calcule le résidu $r_k$ au moyen de la relation :

   $$r_k = \mu_u(Az_{\alpha_k}, u_\delta).$$

4. Pour valeur de $\alpha$, on choisit celle pour laquelle on vérifie l'égalité :

   $$\mu_u(Az_{\alpha_k}, u_\delta) = \delta.$$

5. La méthode donne l'idée concernant la façon dont il faut choisir $\alpha_0$. En programmation, il est commode de procéder par tâtonnements de façon conversationnelle, afin de déceler plus rapidement les ordres de grandeur.

## 5. Application à la résolution approchée des équations intégrales de Fredholm de première espèce

Revenons à l'équation de Fredholm de première espèce :

$$\int_a^b K(x,t)z(t)\,\mathrm{d}t = u(x)$$

pour $x$ appartenant à $(c, d)$, et $u(x)$ appartenant à l'espace $L^2$ des fonctions de carré sommable. Pour obtenir la solution, nous allons utiliser un stabilisateur du premier ordre dans l'espace de Sobolev $W_2^1$. Nous cherchons la solution $z_\alpha$ qui minimise la fonctionnelle :

$$M_\alpha(z,u) = \int_c^d \left\{ \int_a^b K(x,t)z(t)\,\mathrm{d}t - u(x) \right\}^2 \mathrm{d}x + \alpha \int_a^b \left\{ q_0(t)z^2(t) + q_1(t) \left( \frac{\mathrm{d}z(t)}{\mathrm{d}t} \right)^2 \right\} \mathrm{d}t.$$

Il est possible de montrer que la solution régularisée est donnée par la résolution de l'équation intégro-différentielle :

$$\int_a^b H(x,t)z(t)\,\mathrm{d}t - \alpha\left\{\frac{\mathrm{d}}{\mathrm{d}x}\left[q_1(x)\frac{\mathrm{d}z(x)}{\mathrm{d}x}\right] - q_0(x)z(x)\right\} = b(x),$$

avec :

$$H(x,t) = \int_c^d K(y,x)K(y,t)\,\mathrm{d}y,$$

$$b(x) = \int_c^d K(y,x)u(y)\,\mathrm{d}y,$$

expressions auxquelles il faut ajouter la condition :

$$q_1(x)z'(x)v(x)\big|_a^b = 0,$$

où $v(x)$ est une variation arbitraire de $z(x)$ telle que $z(x) + v(x)$ puisse appartenir à la classe des fonctions admissibles.

On peut satisfaire cette condition de deux manières :

**a.** on connaît les valeurs de la solution aux bornes de l'intervalle $(a,b)$ et il faut que la fonction $v(x)$ s'annule à ces extrémités ;

**b.** si ces valeurs ne sont pas connues, on peut alors poser $z'(a) = z'(b) = 0$, ce qui constitue deux contraintes à imposer à la solution approchée.

## 5.1. Résolution numérique

Comme chaque fois en pareil cas, on procède à un maillage du domaine sur lequel on désire effectuer la résolution de l'équation intégro-différentielle. Généralement on adopte un maillage uniforme de pas $h$ sur l'intervalle $(a,b)$, ainsi la solution $z(x)$ est calculée pour les valeurs en progression arithmétique $x_k = a + kh$ avec $k = 0, 1, 2, \ldots, n$ et nous notons $z_k$ la valeur inconnue $z(x_k)$. Nous pouvons également écrire quelques expressions utiles pour la formulation du problème transposé sur le plan numérique :

$$\int_a^b K(x_k,t)z(t)\,\mathrm{d}t = h\sum_{\nu=0}^{n} K_{k,\nu}z_\nu,$$

l'expression du second membre représentant formellement une technique d'intégration usuelle (trapèze, rectangle, Simpson...).

Puis, en utilisant les opérateurs de différences première et deuxième :

$$\begin{aligned}\alpha\frac{\mathrm{d}}{\mathrm{d}x}\left[q_1(x)\frac{\mathrm{d}z(x)}{\mathrm{d}x}\right] &= \frac{\alpha}{h^2}\left[q_1(x_k)(z_{k+1}-z_k) - q_1(x_{k-1})(z_k - z_{k-1})\right]\\ &= \frac{\alpha}{h^2}\left[q_{1,k}z_{k+1} + q_{1,k-1}z_{k-1} - z_k(q_{1,k}+q_{1,k-1}) - z_{k-1})\right],\\ \text{et}\quad \alpha q_0(x)z(x) &= \alpha q_{0,k}z_k.\end{aligned}$$

Ainsi, nous pouvons écrire un système d'équations aux différences finies :

$$h\sum_{\nu=0}^{\nu} K_{k,\nu}z_\nu - \frac{\alpha}{h^2}\left[q_{1,k}z_{k+1} + q_{1,k-1}z_{k-1} - z_k(q_{1,k}+q_{1,k-1}) - z_{k-1}\right] + \alpha q_{0,k}z_k = b_k$$

pour $k = 1, 2, \ldots, n-1$. On obtient donc un système de $n-1$ équations à $n-1$ inconnues à partir du moment où l'on connaît $z_0$ et $z_n$.

Si l'on ne connaît pas $z_0$ et $z_n$, on pose $z'(a) = z'(b) = 0$, et ces conditions seront imposées en écrivant le système pour $k = 0, 1, 2, \ldots, n$, et l'on posera $z_{-1} = z_0$ et $z_{n+1} = z_n$. Ceci assure que les dérivées numériques en $a$ et $b$ sont nulles.

Ici encore, il faut résoudre chaque fois un système linéaire pour une valeur de $\alpha$. Pour déterminer la solution, on peut utiliser la méthode du résidu.

### 5.2. Exemple numérique

On se propose de résoudre numériquement l'équation de Fredholm de première espèce suivante :

$$\int_{-1}^{+1} K(x,y)z(y)\,\mathrm{d}y = u(x)$$

$$\begin{aligned} \text{avec}\quad K(x,y) &= (1.0-x)y \qquad \text{si } 0 \le y \le x \\ K(x,y) &= (1.0-y)x \qquad \text{si } x \le y \le 1 \\ \text{et}\quad u(x) &= [\sin(\pi x)]^3 \qquad \text{sur l'intervalle } (a,b) = (-1,1). \end{aligned}$$

La solution est connue analytiquement, et l'on la comparera aux résultats numériques :

$$u(y) = \frac{3}{4}\pi^2[\sin(\pi x) - 3\sin(3\pi x)].$$

On trouvera sur le Web(*) le programme `fredh_1.c` qui résout ce problème dans lequel les intégrales sont évaluées au moyen de la méthode de Simpson.

## 6. Résolution d'un système linéaire mal conditionné

Soit $AX = B$ un système linéaire le plus général. Nous considérons ici les systèmes d'équations algébriques pour lesquels de petites perturbations du second membre peuvent induire des variations inacceptables de la solution. Il se peut également que la matrice $A$ ait un déterminant très proche de zéro, dans ce cas, comme les calculs sont effectués avec une précision finie, nous ne serons pas en mesure de savoir si le système est réellement dégénéré ou non. Quoi qu'il en soit, il existe une classe de systèmes linéaires qui sont indiscernables entre eux pour un niveau d'erreur connu.

On va rechercher la solution, appelée solution normale, parmi les vecteurs $X$ qui vérifient l'inégalité :

$$[AX - u_{ap}] \le \delta,$$

$$\text{avec :} \qquad [Z] = \left\{\sum_{j=1}^{n} z_j^2\right\}^{1/2}.$$

* `http://www.edpsciences.com/guilpin/`

L'application de la méthode des régularisants nous conduit à minimiser la fonctionnelle lissante :

$$M_\alpha(X, u_{ap}, A) = [AX - u_{ap}]^2 + \alpha[X]^2,$$

$\alpha$ se déterminant par la condition :

$$[AX_\alpha - u_{ap}] = \delta.$$

Sans entrer dans les détails, on obtient le système linéaire :

$$(A^T A + \alpha I) = A^T u_{ap}$$

expression dans laquelle $A^T$ est la matrice transposée de $A$, et $I$ la matrice unité d'ordre identique à celui de $A$. Généralement, ici encore, $\alpha$ se trouve déterminé par le résidu.

***Remarque*** : On peut traiter d'un système linéaire surdéterminé mais dont les équations normales sont mal conditionnées. Nous avons rencontré un tel problème lors du lissage de diagrammes de phase binaires dont les données sont expérimentales.

On trouvera sur le Web(*) le programme `regul.c` qui résout les systèmes linéaires mal conditionnés par la méthode des régularisants.

## 7. Résolution des équations de convolution

C'est une équation de toute première importance en physique dans la mesure où le signal observé $u(t)$ est le produit de convolution du signal émis $z(t)$ et de la fonction d'appareil $K(t)$. Ce sont des problèmes classiques rencontrés en spectroscopie, en astronomie etc. Nous nous proposons de résoudre l'équation :

$$\int_{-\infty}^{+\infty} K(x-t)z(t)\,\mathrm{d}t = u(x),$$

c'est-à-dire de calculer $z(t)$ connaissant $K(t)$ et $u(x)$. On peut encore écrire :

$$Az = u.$$

Le problème se résout très facilement en termes de transformées de Fourier sous réserve que les fonctions appartiennent à l'espace $L^1$ ou $L^2$. Le produit de convolution devient un produit simple dans l'espace de Fourier :

$$K(t)^* z(t) = u(t) \quad \xrightarrow{TF} \quad K(\omega)\cdot\zeta(\omega) = v(\omega)$$

d'où

$$\zeta(\omega) = v(\omega)/K(\omega),$$

$$\text{et} \quad z(t) \quad \xleftarrow{TF^{-1}} \quad \zeta(\omega),$$

$$\text{avec :} \quad \Phi(\omega) = \int_{-\infty}^{+\infty} f(t)\exp(2\pi j\omega t)\,\mathrm{d}t.$$

* `http://www.edpsciences.com/guilpin/`

Sur le plan formel, pourvu que les fonctions appartiennent aux bons espaces, il n'y a rien à reprocher à une telle façon de faire ; en revanche, si l'on doit s'attaquer à la résolution numérique d'un tel problème, on va se heurter à un problème mal posé.

Il est facile de s'en convaincre en choisissant deux fonctions arbitraires, une fonction gaussienne $g(t)$ et une fonction triangle $tri(t)$ par exemple, dont on réalise le produit de convolution $u(t)$, le sup de chacune des fonctions étant égal à un. Si l'on utilise d'une part une machine dont la précision est de 10 chiffres significatifs et d'autre part la transformation de Fourier portant sur 128 points, la résolution de l'équation de convolution est parfaitement réalisée. En revanche, si l'on introduit dans le second membre une erreur relative gaussienne de l'ordre de un pour mille — il s'agit d'un bruit très faible noté $\varepsilon(t)$ — alors les résultats deviennent complètement aberrants : au lieu de trouver des résultats entre 0 et 1 approximativement, on obtient des valeurs oscillantes entre $-26$ et $+34$ et une quinzaine d'oscillations...

Ceci montre que la résolution numérique des équations de convolution ne peut pas donner de résultats corrects si l'on ne prend pas la bonne méthode d'analyse car les données expérimentales sont toujours entachées d'erreur.

L'origine des ennuis est relativement simple à identifier : elle est due aux «hautes fréquences ». Désignons par $\theta(\omega)$ la transformée de Fourier de $\varepsilon(t)$, la transformation inverse de $\theta(\omega)/K(\omega)$ peut ne pas exister, c'est-à-dire que cette intégrale peut diverger, ceci étant essentiellement dû au rôle des hautes fréquences. Cependant, quand bien même la transformation inverse existerait notée $\omega(t)$, l'écart de $\omega(t)$ par rapport à zéro peut être aussi grand que l'on veut dans la métrique $L^2$ ou $L^\infty$.

## 7.1. Présentation d'un opérateur régularisant

Pour neutraliser l'influence des hautes fréquences, on peut multiplier la fonction à inverser $v(\omega)/K(\omega)$ par un facteur stabilisant $f(\omega, \alpha)$. Ainsi, on admettra sans démonstration que l'opérateur suivant est régularisant :

$$\mathbf{R}[u_{ap}(t), \alpha(\delta)] = \int_{-\infty}^{+\infty} \frac{f(\omega, \alpha) v(\omega)}{K(\omega)} \exp(-2\pi j \omega t)\, \mathrm{d}\omega,$$

pourvu que la fonction $f(\omega, \alpha)$ obéisse aux conditions suivantes :

1. $f(\omega, \alpha)$ est définie quel que soit $\alpha \geq 0$ et quel que soit $\omega \in (-\infty, +\infty)$.
2. $0 \leq f(\omega, \alpha) \leq 1$ quels que soient $\alpha \geq 0$ et $\omega \in (-\infty, +\infty)$.
3. $f(\omega, 0) \equiv 1$.
4. $f(\omega, \alpha)$ appartient à l'espace $L^2$, et quel que soit $\alpha \geq 0$, $f(\omega, \alpha)$ est une fonction paire de $\omega$.
5. $f(\omega, \alpha)$ tend vers zéro quand $|\omega| \to \infty$, quel que soit $\alpha \geq 0$.
6. $\dfrac{f(\omega, \alpha) v(\omega)}{K(\omega)}$ appartient à l'espace $L^2$, quel que soit $\alpha \geq 0$.
7. $f(\omega, \alpha)$ tend vers zéro quand $\alpha \to \infty$, pour tout $\omega$ différent de zéro.

## 7.2. Un exemple de fonction f($\omega, \alpha$)

$$f(\omega, \alpha) = \frac{K(-\omega) K(\omega)}{K^2(\omega) + \alpha M(\omega)} = \frac{K^2(\omega)}{K^2(\omega) + \alpha M(\omega)}$$

où $\alpha \geq 0$ et où $M(\omega)$ est une fonction paire, continue, strictement positive pour $\omega \neq 0$ et positive ou nulle si $\omega = 0$.

Il est possible de montrer que le résidu de la fonction régularisée est une fonction strictement croissante de la variable $\alpha$, ce qui entraîne qu'il existe un nombre unique $\alpha$ tel que :

$$\mu_u(Az_\alpha, u_{ap}) = \delta$$

à condition toutefois que $\delta < \mu_u(u_{ap})$.

On peut choisir pour $M(\omega)$ une forme polynomiale que l'on écrit :

$$M(\omega) = \sum_{j=0}^{p} \gamma_j \omega^{2j},$$

expression dans laquelle les $\gamma_j$ sont des constantes positives (éventuellement nulles sauf $\gamma_p$). Cette forme introduit des stabilisateurs d'ordre $p$. En définitive, la solution régularisée s'écrit :

$$z_\alpha(t) = \int_{-\infty}^{+\infty} \frac{K(-\omega)v(\omega)}{K^2(\omega) + \alpha M(\omega)} \exp(-2\pi j\omega t)\, \mathrm{d}\omega.$$

ici encore, $\alpha$ pouvant être déterminé par le résidu.

On trouvera sur le Web(*) le programme `dconvol.c` qui réalise cet algorithme.

## 8. Bibliographie

H. ENGL et C.W. GROETSCH (1987) *Inverse and ill-posed problems*, Academic Press.

J. HADAMARD (1902) *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton University Bulletin, p. 49–52.

J. HADAMARD (1932) *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*, Hermann.

A. TIKHONOV et V. ARSÉNINE (1976) *Méthodes de résolution des problèmes mal posés au sens de Hadamard*, Éditions MIR, Moscou.

* `http://www.edpsciences.com/guilpin/`

# 18 Introduction aux méthodes de Monte-Carlo

Les méthodes de Monte-Carlo sont apparues comme moyen de recherche durant la Seconde Guerre Mondiale afin de simuler le comportement des neutrons dans les matériaux fissiles. La paternité de cette idée revient à J.E. Mayer qui l'utilisa pour la première fois en physique statistique.

Avec le développement des calculateurs arithmétiques, il est très facile de réaliser quelques simulations de ce type à l'aide de générateurs de nombres pseudo-aléatoires. Après avoir présenté le célèbre problème de Buffon (1707–1788) comme illustration historique d'une méthode de Monte-Carlo, nous donnerons deux générateurs de nombres pseudo-aléatoires que nous utiliserons dans quelques applications classiques à savoir :

- Calcul d'une intégrale définie.
- Résolution locale de l'équation de Laplace.
- Inversion d'une matrice carrée.
- Recuit simulé (minimum d'une fonction).

## 1. Le problème de Buffon

Sur une feuille de papier nous traçons des parallèles équidistantes séparées par la distance $2a$. Nous plaçons cette feuille sur un plan parfaitement horizontal, et nous jetons dessus une aiguille de longueur $2l$ telle que $l < a$. Le problème consiste à calculer la probabilité que l'aiguille rencontre une des parallèles (*cf.* Fig. 18.1, page suivante). Ce problème s'appelle aussi problème du drapeau américain, car durant la guerre de Sécession (1861–1865) un officier américain, lors de sa convalescence après blessure, s'était posé le même problème, la feuille de papier étant remplacée par le drapeau américain qui comporte des bandes équidistantes.

Désignons par $M$ le milieu de l'aiguille et par $EF$ la perpendiculaire aux parallèles passant par $M$. Soit $I$ le milieu de $EF$. On pose $EM = x$ que l'on suppose inférieure à $FM$ ($M$ compris entre $E$ et $I$).

Pour que l'aiguille coupe la parallèle $AB$, il faut premièrement que $x < l$. Alors, la probabilité pour que $M$ sur $EI$ soit compris entre $x$ et $x + \mathrm{d}x$ est :

$$\mathrm{d}p_1 = \frac{\mathrm{d}x}{a} \cdot$$

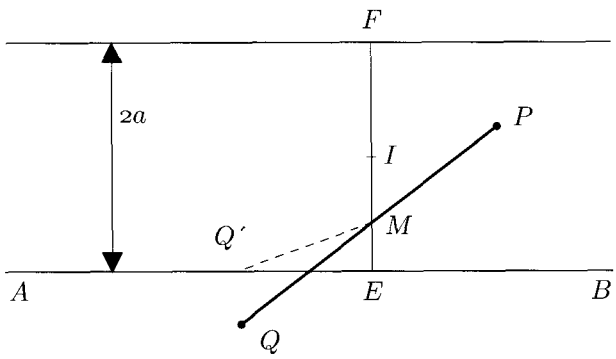


**Figure 18.1.** *Le problème de Buffon.*

Il faut deuxièmement que l'angle aigu $EMQ$ soit plus petit que l'angle aigu $EMQ'$ où $Q'$ est le point de $AB$ tel que $Q'M = l$, soit :

$$EMQ' = \arccos\left(\frac{x}{l}\right) \cdot$$

La probabilité pour que $EMQ$ soit plus petit que $EMQ'$ est alors :

$$p_2 = \frac{2}{\pi} \arccos\left(\frac{x}{l}\right) \cdot$$

$\frac{2}{\pi}$ est le coefficient de normalisation car $EMQ'$ est compris entre 0 et $\frac{\pi}{2}$. Comme $\mathrm{d}p_1$ et $p_2$ sont des probabilités indépendantes, la probabilité $\mathrm{d}p$ de réaliser l'événement recherché est donnée par le produit des probabilités. Il reste à intégrer le long de $EI$ c'est-à-dire de 0 à $l$, pour obtenir la probabilité $p$. Nous pouvons écrire :

$$p = \int_0^l \frac{2}{\pi} \arccos\left(\frac{x}{l}\right) \frac{\mathrm{d}x}{a} = \frac{2l}{a\pi} \cdot$$

Dans le cas où l'on choisit $l = a/2$, la relation précédente se simplifie :

$$p = \frac{1}{\pi} \cdot$$

Le nom de Buffon est resté attaché à ce problème car ce fut le premier à en donner la solution correcte. Au Palais de la Découverte à Paris, cette expérience est effectivement montée et, comme on mesure $p$ qui est le rapport du nombre $n$ de cas où l'aiguille coupe une droite et du nombre total de lancements, on en déduit une détermination expérimentale de $\pi$. La première détermination selon ce procédé remonte à 1850 et fut réalisée par Wolff qui, avec 5 000 lancers de l'aiguille, obtint la valeur 3,159 6.

Il est très simple de simuler ce genre de problème avec une machine arithmétique à condition de pouvoir disposer de **nombres aléatoires obéissant à une distribution donnée**.

## 2. Générateurs de nombres pseudo aléatoires à distribution uniforme

On peut obtenir de bonnes tables de nombres aléatoires à distribution rectangulaire (appelée aussi distribution uniforme) en prélevant dans le numéro de téléphone des abonnés pris dans

l'ordre alphabétique la suite des chiffres à condition toutefois d'en ôter le préfixe. Ces chiffres compris entre 0 et 9 sont approximativement aléatoires à distribution rectangulaire.

On peut obtenir des résultats semblables en prélevant les chiffres successifs figurant sur les plaques minéralogiques des automobiles immatriculées en France pourvu que l'on ne s'intéresse qu'à ceux précédant les lettres.

On peut également obtenir des suites intéressantes en choisissant les chiffres de nombres transcendants tels que $e$ et $\pi$. Le procédé est coûteux car le calcul d'un grand nombre de chiffres demande un énorme travail préalable, y compris la transcription qui doit être sans faille. Il est important de noter que les chiffres des nombres transcendants tels que $e$ et $\pi$ n'ont aucune périodicité, mais, de plus, ils sont répartis aléatoirement selon la loi rectangulaire. Par ailleurs, on sait parfaitement calculer de très longues suites de chiffres exacts de ces nombres. En quoi ces suites générées sont-elles aléatoires? Elles le sont dans la mesure où connaissant uniquement les $q$ premiers chiffres de la suite, il n'est pas possible de déterminer le $q+1^{e}$ quel que soit $q$.

Avec l'apparition des calculateurs modernes, la nécessité de pouvoir générer des suites de **nombres pseudo-aléatoires** s'est très vite manifestée. Il s'agit alors de créer une suite de nombres telle que la périodicité générée soit très grande devant la taille de l'échantillon à prélever.

Une très bonne méthode due à Lehmer-Greenberger (1951 et 1961) consiste à former la suite à partir de la relation de récurrence :

$$k_{i+1} = ak_i + b \quad \text{modulo} \quad m$$

avec les contraintes suivantes :

1. $b$ et $m$ sont premiers entre eux,
2. $a \equiv 1$ modulo $p$ pour chaque facteur premier $p$ de $m$,
3. $a \equiv 1$ modulo 4 si $m$ est un multiple de 4.

**Ainsi la période de la suite générée est $m$.**

Ces conditions sont très faciles à réaliser sur un calculateur binaire en **arithmétique entière.**

Prenons le cas classique de la représentation sur quatre octets. Alors on sait que l'arithmétique entière est effectuée modulo $m = 2^{31}$. Attention en machine ce nombre est négatif, car le premier bit contient un 1. Ainsi le seul facteur premier de $m$ est 2. Maintenant, pour satisfaire les autres conditions, il suffit de choisir $b$ impair et $a = 4k+1$, $k$ entier quelconque. Pour que le modulo $m$ fonctionne dès les premières récurrences, il faut choisir des nombres $k_0$, $a$ et $b$ de taille convenable. Nous avons retenu :

$$\begin{aligned} k_0 &= 223 \\ a &= 125\,661 \\ b &= 95\,783. \end{aligned}$$

***Remarque 1*** : Il y a deux façons d'utiliser le générateur : en arithmétique flottante, on divise les $k_i$ générés par $(-2^{31})$ et l'on obtient des nombres compris entre $-1$ et $+1$. On peut aussi ne former que des nombres positifs compris entre 0 et 1, ainsi quand on rencontre un $k_i$ négatif, il suffit alors de lui ajouter $(-2^{31})$ pour le ramener dans les entiers positifs. Ensuite chaque $k_i$ est divisé par $(-2^{31})$.

***Remarque 2*** : Nous avons vérifié expérimentalement que la période était bien $2^{31}$.

***Remarque 3*** : En utilisant le test du $\chi^2$, nous avons vérifié que l'hypothèse de la distribution uniforme n'était pas contredite par les données générées.

Une autre technique pour obtenir une suite de nombres pseudo-aléatoires consiste à générer la suite suivante :

$$x_{i+1} = x_i + x_{i-1} \quad \text{modulo } 4$$

$$\text{avec} \quad x_0 = e = 2{,}718\,281\,828\,459\,045\,235\ldots$$
$$x_1 = \pi = 3{,}141\,592\,653\,589\,793\,238\ldots$$

Elle donne des résultats tout à fait semblables à ceux fournis par le générateur de Lehmer-Greenberger. On obtient des nombres pseudo-aléatoires à distribution uniforme sur l'intervalle $(0, 1)$ en multipliant $x_i$ par 0,25.

## 3. Calcul de $\pi$

Il est très facile d'obtenir une valeur approchée de $\pi$ en effectuant une simulation analogue à celle liée au problème de Buffon. Pour cela, il suffit de tirer deux nombres aléatoires indépendants à distribution rectangulaire sur $(-1, +1)$ que l'on note $\xi_1$ et $\xi_2$. Dans le plan, le point aléatoire $A$ de coordonnées $\xi_1$ et $\xi_2$ obéit à une distribution uniforme dans le carré de sommets $(-1, 1)$ et $(1, -1)$, elle vaut 0,25. Il est tout à fait simple de dénombrer les points $A$ qui tombent dans le cercle de centre $O$ et rayon unité. Si $N$ est le nombre de tirages de couples $\xi_1$ et $\xi_2$, et si $Q$ est le nombre de points $A$ situés à l'intérieur du cercle unité, on voit que la probabilité de tomber dans le cercle est donnée par :

$$p = \frac{\text{surface du cercle}}{\text{surface du carré}} = \frac{\pi}{4} = \frac{Q}{N}\,.$$

Comme pour $N$ donné on sait dénombrer $Q$, on obtient aisément une valeur approximative de $\pi$ au moyen de la relation :

$$\pi = \frac{4Q}{N}\,.$$

Il ne faut pas se méprendre sur la portée d'une telle méthode, seul son intérêt pédagogique est à prendre en considération car la précision relative varie en $1/\sqrt{N}$. On trouvera sur le Web [*] le programme `calculpi.c` qui effectue le calcul de $\pi$ selon cet algorithme. Pour $N = 1\,000\,000$, on trouve $\pi = 3{,}140\,2$, pour $N = 100\,000\,000$, on trouve $\pi = 3{,}141\,507$, et pour $N = 1\,000\,000\,000$, on trouve $\pi = 3{,}141\,603\,5$. On vérifie bien que pour diminuer l'erreur d'un ordre de grandeur, il faut multiplier le nombre d'épreuves par 100... autrement dit, on vérifie bien sur cet exemple que la précision varie comme $1/\sqrt{N}$ et que le générateur de nombres pseudo-aléatoires se comporte bien.

## 4. Calcul d'une intégrale définie

On désire calculer l'intégrale d'une fonction $f(x)$ entre les bornes finies $a$ et $b$. On suppose que $f(x)$ est une fonction continue par morceaux.

* `http://www.edpsciences.com/guilpin/`

## 4.1. Première approche : simulation géométrique (intérêt strictement pédagogique)

On note $X_{\max}$ et $X_{\min}$ les extremums de $f(x)$ sur $(a, b)$. On désigne par $\alpha$ et $\beta$ deux nombres tels que $\alpha \geq X_{\max}$ et $\beta \leq X_{\min}$.

On tire deux nombres aléatoires indépendants à distribution uniforme l'un sur l'intervalle $(a, b)$, l'autre sur l'intervalle $(\alpha, \beta)$. On note $\xi$ et $\eta$ ces deux nombres. Le point figuratif $A$ de coordonnées $(\xi, \eta)$ est uniformément réparti dans le rectangle de sommets $(a, \alpha)$ et $(b, \beta)$. $N$ est le nombre de tirages de couples $\xi$ et $\eta$, et $Q$ est le nombre de points $A$ situés entre le graphe de $f(x)$ et l'axe des $x$. Attention, il s'agit d'effectuer la somme algébrique des aires, et si $\eta$ est négatif et qu'il figure entre le graphe de $f(x)$ et l'axe des $x$, il convient ôter une unité à $Q$. Si l'on ne réalise pas cette opération, on effectue le calcul de l'intégrale de $|f(x)|$. Si $\beta$ est positif aucun problème ne se pose.

***Remarque*** : Pour tirer des nombres pseudo-aléatoires entre $a$ et $b$, on part d'un nombre $\xi$ compris entre 0 et 1. Ensuite, on effectue la transformation :

$$\eta = a + (b - a)\xi.$$

Si l'on part d'un nombre $\xi$ compris entre $-1$ et 1, on effectue la transformation :

$$\eta = a + (b - a)(\xi + 1)/2{,}0.$$

On voit que la probabilité de tomber dans la surface algébrique de la fonction $f(x)$ est donnée par :

$$p = \frac{\text{aire algébrique}}{\text{surface du rectangle}} = \frac{\int_a^b f(x)\,\mathrm{d}x}{(b-a)(\beta-\alpha)} = \frac{Q}{N}.$$

On aboutit en définitive à l'expression :

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{Q}{N}(b-a)(\beta-\alpha),$$

qui constitue une approximation de l'intégrale.

## 4.2. Deuxième approche

La méthode que nous venons de présenter est très sommaire. À présent nous allons présenter un autre estimateur de l'intégrale qui exploite des connaissances plus fines. Soit $\xi$ une variable aléatoire distribuée uniformément sur $(a,\ b)$. Nous effectuons $N$ tirages que nous notons $\xi_i$ avec $i = 1, \ldots, N$. Si $N$ est un nombre grand devant l'unité, la grandeur :

$$f_0 = \frac{1}{N}\sum_{i=1}^{N} f(\xi_i)$$

est un estimateur de la moyenne de la fonction $f(x)$ sur l'intervalle $(a, b)$, ce qui a pour conséquence que l'intégrale est estimée par la quantité $(b - a)f_0$. On peut donc écrire :

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{(b-a)}{N}\sum_{i=1}^{N} f(\xi_i),$$

cette technique est beaucoup plus efficace que la précédente dans la mesure où elle fait appel à beaucoup moins d'opérations.

Sur le Web(*), on trouvera le programme `intmonte.c` qui réalise ces divers calculs sur des exemples simples.

## 5. Intégration de l'équation de Laplace en un point

On souhaite intégrer dans le domaine $D$ l'équation de Laplace :

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0$$

dans un repère cartésien orthonormé tridimensionnel dans le cadre des conditions de Dirichlet (on connaît $V$ sur la frontière de $D$). Comme nous l'avons déjà fait au cours du chapitre 14, on superpose au domaine $D$ un maillage cubique dont les nœuds des mailles sont notés par les triplets $(i, j, k)$. Une méthode de Monte-Carlo pour calculer le potentiel $V$ au point particulier $(I, J, K)$ consiste à décrire une suite de $N$ chemins tous issus du point $(I, J, K)$ chacun de ces chemins se terminant tôt ou tard sur la frontière du domaine. Dès lors que l'on a atteint un point de la frontière, on note le potentiel $V_i$ de ce point et l'on recommence un nouveau chemin issu de $(I, J, K)$. Il est possible de montrer que le potentiel en $(I, J, K)$ est la moyenne des $V_i$, soit encore :

$$V(I, J, K) \approx \frac{1}{N} \sum_{i=1}^{N} V_i.$$

À présent, il reste à déterminer la façon dont on décrit un chemin dans le maillage. On doit pouvoir atteindre chacun des plus proches voisins d'un point donné quelconque hors la frontière. Avec le maillage que nous avons retenu il existe six plus proches voisins. Pour réaliser un chemin, il suffit de procéder à un tirage au sort de six nombres indépendants 1, 2, 3, 4, 5 et 6 qui sont équirépartis, et à chaque nombre on affecte arbitrairement un déplacement dans une des six directions. Par exemple, on peut convenir que :

1 déplacement vers la gauche,
2 déplacement vers la droite,
3 déplacement vers le bas,
4 déplacement vers le haut,
5 déplacement vers l'avant,
6 déplacement vers l'arrière.

Il est aisé de fabriquer ces nombres pseudo-aléatoires. Soit $\xi$ un nombre pseudo-aléatoire obtenu par l'algorithme de Lehmer-Greenberger. Il est compris entre 0 et 1. Si on le multiplie par 6, il appartiendra à une répartition uniforme entre 0 et 6, 6 étant exclu. Il suffira donc d'en prendre la partie entière et d'y ajouter 1 pour obtenir la suite des points du chemin.

$$\eta = \text{entier } \{6\xi\} + 1.$$

Nécessairement au bout d'un nombre fini d'opérations, on parviendra à la limite du domaine.

* `http://www.edpsciences.com/guilpin/`

## 6. Inversion d'une matrice carrée d'ordre n

À y regarder de près, la résolution de l'équation de Laplace par les techniques du maillage est équivalente à la résolution d'un système linéaire. En effet, il suffit d'écrire la loi des mailles sur un ensemble convenablement choisi de points pour obtenir un système linéaire de $n$ équations à $n$ inconnues, $n$ étant le nombre de nœuds du maillage. Puisque l'on sait résoudre l'équation de Laplace par une méthode de Monte-Carlo, on est en droit de penser qu'il est possible d'inverser une matrice également par une méthode de Monte-Carlo.

Nous allons simplement donner la technique de résolution dans un cas particulier de matrices $A$ d'ordre $n$ qui sont telles que les éléments de la matrice $B = I - A$ (où $I$ est la matrice unité) obéissent aux inégalités suivantes :

$$b_{ij} \geq 0 \quad \text{et} \quad \sum_{j=1}^{n} b_{ij} < 1,$$

c'est-à-dire que les $b_{ij}$ peuvent être interprétés comme des probabilités.

Sans entrer dans le détail, nous allons voir qu'il y a moyen de décrire des chemins (selon un processus markovien) dans les lignes et les colonnes de la matrice $B$ qui nous permet de calculer successivement les lignes de $A^{-1}$. Pour cela nous allons modifier toutes les lignes en remplaçant chaque élément par la somme des précédents sur chacune des lignes, soit pour la ligne $j$ :

$$b_{j0} \qquad b_{j0} + b_{j1} \qquad b_{j0} + b_{j1} + b_{j2} \qquad \cdots \qquad b_{j0} + b_{j1} + b_{j2} + \cdots + b_{jn}.$$

Dorénavant, on désignera par $\beta_{lk}$ ces nouveaux éléments. À présent on peut décrire le processus qui va conduire à l'inversion de la matrice.

**a.** Souhaitant trouver l'inverse de la ligne $j$ de $A$, on commence par tirer un nombre aléatoire $\xi_0$ à distribution rectangulaire sur $(0, 1)$. De deux choses l'une : ou bien ce nombre est inférieur ou égal à l'élément $\beta_{jn}$ ou bien il lui est supérieur.
**b.** S'il lui est supérieur on ajoute une unité au compteur de la ligne $j$.
**c.** S'il lui est inférieur ou égal on cherche la première colonne où l'élément lui est strictement supérieur ou égal, soit $m$ cette colonne.
**d.** On tire un autre nombre aléatoire $\xi_1$ indépendant du premier, et l'on opère avec la ligne $m$ de la même façon que précédemment. On répète ces opérations jusqu'à ce que l'on trouve un nombre $\xi_p$ correspondant à la ligne $r$ tel que $\xi_p > \beta_{rn}$ auquel cas on ajoutera 1 au compteur de la ligne $r$.

On effectue les opérations de (a) à (d) $N$ fois. Désignons par $p_q$ les compteurs de lignes correspondant donc à l'inversion de la ligne $j$. Il est possible de montrer que l'on a :

$$[A^{-1}]_{jk} \approx \frac{1}{N} \frac{p_k}{1 - \beta_{rn}}.$$

Pour terminer, on traite toutes les lignes selon cette procédure ce qui nous fournit l'inverse (approchée) de la matrice $A$.

Dans le cas général, la matrice $A$ ne se présente pas sous la forme indiquée. Alors on choisit des coefficients $\gamma_{ij}$ tels que :

$$b_{ij} = \gamma_{ij}\pi_{ij} \quad \text{avec} \quad \pi_{ij} \geq 0 \quad \text{et} \quad \sum_{j=1}^{n} \pi_{ij} < 1.$$

On trouvera dans l'ouvrage de Yu.A. Shreider cité en bibliographie le développement de cette technique.

Sur le Web (*), le programme `matmonte.c` réalise l'inversion des matrices du type $A$ particulier selon la technique développée dans ce paragraphe.

## 7. Méthode du recuit simulé : recherche du minimum absolu d'une fonction

On recherche le minimum minimorum d'une fonction $f(x, y, z, \dots)$ dans un certain domaine $D$. Pour des raisons de commodité d'écriture, on représente par $\vec{X}_i$ le vecteur de composantes $x_i$, $y_i$, $z_i \dots$

L'algorithme se présente sous la forme suivante :

**a.** On tire un premier vecteur (initial) aléatoire $\vec{X}_0$ au moyen du générateur de Lehmer-Greenberger tel que $\vec{X}_0 \in D$.

**b.** On tire un vecteur $\vec{X}_j$ par le même procédé puis on calcule la probabilité de transition :

$$p = \exp(-\beta_0 \Delta F) \quad \text{avec} \quad \Delta F = f(\vec{X}_j) - f(\vec{X}_{j-1}).$$

Si $p > \text{Seuil}_1$, on conserve le dernier état, et l'on poursuit les opérations $N$ fois (boucle sur $b$).

**c.** Ensuite, on augmente $\beta$ qui suit une loi du genre : $\beta_n = \beta_{n-1}\beta_0$ et l'on effectue $N$ calculs identiques à $b$.

**d.** L'arrêt des calculs s'effectue lorsque $\left|\dfrac{\partial f}{\partial \beta}\right| < \text{Seuil}_2$.

Cette procédure appelle quelques remarques :

***Remarque 1*** : Au départ, il faut choisir $\beta_0$ « petit ».

***Remarque 2*** : Pour passer de $\vec{X}_j$ à $\vec{X}_{j+1}$, on a intérêt à ne changer d'état que sur une seule composante à la fois. On opère alors successivement sur les variables dans l'ordre $x, y, z, \dots$

***Remarque 3*** : $N$ et $\beta_0$ dépendent, pour ce qui concerne leur choix, des erreurs propagées dans les calculs, aussi bien lors de la génération des nombres pseudo-aléatoires que dans le calcul de $\Delta f$ et $f(\vec{X})$.

***Remarque 4*** : Il faut veiller à ne pas « descendre trop vite », sinon, nous risquons de rester dans un minimum local dont on ne peut plus sortir ($p < \text{Seuil}_1$). Toutefois, une « descente trop lente » est source d'instabilité. Il convient de choisir un compromis pour choisir $\beta_0$ et $N$ selon la fonction à minimiser.

On trouvera sur le Web (*) le programme `recuit.c` qui réalise cet algorithme sur une fonction présentant un minimum absolu. Ici, il s'agit de la résolution d'un système de deux équations non linéaires à deux inconnues $f(x, y) = 0$ et $g(x, y) = 0$. Notons qu'il revient au même de minimiser une forme quadratique :

$$F(x, y) = f^2(x, y) + g^2(x, y).$$

* `http://www.edpsciences.com/guilpin/`

# 8. Simulation d'autres lois de distribution

Il est assez simple d'obtenir d'autres lois de distribution à partir de la distribution uniforme. Supposons que la variable aléatoire $\xi$ admette la loi de répartition $F(x)$. La variable $\eta = F(\xi)$ est uniformément répartie, donc pour obtenir $\xi$ il suffit d'inverser la relation précédente à condition toutefois que cela soit possible. Si tel est le cas, on écrit :

$$\xi = F^{-1}(\eta).$$

## 8.1. Distribution sinusoïdale

Par exemple, on peut générer une suite de nombres pseudo-aléatoires à distribution sinusoïdale en utilisant la relation :

$$\xi = \arcsin(\eta),$$

$\xi$ est évidemment entre $-1$ et $+1$.

## 8.2. Distribution exponentielle

De la même façon, on peut générer des nombres pseudo-aléatoires à distribution exponentielle qui peuvent alors permettre la simulation du comportement des neutrons dans une réaction nucléaire (distance entre deux chocs successifs). La loi de répartition est :

$$\begin{aligned} F(x) &= 1 - \exp(-\lambda^2 x) \quad \text{pour } x \geq 0 \\ F(x) &= 0 \qquad\qquad\qquad\quad \text{pour } x < 0. \end{aligned}$$

D'après ce qui précède, on écrit alors : $\log_e(1-\xi) = -\lambda^2\eta$ soit encore :

$$\eta = -\frac{1}{\lambda^2}\log_e(1-\xi).$$

Si l'on remarque que la distribution de $1-\xi$ est la même que celle de $\xi$ puisque $\xi$ appartient à $(0,1)$, alors l'expression se réduit à :

$$\eta = -\frac{1}{\lambda^2}\log_e(\xi) \quad \xi \text{ appartenant à } (0,1).$$

Il est donc très facile d'obtenir des nombres pseudo-aléatoires à distribution exponentielle.

## 8.3. Distribution normale

Eu égard à l'importance de cette distribution, nous allons étudier diverses techniques permettant de la générer.

***Somme de n nombres à répartition uniforme sur* $(-0{,}5, 0{,}5)$ *:* $\eta = \sum_1^n \xi_i$** – Pour obtenir la distribution de $\eta$, nous utilisons la notion de fonction caractéristique (voir le chapitre 21 consacré à cet effet). Soit $\Phi(t)$ la fonction caractéristique de chaque $\xi_i$, elle s'écrit :

$$\Phi(t) = \int_{-0,5}^{0,5} \exp(2\pi j t x)\,\mathrm{d}x = \frac{\sin(\pi t)}{\pi t}.$$

La fonction caractéristique de $\eta$ est donc $\Phi(t)^n$, elle s'écrit :

$$f(t) = \Phi(t)^n = \left\{ \frac{\sin(\pi t)}{\pi t} \right\}^n ,$$

en passant aux logarithmes, on obtient :

$$\log_e[f(t)] = n \log_e \left[ \frac{\sin(\pi t)}{\pi t} \right]$$

comme $\dfrac{\sin(x)}{x}$ est une fonction qui tend vers zéro quand $x$ croit indéfiniment, expression précédente se comporte comme :

$$\log[f(t)] \simeq n \log \left[ 1 - \frac{\pi^2 t^2}{6} \right] \simeq - \frac{\pi^2 t^2 n}{6},$$

de là on tire :

$$f(t) = A \exp \left( - \frac{\pi t^2}{a^2} \right) \qquad \text{avec } a^2 = \frac{12}{2\pi n} \cdot$$

La transformation inverse donne la fonction densité de probabilité, à savoir :

$$g(x) = B \exp \left( -\pi x^2 a^2 \right) = B \exp \left( - \frac{6x^2}{n} \right),$$

soit encore :

$$g(x) = B \exp \left( - \frac{x^2}{2\sigma^2} \right) \qquad \text{avec } \sigma^2 = \frac{n}{12} \cdot$$

En définitive, on obtient :

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left( - \frac{x^2}{2\sigma^2} \right) \qquad \text{avec } \sigma^2 = \frac{n}{12} \cdot$$

On voit aisément que si l'on effectue la somme de 12 nombres aléatoires on obtient la loi normale réduite. Soit $\eta_0$ un nombre aléatoire généré par ce procédé donc obéissant à la loi normale réduite. Il est aisé de voir que le nombre $\zeta = \eta_0 \sigma + m$ obéit à la loi de distribution :

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ - \frac{(x-m)^2}{2\sigma^2} \right],$$

par conséquent, il est facile de générer des nombres à distribution gaussienne avec une moyenne et un écart type choisis à l'avance.

Si les nombres pseudo-aléatoires $\xi$ appartiennent à l'intervalle $(0, 1)$, il suffit d'effectuer le changement $\eta = \xi - 0{,}5$ pour se ramener au problème précédent.

Au cas où les queues de distribution de la gaussienne jouent un rôle prépondérant dans la simulation, on doit adopter un autre procédé pour générer notre suite de nombres gaussiens.

### 8.4. Distribution gaussienne à queues soignées

On part de l'idée suivante : soient deux nombres aléatoires indépendants $\eta_1$ et $\eta_2$ obéissant chacun à la **loi normale réduite**. La somme des carrés de ces variables, soit $\zeta = \eta_1^2 + \eta_2^2$, n'est rien d'autre que la variable $\chi^2$ à deux degrés de liberté (chapitre 22) dont la loi de distribution s'écrit :

$$P(\zeta < x) = 1 - \exp\left(-\frac{x}{2}\right)$$

c'est la loi exponentielle de coefficient $\lambda^2 = 0{,}5$ dont nous venons de simuler la répartition (voir aussi la loi du $\chi^2_m$).

Par ailleurs, $\eta_1$ et $\eta_2$ peuvent être considérés comme les coordonnées, dans le plan, d'un point $M$ d'argument $\phi = (Ox, OM)$. Cet argument $\phi$ suit une loi uniforme indépendante de la longueur de $OM$.

Nous allons exploiter ces deux remarques pour calculer les nombres indépendants $\eta_1$ et $\eta_2$ à partir des deux nombres indépendants distribués uniformément $\xi_1$ et $\xi_2$. En se référant aux propos de l'alinéa *b* du paragraphe 8.2 (concernant la distribution exponentielle), il suffit alors de poser :

$$\begin{aligned} \eta_1^2 + \eta_2^2 &= -2\log_e(\xi_1) && \text{avec } \xi_1 \text{ appartenant à } (0,1) \\ \frac{\eta_1}{\eta_2} &= \tan(\phi) = \tan(2\pi\xi_2) && \text{soit encore : } \eta_1^2\cos^2(2\pi\xi_2) = \eta_2^2\sin^2(2\pi\xi_2), \end{aligned}$$

avec $\xi_2$ appartenant à $(0,1)$, d'où l'on tire :

$$\begin{aligned} \eta_1 &= \sqrt{-2\log_e(\xi_1)}\sin(2\pi\xi_2) \\ \eta_2 &= \sqrt{-2\log_e(\xi_1)}\cos(2\pi\xi_2). \end{aligned}$$

Il suffit de tirer deux nombres aléatoires à distribution rectangulaire $\xi_1$ et $\xi_2$ sur l'intervalle $(0,1)$ indépendants pour obtenir deux nombres aléatoires à distribution gaussienne $\eta_1$ et $\eta_2$ indépendants (distribution normale réduite).

## 9. Éléments de bibliographie

E. DURAND (1961) *Solutions numériques des équations algébriques*, Tome II, Éditions Masson.
G. FISHMAN (1996) *Monte-Carlo : concepts, algorithms and applications*, Springer-Verlag.
J.M. HAMMERSLEY et D.C. HANDSCOMB (1967) *Les méthodes de Monte-Carlo*, Éditions Dunod.
A. SCHREIDER (1966) *The Monte-Carlo Method, The method of the statistical trials*, Pergamon Press.

# 19 Éléments de calcul des probabilités

Même au niveau très élémentaire, le calcul des probabilités est toujours truffé de pièges et de colles qui font que cette discipline est souvent redoutée. La lecture de l'ouvrage de J. Bertrand (1822–1900) cité en bibliographie est fortement recommandée tant elle est agréable. Il s'agit du fac-similé de la seconde édition française de 1907. C'est un ouvrage de référence, particulièrement connu à l'étranger, connu aussi pour sa rigueur épistémologique. Il est d'une lecture très aisée et ne nécessite en mathématiques que les connaissances correspondant au niveau du DEUG première année.

## 1. Introduction et notions fondamentales

Nous nous proposons d'examiner de quelle manière on peut étudier les phénomènes présentant un certain caractère dû au **hasard** (mot d'origine arabe ayant trait au jeu de dé). Une façon classique de présenter les problèmes consiste à examiner les fluctuations d'une grandeur autour d'une moyenne, en voici des exemples : la masse d'un corps au moyen de $n$ mesures, la recette journalière d'un parcmètre, la trajectoire d'un navire entre deux ports, la trajectoire d'un avion entre deux aéroports.

Dans notre présentation, il n'est pas utile de rechercher une définition très subtile du hasard, et il nous suffira de dire qu'il s'agit d'une **causalité fictive**, laquelle pourra évoluer en fonction des connaissances supplémentaires que nous acquerrons sur le système étudié (*cf.* les probabilités conditionnelles, les ensembles canoniques en mécanique statistique).

Il faut bien insister sur le fait que l'étude des phénomènes aléatoires (mot latin ayant trait au jeu de dé) ne présuppose pas l'abandon du déterminisme. Certains problèmes, examinés sous l'angle de la physique traditionnelle, exigent une modélisation très complexe qui, dans les meilleurs des cas, peut conduire à un ensemble d'équations. Ces équations ne sont jamais justiciables d'un traitement analytique et seul l'aspect calcul numérique pourrait être envisagé ; on doit se rendre à l'évidence : le temps de calcul est prohibitif, l'espace mémoire démentiel... bref, le problème est bel et bien irréaliste. C'est ainsi que l'on préfère envisager l'usage d'un autre procédé d'analyse qui sera sans doute beaucoup moins fin mais qui fournira les résultats globaux que nous attendons. Pour se convaincre de cette réalité, il suffit de considérer le simple jeu de pile ou face.

Si l'on fait abstraction de l'hypothèse selon laquelle la pièce pourrait rester en équilibre sur la tranche, ce qui nous intéresse c'est savoir quelle est la face présentée par la pièce à la fin de l'expérience. *A priori*, cette face peut être déterminée à partir du moment où l'on connaît la

position initiale de la pièce, sa forme, sa densité, les forces mises en jeu pour la lancer, plus un certain nombre de renseignements liés à l'atmosphère dans laquelle elle se déplace, c'est-à-dire la densité de l'air, la résistance de l'air, etc. et d'autres renseignements liés à la surface sur laquelle est censée s'immobiliser la pièce par exemple nature du matériau, rugosité etc. Ainsi, le problème présente toutes les qualités pour être modélisé selon les règles classiques et donc de constituer un problème déterministe ordinaire. Point besoin d'être un grand clerc pour se rendre compte que cette conception conduit inexorablement à une impasse, car le problème ainsi envisagé devient immédiatement inextricable. Par conséquent, force est d'abandonner cette méthode et d'utiliser les voies offertes par la théorie des probabilités. Peu importe la façon dont la pièce est parvenue sur l'une des faces, seul le résultat nous intéresse, et cela constitue un **événement**. Dans le cas présent, deux événements sont possibles : la pièce peut présenter face ou pile.

À l'événement, est associée la **probabilité de cet événement**. Sans entrer dans les querelles de définition, on peut dire que la probabilité est la **mesure numérique de la possibilité** de la réalisation effective de l'événement. Cette notion va s'éclairer dans les paragraphes suivants.

Le problème essentiel de la théorie des probabilités repose sur la mesure effective de la probabilité, et il y a essentiellement deux façons de traiter le problème :

**a.** on est en mesure d'effectuer ***a priori*** **une évaluation directe des probabilités** (calcul direct) ;
**b.** on effectue une mesure de la probabilité au moyen d'expériences sur une population, on parlera alors de **probabilité empirique d'un événement** (ou probabilité statistique).

## 2. Évaluation de la probabilité

### 2.1. Calcul direct des probabilités

C'est le cas que l'on rencontre lors de l'étude de systèmes qui peuvent être décrits *a priori* et qui possèdent les propriétés suivantes :

1. Un événement et un seul se manifeste à la fois. Il s'agit par exemple du tirage d'une carte dans un jeu de trente-deux cartes.
2. Les événements sont incompatibles entre eux : ils ne peuvent pas apparaître ensemble. Cela signifie qu'un as de cœur ne peut pas être un roi de pique.
3. Les événements sont équiprobables *a priori*, ou encore les probabilités de chaque événement sont *a priori* égales. Cela veut dire que, pour des raisons de symétrie, aucun des événements ne se manifeste plus que les autres ; il n'y a aucune raison pour que les événements aient des poids différents.

Lorsque la propriété 1 est vérifiée, on dit que le **système est complet**, lorsque les trois propriétés sont vérifiées, on dit que le **système est exhaustif**.

L'intérêt de définir un système exhaustif repose sur la possibilité d'étudier tous les problèmes classiques de jeu de hasard où l'équiprobabilité des événements est assurée d'avance.

Par définition, on dit qu'un **cas est favorable** si son apparition entraîne la réalisation de l'événement en question. Il s'ensuit que la probabilité de l'événement $A$ que nous notons $P(A)$ est alors le rapport du nombre de cas favorables $m$ et du nombre de cas possibles $n$, soit :

$$P(A) = \frac{m}{n} \cdot \tag{19.1}$$

On voit apparaître la propriété « évidente » que $P(A)$ est un nombre compris entre 0 et 1.

### 2.2. Probabilité empirique d'un événement ou mesure de la fréquence relative

Plus simplement il s'agit d'envisager l'étude des cas où la propriété de symétrie a disparu ; c'est le cas, par exemple, du dé pipé. C'est l'expérience qui révèle l'absence de symétrie, ou plutôt une série de $n$ expériences. On appelle **fréquence de $A$** le nombre d'expériences où l'événement $A$ est apparu divisé par le nombre total d'expériences effectuées. On réalise ainsi une mesure expérimentale de la probabilité, et l'on parle dans ce cas de probabilité empirique ou statistique.

À présent apparaît un sérieux problème : la probabilité statistique change d'une série d'expériences à l'autre. Cette difficulté se trouve réglée par l'intermédiaire du **théorème de Bernoulli** :

Soit un nombre $\eta > 0$ aussi petit que l'on veut, la probabilité pour que la différence entre la fréquence d'apparition de l'événement et la probabilité de cet événement soit inférieure à $\eta$, tend vers un lorsque le nombre $n$ des épreuves augmente indéfiniment.

Ce théorème que nous démontrerons plus loin est l'expression de ce qui est appelé **la loi des grands nombres**.

À l'occasion de l'énoncé du théorème de Bernoulli, nous venons d'introduire une notion nouvelle de la convergence : **la convergence en probabilité** :

On dit que $X_n$ tend vers $X_0$ en probabilité si, quel que soit $\varepsilon > 0$ aussi petit que l'on veut, la probabilité de vérifier l'inégalité

$$|X_n - X_0| < \varepsilon$$

tend vers 1 quand $n$ tend vers l'infini.

Notons qu'un écart important peu exister mais il est peu probable, et d'autant plus improbable que $n$ le nombre d'épreuves est grand.

## 3. Notion de variable aléatoire

Par définition, toute variable dont la valeur est fixée par la réalisation d'une épreuve appartenant à une catégorie déterminée est appelée **variable aléatoire.**

Il ne faut pas confondre la variable aléatoire avec les valeurs effectives que cette variable est susceptible de prendre.

Par exemple, la variable aléatoire $X$ désigne le nombre présenté par la face d'un dé normalement numéroté. Ainsi, les valeurs possibles de la variable aléatoire $X$ sont 1, 2, 3, 4, 5 et 6.

Nous venons de définir une variable aléatoire $X$ qui est discrète. Il est tout à fait possible de définir une variable aléatoire **continue** en disant qu'elle peut prendre toutes les valeurs à l'intérieur d'un intervalle $D$ fini ou infini. Par exemple, il suffit de penser à la longueur de la trajectoire suivie par un avion entre Paris et New-York : toutes les valeurs sont possibles à l'intérieur du domaine $D$.

## 4. Somme et produit d'événements. Théorèmes fondamentaux

Ces notions et théorèmes vont nous permettre de réduire une expérience relativement complexe en sommes et produits d'événements à partir desquels il sera aisé d'effectuer des calculs. Il s'agit de mettre en œuvre une méthode indirecte qui peut s'appliquer à tout un ensemble de problèmes que la probabilité soit connue directement ou empiriquement, que les variables aléatoires soient discrètes ou continues.

## 4.1. Définitions

**a.** La somme de deux événements $A$ et $B$ notée $(A+B)$ est un événement $C$ tel que l'événement $A$ se réalise, ou bien l'événement $B$ ou encore les événements $A$ et $B$.
Il s'agit du OU inclusif. Par exemple, le tirage d'un as en deux coups dans un jeu de trente-deux cartes : on peut tirer un as le premier coup, ou bien un as le second coup, ou encore tirer un as le premier coup et un as le second coup.

**b.** Le produit de deux événements $A$ et $B$ noté $(A \cdot B)$ est un événement $C$ consistant en la réalisation simultanée des événements $A$ et $B$.
Il s'agit du ET. Par exemple, le tirage de deux boules blanches dans une urne contenant $n_1$ boules blanches et $n_2$ boules noires (avec $n_1 \geq 2$ et $n_2 \neq 0$), en effectuant deux tirages consécutifs en remettant la première boule tirée dans l'urne. (Rien n'est changé si on ne remet pas la boule dans l'urne.)

## 4.2. Exemples d'application

Nous allons utiliser ces deux notions pour décomposer en « éléments simples » quelques problèmes de probabilité relativement complexes.

On considère le problème suivant : on joue trois fois de suite au jeu de pile ou face et l'on note le résultat $P$ ou $F$ selon que l'on a obtenu pile ou face. On numérote les résultats en donnant un indice à $P$ et $F$. On peut tout de suite remarquer que, dans ce jeu, l'événement contraire à $P$ que l'on note $\overline{P}$ est $F$, et cela peut aider l'écriture des événements.

***a – Problème n° 1 –*** Sur les trois lancements de la pièce, on désire obtenir exactement une seule fois pile.

L'événement qui nous intéresse est l'événement $E$ ainsi défini :

$$E = P_1 \cdot \overline{P_2} \cdot \overline{P_3} + \overline{P_1} \cdot \overline{P_2} \cdot P_3 + \overline{P_1} \cdot P_2 \cdot \overline{P_3}.$$

***b – Problème n° 2 –*** Sur les trois lancements de la pièce, on désire obtenir au moins deux fois pile.

L'événement se note :

$$E = P_1 \cdot P_2 \cdot P_3 + P_1 \cdot P_2 \cdot \overline{P_3} + P_1 \cdot \overline{P_2} \cdot P_3 + \overline{P_1} \cdot P_2 \cdot P_3.$$

À partir de cette décomposition, il sera facile de calculer les probabilités correspondant à la réalisation des événements étudiés dans la mesure où, toutefois, nous savons calculer les probabilités correspondant à la somme et au produit de deux événements. Avant d'aborder cet aspect du problème, il convient de rappeler deux définitions :

***c – Événements incompatibles –*** Ils ne peuvent se réaliser simultanément, par exemple, si l'on tire une carte dans un jeu traditionnel, cette carte ne peut être à la fois un cœur et un pique, en revanche, l'extraction et d'un carreau et d'un roi constitue deux événements compatibles.

***d – Événements indépendants –*** La réalisation de $A$ ne dépend pas de la réalisation de $B$ ou de sa non-réalisation. On dit que deux événements sont indépendants si la réalisation de l'un ne change pas la probabilité de réalisation de l'autre.

### 4.3. Théorème des probabilités totales

La probabilité de réaliser l'événement $E$ constitué par la somme des événements incompatibles $A$ et $B$ est égale à la somme des probabilités de ces événements ; on notera :

$$P(E) = P(A + B) = P(A) + P(B) \ ;$$

s'ensuivent deux corollaires pratiquement évidents :

***a – Corollaire 1*** – Si les événements $A_1, A_2, \ldots, A_n$ forment un système complet d'événements incompatibles, la somme de leurs probabilités est égale à l'unité, soit :

$$\sum_{i=1}^{n} P(A_i) = 1.$$

***b – Corollaire 2*** – La somme des probabilités d'un événement et de son contraire est égale à un.

### 4.4. Théorème des probabilités composées

La probabilité du produit $E$ de deux événements est égale au produit de la probabilité de l'un d'eux et de la probabilité conditionnelle du second, calculée sous la condition que le premier ait lieu :

$$P(E) = P(A \cdot B) = P(A) \cdot P(B/A).$$

Commentons cette proposition. Rappelons que deux événements sont **indépendants**, si la probabilité de $A$ ne dépend pas de la réalisation de $B$, sinon, ils sont **dépendants**. Prenons un exemple : une urne contient deux boules blanches et une boule noire. On considère les événements suivants :

L'événement $A$ : la première boule tirée est blanche.

L'événement $B$ : la deuxième boule tirée est blanche.

On réalise l'expérience suivante : on tire une première boule que l'on place dans la boîte n° 1 sans en prendre connaissance, puis on tire la seconde boule que l'on place dans la boîte n° 2 sans en prendre connaissance. À présent, on peut prendre éventuellement connaissance de la boule contenue dans la boîte n° 2, en ignorant le contenu de la boîte n° 1.

La probabilité de tirer une boule blanche, dans le cas de l'événement $A$, ne connaissant pas l'événement $B$, est $P(A) = 2/3$. En revanche, si l'on sait que l'événement $B$ s'est réalisé, alors la probabilité est $P(A) = 1/2$.

On appelle **probabilité conditionnelle** de $A$ relative à $B$, et l'on note $P(A/B)$ la probabilité de $A$ calculée sachant que $B$ s'est réalisé. On rencontre aussi d'autres façons d'écrire : $P(A|B)$ et $P(A \text{ si } B)$. La notation devient alors :

$$P(A) = 2/3 \quad \text{et} \quad P(A/B) = 1/2.$$

Dans le cas où $A$ est indépendant de $B$, on a : $P(A/B) = P(A)$ Il s'ensuit les corollaires suivants :

***a – Théorème*** – La probabilité de deux événements indépendants est égale au produit des probabilités de ces deux événements.

***b – Théorème*** – Si $A$ est indépendant de $B$, alors $B$ est indépendant de $A$.

## 4.5. Formule des probabilités totales

Il s'agit de calculer la probabilité d'un événement $A$ pouvant avoir lieu simultanément avec l'un des événements $H_1, H_2, \ldots, H_n$ lesquels forment un système complet d'événements incompatibles, on obtient :

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A/H_i).$$

## 4.6. Formule de Bayes (1702–1761) (théorème des causes ou des antécédents)

Envisageons à présent que l'événement $A$ se produise obligatoirement avec l'un des événements $H_1, H_2, \ldots, H_n$ qui sont antérieurs à l'événement $A$ et qui constituent un système complet d'événements incompatibles. L'ensemble des événements $\{H_k\}$ peut être considéré comme la « cause » de $A$. Le mot cause tient à la formulation historique et non à sa nature philosophique, et J. Bertrand dit à ce sujet : « Les causes sont pour nous des accidents qui ont accompagné ou précédé un événement observé ». C'est pourquoi, il serait plus raisonnable de parler d'antécédent que de cause, cependant, par respect de la tradition nous utilisons le mot cause.

Le problème que l'on pose alors est le suivant : sachant que $A$ s'est produit, quelle est la probabilité pour que ce soit précisément la cause $H_i$ qui l'ait produit ? On suppose que l'on connaît les probabilités d'intervention de chaque cause $H_k$. *A priori* ces probabilités sont connues, et on cherche a posteriori la probabilité sachant que l'événement $A$ s'est effectivement réalisé.

Donnons un exemple d'après J. Bertrand. Un joueur prend le pari d'amener avec deux dés un point supérieur à 10. L'événement $A$ est : il a gagné. Ceci peut être réalisé en amenant le 11 et le 12 : telles sont les causes de succès.

Avant l'expérience, les probabilités des causes sont $P(H_1), P(H_2), \ldots, P(H_n)$. Comment ces probabilités se trouvent-elles modifiées par la réalisation de l'événement $A$ ? Il nous faut donc trouver $P(H_k/A)$ pour chaque cause.

La démonstration est simple, nous pouvons écrire :

$$P(H_k \cdot A) = P(A \cdot H_k) = P(A)P(H_k/A) = P(H_k)P(A/H_k) \quad \text{pour } k = 1, 2, \ldots, n,$$

soit encore en tenant compte de la dernière égalité :

$$P(H_k/A) = \frac{P(H_k)P(A/H_k)}{P(A)};$$

enfin, en utilisant la formule des probabilités totales, on obtient :

$$P(H_k/A) = \frac{P(H_k)P(A/H_k)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)}.$$

***Exemple*** – Deux étudiants passent l'examen de MPA. Le premier a la probabilité $p_1 = 0{,}7$ d'obtenir l'examen, le second la probabilité $p_2 = 0{,}6$. Un étudiant a été reçu. Trouver la probabilité que ce soit le premier.

*A priori*, on retient les causes (antécédents) suivantes :

$H_1$ : aucun étudiant n'est reçu ;

$H_2$ : les deux étudiants sont reçus ;

$H_3$ : seul le premier étudiant est reçu ;

$H_4$ : seul le second étudiant est reçu.

Il n'y a pas de difficulté à calculer les probabilités attachées à ces causes :

$$P(H_1) = (1-p_1)(1-p_2) = 0{,}3 \times 0{,}4 = 0{,}12$$
$$P(H_2) = p_1 p_2 = 0{,}7 \times 0{,}6 = 0{,}42$$
$$P(H_3) = p_1(1-p_2) = 0{,}7 \times 0{,}4 = 0{,}28$$
$$P(H_4) = (1-p_1)p_2 = 0{,}3 \times 0{,}6 = 0{,}18$$

***Remarque*** : On vérifie bien que la somme des probabilités est égale à 1.

Les probabilités conditionnelles de l'événement observé $A$ (un étudiant est reçu) liées aux causes $H_k$ s'écrivent :

$$P(A/H_1) = 0{,}0$$
$$P(A/H_2) = 0{,}0$$
$$P(A/H_3) = 1{,}0$$
$$P(A/H_4) = 1{,}0$$

Une fois l'événement $A$ réalisé, les deux premières causes deviennent impossibles, et les probabilités des causes deviennent :

$$P(H_3/A) = \frac{P(H_3)P(A/H_3)}{P(H_3)P(A/H_3) + P(H_4)P(A/H_4)} = \frac{0{,}28}{0{,}28 + 0{,}18} = 0{,}61,$$
$$P(H_4/A) = \frac{P(H_4)P(A/H_4)}{P(H_3)P(A/H_3) + P(H_4)P(A/H_4)} = \frac{0{,}18}{0{,}28 + 0{,}18} = 0{,}39.$$

Il faut bien comprendre ce que signifient ces calculs, et ils peuvent être obtenus par une autre voie. Seules les deux hypothèses $H_3$ et $H_4$ sont compatibles avec la réalisation de l'événement $A$, $P(H_3) + P(H_4)$ n'est plus égal à 1 puisqu'on tient un renseignement de plus : l'événement $A$ s'est réalisé. Les probabilités conditionnelles restent proportionnelles entre elles, il suffit donc de calculer $\alpha$ tel que $\alpha[P(H_3) + P(H_4)] = 1$, donc :

$$\alpha = \frac{1}{P(H_3) + P(H_4)},$$

$$\text{puis } P(H_3/A) = \alpha P(H_3) \quad \text{et} \quad P(H_4/A) = \alpha P(H_4).$$

## 5. Lois de répartition des variables aléatoires

La description complète d'une variable aléatoire, qu'elle soit discrète ou continue, impose la connaissance de l'ensemble des valeurs susceptibles d'être prises par la variable aléatoire mais aussi du poids attaché à chacune des valeurs (ou probabilité).

On se trouve alors dans le cadre d'un système complet d'événements incompatibles.

## 5.1. Cas de la variable discrète

La somme des probabilités attachées à tous les événements est égal à 1. La variable aléatoire discrète $X$ prend ses valeurs sur l'ensemble fini ou infini $(x_1, x_2, x_3, \ldots, x_n)$ — $n$ fini ou non — chaque valeur $x_k$ ayant la probabilité $p_k$ de se réaliser. On a alors :

$$\sum_{k=1}^{n} p_k = 1.$$

La connaissance de la loi entre le poids attaché à chacune des valeurs possibles de la variable aléatoire et ces mêmes valeurs possibles constitue toute l'information nécessaire à la description de la variable aléatoire. Souvent, il est commode d'employer une autre description en utilisant la **fonction de répartition** ou **répartition**.

## 5.2. Définition

La fonction de répartition est la probabilité de l'événement $X < x$. On note :

$$P(X < x) = F(x).$$

C'est une caractéristique universelle de la loi entre les $\{x_k\}$ et les $\{p_k\}$ que les variables soient discrètes ou continues. Ses propriétés essentielles sont :

**a.** $F(x)$ est une fonction non décroissante de $x$.
**b.** $F(-\infty) = 0$.
**c.** $F(+\infty) = 1$.

La répartition permet de connaître la probabilité pour qu'une variable aléatoire tombe dans un intervalle donné.

## 5.3. Cas de la variable continue

La répartition demeure une approche classique, mais on utilise aussi la **densité de probabilité** ou **distribution**, laquelle en première approximation est la dérivée de la fonction de répartition.

La rigueur voudrait que la répartition soit introduite au moyen de l'intégrale de Stieltjes (1856–1894), mais cela n'est point nécessaire pour le propos qui nous préoccupe. Par conséquent le lien entre la répartition $F(x)$ et la densité de probabilité $f(y)$ est réalisé par l'intégrale de Riemann :

$$F(x) = \int_{-\infty}^{x} f(y)\, \mathrm{d}y.$$

Dans les cas ordinaires, la connaissance de $F(x)$ ou de $f(y)$ suffit à caractériser la variable aléatoire étudiée.

En pratique, on aime bien utiliser un certain nombre de grandeurs associées à la variable aléatoire telles que les caractéristiques de position et les moments.

## 5.4. Les caractéristiques de position

***a – La valeur moyenne*** – Encore appelée espérance mathématique d'une variable aléatoire. Elle est définie ainsi :

$$M(X) = \sum_{k=1}^{n} p_k x_k \quad \text{ou} \quad M(X) = \int_{-\infty}^{+\infty} x f(x)\, \mathrm{d}x$$

selon que la variable est discrète ou continue. Comme on le verra plus loin c'est aussi le moment non centré du premier ordre.

***b – Le mode*** – C'est la valeur la plus probable de la variable aléatoire, soit encore le maximum de la densité de probabilité dans le cas l'une répartition unimodale d'une variable aléatoire continue.

***c – La médiane*** – C'est la valeur $m_e$ de la variable aléatoire telle que :

$$P(X < m_e) = P(X > m_e) = 0{,}5.$$

## 5.5. Les moments

Par définition, on appelle **moments non centrés** d'ordre $k$ l'espérance mathématique de la variable $X^k$. On écrit alors :

$$m_k(X) = M[X^k] = \sum_{i=1}^{n} p_i x_i^k \quad \text{ou} \quad = \int_{-\infty}^{+\infty} x^k f(x)\, \mathrm{d}x.$$

Il s'agit de moments tout à fait semblables à ceux de la mécanique, en assimilant $p_i$ à un poids ou $f(x)$ à une densité massique (volumique, surfacique, linéique).

Il est plus utile de définir les **moments centrés** d'ordre $k$ de la variable aléatoire centrée $X^*$. Par définition, la variable aléatoire centrée $X^*$ est la variable aléatoire $X$ qui a subi un décalage d'origine égal à son espérance mathématique $m_1$ :

$$X^* = X - m_1.$$

Les moments centrés s'écrivent :

$$\mu_k = m_k(X^*) = M[X^{*k}] = \sum_{i=1}^{n} p_i (x_i - m_1)^k \quad \text{ou} \quad = \int_{-\infty}^{+\infty} (x - m_1)^k f(x)\, \mathrm{d}x.$$

On vérifie sans peine que le moment centré d'ordre 1 est nul : $\mu_1 = 0$. Le moment centré d'ordre deux s'appelle **variance** $D(X) = \mu_2$, c'est l'espérance mathématique du carré de la variable centrée. Cette grandeur est attachée à la mesure de la **dispersion**, et l'on appelle **écart quadratique moyen** ou **écart type** la quantité $\sigma(X) = \sqrt{D(X)}$.

Le moment d'ordre 3, $\mu_3$, caractérise l'asymétrie tandis que $\mu_4$ caractérise l'aplatissement. Ces deux dernières grandeurs permettent de comparer une loi de distribution donnée à la loi normale (Gauss-Laplace).

## 6. L'inégalité de Bienaymé (1796–1878) - Tchebycheff

On considère une variable aléatoire $X$ de moyenne $m$ et d'écart type $\sigma$. Par ailleurs, on se donne une valeur $t > 0$ arbitraire et l'on se propose de calculer la probabilité $P$ que le module de la variable $(X - m)$ soit supérieur à $t$ fois la grandeur $\sigma$. On recherche donc $P(|X - m| \geq t\sigma)$.

Sur la droite $Ox$, on porte le point $M$ d'abscisse $m$, le point $A$ d'abscisse $(m - t\sigma)$ et le point $B$ d'abscisse $(m + t\sigma)$. Sur $Oy$ on porte la densité de probabilité (*cf.* Fig. 19.1).

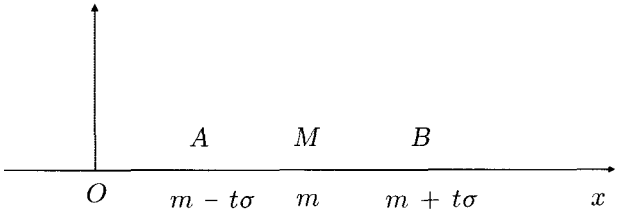


**Figure 19.1.**

En d'autres termes, la probabilité cherchée est celle pour laquelle $X$ n'appartient pas au segment $AB$. Écrivons $D(X) = \sigma^2$ :

$$\sigma^2 = D(X) = \sum_{i=1}^{n} p_i (x_{i-m})^2 \quad \text{ou} \quad \sigma^2 = \int_{-\infty}^{+\infty} (x - m)^2 f(x)\, \mathrm{d}x.$$

On peut encore écrire la décomposition suivante :

$$\sigma^2 = \sum_{X \in (\infty, B)} + \sum_{X \in (A,B)} + \sum_{X \in (A, -\infty)}$$

ou encore

$$\sigma^2 = \int_{-\infty}^{A} + \int_{A}^{B} + \int_{B}^{\infty};$$

quelle que soit l'expression, on peut écrire :

$$\sigma^2 = A + B + C \quad \text{d'où} \quad \sigma^2 \geq A + C.$$

Par ailleurs, hors le segment $AB$ :

$$(x_i - m)^2 > t^2 \sigma^2$$

et l'on peut écrire que :

$$A + C > t^2 \sigma^2 \sum{}' p_i,$$

avec $\sum' p_i$ signifiant que la sommation porte sur les valeurs extérieures au segment $AB$. Mais c'est aussi la probabilité recherchée que l'on désigne par $P$. En définitive, on obtient :

$$\sigma^2 \geq P t^2 \sigma^2 \quad \text{soit} \quad P \leq \frac{1}{t^2}.$$

En conclusion, on voit que la probabilité d'un écart supérieur à $t\sigma$ décroît comme $1/t^2$, et cela quelle que soit la loi de distribution. Cette inégalité de Bienaymé-Tchebycheff trouve une belle application dans la démonstration du théorème de Bernoulli.

## 7. Le théorème de Bernoulli

On se propose donc d'étudier la convergence en probabilité de la moyenne empirique $\nu/n$ vers la moyenne théorique $p$. On verra au chapitre 20 que des considérations sur la loi binomiale donnent les résultats suivants :

$$E\left(\frac{\nu}{n}\right) = p \quad \text{et} \quad D\left(\frac{\nu}{n}\right) = \frac{pq}{n} \quad \text{avec } p+q=1.$$

On utilise l'inégalité de Bienaymé-Tchebycheff en posant :

$$t = \varepsilon\sqrt{\frac{n}{pq}} \quad \text{avec } \varepsilon > 0 \quad \text{arbitraire.}$$

D'où

$$P\left(\left|\frac{\nu}{n} - m\right| \geq \varepsilon\right) \leq \frac{pq}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

Soit $\delta$ un nombre positif tel que $n > \frac{1}{4\delta\varepsilon^2}$. $n$ croissant autant que l'on veut, $\delta$ sera aussi petit que l'on veut, et l'on aura :

$$P\left(\left|\frac{\nu}{n} - m\right| \geq \varepsilon\right) \leq \delta.$$

En définitive, la probabilité pour que la fréquence $\nu/n$ diffère de sa valeur moyenne d'une quantité dont le module est au moins égal à $\varepsilon$, tend vers zéro quand $n$ tend vers l'infini quel que soit $\varepsilon > 0$. Il s'agit, rappelons-le, de la convergence en probabilité.

## 8. Éléments de bibliographie

J. BERTRAND, *Calcul des probabilités*, Chelsea, New-York fac-similé de l'édition française de 1907.

E. BOREL, R. DELTEIL et R. HURON (1962) *Probabilités, Erreurs*, Armand Colin, Paris, 12e édition.

P. JAFFARD (1996) *Initiation aux méthodes de la statistique et du calcul des probabilités*, Masson.

H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 20 La loi binomiale, la loi de Poisson et la loi de Gauss-Laplace

Les fondements de la statistique reposent sur la connaissance fine de quelques lois de distribution dont la principale est la loi de Gauss-Laplace. Cette dernière découle directement de l'étude de la loi binomiale : elle en exprime alors le comportement asymptotique lorsque le nombre d'épreuves devient extrêmement grand (supérieur à 100 pour fixer les idées), autrement dit, on effectue un passage de la variable discrète à la variable continue.

Dans le cas où la probabilité dans la loi binomiale est très petite devant l'unité, sans pour autant que le nombre d'épreuves soit très grand — le produit $\lambda$ de ces deux nombres est proche de l'unité — on débouche sur la loi de Poisson. Celle-ci admet naturellement la loi de Gauss-Laplace comme loi asymptotique dans le cas où le nombre d'épreuves devient très grand, $\lambda$ étant très nettement plus grand que l'unité. Cette loi est utilisée lors de l'étude des « événements rares » qui sont, par exemple, les phénomènes observés par les compagnies d'assurance (sinistres) ; elle entre aussi pour une part active dans l'étude des files d'attente et de la fiabilité.

## 1. La loi binomiale, schéma de Bernoulli

On effectue une série d'épreuves qui donnent lieu chacune à une éventualité parmi deux éventualités, chacune des éventualités ayant une probabilité constante.

Par exemple, on considère une urne qui contient $P$ boules blanches et $Q$ boules noires. On tire une boule, on note sa couleur, puis on la remet dans l'urne et ainsi de suite. Nous avons alors effectué des **tirages non exhaustifs dans une urne à deux catégories**, on dit aussi **tirages avec remise dans une urne à deux catégories, c'est ce qu'on désigne par schéma de Bernoulli.**

Choisissons comme éventualité favorable le tirage d'une boule blanche dont la probabilité est $p$. On réalise $n$ épreuves consécutives. On note d'abord que les tirages sont indépendants les uns des autres. La probabilité pour que $k$ d'entre eux soient favorables, donc $(n-k)$ soient défavorables est donc :

1. $p^k q^{n-k}$ si l'on spécifie l'ordre dans lequel les cas favorables et défavorables doivent se succéder ;
2. $C_n^k p^k q^{n-k}$ si l'ordre n'a pas d'importance.

C'est cette seconde forme qui va faire l'objet de notre attention. Ainsi la probabilité est donnée par l'expression :

$$P_k = \frac{n!}{k!(n-k)!} p^k q^{n-k},$$

c'est aussi le terme de rang $k$ du développement du binôme de Newton :

$$(p+q)^n = q^n + nq^{n-1}p + \cdots + C_n^k q^{n-k}p^k + \cdots p^n.$$

## Calcul des moments

On écrit le binôme sous la forme légèrement modifiée :

$$\Phi(t) = (q+pt)^n = \sum_{k=0}^{n} P_k t^k,$$

où $\Phi(t)$ s'appelle **la fonction génératrice des moments** ; elle constitue un artifice commode pour effectuer le calcul des caractéristiques de position. Le moment d'ordre zéro s'écrit en fonction de $\Phi(t)$ de la manière suivante :

$$\Phi(1) = \sum_{k=0}^{n} P_k.$$

Si l'on dérive les deux membres donnant $\Phi(t)$ par rapport à $t$, on obtient :

$$\Phi'(t) = np(q+pt)^{n-1} = \sum_{k=0}^{n} kP_k t^{k-1}.$$

En faisant $t = 1$, on trouve :

$$\Phi'(1) = np = \sum_{k=0}^{n} kP_k = M(X) = m_1,$$

c'est-à-dire le moment d'ordre un.

Les dérivations successives donnent des expressions fournissant tous les moments successifs après avoir fait $t = 1$. Seul le deuxième moment va retenir notre attention :

$$\Phi''(t) = n(n-1)p^2(q+pt)^{n-2} = \sum_{k=0}^{n} k(k-1)P_k t^{k-2} = \sum_{k=0}^{n} k^2 P_k t^{k-2} - \sum_{k=0}^{n} kP_k t^{k-2}$$

Faisons $t = 1$ :

$$\Phi''(1) = n(n-1)p^2 = m_2 - m_1.$$

De là on déduit :

$$m_2 = n(n-1)p^2 + np = npq + n^2p^2.$$

Or le moment centré d'ordre deux s'écrit :

$$\mu_2 = m_2 - m_1^2 = npq,$$

l'écart type a pour expression : $\sigma = \sqrt{npq}$.

Il s'agit là des deux résultats essentiels de la loi binomiale ; et comme on va le voir, seuls ces deux premiers moments figurent dans la loi de Gauss-Laplace. Avant d'entreprendre l'étude de cette loi, étudions la loi de Poisson qui, elle aussi, est une loi asymptotique de la loi binomiale avant d'admettre elle-même la loi de Gauss-Laplace pour asymptote.

## 2. Loi de Poisson

Nous envisageons le cas où $p$ (ou $q$) est très petit devant l'unité et où $n$ est très grand sans toutefois l'être suffisamment pour que l'on retrouve directement la loi normale, le produit $pn$ doit être de l'ordre de grandeur de l'unité. Reprenons l'expression de $P_k$ :

$$P_k = \frac{n!}{k!(n-k)!}p^k q^{n-k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!(1-p)^k}p^k(1-p)^n,$$

soit encore :

$$P_k = \frac{np(np-p)(np-2p)\dots(np-kp+p)}{k!(1-p)^k}(1-p)^n,$$

ou bien :

$$P_k = \frac{np(np-p)(np-2p)\dots(np-kp+p)}{k!(1-p)^k}\left(1-\frac{np}{n}\right)^n.$$

Posons $\lambda = np$, il vient :

$$P_k = \frac{\lambda(\lambda-p)(\lambda-2p)\dots(\lambda-kp+p)}{k!(1-p)^k}\left(1-\frac{\lambda}{n}\right)^n.$$

Comme $\lambda = np$ est de l'ordre de l'unité, il est grand devant $p$ qui lui-même est petit devant l'unité ; ainsi le numérateur du second membre tend vers $\lambda^k$ et $(1-p)^k$ tend vers un. Il s'ensuit que, puisque $n$ est grand devant $k$, nous pouvons écrire :

$$P_k = \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n.$$

Or $\left(1-\frac{\lambda}{n}\right)^n$ tend vers $\exp(-\lambda)$ quand $n$ tend vers l'infini (résultat classique d'analyse). D'où l'expression de la loi de Poisson :

$$P_k = \frac{\lambda^k}{k!}\exp(-\lambda).$$

La loi de Poisson donne la distribution d'une variable aléatoire discrète qui peut prendre les valeurs entières non négatives $(0, 1, 2, \dots, n, \dots)$.

On dit que la variable $X$ est répartie selon la loi de Poisson si sa probabilité est donnée par l'expression de $P_k$, quel que soit le paramètre $\lambda$.

### 2.1. Propriétés essentielles de la loi de Poisson

Ayant effectué un certain nombre d'opérations sur la loi binomiale, il est indispensable de vérifier que la loi de distribution proposée est bien normalisée, c'est-à-dire : $\sum_{k=0}^{\infty} P_k = 1$. En effet :

$$\sum_{k=0}^{\infty} P_k = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\exp(-\lambda) = \exp(-\lambda)\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \exp(-\lambda)\exp(+\lambda) = 1.$$

### a – Moment du premier ordre

$$m_1 = M(X) = \sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}\exp(-\lambda) = \exp(-\lambda)\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}$$

$$m_1 = \lambda\exp(-\lambda)\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

La moyenne de la variable aléatoire $X$ qui prend les valeurs entières non négatives et qui suit une loi de Poisson est égale au paramètre $\lambda$.

### b – Moment du deuxième ordre

$$m_2 = \sum_{k=0}^{\infty} k^2 P_k = \lambda\exp(-\lambda)\sum_{k=1}^{\infty}\frac{k\lambda^{k-1}}{(k-1)!}\cdot$$

Par le même procédé que celui que nous venons d'utiliser pour calculer le moment du premier ordre, après avoir posé $K = k - 1$, nous obtenons :

$$m_2 = \lambda\exp(-\lambda)\sum_{K=0}^{\infty}\frac{(K+1)\lambda^K}{K!} = \lambda\exp(-\lambda)\left[\sum_{K=0}^{\infty}\frac{K\lambda^K}{K!} + \sum_{K=0}^{\infty}\frac{\lambda^K}{K!}\right].$$

Les deux sommes dans la dernière expression ont été calculées, il suffit donc de reporter leurs valeurs :

$$m_2 = \lambda\exp(-\lambda)(1+\lambda)\exp(+\lambda) = \lambda^2 + \lambda.$$

Comme la variance $D(X) = m_2 - m_1^2$, on en déduit que :

$$D(X) = \sigma^2 = \lambda.$$

Donc le moment centré du deuxième ordre est égal à la moyenne, tous deux étant égaux au paramètre $\lambda$.

***Remarque 1*** : On dit souvent que la loi de Poisson est la loi qui régit les événements rares, et l'on trouve une application importante dans l'étude des séries chronologiques, l'étude des files d'attente ainsi que l'étude de la fiabilité.

***Remarque 2*** : Soit $R_k$ la probabilité pour que $X \geq k$, elle s'écrit simplement :

$$R_k = \sum_{m=k}^{\infty} P_m = 1 - \sum_{j=0}^{k-1} P_j.$$

***c – Processus de Poisson*** – On considère une suite d'événements $\{E_j\}$ qui se succèdent dans le temps. Durant un intervalle fini de temps $\tau$, le nombre $k$ d'événements est aléatoire, et l'on désigne par $\pi_k(\tau)$ la probabilité d'observer $k$ événements pendant l'intervalle de temps $\tau$.

La variable aléatoire discrète $X$ attachée aux événements $\{E_j\}$ est construite de la façon suivante :

**a.** À $t = 0$, $X = 0$.
**b.** $X$ augmente d'une unité lors de l'apparition d'un événement.
**c.** $X$ conserve sa valeur tant qu'aucun événement ne se manifeste.

On définit alors un processus de Poisson de la manière suivante :

1. $\pi_k(\tau)$ ne dépend que de $\tau$ et non de l'origine des temps $\tau_0$.
2. Durant un intervalle de temps infiniment petit $\mathrm{d}\tau$, la probabilité d'apparition d'un événement $E$ est proportionnelle à $\mathrm{d}\tau$, et l'on désigne par $\mathrm{d}\pi_1(d\tau) = r\,\mathrm{d}\tau$ cette probabilité.
3. La probabilité de voir apparaître deux événements ou plus dans l'intervalle de temps infiniment petit $\mathrm{d}\tau$ est infiniment petite par rapport à $\mathrm{d}\tau$.
4. Si $\tau_1$ et $\tau_2$ sont des intervalles de temps sans intersection commune, alors les probabilités $\pi_k(\tau_1)$ et $\pi_k(\tau_2)$ sont indépendantes.

Compte tenu de ces conditions, la probabilité de n'observer aucun événement durant le temps infiniment petit $\mathrm{d}\tau$ est :

$$\mathrm{d}\pi_0(\,\mathrm{d}\tau) = 1 - r\,\mathrm{d}\tau.$$

À présent, on se propose de trouver la loi de distribution $\pi_k(\tau)$ de la variable aléatoire $X$, c'est-à-dire la probabilité d'obtenir exactement $k$ événements durant l'intervalle de temps fini $\tau$.

Découpons l'intervalle de temps $\tau$ en $n$ sous-intervalles égaux de telle sorte que $\Delta\tau = \frac{\tau}{n}$ soit de l'ordre de $\mathrm{d}\tau$.

La probabilité d'observer un événement dans chaque intervalle est donc $r\Delta\tau$ et la probabilité de n'en observer aucun est $1 - r\Delta\tau$. Comme les événements sont indépendants dans les intervalles de temps disjoints de longueur $\Delta\tau$, on peut considérer les $n$ intervalles élémentaires comme $n$ expériences indépendantes pour chacune desquelles un intervalle élémentaire a la probabilité $r\Delta\tau$ de voir apparaître effectivement un événement. La probabilité pour que $k$ des $n$ sous-intervalles aient vu se réaliser un événement est par conséquent donnée par la loi binomiale :

$$\pi_k(\tau) = C_n^k \left(\frac{r\tau}{n}\right)^k \left(1 - \frac{r\tau}{n}\right)^{n-k}.$$

On pose $a = r\tau$, on obtient alors :

$$\pi_k = C_n^k \left(\frac{a}{n}\right)^k \left(1 - \frac{a}{n}\right)^{n-k}.$$

Il nous reste à étudier le comportement de $\pi_k(\tau)$ quand $n$ tend vers l'infini. Pour cela, il suffit de reprendre le calcul déjà effectué lors de l'établissement de la loi de Poisson, et avec la même technique, on trouve :

$$\pi_k = \frac{a^k e^{-a}}{k!}.$$

Donc, la variable aléatoire $X$ suit une loi de Poisson de paramètre $a$.

***Remarque*** : Au lieu de considérer le temps $\tau$, il est possible de s'intéresser à une distribution de points sur la droite ou dans le plan qui obéiraient aux mêmes propriétés que celles qui ont été définies à propos du temps. Il en est de même des points répartis dans l'espace.

***d – Application de la loi de Poisson*** – Une petite route de campagne est fréquentée par 120 voitures par heure en moyenne.

1. Calculer la probabilité $R_1$ pour que la route soit fréquentée douze fois en cinq minutes.
2. Calculer la probabilité $R_2$ pour que la route soit fréquentée au moins une fois en cinq minutes.
3. Calculer la probabilité $R_3$ pour que la route soit fréquentée au moins sept fois en cinq minutes.

Le calcul de $\lambda$ s'effectue sur cinq minutes, on trouve alors :

$$\lambda = \frac{120}{60} \times 5 = 10.$$

Il s'ensuit que :

$$R_1 = \frac{\exp(-10)10^{12}}{12!} = 9{,}48\,10^{-2};$$
$$R_2 = 1 - P_0 = 1 - \exp(-10) = 0{,}999\,954\,6;$$
$$R_3 = 1 - P_0 - P_1 - P_2 - P_3 - P_4 - P_5 - P_6;$$
$$R_3 = 1 - \exp(-10)\left[10 + \frac{10^2}{2!} + \frac{10^3}{3!} + \frac{10^4}{4!} + \frac{10^5}{5!} + \frac{10^6}{6!}\right] = 0{,}869\,90.$$

## 3. Loi de Gauss-Laplace

Il nous faut rechercher la valeur asymptotique de $P_k$ lorsque $n$ devient infiniment grand. C'est la formule de Stirling qui va nous permettre de développer les factorielles, donc la fonction à laquelle nous allons nous intéresser est $\log_e(P_k)$. Il y a alors deux façons de parvenir à nos fins. On développe $\log_e(P_k)$ au deuxième ordre ce qui conduit à des longs calculs un peu fastidieux. Il est alors nécessaire d'user de la formule de Stirling très complète :

$$n! = n^n \exp(-n)\sqrt{2\pi n}(1 + \varepsilon_n) \qquad \text{avec} \qquad n\varepsilon_n \longrightarrow \frac{1}{12}\cdot$$

Un second procédé, moins rigoureux mais plus simple, consiste à choisir le maximum de $\log_e(P_k)$ comme point autour duquel on effectue le développement en série de Taylor, ce qui a pour but de faire disparaître la dérivée première dudit développement. Dans ce cas, l'approximation :

$$\log_e(n!) \approx n\log_e(n) - n$$

sera suffisante. C'est cette méthode que nous présentons. On part donc de la relation :

$$\log_e(P_k) = n\log_e(n) - n - k\log_e(k) + k - (n-k)\log_e(n-k) + (n-k) + k\log_e(p) + (n-k)\log_e(q).$$

À présent, on cherche le maximum de cette expression, il est donné par :

$$\frac{\mathrm{d}}{\mathrm{d}k}\log_e(P_k) = 0 = -\log_e(k) - 1 + 1 + \log_e(n-k) + 1 - 1 + \log_e\left(\frac{p}{q}\right),$$

on obtient alors :

$$(n-k)p = kq,$$
$$\text{d'où :} \quad k = np.$$

C'est la valeur la plus probable, mais on a déjà vu que c'était aussi la moyenne $m_1$ de la loi binomiale. Calculons maintenant la dérivée deuxième :

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}\log_e(P_k) = \frac{-1}{n-k} - \frac{1}{k} = \frac{-n}{k(n-k)} = \frac{-1}{\frac{k}{n}\left(1 - \frac{k}{n}n\right)} = \frac{-1}{npq} = -\frac{1}{\sigma^2},$$

d'où le développement en série de Taylor :

$$\log_e(P_k) = \log_e(P_{\max}) + O(k-np) - \frac{1}{2npq}(k-np)^2,$$

il s'ensuit que :

$$P_k = P_{\max} \exp\left(-\frac{(k-np)^2}{2npq}\right) = P_{\max} \exp\left(-\frac{(k-m_1)^2}{2\sigma^2}\right)$$

ou encore, si $k$ devient une variable aléatoire continue, la probabilité élémentaire $\mathrm{d}P = p(x)\,\mathrm{d}x$ que la variable $x$ soit comprise dans l'intervalle $(x, x+\mathrm{d}x)$, est donnée par l'expression :

$$p(x)\,\mathrm{d}x = P_{\max} \exp\left(-\frac{(k-m_1)^2}{2\sigma^2}\right)\,\mathrm{d}x.$$

La probabilité pour que la variable $x$ soit plus petite que la valeur $X$ s'écrit :

$$P(x<X) = P_{\max} \int_{-\infty}^{X} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right)\,\mathrm{d}x_p tf$$

Il reste à normer la distribution :

$$1 = \int_{-\infty}^{+\infty} P_k\,\mathrm{d}k = P_{\max} \int_{-\infty}^{+\infty} \exp\left(-\frac{(k-m_1)^2}{2\sigma^2}\right)\,\mathrm{d}k = \sigma\sqrt{2\pi}P_{\max}.$$

De là on tire :

$$P_{\max} = \frac{1}{\sigma\sqrt{2\pi}}.$$

En définitive, on obtient :

$$P_k = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(k-m_1)^2}{2\sigma^2}\right) \qquad \text{pour le cas discret, et}$$

$$P(x<X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{X-m_1}{\sigma}} \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x \qquad \text{pour le cas continu.}$$

Cette loi joue un rôle privilégié dans la mesure où de nombreuses distributions admettent la distribution de Gauss-Laplace comme loi asymptotique.

## 3.1. Propriétés essentielles de la loi de Gauss-Laplace

La loi est normalisée puisque nous venons de faire en sorte qu'il en soit ainsi.

***a – Moment du premier ordre***

$$m_1 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \mathrm{d}x = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x+m) \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x$$
$$= \frac{m}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x + \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x.$$

La première intégrale vaut $m$ et la seconde est nulle puisque l'expression sous le signe somme est impaire. Donc :

$$m_1 = m.$$

***b – Calcul de la variance***

$$D(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m)^2 \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \mathrm{d}x = \frac{4\sigma^2}{\sqrt{\pi}} \int_{0}^{+\infty} x^2 \exp(-x^2)\, \mathrm{d}x.$$

Nous allons intégrer par parties cette expression en posant :

$$u = x \qquad \mathrm{d}u = \mathrm{d}x$$
$$\mathrm{d}v = x\exp(-x^2) \qquad v = -\frac{1}{2}\exp(-x^2)$$

donc :

$$D(x) = \frac{2\sigma^2}{\sqrt{\pi}} \left[x\exp(-x^2)\right]_0^{+\infty} + \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{+\infty} \exp(-x^2)\, \mathrm{d}x = \sigma^2.$$

***Remarque*** : Dans les problèmes d'évaluation de la précision, on définit la grandeur $h = \frac{1}{\sigma\sqrt{2}}$ qui est appelée mesure de la précision (*cf.* sur ce sujet, la théorie des erreurs ou incertitudes).

***c – Relation de récurrence entre les moments centrés*** – On pose :

$$m_q = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m)^q \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \mathrm{d}x.$$

On effectue le changement de variable suivant :

$$t = \frac{x-m}{\sigma\sqrt{2}}$$

ce qui donne :

$$m_q = \frac{(\sigma\sqrt{2})^q}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t^q \exp(-t^2)\, \mathrm{d}t = K_q \int_{-\infty}^{+\infty} \tau^q \exp(-t^2)\, \mathrm{d}t.$$

L'intégration par parties permet d'écrire :

$$m_q = K_q \int_{-\infty}^{+\infty} t^{q-1} t \exp(-t^2)\, \mathrm{d}t$$

$$= K_q \left[ \left\{ -\frac{1}{2} t^{q-1} \exp(-t^2) \right\}_{-\infty}^{+\infty} + \frac{q-1}{2} \int_{-\infty}^{+\infty} t^{q-2} \exp(-t^2)\, \mathrm{d}t \right] ;$$

$$\text{donc :} \quad m_q = \frac{(q-1)(\sigma\sqrt{2})^q}{2\sqrt{\pi}} \int_{-\infty}^{+\infty} t^{q-2} \exp(-t^2)\, \mathrm{d}t.$$

Si l'on remarque que :

$$m_{q-2} = \frac{(\sigma\sqrt{2})^{q-2}}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t^{q-2} \exp(-t^2)\, \mathrm{d}t,$$

on trouve en définitive :

$$m_q = (q-1)\sigma^2 m_{q-2}.$$

Comme on connaît les deux premiers moments centrés $m_0 = 1$ et $m_1 = 0$, il est aisé de former la suite des moments centrés :

$$m_2 = \sigma^2,$$
$$m_4 = 3\sigma^4,$$
$$m_6 = 15\sigma^6,$$

tous les moments centrés d'ordre impair étant nuls.

***d – La dissymétrie*** – Elle est donnée par l'expression :

$$\delta = \frac{m_3}{m_2^{3/2}} = \frac{M[x - M(x)]^3}{[D(x)]^{3/2}} \cdot$$

L'étalon étant la courbe de Gauss-Laplace, $\delta = 0$ pour cette dernière.

***e – L'aplatissement*** – Il est donné par une expression qui est telle que la loi de Gauss-Laplace ait un aplatissement nul :

$$a = \frac{m_4}{m_2^2} - 3 \cdot$$

***f – Fonction de répartition de la loi normale*** – Il s'agit de calculer la probabilité de trouver une variable $x$ dans l'intervalle $(a, b)$, sachant que sa distribution obéit à une loi gaussienne :

$$P(a < x < b) = \Psi(b) - \Psi(a) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp\left( -\frac{(t-m)^2}{2\sigma^2} \right) \mathrm{d}t.$$

Comme cette expression dépend de $\sigma$ et $m$, on préfère utiliser la variable centrée réduite, ce qui conduit à la loi normale réduite. On pose alors :

$$y = \frac{x-m}{\sigma} \Longrightarrow \mathrm{d}y = \frac{\mathrm{d}t}{\sigma}$$

on aboutit alors à la relation :

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t-m}{\sigma}} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y.$$

On tabule la loi réduite ($m = 0$, $\sigma = 1$), soit :

$$\Psi(x) = \Phi\left(\frac{x-m}{\sigma}\right);$$

on peut encore écrire :

$$P(a < x < b) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$$

De plus on a :

$$\Phi(-\infty) = 0,$$
$$\Phi(+\infty) = 1.$$

En outre, on vérifie que $\Phi(x)$ est une fonction non décroissante car la fonction de distribution est définie positive.

On trouvera sur le Web (*) le sous-programme `gaussleg.h` qui permet de tabuler la loi normale réduite.

***g – Écart le plus probable*** – On considère un intervalle de largeur $2E$ centré sur la valeur moyenne, tel qu'il y ait la probabilité 0,5 que l'événement étudié tombe dans cet intervalle. On appelle **écart probable**, l'écart ayant la largeur $E$.

Dans le cas de la loi normale, on trouve :

$$P(|X-m| < E) = 0{,}5 = 2\Phi\left(\frac{E}{\sigma}\right) - 1.$$

$$\text{car :} \quad \Phi(-y) = 1 - \Phi(y);$$
$$\text{donc :} \quad \Phi\left(\frac{E}{\sigma}\right) = 0{,}75$$

ce qui donne $E = 0{,}674\sigma$.

***h – Application numérique*** – Lorsque $p$ et $q$ ne sont pas trop proches de zéro ou de un, et à la condition que $n$ soit grand, on peut utiliser la loi de Gauss-Laplace pour évaluer les probabilités des « jeux » relevant du schéma de Bernoulli.

* `http://www.edpsciences.com/guilpin/`

***Exemple n° 1*** – On joue 800 fois à pile ou face, et l'on demande quelle est la probabilité pour que pile sorte moins de 360 fois (10 % inférieur à la moyenne).

$$n = 800\,;\quad p = q = 0{,}5\,;\quad m = np = 400\,;\quad \sigma = \sqrt{npq} = 10\sqrt{2}\,;$$

soit $\xi$ la variable réduite :

$$\xi = \frac{360 - 400}{10\sqrt{2}} = -2{,}828\,427\,125$$

il s'ensuit que (attention, ici, $\xi$ est négatif) :

$$P(x < \xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\xi} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = 2{,}345\,10^{-3}.$$

***Exemple n° 2*** – On réalise 1 000 parties d'écarté, quelle est la probabilité de retourner le roi plus de 140 fois?

$$n = 1000\,;\quad p = \frac{4}{32} = \frac{1}{8}\,;\quad q = \frac{7}{8}\,;\quad m = np = 125\,;\quad \sigma = \sqrt{npq} = 10{,}458\,;$$

soit $\xi$ la variable réduite :

$$\xi = \frac{140 - 125}{10{,}458} = 1{,}434\,274\,331$$

il s'ensuit que :

$$P(x < \xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y = 0{,}076,$$

grosso modo entre 7 et 8 chances sur cent.

***i – Une application de la loi des grands nombres*** – La probabilité d'amener une face bien déterminée d'un dé non pipé est *a priori* $p = 1/6$. On étudie le rapport entre le nombre de fois $n$ où sort la face choisie et le nombre d'épreuves $N$.

Combien faut-il d'épreuves pour que la différence entre ce rapport et la probabilité *a priori* soit inférieure à $\varepsilon = 10^{-3}$ en valeur absolue avec une probabilité supérieure à $\pi = 1 - 10^{-6}$?

La condition à remplir est donc :

$$P\left(\left|p - \frac{n}{N}\right| \leq \varepsilon\right) \geq \pi.$$

***1re approche*** – L'inégalité de Bienaymé-Tchebycheff valable quelle que soit la distribution permet de répondre à la question :

$$P(|X - m| \geq t\sigma) \leq \frac{1}{t^2},$$

$$\text{soit encore :}\quad P(|X - m| \leq t\sigma) \geq 1 - \frac{1}{t^2}\cdot$$

Dans le cas précis, $p = 1/6$, $q = 5/6$, $\sigma = \sqrt{\frac{pq}{N}}$, $\frac{1}{t^2} = 10^{-6}$, puis :

$$\varepsilon = t\sigma = t\sqrt{\frac{pq}{N}} \quad \Longrightarrow \quad t^2 = \frac{N\varepsilon^2}{pq} = 10^6$$

$$\text{d'où} \quad N = \frac{pqt^2}{\varepsilon^2} = 1{,}4\,10^{11} \text{ épreuves.}$$

L'inégalité de Bienaymé-Tchebycheff donne un nombre d'épreuves obtenu par une majoration excessive puisque cette relation ne tient pas compte de la distribution effective de la variable aléatoire. Dans le cas où la loi de distribution est connue, le nombre d'épreuves sera très inférieur à cette première détermination.

***2e approche*** – Compte tenu de l'ordre de grandeur du nombre d'épreuves, il est tout à fait légitime de faire usage de la loi de Gauss-Laplace. On calcule alors l'écart réduit :

$$\frac{p - \dfrac{n}{N}}{\sigma} \leq \frac{\varepsilon}{\sigma} = \frac{\varepsilon\sqrt{N}}{\sqrt{pq}} = \beta,$$

on a alors :

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\beta}^{+\beta} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = 2\Psi(\beta) = 1 - 10^{-6}$$

$$\text{d'où} \quad \Psi(\beta) = 0{,}499\,999\,5 \quad \text{et} \quad \beta = 4{,}5.$$

On obtient : $N = 2{,}8\,10^6$ épreuves.

## 4. Changement de variable aléatoire dans les lois de répartition

Soit $y = \phi(x)$ une fonction quelconque à dérivée $\phi'(x)$ positive et continue par morceaux sur le segment $(0, L)$, à valeurs sur $(a, b)$ (*cf.* Fig. 20.1, page ci-contre).

Soit $\eta = \phi(\xi)$ une variable aléatoire, elle-même fonction d'une autre variable aléatoire $\xi$. On se pose la question de connaître la probabilité de trouver dans l'intervalle $(y', y'')$ sachant que l'on connaît

$$P(x' \leq \xi \leq x'') = \int_{x'}^{x''} p_\xi(x)\, \mathrm{d}x.$$

Soit $x = \Psi(y)$ la fonction inverse de $y = \phi(x)$, sur l'intervalle $(a, b)$ avec $0 \leq x \leq L$. On a l'identité des événements :

$$\begin{aligned} & E\{y' \leq \eta \leq y''\} \\ & E\{\Psi(y') \leq \xi \leq \Psi(y'')\} \\ \text{et} \quad & E\{x' \leq \xi \leq x''\}. \end{aligned}$$

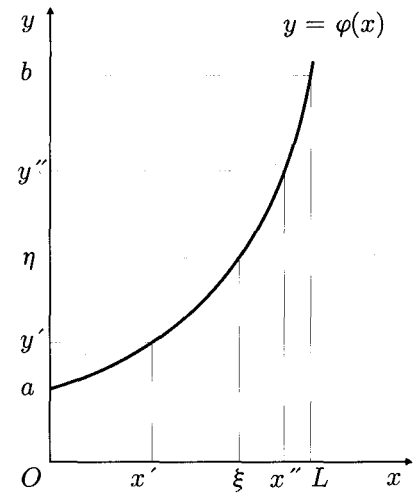


**Figure 20.1.** *Changement de variable dans les lois de distribution.*

La probabilité de l'événement $\xi$ est donnée par :

$$P = \int_{x'}^{x''} p_\xi(x)\,\mathrm{d}x = \int_{\Psi(y')}^{\Psi(y'')} p_\xi(x)\,\mathrm{d}x = P(x' \leq \xi \leq x'') = P(y' \leq \eta \leq y'').$$

Avec le changement de variable $x = \Psi(y)$, on obtient la probabilité attachée à la variable $\eta$. En tenant compte du fait que $\mathrm{d}x = \Psi'(y)\,\mathrm{d}y$, on peut écrire :

$$P(y' \leq \eta \leq y'') = \int_{y'}^{y''} p_\xi(x)\,\mathrm{d}x = \int_{y'}^{y''} p_\xi[\Psi(y)]\Psi'(y)\,\mathrm{d}y = \int_{y'}^{y''} p_\eta(y)\,\mathrm{d}y,$$

$$\text{car} \quad \phi'(x) = \frac{1}{\Psi'(y)} \quad \text{et} \quad \Psi'(y) = \frac{1}{\phi'(x)} = \frac{1}{\phi'[\Psi(y)]}.$$

En définitive, on obtient la distribution désirée :

$$p_\eta(y) \begin{cases} = p_\xi[\Psi(y)]\dfrac{1}{\phi'[\Psi(y)]} & \text{si } a \leq y \leq b \\ = 0 & \text{si } y < a \text{ et } y > b. \end{cases}$$

***Remarque*** : Pour éviter les omissions lors des changements de variable dans les fonctions de distribution, il suffit de penser à manipuler non pas la distribution $f(x)$ mais la probabilité élémentaire $\mathrm{d}P$, c'est-à-dire $\mathrm{d}P = f(x)\,\mathrm{d}x$ pour que $x$ soit compris entre $x$ et $x + \mathrm{d}x$, ainsi, en effectuant le changement de variable $x = g(y)$, on obtient :

$$\mathrm{d}P = f(x)\,\mathrm{d}x = f[g(y)]\frac{\mathrm{d}x}{\mathrm{d}y}\,\mathrm{d}y = f[g(y)]g'(y)\,\mathrm{d}y = h(y)\,\mathrm{d}y.$$

### Exemple

On effectue le changement de variable $y = x^2$ dans la loi normale réduite :

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right).$$

On a donc : $x = \sqrt{y}$, et l'on note la fonction inverse $\Psi = \sqrt{x}$. On écrit alors :

$$\frac{\mathrm{d}\Psi}{\mathrm{d}y} = \Psi'(y) = \frac{1}{2\sqrt{y}}$$

$$\text{il s'ensuit que :} \quad \pi(y) = \frac{1}{2\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right).$$

Attention, la loi de probabilité que nous venons de trouver est deux fois trop petite car :

$$P(y' \leq \eta \leq y'') = P(-x' \leq \xi \leq -x'') + P(x' \leq \xi \leq x'') = \int_{-x'}^{-x''} p_\xi(x)\,\mathrm{d}x + \int_{x'}^{x''} p_\xi(x)\,\mathrm{d}x$$

$$= \int_{x'}^{x''} [p_\xi(-x) + p_\xi(x)]\,\mathrm{d}x = \int_{-x'}^{-x''} \frac{1}{2\sqrt{y}} [p_\xi(-\sqrt{y}) + p_\xi(\sqrt{y})]\,\mathrm{d}y.$$

En définitive, la densité de probabilité s'écrit :

$$\pi(y) = \frac{1}{\sqrt{2\pi y}} \exp(-\frac{y}{2}) \quad \text{si } y \geq 0,$$

$$\text{et} \quad \pi(y) = 0 \quad \text{si } y < 0.$$

Cet exemple nous sera de quelque utilité lors de l'étude de la distribution du $\chi^2_m$ à $m$ degrés de liberté.

## 5. Éléments de bibliographie

S. AÏVAZIAN (1970) *Étude statistique des dépendances*, Moscou.
D.R. COX et P.A.W. LEWIS (1969) *L'analyse statistique des séries d'événements*, Éditions Dunod.
V. ROTHSCHILD et N. LOGOTHETIS (1986) *Probability distributions*, Wiley.
Y. ROZANOV (1975) *Processus aléatoires*, Éditions MIR, Moscou.
A.W. VAART et J.A. WELLNER (1996) *Weak convergence and empirical process*, Springer.
H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 21 | La fonction caractéristique

Certaines opérations mathématiques sont plus commodes à réaliser dans l'espace réciproque de Fourier que dans l'espace direct (propriétés des espaces duals). En théorie des probabilités, il peut être plus simple de traiter ou bien de la fonction de distribution ou bien de la **fonction caractéristique**. Cela dit, il ne faut pas fonder trop d'espoir sur cette conception, c'est une technique parmi d'autres techniques. En effet, on se persuade des limitations pratiques de cette méthode en consultant une table d'intégrales de Fourier : il y a peu de cas de figures. Quoi qu'il soit, la méthode offre un indiscutable intérêt tant théorique que pratique, aussi allons-nous y consacrer un chapitre.

## 1. Définition et propriétés

Soit $\xi$ une variable aléatoire, la fonction caractéristique est donnée par l'expression :

$$\Phi(t) = M[\exp(jtx)]$$

où $M[\ ]$ est l'opérateur moyenne.

Si la variable aléatoire est discrète ($k$ entier appartenant $(-\infty; +\infty)$), on écrit :

$$\Phi(t) = \sum_{k=-\infty}^{+\infty} p(k) \exp(jtk).$$

Il ne s'agit simplement que du développement en série de Fourier de la probabilité. $\Phi(t)$ est une fonction périodique de période unité.

Si la variable aléatoire est continue sur un intervalle fini ou infini, et si l'on désigne par $p(x)$ la densité de probabilité, on écrit :

$$\Phi(t) = \int_{-\infty}^{+\infty} p(x) \exp(jtx)\, \mathrm{d}x.$$

Il s'agit de la transformée de Fourier de la densité de probabilité.

En général, on peut écrire la transformation inverse :

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(t) \exp(-jtx)\, \mathrm{d}t.$$

### 1.1. Théorème (sans démonstration)

Une distribution de probabilité est déterminée d'une manière univoque par sa fonction caractéristique.

### 1.2. Calcul des moments

Considérons le développement de $\exp(jtx)$ :

$$\exp(jtx) = \sum_{k=0}^{n} j^k \frac{x^k}{k!} t^k + \Theta \frac{x^{n+1}}{(n+1)!} t^{n+1} \qquad \text{où } |\Theta| \leq 1.$$

Si les moments $\mu_k = M[x^k]$ existent pour $k = 0, 1, 2, \ldots, n+1$, alors, compte tenu de la relation :

$$M[\exp(jt\xi)] = \sum_{k=0}^{n} j^k \frac{M[x^k]}{k!} t^k + M[\Theta x^{n+1}] \frac{t^{n+1}}{(n+1)!},$$

on peut alors écrire :

$$\Phi(t) = \sum_{k=0}^{n} j^k \frac{\mu_k}{k!} t^k + \frac{R_n}{(n+1)!} t^{n+1}$$

avec :

$$|R_n| \leq M[|x^{n+1}|].$$

Ainsi, on trouve :

$$\mu_0 = \Phi(0) = 1,$$
$$\mu_k = j^{-k} \Phi^{(k)}(0), \qquad \text{pour } k = 1, 2, \ldots, n.$$

On peut donc calculer les moments $\mu_k = M[x^k]$ à partir des dérivées de la fonction caractéristique de la variable aléatoire.

### 1.3. Propriété de la fonction caractéristique

La somme des variables **indépendantes** :

$$\xi = \xi_1 + \xi_2 + \cdots + \xi_n$$

a une fonction caractéristique de la forme :

$$\Phi(t) = \Phi_1(t) \cdot \Phi_2(t) \ldots \Phi_n(t)$$

qui est le produit des fonctions caractéristiques de chacune des variables aléatoires indépendantes $\xi_1, \xi_2, \ldots, \xi_n$. Cela découle de la relation de définition :

$$M[\exp(jtx)] = M[\exp jt(x_1 + x_2 + \cdots + x_n)] = \prod_{k=1}^{n} M[\exp(jtx_k)].$$

### 1.4. Cas particulier de la somme de deux variables indépendantes

La fonction caractéristique de la somme de deux variables indépendantes est donc le produit de chacune des deux fonctions caractéristiques, soit :

$$\Phi(t) = \Phi_1(t) \cdot \Phi_2(t).$$

La transformée de Fourier inverse va transformer ce produit simple en un produit de convolution, nous obtenons :

$$p(x) = \int_{-\infty}^{+\infty} \prod_1(x-t) \cdot \prod_2(t)\, \mathrm{d}t$$

expression dans laquelle $\prod_k(t)$ est la transformée de Fourier de $\Phi_k(t)$, $k = 1, 2$.

### 1.5. Quelques autres propriétés importantes

Soit une variable $\xi$ de densité de probabilité $p(x)$. La fonction caractéristique de cette variable s'écrit donc :

$$\Phi(t) = \int_{-\infty}^{+\infty} p(x) \exp jxt \, \mathrm{d}x.$$

La variable $\xi/n$, ayant la densité de probabilité $np(ny)$, admettra comme fonction caractéristique (après changement de variable) :

$$\int_{-\infty}^{+\infty} p(x) \exp j\frac{x}{n}t \, \mathrm{d}x = \int_{-\infty}^{+\infty} p(x) \exp j\frac{t}{n}x \, \mathrm{d}x = \Phi\left(\frac{t}{n}\right),$$

ce résultat étant obtenu par application directe de la définition. Dans le même esprit, la moyenne $\xi$ d'une somme de variables indépendantes de même densité de probabilité $p(x)$ :

$$x = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

admet comme fonction caractéristique :

$$\left[\Phi\left(\frac{t}{n}\right)\right]^n.$$

Dans bien des cas, on s'intéresse aux logarithmes de la fonction caractéristique :

$$\Psi(t) = \log_e[\Phi(t)].$$

En particulier, dans le cas de la dernière variable $\xi$, la fonction $\Psi(t)$ s'écrit :

$$\Psi(t) = n \log_e \left[\Phi\left(\frac{t}{n}\right)\right].$$

Si par hasard $\Phi\left(\frac{t}{n}\right)$ admet un développement du genre $1 - \frac{t^2}{2n}$ lorsque $n$ est grand devant $t$, on aura toutes les chances de montrer que $\xi$ obéit à une loi gaussienne puisque la fonction caractéristique sera gaussienne...

### 1.6. Exercice

**a.** Soit deux variables aléatoires indépendantes $\xi_1$ et $\xi_2$, qui sont toutes les deux normales (gaussiennes) centrées et réduites. Calculer la fonction caractéristique de chacune d'elle, puis la fonction caractéristique de la variable $\eta = \xi_1 + \xi_2$. En déduire la fonction de distribution de $\eta$.

**b.** Les variables indépendantes sont toujours gaussiennes mais ne sont plus centrées réduites. On désigne par $m_1$ et $m_2$ les moyennes respectives et par $\sigma_1$ et $\sigma_2$ les écarts quadratiques moyens. Donner la fonction de distribution de $\eta = \xi_1 + \xi_2$.

**c.** Généraliser le calcul précédent du §$b$ en considérant la variable

$$\eta = \sum_{k=1}^{n} x_k,$$

puis en déduire le théorème d'addition des moyennes et le théorème d'addition des variances (écart type élevé au carré).

## 2. La distribution du $\chi^2$

On se propose de rechercher la fonction caractéristique, puis la distribution, de la variable aléatoire :

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2$$

qui est la somme des carrés de variables aléatoires, indépendantes, gaussiennes centrées réduites. Chacune de ces variables gaussiennes variables à une densité de probabilité de la forme :

$$\pi_j(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad \text{avec } x \in (-\infty; +\infty).$$

Il s'ensuit que les variables $\xi_j^2$ ont une densité de probabilité de la forme :

$$p_j(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right) \qquad \text{si } y \in (0; +\infty)$$
$$p_j(y) = 0 \qquad \text{si } y \in (-\infty; 0).$$

Compte tenu de ce que :

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} \exp(-t)\, \mathrm{d}t \quad \text{(fonction gamma)},$$

on obtient la fonction caractéristique de la variable $\xi_j^2$ :

$$\frac{1}{\sqrt{2\pi}} \int_0^{+\infty} p_k(x) \exp(jtx)\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} \frac{\mathrm{d}x}{\sqrt{x} \exp\left[(1-2jt)\frac{x}{2}\right]}$$
$$= \frac{1}{\sqrt{\pi(1-2jt)}} \int_0^{+\infty} x^{-1/2} \exp(-x)\, \mathrm{d}x = \Gamma(1/2) \frac{1}{\sqrt{\pi(1-2jt)}} = (1-2jt)^{-1/2}.$$

Il s'ensuit que la fonction caractéristique du $\chi^2_n$ à $n$ degrés de liberté :

$$\Phi(t) = (1 - 2jt)^{-n/2} \quad \text{pour } t \in (-\infty; +\infty).$$

Pour obtenir la fonction de distribution, il n'est pas utile de calculer la transformée de Fourier inverse, pour cela il suffit de remarquer que :

$$A \int_0^{+\infty} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \exp(jtx)\, \mathrm{d}x = (1 - 2jt)^{-n/2}$$

$A$ étant un coefficient de normalisation. Donc la densité de probabilité du $\chi^2_n$ s'écrit :

$$p(x) = Ax^{n/2-1} \exp\left(-\frac{x}{2}\right) \quad \text{pour } x \in (0; +\infty),$$

avec

$$A = \frac{1}{2^{n/2}\Gamma(n/2)} \cdot$$

d'où l'expression de la distribution du $\chi^2$ à $n$ degrés de liberté :

$$p(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \cdot$$

## 3. Éléments de bibliographie

M. FISZ (1980) *Probability theory and Mathematical Statistics*, Robert E. Krieger publishing company.
E. LUKACS (1964) *Fonctions caractéristiques*, Dunod.
R. PETIT (1995) *L'outil mathématique*, Masson.
Y. ROZANOV (1975) *Processus aléatoires*, Éditions MIR, Moscou.
H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 22 La loi du $\chi^2$ et la loi de Student

L'étude des critères de conformité va nous conduire à étudier la distribution d'une somme des carrés de $n$ variables gaussiennes indépendantes (centrées réduites), il s'agit alors de la loi du $\chi^2_n$ à $n$ degrés de liberté.

Lorsqu'on aborde l'étude de la corrélation et de la régression, l'évaluation des intervalles de confiance conduit inévitablement à l'étude de la distribution de Student qui est la distribution d'une variable gaussienne (centrée réduite) divisée par la racine carrée de la moyenne d'un $\chi^2_m$ à $m$ degrés de liberté, ces deux variables étant indépendantes.

Ce chapitre est donc destiné à la présentation desdites lois qu'il convient de bien manipuler pour exploiter correctement les données expérimentales.

## 1. La loi du $\chi^2_n$

Nous avons établi au précédent chapitre la loi de distribution d'une somme de $n$ variables gaussiennes centrées réduites, indépendantes, élevées au carré et nous nous proposons d'en étudier les principales propriétés. Auparavant, nous allons effectuer quelques rappels concernant la fonction $\Gamma(z)$ (fonction gamma ou fonction eulérienne de deuxième espèce) qui est étroitement liée à la loi du $\chi^2_n$ :

$$\begin{aligned} \Gamma(m) &= (m-1)! \\ \Gamma(1/2) &= \sqrt{\pi} \\ \Gamma(n+1/2) &= \frac{1 \times 3 \times 5 \times \cdots (2n-1)}{2^n}\sqrt{\pi} \\ \text{et :}\quad \Gamma(z) &= \int_0^{+\infty} t^{z-1}\exp(-t)\,\mathrm{d}t. \end{aligned}$$

La densité de probabilité de la loi du $\chi^2_n$ à $n$ degrés de liberté est donnée par l'expression :

$$p_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}\exp\left(-\frac{x}{2}\right) \qquad \text{pour } x \in (0,+\infty),$$

tandis que la loi de répartition s'écrit :

$$L_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^x y^{n/2-1}\exp\left(-\frac{y}{2}\right)\,\mathrm{d}y \qquad \text{pour } x \in (0,+\infty).$$

### 1.1. Vérification de la normalité de la loi

Il s'agit de calculer $L_n(\infty)$, on a donc :

$$L_n(\infty) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty y^{n/2-1} \exp\left(-\frac{y}{2}\right) \, \mathrm{d}y.$$

Effectuons un changement de variable $u = y/2$, on obtient :

$$L_n(\infty) = \frac{2}{2^{n/2}\Gamma(n/2)} \int_0^\infty (2u)^{n/2-1} \exp(-u) \, \mathrm{d}u = \frac{1}{\Gamma(n/2)} \int_0^\infty u^{n/2-1} \exp(-u) \, \mathrm{d}u = 1.$$

### 1.2. Calcul de la valeur moyenne

Nous voulons calculer la moyenne de la variable $\chi_n^2$ :

$$M\left(\chi_n^2\right) = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty y y^{n/2-1} \exp\left(-\frac{y}{2}\right) \, \mathrm{d}y,$$

le même changement de variable que celui effectué dans le paragraphe précédent permet d'écrire :

$$M\left(\chi_n^2\right) = \frac{2}{\Gamma(n/2)} \int_0^\infty u^{n/2} \exp(-u) \, \mathrm{d}u,$$

d'où :

$$M\left(\chi_n^2\right) = \frac{2\Gamma(n/2+1)}{\Gamma(n/2)} = n.$$

### 1.3. Calcul de la variance D $(\chi_n^2)$

Au préalable, on calcule le moment d'ordre deux que l'on note $m_2$ :

$$\begin{aligned} m_2\left(\chi_n^2\right) &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty y^2 y^{n/2-1} \exp\left(-\frac{y}{2}\right) \, \mathrm{d}y = \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty y^{n/2+1} \exp\left(-\frac{y}{2}\right) \, \mathrm{d}y \\ &= \frac{1}{2^{n/2}\Gamma(n/2)} 2^{n/2+2}\Gamma(n/2+2) = (m+2)m. \end{aligned}$$

Comme on a :

$$D\left(\chi_n^2\right) = m_2\left(\chi_n^2\right) - M^2\left[\chi_n^2\right],$$

on en déduit :

$$D\left(\chi_n^2\right) = 2n.$$

On donne sur le Web (*) le sous-programme `khi2.h` qui permet de tabuler la loi du $\chi_m^2$ à $m$ degrés de liberté.

* `http://www.edpsciences.com/guilpin/`

## 2. Distribution d'une somme de deux variables aléatoires indépendantes obéissant chacune à une distribution du $\chi^2_n$

On considère une variable $\Theta$ qui est la somme de deux variables aléatoires indépendantes obéissant à une loi du $\chi^2$ l'une à $n$ degrés de liberté, l'autre à $m$ degrés de liberté :

$$\Theta = \chi^2_n + \chi^2_m.$$

La fonction caractéristique de $\Theta$ est donc :

$$\Phi(t) = (1-2jt)^{-n/2}(1-2jt)^{-m/2} = (1-2jt)^{-(n+m)/2},$$

alors, en faisant $n = n+m$ dans les relations établissant la loi du $\chi^2_n$ dans le paragraphe précédent, on en déduit :

$$A\int_0^{+\infty} y^{(n+m)/2-1}\exp\left(-\frac{y}{2}\right)\exp(jtx)\,\mathrm{d}x = (1-2jt)^{-(n+m)/2}$$

avec

$$A = \frac{1}{2^{(n+m)/2}\Gamma[(n+m)/2]}\cdot$$

Il s'ensuit que la fonction de distribution s'écrit :

$$p(y) = \frac{1}{2^{(n+m)/2}\Gamma[(n+m)/2]}y^{(n+m)/2-1}\exp\left(-\frac{y}{2}\right),$$

on en conclut que la distribution d'une somme de deux variables indépendantes obéissant chacune à une distribution du $\chi^2$ respectivement à $n$ et $m$ degrés de liberté est encore une distribution du $\chi^2$ mais à $m+n$ degrés de liberté. Il est aisé de généraliser :

***Théorème*** – Une variable qui est la somme de $m$ variables aléatoires indépendantes obéissant chacune à une distribution du $\chi^2$ obéit encore une distribution du $\chi^2$.

## 3. La loi du $\chi^2_m$ à m degrés de liberté tend asymptotiquement vers la loi de Gauss quand m tend vers l'infini

Exprimons la loi du $\chi^2_m$ à $m$ degrés de liberté en variable centrée réduite, en rappelant que $M[x] = m$ et $D[x] = 2m$ $(\sigma = \sqrt{2m})$ :

$$y = \frac{x-m}{\sqrt{2m}},$$

on obtient :

$$p_m(y) = \frac{\sqrt{2m}}{2^{m/2}\Gamma(m/2)}\left(\sqrt{2m}y+m\right)^{m/2-1}\exp\left[-\frac{(\sqrt{2m}y+m)}{2}\right].$$

Passons aux logarithmes :

$$\log_e[p_m(y)] = \text{Cste} + \frac{1}{2}\log_e(m) - \frac{m}{2}\log_e(2) - \frac{m-2}{2}\log_e\left(\frac{m-2}{2}\right) + \frac{m-2}{2}$$
$$- \frac{1}{2}\log_e\left(\frac{m-2}{2}\right) + \frac{m-2}{2}\log_e(m) + \frac{m-2}{2}\log_e\left(1 + \frac{y}{\sqrt{\frac{m}{2}}}\right) - y\sqrt{\frac{m}{2}} - \frac{m}{2},$$

en remarquant que :

$$\log_e\left(1 + \frac{y}{\sqrt{\frac{m}{2}}}\right) = \frac{y}{\sqrt{\frac{m}{2}}} - \frac{y^2}{m} + \cdots$$

tous calculs faits, quand $m$ tend vers l'infini :

$$\log_e[p_m(y)] = \text{Cste} - \frac{y^2}{2},$$

d'où :

$$p_m(y) = A\exp\left(-\frac{y^2}{2}\right) \qquad \text{avec } A = \frac{1}{\sqrt{2\pi}}\cdot$$

On reconnaît la loi normale réduite.

## 4. Distribution d'une variable aléatoire fonction de deux variables aléatoires indépendantes

On considère la variable aléatoire $\tau$ qui est une fonction de deux variables aléatoires indépendantes $\chi$ et $\eta$. On désigne par $p_\chi(x)$ et $p_\eta(y)$ les fonctions de distribution de ces deux variables. *A priori* et sans nuire à la généralité, on peut convenir que ces fonctions sont définies sur $(-\infty, +\infty)$. Nous avons donc :

$$\tau = f(x, y),$$

et nous voulons obtenir la distribution $p_\tau(t)$ de la variable aléatoire $\tau$. Comme les variables sont indépendantes nous avons :

$$p_\tau(t) = p_\xi(x) \cdot p_\eta(y)$$

On cherche à expliciter $p_\tau(t)$. Pour cela on va rechercher la répartition $P_\tau(\alpha)$ pour que $f(x, y) < \alpha$. Cette répartition ne sera rien d'autre que l'intégrale de $p_\xi(x) \cdot p_\eta(y)$ étendue au domaine $f(x, y) < \alpha$. Soit :

$$P_\tau(\alpha) = \iint\limits_{f(x,y)<\alpha} p_\xi(x) \cdot p_\eta(y)\, \mathrm{d}x\, \mathrm{d}y.$$

### 4.1. Cas particulier n° 1

Dans le cas où l'on peut écrire $x = \alpha\Phi(y)$, alors on effectue un changement de variables très intéressant à savoir :

$$v = \Phi(y) \quad \text{et} \quad x = uv.$$

Le jacobien de la transformation s'écrit :

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} v & u \\ 0 & \dfrac{1}{\mathrm{d}v/\,\mathrm{d}y} \end{pmatrix} = v\frac{1}{\mathrm{d}v/\,\mathrm{d}y}\cdot$$

Ainsi, $P_\tau(\alpha)$ devient :

$$P_t(\alpha) = \int_{-\infty}^{\alpha} \mathrm{d}u \int_{-\infty}^{+\infty} p_\xi(u\cdot v)\cdot p_\eta\left[\Phi^{-1}(v)\right]\frac{\partial(x,y)}{\partial(u,v)}\,\mathrm{d}v = \int_{-\infty}^{\alpha} p_t(u)\,\mathrm{d}u.$$

***Application au rapport de deux variables indépendantes*** – La variable $\alpha$ est donc un rapport :

$$\alpha = x/y,$$

et l'on effectue le changement de variables :

$$y = v \quad \text{et} \quad y = u\cdot v$$

dans ce cas, le jacobien de la transformation devient :

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

d'où :

$$P_\tau(\alpha) = \int_{-\infty}^{\alpha} \mathrm{d}u \int_{-\infty}^{+\infty} p_\xi(u\cdot v)\cdot p_\eta(v)v\,\mathrm{d}v.$$

***Remarque importante*** – Il convient de prendre garde aux signes de la distribution au cas où la variable $\eta$ de distribution $p_\eta(y)$ est définie sur $(-\infty, +\infty)$ :

$$\prod_\tau(u) = \int_0^{+\infty} p_\xi(u\cdot v)\cdot p_\eta(v)v\,\mathrm{d}v - \int_{-\infty}^{0} p_\xi(u\cdot v)\cdot p_\eta(v)v\,\mathrm{d}v.$$

En effet une densité de probabilité ne peut pas être négative et, seule, dans ces expressions, la variable $v$ change de signe en franchissant la valeur zéro.

### 4.2. Cas particulier n° 2

Dans le cas où l'on peut écrire $x = \alpha + \Phi(y)$, alors on effectue un changement de variables très intéressant à savoir :

$$v = \Phi(y) \quad \text{et} \quad x = u + v.$$

Le jacobien de la transformation s'écrit :

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} 1 & 1 \\ 0 & \dfrac{1}{\dfrac{\mathrm{d}v}{\mathrm{d}y}} \end{vmatrix} = \frac{1}{\dfrac{\mathrm{d}v}{\mathrm{d}y}}.$$

Ainsi, $P_\tau(\alpha)$ devient :

$$P_\tau(\alpha) = \int_{-\infty}^{\alpha} \mathrm{d}u \int_{-\infty}^{+\infty} p_\xi(u+v) \cdot p_\eta\left[\Phi^{-1}(v)\right] \frac{1}{\dfrac{\mathrm{d}v}{\mathrm{d}y}}\, \mathrm{d}v = \int_{-\infty}^{\alpha} p_\tau(u)\, \mathrm{d}u.$$

***Application à la somme de deux variables indépendantes*** – Donc la variable $\alpha$ est une somme, on a alors :

$$\alpha = x + y$$

on effectue alors le changement de variables :

$$y = v \quad \text{et} \quad x = u - v$$

dans ce cas, le jacobien de la transformation devient :

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1$$

d'où :

$$P_\tau(\alpha) = \int_{-\infty}^{\alpha} \mathrm{d}u \int_{-\infty}^{+\infty} p_\xi(u-v) \cdot p_\eta(v)\, \mathrm{d}v.$$

On retrouve le résultat que nous avions établi à l'aide de la fonction caractéristique.

## 5. La distribution de Student (W. Gosset) (1876–1937)

C'est la distribution de la variable aléatoire :

$$t = \frac{\xi}{\sqrt{\dfrac{\chi^2_m}{m}}} = \frac{\xi}{\eta}$$

expression dans laquelle $\xi$ est une variable gaussienne centrée réduite (indépendante de $\chi^2_m$) de loi de distribution :

$$p_\xi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right).$$

Il sera aisé d'appliquer les résultats du paragraphe précédent lorsque nous connaîtrons la distribution $p_\eta(y)$ de la variable $\eta = \sqrt{\dfrac{\chi^2_m}{m}}$. Il convient pour ce faire d'effectuer un changement de variable :

$$y = \sqrt{\frac{x}{m}} \quad \Longrightarrow \quad x = my^2 \quad \text{pour } x \in (0, +\infty).$$

On peut écrire :

$$y = \Phi(x) \qquad x = \Psi(y) \qquad \Psi'(y) = 2my.$$

La distribution devient donc $p_\eta(y) = p[\Psi(y)]\Psi'(y)$ soit :

$$p_\eta(y) = \frac{1}{2^{m/2}\Gamma(m/2)}(my^2)^{m/2-1}\exp\left(-\frac{my^2}{2}\right)2my = \frac{m^{m/2}}{2^{m/2-1}\Gamma(m/2)}y^{m-1}\exp\left(-\frac{my^2}{2}\right)$$
pour $x \in (0, +\infty)$.

La distribution $p_m(t)$ de la variable $t$ s'écrit alors :

$$p_m(t) = \int_0^{+\infty} p_\xi(t\cdot v)\cdot p_\eta(v)v\,\mathrm{d}v = \frac{m^{m/2}}{\sqrt{2\pi}2^{m/2-1}\Gamma(m/2)}\int_0^{+\infty} y^{m-1}\exp\left(-\frac{my^2}{2}\right)$$
$$\times \exp\left(-\frac{y^2t^2}{2}\right)y\,\mathrm{d}y = \frac{m^{m/2}}{\sqrt{2\pi}2^{m/2-1}\Gamma(m/2)}\int_0^{+\infty} y^m\exp\left[-\frac{my^2}{2}\left(1+\frac{t^2}{m}\right)\right]\mathrm{d}y,$$

on pose :

$$\frac{my^2}{2}\left(1+\frac{t^2}{m}\right) = u \quad \Longrightarrow \quad \mathrm{d}u = \left(1+\frac{t^2}{m}\right)my\,\mathrm{d}y\,;$$

soit

$$y = \sqrt{\frac{2u}{m\left(1+\frac{t^2}{m}\right)}}$$

on trouve alors :

$$p_m(t) = \frac{m^{m/2}}{m\sqrt{2\pi}2^{m/2-1}\Gamma(m/2)}\int_0^{+\infty}\left[\frac{2u}{m\left(1+\frac{t^2}{m}\right)}\right]^{m/2}\exp(-u)\frac{1}{\left(1+\frac{t^2}{m}\right)}\,\mathrm{d}u$$
$$= \frac{m^{m/2}2^{(m-1)/2}}{m\sqrt{2\pi}2^{m/2-1}\Gamma(m/2)m^{(m-1)/2}}\left[1+\frac{t^2}{m}\right]^{-(m+1)/2}\int_0^{+\infty}u^{(m-1)/2}\exp(-u)\,\mathrm{d}u,$$

la quantité sous le signe $\int$ n'est rien d'autre que $\Gamma[(m+1)/2]$, en définitive, on trouve :

$$p_m(t) = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)}\left[1+\frac{t^2}{m}\right]^{-(m+1)/2} \qquad \text{pour } t \in (-\infty, +\infty).$$

## 5.1. Vérification de la normalité de la loi

Il s'agit de calculer $\Lambda_m(\infty)$, on a donc :

$$\Lambda_m(\infty) = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)}\int_{-\infty}^{+\infty}\left[1+\frac{t^2}{m}\right]^{-(m+1)/2}\mathrm{d}t.$$

Posons $y = \sqrt{m}t$, nous obtenons :

$$\Lambda_m(\infty) = 2\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \int_0^{+\infty} \left[1+y^2\right]^{-(m+1)/2} \, \mathrm{d}y$$

$$= 2\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \int_0^{+\infty} y^{-1} y \left[1+y^2\right]^{-(m+1)/2} \, \mathrm{d}y$$

maintenant, il suffit de poser $u = y^2$, ainsi $\mathrm{d}u = 2y \, \mathrm{d}y$ et l'intégrale se transforme :

$$= \frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \int_0^{+\infty} \frac{1}{\sqrt{u}[1+u]^{(m+1)/2}} \, \mathrm{d}u = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} B\left(\frac{1}{2}, \frac{m+1}{2} - \frac{1}{2}\right),$$

expression dans laquelle $B(x,y)$ est la fonction eulérienne de première espèce qui s'écrit :

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^{+\infty} \frac{u^{x-1}}{[1+u]^{x+y}} \, \mathrm{d}u = 2\int_0^{+\infty} \frac{v^{2x-1}}{[1+v^2]^{x+y}} \, \mathrm{d}v$$

donc :

$$\Lambda_m(\infty) = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \cdot \frac{\Gamma(1/2)\Gamma(m/2)}{\Gamma[(m+1)/2]} = 1.$$

### 5.2. Calcul du moment du premier ordre

Le résultat est nul car la fonction sous le signe somme est impaire.

### 5.3. Calcul du moment du deuxième ordre

Il nous faut calculer :

$$D[t] = 2\frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)} \int_0^{+\infty} \frac{t^2}{\left[1+\frac{t^2}{m}\right]^{(m+1)/2}} \, \mathrm{d}t.$$

Les mêmes changements de variable utilisés pour vérifier la normalité de la loi conduisent aux résultats suivants :

$$D[t] = 2m\sqrt{m}\frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)} \int_0^{+\infty} \frac{y^2}{[1+y^2]^{(m+1)/2}} \, \mathrm{d}y$$

$$D(x) = m\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} B\left(\frac{3}{2}, \frac{m-2}{2}\right) = m\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \cdot \frac{\Gamma(3/2)\Gamma(m/2-1)}{\Gamma[(m+1)/2]}$$

$$= \frac{m}{2}\frac{1}{\frac{m}{2}-1} = \frac{m}{m-2}.$$

On en conclut que :

$$\sigma = \sqrt{\frac{m}{m-2}}.$$

On donne sur le Web (*) le sous-programme `student.h` qui calcule les valeurs de la distribution de Student quel que soit le degré de liberté $m$.

## 6. La distribution d'une somme de deux variables aléatoires indépendantes obéissant à une distribution de Student est-elle encore une distribution de Student ?

La réponse est non. Soit $\tau = \xi_m + \eta_n$ une variable aléatoire somme de deux variables aléatoires indépendantes obéissant chacune à une distribution de Student respectivement à $m$ et $n$ degrés de liberté. Soit $\Phi(t)$ la fonction caractéristique de $\xi_m$. Nous avons (en prenant la transformée de Fourier réelle car la distribution de Student est symétrique) :

$$\Phi(t) = 2\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \int_0^{+\infty} \frac{\cos(ty)}{[1+y^2]^{(m+1)/2}}\, dy$$

$$\Phi(t) = 2\frac{\Gamma[(m+1)/2]}{\sqrt{\pi}\Gamma(m/2)} \cdot \frac{1}{2} \cdot \frac{x\sqrt{\pi}}{\Gamma\left(\frac{m+1}{2}\right)} K_{m/2}(t),$$

expression dans laquelle $K_{m/2}(x)$ est la fonction modifiée de Bessel de troisième espèce. La fonction caractéristique de $\eta_n$ s'écrit :

$$\Psi(t) = 2\frac{\Gamma[(n+1)/2]}{\sqrt{\pi}\Gamma(n/2)} \cdot \frac{1}{2} \cdot \frac{x\sqrt{\pi}}{\Gamma\left(\frac{n+1}{2}\right)} K_{n/2}(t),$$

le produit des fonctions caractéristiques donne une fonction que l'on peut écrire d'une façon simplifiée : $A(m,n)K_{m/2}(x)K_{n/2}(x)$. Malheureusement, le produit de deux fonctions modifiées de Bessel n'est ni une fonction modifiée de Bessel ni une somme de deux fonctions modifiées de Bessel, il s'ensuit que la transformée de Fourier inverse n'est pas une distribution de Student.

## 7. La loi de Student à m degrés de liberté tend asymptotiquement vers la loi de Gauss quand m tend vers l'infini

Comme la loi est symétrique, il suffira d'effectuer un développement de MacLaurin (au voisinage de zéro). La densité de probabilité s'écrit :

$$\vartheta_m(x) = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)} \cdot \frac{1}{\left[1+\frac{x^2}{m}\right]^{(m+1)/2}}.$$

En passant aux logarithmes, on trouve :

$$\log_e(\vartheta_m(x)) = \text{Cste} + \frac{m-1}{2}\log_e\left(\frac{m-1}{2}\right) - \frac{m-1}{2} + \frac{1}{2}\log_e\left(\frac{m-1}{2}\right) - \frac{1}{2}\log_e(m)$$
$$- \frac{m-2}{2}\log_e\left(\frac{m-2}{2}\right) + \frac{m-2}{2} - \frac{1}{2}\log_e\left(\frac{m-2}{2}\right) - \frac{m+1}{2}\log_e\left(1+\frac{x^2}{m}\right)$$

* `http://www.edpsciences.com/guilpin/`

soit encore :

$$\begin{aligned}\log_e(\vartheta_m(x)) &= \text{Cste} + \frac{m}{2}\log_e\left(\frac{m-1}{m-2}\right) - \frac{1}{2}\log_e\left(\frac{m-1}{2}\right) + \log_e\left(\frac{m-2}{2}\right) \\ &\quad + \frac{1}{2}\log_e\left(\frac{m-1}{m-2}\right) - \frac{1}{2}\log_e(m) - \frac{m+1}{2}\log_e\left(1+\frac{x^2}{m}\right) \\ &= \text{Cste} + \frac{m+1}{2}\log_e\left(\frac{m-1}{m-2}\right) - \frac{1}{2}\log_e\left[\frac{2m(m-1)}{(m-2)^2}\right] - \frac{m+1}{2}\log_e\left(1+\frac{x^2}{m}\right).\end{aligned}$$

Quand $m$ est grand devant l'unité, on peut écrire :

$$\log_e(\vartheta_m(x)) = \log_e(K) - \frac{x^2}{2},$$

expression dans laquelle $K$ est une constante numérique que l'on détermine par normalisation de la distribution. On en déduit :

$$\vartheta_m(x) \quad \Longrightarrow \quad K\exp\left(-\frac{x^2}{2}\right).$$

On reconnaît la loi normale et par conséquent la constante $K$ vaut $\dfrac{1}{\sqrt{2\pi}}$. Quand le nombre de degrés de liberté croît, la loi de Student tend vers la loi normale. On pourra aussi s'en convaincre en consultant directement les tables numériques.

## 8. Éléments de bibliographie

S. AÏVAZIAN (1970) *Étude statistique des dépendances*, Éditions MIR, Moscou.
E. LUKACS (1964) *Fonctions caractéristiques*, Dunod.
V. ROTHSCHILD et N. LOGOTHETIS (1986) *Probability distributions*, Wiley.
Y. ROZANOV (1975) *Processus aléatoires*, Éditions MIR, Moscou.
H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 23 Systèmes à plusieurs variables aléatoires

## 1. Généralités

Il s'agit de généraliser un certain nombre de notions afin de les étendre aux systèmes de plusieurs variables aléatoires et plus spécialement aux systèmes de deux variables aléatoires qui vont retenir notre attention lors de l'étude de l'analyse de corrélation-régression.

L'apport essentiel de ce chapitre sera l'étude de la matrice de covariance et de la matrice de corrélation.

## 2. Système de variables aléatoires, fonction de répartition

On considère deux variables aléatoires $X$ et $Y$ qui pourront éventuellement figurer l'abscisse et l'ordonnée d'un point du plan. Il est alors tentant de penser que les propriétés du système dépendent uniquement de celles de chacune des variables ; il n'en est rien et il convient de tenir compte de leur **interdépendance**. Bien entendu, ces propos se généralisent au cas de $n$ variables aléatoires $X_1, X_2, \ldots, X_n$ dont le point figuratif appartiendra à un espace à $n$ dimensions.

### 2.1. Définition

On appelle fonction de répartition d'un système de deux variables aléatoires $X$ et $Y$, la probabilité de vérifier simultanément les deux inégalités :

$$X < \xi \quad \text{et} \quad Y < \eta.$$

Elle est définie par l'expression :

$$F(\xi, \eta) = P(X < \xi, Y < \eta)$$

(*cf.* Fig. 23.1, page suivante) et cela revient à dire que $F(\xi, \eta)$ est la probabilité de se trouver dans le quadrant hachuré de la figure ci-contre. Désignons par $\phi(\xi)$ la fonction de répartition de la variable aléatoire $X$ toute seule, et par $\psi(\eta)$ la fonction de répartition de la variable aléatoire $Y$ toute seule.

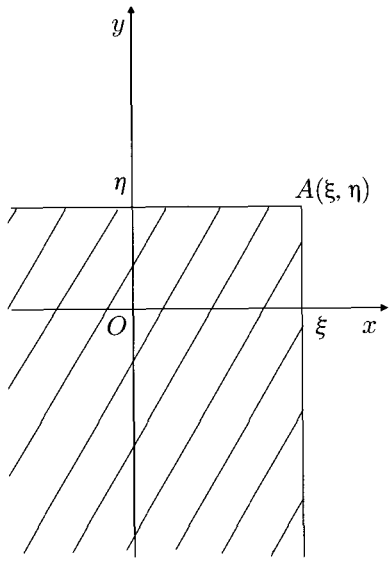


**Figure 23.1.** *Fonction de répartition à deux dimensions.*

## 2.2. Propriété de $F(\xi, \eta)$

1. C'est une fonction non décroissante des deux arguments :

$$\begin{aligned} \text{si } \xi \geq \alpha &\implies F(\xi,\eta) \geq F(\alpha,\eta) \\ \text{si } \eta \geq \beta &\implies F(\xi,\eta) \geq F(\eta,\beta). \end{aligned}$$

2. $F(\xi, -\infty) = F(-\infty, \eta) = F(-\infty, -\infty) = 0.$
3. $F(\xi, +\infty) = \phi(\xi)$
4. $F(+\infty, \eta) = \psi(\eta)$
5. $F(+\infty, +\infty) = 1.$

## 2.3. Probabilité pour qu'un point aléatoire $(\alpha, \beta)$ appartienne à un domaine rectangle parallèle aux axes

Soient $A, B, C$ et $D$ les quatre sommets du rectangle dont les côtés sont parallèles aux axes.

$$A = (a, d) \qquad B = (b, d) \qquad C = (b, c) \qquad D = (a, c).$$

Il s'agit en fait de trouver $a < \alpha < b$ et $c < \beta < d$ qui représente la probabilité de tomber dans le rectangle :

$$P(\alpha, \beta) = F(b, d) - F(a, d) - F(b, c) + F(a, c).$$

Cette probabilité peut s'exprimer en fonction de $F(x, y)$ qui à son tour s'écrit comme une intégrale double :

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\alpha, \beta)\, \mathrm{d}\alpha\, \mathrm{d}\beta.$$

## 3. Variables aléatoires liées et indépendantes

Revenons au cas de deux variables aléatoires $X$ et $Y$. Le problème fondamental est de savoir quelle est la dépendance mutuelle des deux variables (ou l'absence de dépendance).

On dit que la variable aléatoire $Y$ est indépendante de la variable aléatoire $X$ si la loi de répartition de $Y$ ne dépend pas de la valeur prise par $X$. Soit :

$$f(y|x) = \psi(y) \quad \text{quel que soit } x.$$

En revanche, si $Y$ dépend de $X$, on a :

$$f(y|x) \neq \psi(y) \quad \text{pour tout } x.$$

La dépendance de $X$ et de $Y$ est une propriété mutuelle des deux variables. Autrement dit, si :

$$f(y|x) = \psi(y) \quad \text{cela entraîne } f(x|y) = \phi(x).$$

Dans ce dernier cas, on dit que les variables aléatoires sont **indépendantes**, car la loi de répartition de chacune des variables ne dépend pas des valeurs prises par l'autre. Dans le cas contraire on dit qu'elles sont **dépendantes** ou **liées**.

Quand $X$ et $Y$ sont des variables indépendantes, la fonction de répartition $f(x,y)$ est le produit des fonctions de répartition de chacune des variables $X$ et $Y$ ; soit encore :

$$f(x,y) = \phi(x)\psi(y).$$

Réciproquement, si $f(x,y)$ est le produit de deux fonctions indépendantes $\phi(x)$ et $\psi(y)$, alors les variables aléatoires $X$ et $Y$ sont indépendantes et cela constitue un critère pour juger de l'indépendance des variables.

On peut dire que deux variables sont dépendantes quand elles ont une tendance plus ou moins marquée à varier simultanément.

## 4. Caractéristiques numériques, covariance, coefficient de corrélation

Par définition le moment non centré d'ordre $(k,s)$ du système de deux variables $X$ et $Y$ est l'espérance mathématique du produit $x^k y^s$, et l'on écrit :

$$m_{k,s} = M[x^k y^s].$$

Le moment centré s'écrit :

$$\mu_{k,s} = M[x_0^k y_0^s],$$

en désignant par $x_0$ et $y_0$ les variables centrées :

$$x_0 = x - m_{1,0}$$
$$y_0 = y - m_{0,1}.$$

On peut encore écrire :

$$m_{k,s} = \sum_i \sum_j x_i^k y_j^s p_{ij}$$

$$\text{et} \quad \mu_{k,s} = \sum_i \sum_j (x_i - m_{1,0})^k (y_j - m_{0,1})^s p_{ij}$$

expressions dans lesquelles $p_{ij} = P(x = x_i, y = y_j)$. S'il s'agit de variables aléatoires continues, nous écrirons :

$$m_{k,s} = \int_{-\infty}^{+\infty}\int x^k y^s f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

$$\text{et} \quad \mu_{k,s} = \int_{-\infty}^{+\infty}\int (x-m_x)^k (y-m_y)^s f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

D'un point de vue strictement pratique, seuls les moments d'ordre 1 et 2 sont utilisés ; soit :

$$m_x = m_{1,0} = M[x^1 y^0] = M[x]$$
$$m_y = m_{0,1} = M[x^0 y^1] = M[y]$$

et

$$D_x = m_{2,0} = M[x_0^2 y_0^0] = M[x_0^2] = D[x] = \sigma_x^2$$
$$D_y = m_{0,2} = M[x_0^0 y_0^2] = M[y_0^2] = D[y] = \sigma_y^2.$$

Pour ce qui concerne les moments mixtes, seul le premier joue vraiment un rôle important :

$$m_{1,1} = M[x^1 y^1]$$

dont la forme centrée s'écrit :

$$\mu_{1,1} = M[x_0^1 y_0^1].$$

ceci est simplement la moyenne du produit des variables aléatoires centrées. Ce moment mixte porte un nom particulier, il s'agit de la **covariance** que l'on écrit :

$$K_{x,y} = \sum_i \sum_j (x_i - m_x)(y_j - m_y) p_{ij}$$

$$\text{ou} \quad K_{x,y} = \int_{-\infty}^{+\infty}\int (x-m_x)(y-m_y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

$K_{x,y}$ décrit la liaison de $X$ et de $Y$. Si $X$ et $Y$ sont des variables aléatoires indépendantes, alors $K_{x,y} = 0$. Autrement dit, puisque $f(x,y) = \phi(x)\psi(y)$, on a alors :

$$K_{x,y} = \int_{-\infty}^{+\infty} (x-m_x)\phi(x)\,\mathrm{d}x \int_{-\infty}^{+\infty} (y-m_y)\psi(y)\,\mathrm{d}y = 0.$$

## 4.1. Remarques importantes

***a*** – Non seulement $K_{x,y}$ caractérise la liaison de $X$ et de $Y$ mais aussi la dispersion de chacune des variables. Ainsi, si une des variables s'écarte peu de sa moyenne (faible variance), il en résulte que $K_{x,y}$ reste faible même si $X$ et $Y$ sont étroitement liées. Pour pallier cet ennui, c'est-à-dire pour caractériser uniquement la liaison des variables, on définit le coefficient de corrélation des

deux variables $X$ et $Y$ en divisant $K_{x,y}$ par l'écart type de chacune des variables, soit $\sigma_x$ et $\sigma_y$. On définit alors :

$$r_{x,y} = \frac{K_{x,y}}{\sigma_x \sigma_y},$$

on notera que ce rapport est adimensionnel. Ici encore, si $X$ et $Y$ sont indépendantes, alors $r_{x,y} = 0$.

Attention, la proposition réciproque est fausse. Si deux variables aléatoires $X$ et $Y$ ont un coefficient de corrélation nul ($r_{x,y} = 0$), cela ne signifie pas que $X$ et $Y$ sont des variables indépendantes.

Le fait que $r_{x,y}$ soit nul signifie non pas que les variables soient indépendantes mais qu'elles ne sont pas **linéairement dépendantes.**

Il est possible de trouver un exemple simple qui montre que des variables $X$ et $Y$ non corrélées sont cependant dépendantes. Considérons deux variables aléatoires $X$ et $Y$ réparties uniformément à l'intérieur d'un cercle de rayon $R$ centré à l'origine des coordonnées. La densité de probabilité s'écrit :

$$\begin{aligned} & f(x,y) = p \quad \text{pour } x^2 + y^2 \leq R^2 \\ \text{et} \quad & f(x,y) = 0 \quad \text{ailleurs.} \end{aligned}$$

La normalisation de la densité de probabilité donne la valeur de $p$ :

$$1 = \iint_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_{-\infty}^{+\infty} p\,\mathrm{d}x\,\mathrm{d}y \quad \Longrightarrow \quad p = \frac{1}{\pi R^2}\,.$$

Les variables sont dépendantes puisque, pour $y$ fixé dans le cercle, $x$ ne peut pas prendre n'importe quelles valeurs, mais seulement celles comprises dans l'intervalle $(-\sqrt{R^2 - y^2}, \sqrt{R^2 - y^2})$ ou encore si $x = R$, $y = 0$. Exprimons la covariance : elle est nulle puisque la moyenne de $x$ et la moyenne de $y$ sont nulles.

$$K_{x,y} = 0.$$

***b*** – Le **coefficient de corrélation caractérise** non pas une relation quelconque entre deux variables, mais **une relation linéaire.** Quand une des variables croît, l'autre suit une loi linéaire au sens des probabilités et le coefficient de corrélation indique le « degré » de cette **dépendance linéaire.** Dans le cas d'une relation fonctionnelle linéaire exacte, on a :

$$r_{x,y} = \pm 1,$$

autrement dit lorsque $Y = aX + b$. Si la loi de dépendance linéaire est quelconque, alors nous avons :

$$-1 \leq r_{x,y} \leq +1,$$

la corrélation étant positive ou négative. Nous allons démontrer ces résultats au paragraphe 6.

## 5. Généralisation au cas de plusieurs variables

Il n'y a pas de difficulté à étendre ces résultats au cas de plusieurs variables aléatoires $X_1, X_2, \ldots, X_n$ ; ainsi, la covariance et la corrélation se calculent pour toutes les variables prises deux à deux. Si les variables sont au nombre de $n$, alors il y a $n^2$ coefficients de covariance et $n^2$ coefficients de corrélation. Dans chacun de ces deux cas, les coefficients sont naturellement les éléments d'une matrice carrée d'ordre $n$.

### 5.1. Matrice de covariance

Notons ses éléments $k_{ij}$. Elle est évidemment symétrique puisque le produit est une opération commutative : $k_{ji} = k_{ij}$.

D'autre part, $k_{ii} = D_i$ qui est la variance de $X_i$.

### 5.2. Matrice de corrélation

Notons ses éléments $r_{ij}$. Elle aussi est évidemment symétrique et :

$$r_{ji} = r_{ij} = \frac{k_{ij}}{\sigma_{ij}} \quad \text{avec} \quad \sigma_{ij} = \sqrt{D_i D_j} \quad \text{et} \quad r_{ii} = 1.$$

## 6. Quelques théorèmes importants

Soient $X$ et $Y$ deux variables aléatoires quelconques ; elles prennent respectivement les valeurs $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n$. Soit $g(x, y)$ une fonction de deux variables offrant les bonnes qualités usuelles de régularité, et $p_{ij}$ la probabilité pour que $X = x_i$ et $Y = y_j$. La moyenne de la fonction $g(x_i, y_j)$ est alors donnée par l'expression :

$$M[g(X,Y)] = \sum_{i=1}^{n} \sum_{j=1}^{n} g(x_i, y_j) p_{ij}.$$

et la forme continue :

$$M[g(X,Y)] = \int_{-\infty}^{+\infty} \!\!\int g(x,y) f(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

expression dans laquelle $f(x, y)$ est la densité de probabilité des deux variables aléatoires $X$ et $Y$.

Nous allons utiliser ces expressions en écrivant :

$$g(X,Y) = X + Y \quad \text{et} \quad g(X,Y) = X \cdot Y.$$

### 6.1. Moyenne d'une somme de variables aléatoires

Il nous faut calculer M[X + Y] dont l'expression s'écrit :

$$\begin{aligned} M[X+Y] &= \sum_{i=1}^{n} \left( \sum_{j=1}^{n} [x_i + y_j] p_{ij} \right) = \sum_{i=1}^{n} \left( \sum_{j=1}^{n} x_i p_{ij} \right) + \left( \sum_{i=1}^{n} \sum_{j=1}^{n} y_j p_{ij} \right) \\ &= \sum_{i=1}^{n} \left( x_i \sum_{j=1}^{n} p_{ij} \right) + \sum_{j=1}^{n} \left( y_j \sum_{i=1}^{n} p_{ij} \right). \end{aligned}$$

Or $\sum_{j=1}^{n} p_{ij}$ n'est rien d'autre que la probabilité $p_i$ que la variable $X$ prenne la valeur $x_i$ et $\sum_{i=1}^{n} p_{ij}$ la probabilité $p_j$ que la variable $Y$ prenne la valeur $y_j$. D'où les expressions qui suivent :

$$M[X+Y] = \sum_{i=1}^{n} (x_i p_i) + \sum_{j=1}^{n} (y_j p_j)$$

$$M[X+Y] = M[X] + M[Y].$$

On démontrera évidemment le même théorème en ce qui concerne les variables continues. Insistons sur le fait que cette relation est indépendante du fait que les variables aléatoires soient indépendantes ou non. Ce théorème est susceptible d'une généralisation simple :

$$M\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} M[X_i].$$

## 6.2. Variance d'une somme de variables aléatoires

Nous allons calculer $D[X+Y]$. Après avoir posé $Z = X + Y$, on passe aux variables centrées que l'on indice avec un zéro :

$$Z_0 = X_0 + Y_0.$$

On obtient alors les expressions suivantes :

$$D[X+Y] = D[Z] = M\left[Z_0^2\right] = M\left[X_0^2\right] + M\left[Y_0^2\right] + 2M\left[X_0 Y_0\right] = D[X] + D[Y] + 2K_{x,y}.$$

On note alors que la variance de la somme de deux variables aléatoires est la somme et de leur variance respective augmentée de deux fois la covariance.

Ici encore il est intéressant de généraliser à une somme quelconque :

$$D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} D[X_i] + 2\sum_{i>j} K_{ij},$$

$$\text{soit encore} \quad D\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij},$$

et l'on remarquera que la variance de la somme de variables est égale à la somme des variances de chacune des variables **si celles-ci sont toutes indépendantes**.

## 6.3. Moyenne d'un produit de variables aléatoires

Il nous faut calculer $M[X \cdot Y]$ dont nous obtenons l'expression à partir de la covariance :

$$K_{x,y} = M\left[X_0 \cdot Y_0\right] = M\left[(X - m_x) \cdot (Y - m_y)\right]$$

avec

$$m_x = M[X] \quad \text{et} \quad m_y = M[Y].$$

Le développement de l'équation donne :

$$\begin{aligned} K_{x,y} &= M[X \cdot Y] - m_x M[Y] - m_y M[X] + m_x m_y \\ K_{x,y} &= M[X \cdot Y] - m_x m_y = M[X \cdot Y] - M[X]M[Y] \\ M[X \cdot Y] &= M[X]M[Y] + K_{x,y}. \end{aligned}$$

Si les variables sont indépendantes, $K_{x,y} = 0$, et la moyenne du produit est égale au produit des moyennes. On généralise ce théorème au cas d'un produit de $n$ **variables aléatoires indépendantes** :

$$M\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} M[X_i].$$

### 6.4. Variance d'un produit de variables aléatoires

Il nous faut calculer $D[X \cdot Y]$, pour cela posons $Z = X \cdot Y$ :

$$D[X \cdot Y] = D[Z] = M\left[Z_0^2\right] = M\left[(Z - m_z)^2\right]$$

avec

$$\begin{aligned} m_z &= M[X \cdot Y] = m_x m_y + K_{x,y} \\ D[X \cdot Y] &= M\left[Z^2 + m_z^2 - 2m_z Z\right] = M\left[Z^2\right] + m_z^2 - 2m_z M[Z] = M\left[Z^2\right] - m_3^2 \\ &= M\left[X^2Y^2\right] - 2(m_x m_y + K_{x,y})M[X \cdot Y] + (m_x m_y + K_{x,y})^2 \\ &= M\left[X^2Y^2\right] - 2(m_x m_y + K_{x,y})(m_x m_y + K_{x,y}) + (m_x m_y + K_{x,y})^2 \\ D[X \cdot Y] &= M\left[X^2Y^2\right] - (m_x m_y + K_{x,y})^2 . \end{aligned}$$

Dans cette dernière expression, on voit apparaître la covariance. Dans le **cas particulier où les variables sont indépendantes**, la dernière expression se simplifie. Puisque $X$ et $Y$ sont des variables indépendantes, il en est de même de $X^2$ et $Y^2$ ; alors, en tenant compte du fait que $K_{x,y} = 0$, on peut écrire :

$$\begin{aligned} & D[X \cdot Y] = M[X^2]M[Y^2] - m_x^2 m_y^2 \\ \text{soit :} \quad & D[X \cdot Y] = D[X] \cdot D[Y] + m_x^2 D[Y] + m_y^2 D[X]. \end{aligned}$$

Si les variables sont centrées, on obtient :

$$D[X_0 \cdot Y_0] = D[X_0] \cdot D[Y_0].$$

## 7. Propriétés du coefficient de corrélation (démonstrations)

Si les variables aléatoires $X$ et $Y$ sont liées linéairement :

$$y = ax + b,$$

alors le coefficient de corrélation vaut $\pm 1$. En effet, à partir de la covariance on peut écrire :

$$K_{x,y} = M[X_0.Y_0] = M\left[(X - m_x) \cdot (Y - m_y)\right].$$

En remarquant que $m_y = am_x + b$

$$\begin{aligned} K_{x,y} &= M\left[(X - m_x) \cdot (aX + b - am_x - b)\right] = aM[X^2] - am_x M[X] - am_x M[X] + am_x^2 \\ &= aM[X^2] - am_x^2 = aD[X]. \end{aligned}$$

Reportons cette valeur dans l'expression du coefficient de corrélation :

$$r_{x,y} = \frac{K_{x,y}}{\sigma_x \sigma_y} = \frac{aD[X]}{\sigma_x \sigma_y} .$$

Il reste à exprimer $\sigma_y$ :

$$\sigma_y = \sqrt{D[aX + b]} = \sqrt{a^2 D[X]} = |a|\sigma_x.$$

On en conclut que $r_{x,y} = +1$ si $a$ est positif et $-1$ si $a$ est négatif.

Maintenant, il s'agit de montrer que $|r_{x,y}| \leq 1$ quelle que soit la liaison entre les variables aléatoires $X$ et $Y$. À cette fin, calculons la variance de la variable $Z = aX + bY$, elle s'écrit :

$$D[Z] = D[aX + bY] = a^2 D[X] + b^2 D[Y] + 2abK_{x,y}.$$

Faisons à présent $b = \sigma_x$ puis $a = \sigma_y$ et $a = -\sigma_y$, nous obtenons :

$$D[Z] = 2\sigma_x^2\sigma_y^2 \pm 2\sigma_x\sigma_y K_{x,y} = 2\sigma_x^2\sigma_y^2[1 \pm r_{x,y}].$$

Compte tenu du fait que $D[Z] \geq 0$, nous devons avoir :

$$\sigma_x^2\sigma_y^2 \pm \sigma_x\sigma_y K_{x,y} \geq 0$$

ce qui s'écrit encore :

$$|K_{x,y}| \leq \sigma_x\sigma_y$$

et par conséquent :

$$|r_{x,y}| \leq 1.$$

## 8. Éléments de bibliographie

E. BOREL, R. DELTHEIL et R. HURON (1962) *Probabilités, Erreurs*, Armand Colin, Paris, 12e édition.

A.L. EDWARDS (1984) *An introduction to linear regression and correlation*, W.H. Freeman.

M. FISZ (1980) *Probability theory and mathematical statistics*, Krieger.

H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 24 | Critères de conformité

## 1. Généralités

Ce chapitre a pour objet l'étude de critères permettant de tester la validité des hypothèses formulées lors de l'étude d'une **population parente**. Il nous faut donc définir ce qu'est une population parente. L'étude du caractère stochastique d'un système réel (physique) conduit à l'étude de la loi de distribution de la dispersion aléatoire touchant aux mesures effectuées sur ledit système. La loi est propre au système qui est soumis à un ensemble de conditions réelles d'observations. Par exemple, on peut évoquer la dispersion du tir d'un certain canon alimenté par des munitions d'un certain lot ; bien d'autres conditions peuvent influencer la dispersion : l'expérience du tireur, la météorologie...

Une certaine mesure effectuée sur le système est alors caractérisée par la variable aléatoire $\xi$, accompagnée de sa distribution $p(\xi)$, c'est-à-dire de la loi permettant d'obtenir la probabilité objective de trouver l'événement $x$ dans l'intervalle $\Delta x$.

On appelle population parente l'ensemble de toutes les observations possibles qui pourraient être réalisées pour un système soumis à un ensemble de conditions. Il faut bien noter que l'ensemble de toutes les observations (réalisations) possibles est différent de l'ensemble des valeurs possibles de la variable aléatoire $\xi$. À chaque valeur fixée de $\xi$ correspond plusieurs voire une infinité d'observations possibles ; par exemple, on jette trois dés de couleur différente mais par ailleurs parfaitement identiques, et l'on s'intéresse à la somme présentée par les faces supérieures : on compte zéro si la somme est inférieure à 12 et 1 si la somme est supérieure ou égale à 12. La variable aléatoire $\xi$ attachée au lancement prend les valeurs zéro et un tandis qu'il y a 216 configurations ou observations possibles, 135 configurations donnent une somme inférieures à 12 et 81 configurations une somme supérieure ou égale à 12.

Donc une population parente est définie par des objets sur lesquels on pratique la mesure d'une certaine grandeur. D'une part les objets sont très bien définis (ensemble de conditions) et d'autre part la mesure effectuée est sujette à des fluctuations incontrôlables ; en conséquence, à celle-là il est associé une variable aléatoire. Par exemple, on peut s'intéresser à la longueur des feuilles des arbres d'un parc. Les conditions sont donc constituées par **toutes les feuilles de tous les arbres** du parc. Toutes les feuilles ainsi définies constituent une population parente. Les valeurs possibles de la variable aléatoire $\lambda$ seront, d'un point de vue pratique, en nombre fini ; elles peuvent être mesurées au millimètre près. Une autre population parente pourra être toutes les feuilles de tous les chênes du parc ou encore toutes les feuilles de tous les peupliers...

Généralement les populations parentes étudiées sont finies, mais on est tenté d'imaginer des populations parentes infinies qui sont des extrapolations pour un nombre infini d'observations possibles. Cette notion peut alors se révéler commode dans l'élaboration d'une théorie mathématique.

Un **échantillon** est un prélèvement fini à l'intérieur d'une population parente, et le nombre $n$ d'observations constituant l'échantillon est appelé la **taille** de l'échantillon. Le prélèvement de l'échantillon doit être fait **au hasard** et chaque élément de l'échantillon doit être indépendant des autres, c'est ce qu'on appelle **l'indépendance stochastique**. Au cours de ce chapitre, on verra comment contrôler la qualité de l'échantillon prélevé.

## 2. Représentation des données numériques. Histogramme

On se trouve en présence d'un échantillon de taille $N$ prélevé au hasard dans une certaine population parente, l'échantillon étant caractérisé par un ensemble de mesures que l'on note $\{x_i\}\, i = 1, 2, \ldots, n$. Il est fondamental de connaître la loi de distribution à laquelle obéit l'échantillon. Pour ce faire on va procéder à une représentation graphique des données groupées que l'on obtient en comptabilisant le nombre de données appartenant à un petit intervalle. De façon pragmatique, on opère de la manière suivante :

1. on recherche la borne inférieure $x_{inf}$ et la borne supérieure $x_{sup}$ de l'ensemble des $\{x_i\}$, et cela nous définit l'intervalle de représentation noté $I$ ;
2. on divise l'intervalle $I$ en $k$ sous-intervalles égaux de telle sorte que $k$ soit compris entre 8 et 25. Pour cela, on peut utiliser la règle approchée :

$$k = \log_2 n + 1,$$

3. on compte dans chaque sous-intervalle le nombre $m_k$ d'individus qui y tombent.

La courbe représentative des $m_k$ en fonction de $k$ s'appelle un histogramme, elle donne une représentation de la densité de probabilité à laquelle l'échantillon obéit, ou peut obéir.

À partir de cette représentation, nous allons chercher à déterminer les propriétés réellement stables qui caractérisent le phénomène étudié. Ceci signifie corrélativement que l'on cherchera aussi à éliminer tout le côté aléatoire apparaissant dans la série finie des mesures.

L'aboutissement de la recherche des éléments stables sera donc une **loi** que l'on dira **stochastique** par opposition à **loi causale**. Stochastique est un mot d'origine grecque — ὁ στοχαστής — qui signifie le devin, celui qui conjecture.

On retrouve ici une généralisation de la notion d'incertitude en physique, laquelle accompagne inévitablement la réalisation de la mesure. La dispersion de la mesure obéit à une loi stochastique — la plupart du temps gaussienne.

Une fois obtenue la représentation de la densité de probabilité, on recherche une expression analytique pouvant « raisonnablement » rendre compte de la loi expérimentale obtenue.

Sur le Web(*), on trouve le programme `histog.c` qui réalise l'histogramme d'une série de nombres aléatoires à distribution gaussienne.

* `http://www.edpsciences.com/guilpin/`

## 3. Conformité entre une répartition théorique et une répartition expérimentale (ou répartition statistique)

Après avoir soigneusement examiné la répartition expérimentale, on est en mesure de pressentir une loi analytique susceptible de représenter théoriquement l'ensemble des données de l'échantillon.

Inévitablement, il existe des écarts entre la loi expérimentale et la loi théorique, et le problème de fond consiste à pouvoir dire si ces fluctuations sont dues au hasard ou non, et surtout avec quelle probabilité. En effet, la représentation de la loi expérimentale est intimement liée à l'échantillon étudié. Un autre échantillon donnera une loi expérimentale plus ou moins différente, ainsi qu'un suivant. Donc, le problème concernant la conformité de la loi expérimentale et de la loi théorique repose uniquement sur la taille restreinte de l'échantillon (nombre réduit d'observations). Ce problème devient caduque si l'on étudie toute la population parente, alors, il n'y a plus qu'une seule loi expérimentale possible.

La recherche d'une loi théorique consiste donc à formuler une certaine hypothèse $H$ et l'on veut savoir si cette hypothèse est conforme aux données expérimentales. En fin de compte, on veut savoir si l'on peut ou non adopter cette hypothèse $H$ et avec quel degré de certitude.

Il s'agit donc de définir une **mesure de la conformité** que rien n'empêche dès maintenant de désigner par $U$. Force est de remarquer que $U$ est une variable aléatoire dont la loi de répartition sera *a priori* liée à la distribution de la variable aléatoire étudiée $x$ et de la taille $n$ de l'échantillon.

Si l'hypothèse est correcte, la loi de répartition de $U$ est déterminée par la loi de répartition de $x$ et par le nombre $n$. Désignons par $v$ la valeur particulière prise par la variable aléatoire $U$ dans une expérience bien déterminée. Alors notre problème de conformité consiste à savoir si la valeur calculée $v$ s'explique par le hasard ou encore si la valeur $v$ est trop grande et indique alors une différence essentielle entre les deux répartitions théorique et statistique. Ce dernier cas de figure montrerait alors que l'hypothèse $H$ devrait être rejetée, mais cela n'est pas aussi simple que les apparences le montrent.

Après avoir déterminé la valeur numérique $v$, on calcule la probabilité pour que la variable aléatoire $U$ puisse effectivement dépasser $v$. Soit $P$ cette probabilité. Si $P$ a une valeur « trop petite », on doit alors rejeter l'hypothèse $H$, elle apparaît alors comme peu vraisemblable.

### Choix de la mesure de U

La première difficulté rencontrée pour choisir une mesure de $U$ repose sur le fait que la répartition de $U$ dépend de la répartition de $x$. Il se trouve fort heureusement que, dans certains cas, la loi de répartition de $U$ a des propriétés simples, et, pour $n$ grand devant l'unité, cette loi ne dépend plus de la loi de répartition des $x$. Nous allons voir deux mesures de $U$.

## 4. Le $\chi^2_n$ de Pearson (1857–1936)

On désigne par $p_i$ la probabilité théorique de tomber dans l'intervalle $i$, et par $\bar{p}_i$ la probabilité expérimentale de tomber dans le même intervalle.

Pour la mesure de $U$, on choisit la somme des carrés des différences $(p_i - \bar{p}_i)^2$ que l'on pondère par le coefficient $c_i$. $c_i$ tient compte des valeurs relatives des écarts quadratiques à $\Delta p_i$ égal, l'importance étant plus grande si $\Delta p_i$ est petit. C'est la raison pour laquelle on prend $c_i = \alpha/p_i$, c'est-à-dire proportionnel à $1/p_i$.

Si l'on choisit $\alpha = n$, il est possible de montrer que la fonction de répartition de $U$ ne dépend pratiquement pas de la fonction de répartition $F(x)$, ni de $n$, mais uniquement du nombre d'intervalles de regroupement $k$ qui a servi à construire l'histogramme. Dans ce cas, on dit que $U$ obéit à une loi du $\chi^2$, et l'on a :

$$u = n \sum_{i=1}^{k} \frac{(p_i - \bar{p}_i)^2}{p_i}$$

$u$ est la valeur numérique de la conformité. On trouvera plus loin une démonstration de ces affirmations (*cf.* § 14.1 Annexe H, p. 443).

***Remarque 1*** : La loi théorique dépend de $\mu$ paramètres inconnus que l'on va chercher à déterminer.

***Remarque 2*** : $n$ est le nombre d'observations indépendantes de la variable aléatoire $x$, et $k$ le nombre d'intervalles de regroupement. Il est nécessaire d'obtenir un nombre minimum de plusieurs unités — voire une dizaine — dans chacun des intervalles de regroupement. Si tel n'est pas le cas, il faut regrouper certains intervalles pour qu'il en soit ainsi.

***Remarque 3*** : La variable $U$ obéit à une loi du $\chi^2$ à $m$ degrés de liberté lequel est déterminé par le nombre d'intervalles de regroupement $k$ moins le nombre de contraintes auxquelles est soumise la loi. Ces contraintes sont constituées par le nombre de paramètres inconnus $\mu$ figurant dans la loi théorique auquel il convient d'ajouter la valeur 1 car il y a une **contrainte implicite** qui est la **normalisation de la loi de distribution**. D'où :

$$m = k - \mu - 1.$$

***Remarque 4*** : Il n'y a pas de difficulté à calculer la probabilité $\theta$ de dépasser la valeur numérique $u$, soit en ayant programmé la loi du $\chi^2$ à $m$ degrés de liberté, soit en consultant une table figurant à la fin de tous les manuels. Si cette probabilité $\theta$ est très petite, on rejette l'hypothèse $H$ **avec $\theta$ chances de la rejeter à tort**. Dans le cas contraire, l'hypothèse $H$ est vraisemblable c'est-à-dire qu'elle n'est pas en contradiction avec les données expérimentales.

***Remarque 5*** : Parler de probabilité très petite n'a pas grand sens si l'on ne se fixe un **seuil de signification** $\alpha$. D'un point de vue pratique, si $\theta$ est inférieur à 0,05, il y a lieu de vérifier si possible l'expérience, et si les écarts persistent, il y a lieu de chercher une autre loi. Cela signifie que l'on prend le risque de rejeter à tort l'hypothèse $H$ avec 5 chances sur cent...

***Remarque 6*** : Le test du $\chi^2$ est à rapprocher de la preuve par neuf. La preuve par neuf ne dit pas quand l'opération est juste, elle dit seulement quand elle est fausse. Autrement dit, si la preuve par neuf ne donne pas, lors du processus de calcul, les deux derniers nombres identiques, on est certain que l'opération est fausse. En revanche, si les deux dernières opérations sont identiques cela signifie que l'opération peut être correcte ou encore que, s'il y a une erreur, l'erreur est modulo 9. Il en est de même du $\chi^2$. On peut aisément trouver plusieurs lois tout à fait acceptables qui peuvent rendre compte correctement de la statistique d'un échantillon donné. Le critère du $\chi^2$ ne permet pas de dire qu'une loi est meilleure ou pire qu'une autre. Simplement, nous savons que les données expérimentales ne sont pas en contradiction avec les différentes hypothèses formulées.

## 4.1. Un exemple — Vérification d'un générateur de nombres aléatoires gaussiens

À partir d'un générateur de nombres aléatoires à distribution uniforme, on se propose de fabriquer un générateur de nombres aléatoires gaussiens. Notre hypothèse $H$ est donc la loi de Gauss-Laplace. Nous avons généré 4 096 nombres aléatoires gaussiens en choisissant la

valeur moyenne 6,0 et l'écart type 2,0 (nous avons ajouté chaque fois 12 nombres aléatoires à distribution rectangulaire donnés par la technique de Lehmer). Les résultats obtenus sont présentés dans le tableau 24.1.

**Tableau 24.1.**

| | | | | |
|---|---|---|---|---|
| $k = 0$ | $m = 4$ | $x = -3{,}096$ | $p = 8{,}037\,e - 4$ | $p^* = 9{,}766\,e - 4$ |
| $k = 1$ | $m = 18$ | $x = -2{,}620$ | $p = 3{,}417\,e - 3$ | $p^* = 4{,}395\,e - 3$ |
| $k = 2$ | $m = 48$ | $x = -2{,}144$ | $p = 1{,}169\,e - 2$ | $p^* = 1{,}172\,e - 2$ |
| $k = 3$ | $m = 130$ | $x = -1{,}668$ | $p = 3{,}167\,e - 2$ | $p^* = 3{,}174\,e - 2$ |
| $k = 4$ | $m = 294$ | $x = -1{,}191$ | $p = 6{,}904\,e - 2$ | $p^* = 7{,}178\,e - 2$ |
| $k = 5$ | $m = 481$ | $x = -0{,}7153$ | $p = 1{,}205\,e - 1$ | $p^* = 1{,}174\,e - 1$ |
| $k = 6$ | $m = 667$ | $x = -0{,}2391$ | $p = 1{,}683\,e - 1$ | $p^* = 1{,}628\,e - 1$ |
| $k = 7$ | $m = 765$ | $x = 0{,}2370$ | $p = 1{,}882\,e - 1$ | $p^* = 1{,}868\,e - 1$ |
| $k = 8$ | $m = 724$ | $x = 0{,}7132$ | $p = 1{,}684\,e - 1$ | $p^* = 1{,}768\,e - 1$ |
| $k = 9$ | $m = 498$ | $x = 1{,}189$ | $p = 1{,}207\,e - 1$ | $p^* = 1{,}216\,e - 1$ |
| $k = 10$ | $m = 267$ | $x = 1{,}665$ | $p = 6{,}925\,e - 2$ | $p^* = 6{,}519\,e - 2$ |
| $k = 11$ | $m = 141$ | $x = 2{,}142$ | $p = 3{,}180\,e - 2$ | $p^* = 3{,}442\,e - 2$ |
| $k = 12$ | $m = 41$ | $x = 2{,}618$ | $p = 1{,}169\,e - 2$ | $p^* = 1{,}001\,e - 2$ |
| $k = 13$ | $m = 13$ | $x = 3{,}094$ | $p = 3{,}438\,e - 3$ | $p^* = 3{,}174\,e - 3$ |
| $k = 14$ | $m = 4$ | $x = 3{,}570$ | $p = 8{,}092\,e - 4$ | $p^* = 9{,}766\,e - 4$ |

Dans ce tableau, $k$ est le numéro de l'intervalle de regroupement, $m$ le nombre d'occurrences, $x$ est l'abscisse du début de chaque intervalle, $p$ la probabilité théorique de tomber dans l'intervalle, et $p^*$ la probabilité empirique.

Expérimentalement, on trouve que les nombres pseudo-aléatoires gaussiens ont une moyenne de 6,05 et un écart type de 1,98. Comme le montre le tableau, nous avons effectué les regroupements sur 15 sous-intervalles. Nous avons alors obtenu la valeur de $u = 7{,}601$, comme l'hypothèse à tester est la loi normale, le nombre de degrés de liberté est 12. Ainsi, la probabilité de dépasser la valeur $u$ est 0,815. Il n'y a aucune raison de rejeter la loi de Gauss comme hypothèse $H$, les données expérimentales ne la contredisant pas. On dit aussi que l'hypothèse $H$ est tout à fait vraisemblable.

## 4.2. Remarques sur le seuil de signification

Le choix du seuil de signification $\alpha$ est guidé par les conséquences qui pourraient découler du rejet à tort ou de la conservation à tort de l'hypothèse $H$. En l'absence de toute considération sur les risques encourus, le seuil de signification le plus répandu est 0,05. Cela signifie que l'on choisit de rejeter la loi avec **5 chances sur cent de la rejeter à tort.**

D'un point de vue pratique, après avoir calculé $u$, on opérera selon l'un des cas de figure suivants :

**a.** la valeur de $u$ peut se situer dans deux domaines : le domaine des valeurs possibles $u < \chi^2_{1-\alpha}$ et le domaine des valeurs improbables $u \geq \chi^2_{1-\alpha}$, expressions dans lesquelles $\alpha$ est le seuil de signification. Alors dans cette dernière éventualité, on rejettera l'hypothèse $H$ avec $\alpha$ chances sur cent de la rejeter à tort ;

**b.** on peut avoir des raisons particulières de se méfier des valeurs de $u$ trop petites et le domaine des valeurs possibles devient $u > \chi^2_\alpha$ et le domaine des valeurs improbables $u \leq \chi^2_\alpha$. Dans la dernière éventualité, on rejettera l'hypothèse $H$ avec $\alpha$ chances sur cent de la rejeter à tort ;

**c.** on peut aussi avoir des raisons de se méfier à la fois des valeurs trop petites et trop grandes de $u$, autrement dit on se méfie des valeurs de $u$ qui se situent dans les queues de distribution. Le domaine des valeurs probables devient $\chi^2_{\alpha/2} \leq u \leq \chi^2_{1-\alpha/2}$ et les domaines des valeurs improbables $u < \chi^2_{\alpha/2}$ et $u > \chi^2_{1-\alpha/2}$. Dans la dernière éventualité, on rejettera l'hypothèse $H$ avec $\alpha$ chances sur cent de la rejeter à tort.

## 5. Critère de Kolmogorov (1903–1987)

Il s'agit d'un critère de mesure de la conformité (ou de la non-conformité) qui présente un grand intérêt dans la mesure où il est très simple d'emploi. Ici la mesure de la conformité $v$ est le sup du module de la différence entre les répartitions théorique $F(x)$ et expérimentale $F^*(x)$. Donc :

$$v = \sup |F(x) - F^*(x)| .$$

Bien entendu, $v$ est une variable aléatoire. Là encore, il est possible de montrer que la répartition de $v$ ne dépend pas de $F(x)$ à condition toutefois que la taille de l'échantillon $n$ soit très grand devant l'unité ($n$ est le nombre d'expériences indépendantes). La probabilité de l'inégalité $v\sqrt{n} \geq \lambda$ est donnée par l'expression :

$$p(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \exp\left(-2k^2\lambda^2\right) = 2\sum_{k=1}^{\infty}(-1)^{k+1} \exp\left(-2k^2\lambda^2\right).$$

Voici quelques valeurs de $p(\lambda)$ :

| $\lambda$ | 0,0 | 0,5 | 1,0 | 1,5 | 2,0 |
|---|---|---|---|---|---|
| $p(\lambda)$ | 1,0 | 0,964 | 0,27 | 0,022 | 0,001 |

Pour utiliser le critère de Kolmogorov, il suffit de :

**a.** calculer $v$ puis $v\sqrt{n}$,
**b.** chercher dans la table de $p(\lambda)$ la probabilité correspondante.

C'est la probabilité que l'écart maximal entre les deux répartitions soit non inférieur à la valeur observée, à condition toutefois que les écarts soient dus uniquement à des facteurs aléatoires.

***Remarque*** : On notera l'absence de degré de liberté dans l'expression de ce critère. Il s'ensuit qu'il est plus « optimiste » que le précédent critère.

Le programme `kolmogor.c` sur le Web(*) calcule les valeurs de cette distribution.

## 6. Estimation des paramètres d'une loi inconnue. Estimateurs

La formulation d'une hypothèse $H$ conduit à exprimer une certaine loi dont on recherchera l'éventuelle conformité. Les paramètres intervenant dans la loi devront être **estimés à partir de l'échantillon**. Il faut bien préciser que ces estimations sont des variables aléatoires et nous

* `http://www.edpsciences.com/guilpin/`

n'avons pas accès à la valeur exacte de chacun des paramètres et nous devrons nous contenter de l'estimation **du paramètre considéré**.

Pour obtenir une estimation, il faut employer un **estimateur**, la notion d'estimateur va se préciser dans les lignes qui suivent, toutefois on peut d'ores et déjà dire qu'il s'agit de la formule mathématique utilisée pour effectuer le calcul de l'estimation. Bien sûr, on ne choisit pas les estimateurs n'importe comment et il est souhaitable que quelques contraintes judicieusement choisies nous aident à déterminer des estimateurs les plus «performants » possibles. Par exemple, on peut souhaiter avoir les estimateurs pour lesquels les erreurs sont les plus petites possibles ; à ce sujet on se souvient que si $n$ est élevé, nous avons une très grande chance de trouver la moyenne expérimentale proche de la moyenne théorique (théorème de Bernoulli).

D'une façon générale, on peut considérer une variable aléatoire $X$ qui prend les valeurs $x_1, x_2, x_3, \ldots, x_n$, puis un estimateur $\bar{a}$ du paramètre $a$ qui est une grandeur associée à la variable $X$. Toutes les estimations possibles de $a$ le sont à partir des valeurs $\{x_k\}$. Il s'ensuit que $\bar{a}$ est une fonction des $\{x_k\}$. Par conséquent, $\bar{a}$ est une variable aléatoire dont la loi de répartition dépend de celle de $X$ ; elle dépend aussi du paramètre inconnu a ainsi que du nombre $n$ d'expériences.

Ces remarques ne suffisent pas à concevoir des estimateurs intéressants et utiles, on leur impose donc certaines contraintes dont les trois plus importantes sont les suivantes :

1. l'estimateur $\bar{a}$ doit converger vers $a$ en probabilité. On dit alors que l'estimation est **consistante** ;
2. on souhaite que $M[\bar{a}] = a$ ; lorsque cette condition est remplie, on dit que l'estimateur est **non biaisé**. Le plus souvent le **biais** dépend de $n$, et l'on parvient à corriger l'estimateur afin que sa nouvelle expression ne soit plus biaisée ; mais il convient alors de le vérifier ;
3. souvent, on désire que l'estimation centrée possède la variance la plus petite possible, c'est-à-dire que $D(\bar{a})$ soit minimum.

La plupart du temps, il n'est pas possible de satisfaire simultanément ces trois contraintes, et l'on se contente seulement des deux premières.

## 6.1. Application à l'estimation de la moyenne et de la variance

Revenons à notre variable aléatoire $X$ à valeurs $x_1, x_2, x_3, \ldots, x_n$, et désignons par $m$ et $D$ respectivement la moyenne et la variance, toutes deux inconnues.

***a – Cas de la moyenne*** – Nous définissons la moyenne au moyen de l'expression :

$$M[\bar{m}] = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

En vertu de la loi des grands nombres (théorème de Bernoulli), l'estimation $\bar{m}$ converge en probabilité vers $m$.

À présent, supposons que l'on réalise $n'$ expériences ayant fourni chacune une moyenne $m_j, j = 1, 2, \ldots, n'$. On peut écrire :

$$M[\bar{m}] = \frac{1}{n'}\sum_{j=1}^{n'} m_j = \frac{1}{nn'}\sum_{i=1}^{nn'} x_i (= m)$$

on voit alors que l'estimateur n'est pas biaisé, et que la moyenne de la moyenne est encore la moyenne. En ce qui concerne la variance de m :

$$D[\bar{m}] = \frac{1}{n'}D$$

il faut connaître la loi de répartition de $X$ pour pouvoir conclure. Notamment, il est possible de montrer que si $X$ obéit à une distribution gaussienne, la variance de $\bar{m}$ est alors minimum. Cela n'est généralement pas vrai avec d'autres lois de répartition.

***b – Cas de la variance –*** L'estimation naturelle de la variance est la variance statistique :

$$\bar{D} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{m})^2$$

$$\text{avec} \quad \bar{m} = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Comme $D$ est égal au moment du deuxième ordre non centré moins le carré de la moyenne, on peut encore écrire :

$$\bar{D} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{m}^2.$$

$\frac{1}{n}\sum_{k=1}^{n} x_k^2$ converge en probabilité vers $M[X^2]$ tandis que $\bar{m}^2$ converge en probabilité vers $m^2$. Donc $\bar{D}$ converge en probabilité vers $M[X^2] - \bar{m}^2$ et par conséquent l'estimation converge correctement en probabilité (estimation consistante). Maintenant, nous allons voir que l'estimation est biaisée, pour cela nous allons transformer $\bar{D}$ :

$$\bar{D} = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left[\frac{1}{n}\sum_{k=1}^{n} x_k\right]^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \frac{1}{n^2}\sum_{k=1}^{n} x_k^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_i x_j$$

soit encore :

$$\bar{D} = \frac{n-1}{n^2}\sum_{k=1}^{n} x_k^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_i x_j.$$

À présent, étudions la moyenne de $\bar{D}$ au travers de cette dernière expression dans laquelle on numérote les expériences (on fait $K$ expériences dans lesquelles chaque échantillon contient $n$ individus, les $Kn$ individus sont tirés au hasard indépendamment les uns des autres) :

$$M[\bar{D}] = \frac{1}{K}\sum_{j=1}^{K}\bar{D}_j = \frac{n-1}{n^2}M\left[\sum_{l=1}^{n} x_{kl}^2\right] - \frac{1}{n^2}M\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} x_{ki}x_{kj}\right].$$

Comme la variance est indépendante du choix de l'origine, alors on peut choisir l'origine des abscisses pour chacune des expériences en $\bar{m}_k$. En posant $y_{ki} = x_{ki} - \bar{m}_k$, on obtient alors :

$$M[\bar{D}] = D^* = \frac{1}{K}\frac{n-1}{n^2}\sum_{k,l} y_{kl}^2 - \frac{1}{K}\frac{1}{n^2}\sum_{k,i,j\neq i} y_{ki}y_{kj}$$

la dernière somme affectée d'un signe moins est nulle car les $y_{ki}$ sont des variables aléatoires centrées indépendantes pour lesquelles $j$ est différent de $i$ (il n'y a pas de termes au carré). D'où :

$$M[\bar{D}] = \frac{1}{K}\frac{n-1}{n^2}\sum_{k,l} y_{kl}^2 = \frac{n-1}{n}\cdot\frac{1}{Kn}\sum_{k,l} y_{kl}^2 = \frac{n-1}{n}D.$$

Donc, la moyenne de $\bar{D}$ n'est pas égale à $\bar{D}$, et l'on dira que $\bar{D}$ est un estimateur biaisé puisque l'on commet une erreur systématique en l'utilisant. On corrige aisément en prenant l'opérateur $D^* = \bar{D}\frac{n}{n-1}$. On démontre sans peine que ce nouvel estimateur est consistant puisque $\frac{n}{n-1}$ tend vers 1 quand $n$ tend vers l'infini et qu'il n'est pas biaisé. En revanche, dans le cas général, l'estimation ne donne pas un $\bar{D}$ minimum.

***Remarque*** : Ceci justifie le fait que de nombreux ouvrages donnent :

$$\bar{D} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{m})^2$$

comme estimateur de la variance.

## 7. Éléments de bibliographie


S. AÏVAZIAN (1970) *Étude statistique des dépendances*, Moscou.

E. BOREL, R. DELTHEIL et R. HURON (1962) *Probabilités, Erreurs*, Armand Colin, Paris, 12e édition.

P. DAGNELIE (1973) *Théorie et Méthodes Statistiques*, Tomes 1 et 2, Les Presses Agronomiques de Gembloux.

P. DAGNELIE (1998) *Statistique théorique et appliquée*, de Boeck.

M. FISZ (1980) *Probability theory and mathematical statistics*, Krieger.

R. RISSER et C.E. TRAYNARD (1969) *Les principes de la statistique mathématique*, Gauthier-Villars.

G. SAPORTA (1978) *Théories et méthodes de la statistique*, Publications de l'Institut Français des Pétroles.

H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 25 Étude des dépendances dans le cas linéaire

Notre étude se limite au cas de deux variables aléatoires dont la dépendance est linéaire. Ici encore il n'y a pas de difficultés à généraliser les résultats au cas d'un système dépendant linéairement de $n$ variables puisqu'il suffit d'en effectuer l'étude en combinant deux à deux les variables.

Si par malheur la liaison entre deux variables est de type curviligne, on peut chercher alors à se ramener au cas linéaire par un changement de variables adéquat. Quoi qu'il en soit, le but ultime est l'usage de la méthode des moindres carrés appliquée au calcul des paramètres de la « droite de régression » quel que soit le schéma de dépendance linéaire que l'on étudie. Le risque principal se fonde sur l'utilisation d'estimateurs qui peuvent devenir biaisés lorsque les conditions requises d'applicabilité de la méthode des moindres carrés ne sont pas toutes vérifiées.

## 1. Les types de schémas de dépendance linéaire

### 1.1. Les deux variables X et Y ne sont pas aléatoires

$Y$ est complètement déterminé par la connaissance de $X$. À première vue cela peut paraître en désaccord avec la théorie des incertitudes mais ce schéma se rencontre chaque fois que l'on aborde un problème de dénombrement, à ce moment aucun élément aléatoire ne rentre en jeu. Il s'agit alors d'une dépendance fonctionnelle classique qui prend la forme :

$$Y = aX + b.$$

Voici un exemple simple : l'âge d'un arbre en fonction du nombre de ses anneaux.

### 1.2. Dépendance de la variable aléatoire $\eta$ et de la variable non aléatoire x (schéma de régression)

Ici, on étudie le comportement de la variable aléatoire dépendante $\eta$ en fonction de la variable non aléatoire $x$. Par exemple, la variable aléatoire $\eta$ est une mesure entachée d'incertitude ; elle prend des valeurs qui sont influencées par $x$ mais aussi par des facteurs incontrôlables. La dépendance linéaire s'écrira sous la forme :

$$\eta = (ax + b) + \delta.$$

expression dans laquelle les termes entre parenthèses expriment la dépendance non aléatoire et $\delta$ exprime le **résidu**, c'est l'expression aléatoire de $\eta$. Dans ce schéma, $\delta$ doit vérifier un certain nombre de conditions :

$$M[\delta] = 0$$
$$\text{et} \quad D[\delta] = \sigma^2,$$

alors, on ne recherche qu'un schéma en moyenne pour $x$, c'est-à-dire une loi de variation de la moyenne conditionnelle $M[\eta|x]$ en fonction de $x$. Il s'ensuit que la valeur moyenne de $\eta$ est une fonction linéaire de $x$ :

$$y = \bar{a}x + \bar{b},$$

cela signifie que $y(x)$ est la moyenne de $\eta$ pour une valeur donnée de $x$. La droite ainsi définie porte le nom de **droite de régression** ; cette dénomination doit être prise comme un simple nom générique, elle est historiquement liée aux travaux de Galton (1822–1911) — considéré comme un des fondateurs de la méthode statistique — qui avait constaté que, durant ses premiers travaux, la pente de la droite en question était négative d'où l'idée de régression.

Le coefficient de la pente $\bar{a}$ s'appelle **coefficient de régression** de $y$ en $x$.

**Tableau 25.1.**

| $J'$ | $x_j = J'(J'+1)$ | $y_j = \log_e\left(\frac{I_j}{S_j}\right)$ |
|---|---|---|
| 5 | 30 | 4,196 |
| 11 | 132 | 3,47 |
| 12 | 156 | 3,37 |
| 13 | 182 | 3,333 |
| 14 | 210 | 3,176 |
| 15 | 240 | 3,196 |
| 16 | 272 | 2,980 |
| 17 | 306 | 2,941 |
| 18 | 342 | 2,823 |

***Exemple*** – L'analyse de la structure fine des bandes d'émission d'une molécule diatomique — ici H Al — permet la détermination de la température du plasma émetteur. Pour cela on mesure la surface relative $I_j$ de chaque raie à laquelle un produit de nombres quantiques $J'(J'+1)$ est associé. La théorie permet d'établir que :

$$\log_e\left(\frac{I_j}{S_j}\right) = -\frac{A}{T}J'(J'+1),$$

expression dans laquelle $S_j$ est calculé pour chaque raie $j$, $A$ est un coefficient connu attaché à la molécule étudiée. Sur un graphique, on porte en abscisse $J'(J'+1)$ et en ordonnée la valeur $y_j = \log_e\left(\frac{I_j}{S_j}\right)$. La mesure de la pente de la droite permet de calculer $T$. Sur le tableau 25.1 on a porté les résultats.

La variable $x_j$ n'est pas du tout entachée d'erreur tandis que la variable $y_j$ est sujette à certaines fluctuations incontrôlables qui confère à cette variable un caractère aléatoire.

Ce type de schéma s'applique encore même si la variable $x_j$ fait l'objet de très faibles fluctuations comparées à celles qui affectent la variable $y_j$.

### 1.3. Dépendance de la variable aléatoire $\eta$ et de la variable aléatoire $\xi$ (schéma de corrélation)

Les deux variables étudiées dépendent d'un ensemble de facteurs incontrôlables qui rendent fondamentalement aléatoires les variables étudiées $\eta$ et $\xi$.

Si l'on fait l'hypothèse d'une liaison linéaire entre $\eta$ et $\xi$, la grandeur $\eta$ se décompose en deux parties :

$$\eta = (\bar{a}\xi + \bar{b}) + \delta.$$

expression dans laquelle $\delta$ exprime le **résidu** avec comme caractéristiques :

$$M[\delta] = 0$$
$$\text{et} \quad D[\delta] = \sigma^2.$$

Il y a une contrainte supplémentaire : cette décomposition doit être effectuée de telle sorte que **les variables $\delta$ et $\xi$ soient non corrélées.** Cela impose que $M[\delta\xi] = 0$, en effet $\delta$ est centrée, il n'est donc pas nécessaire que $\xi$ le soit. Si la distribution de ces deux variables $\delta$ et $\xi$ est gaussienne, alors on est assuré de l'indépendance.

Ici encore on s'intéressera à la loi en moyenne et l'on écrira :

$$Y(x) = \bar{a}x + \bar{b}, \qquad \text{soit encore } Y(x) = M[\eta|\xi = x].$$

***Remarque*** **:** Le cas où les variables $\eta$ et $\xi$ sont simplement entachées d'erreur ne rentre pas tout à fait dans ce type de schéma bien que ce soit celui-ci qui soit appliqué. En effet désignons par $\delta_x$ et $\delta_y$ les erreurs qui entachent les variables $x$ et $y$. On peut écrire :

$$\xi = x + \delta_x \qquad \text{et} \qquad \eta = y + \delta_y$$

et l'on ajoute :

$$M[\delta_x] = 0 \qquad \text{et} \qquad M[\delta_y] = 0.$$

Remplaçons $x$ et $y$ dans l'équation de la régression :

$$\eta - \delta_y = \bar{a}(\xi - \delta_x) + \bar{b}$$

qui se transforme :

$$\eta = \bar{a}\xi + \bar{b} + (\delta_y - \bar{a}\delta_x).$$

Dans ce schéma, on voit que le résidu est corrélé avec $\xi$ puisqu'il dépend du paramètre $\bar{a}$ à estimer. Cela peut entraîner des ennuis dans le calcul des estimateurs qui risquent de perdre leurs bonnes propriétés évoquées au chapitre 24.

## 2. Fondements de l'analyse de corrélation-régression

La détermination des coefficients $\bar{a}$ et $\bar{b}$ s'effectue selon la **méthode des moindres carrés.** Pour s'assurer que les estimateurs issus de l'analyse de corrélation-régression sont tout à fait convenables, on doit vérifier que trois conditions suivantes sont remplies :

1. indépendance stochastique des résultats d'observation (l'échantillon a bien été tiré au hasard) ;
2. homogénéité des variances conditionnelles ;
3. les distributions des variables sont gaussiennes.

Nous allons reprendre en détail ces trois points.

### 2.1. Indépendance stochastique des résultats d'observation

Il s'agit de vérifier que les échantillons ont bien été prélevés au hasard. On doit s'assurer que l'observation de l'expérience n° $i$ est bien indépendante de celles qui ont été effectuées auparavant et qu'elle n'influence pas non plus les observations suivantes.

L'étude de l'indépendance stochastique des expériences évoquées dans les schémas typiques de l'analyse de corrélation-régression peut se ramener à l'étude de l'indépendance stochastique des séries de résidus $\{\delta_i\}$. Il importe de vérifier que **l'échantillonnage n'a pas été dirigé.** Désignons par $x_1, x_2, x_3, \ldots, x_n$ l'échantillon extrait de la population parente étudiée dans l'ordre où ils apparaissent. On propose plusieurs méthodes de vérification de l'indépendance stochastique des tirages.

***a – Test de la suite formée à partir de la médiane de l'échantillon –*** À partir de la suite des $x_j$, on forme une autre suite $y_j$ constituée par l'ensemble des $x_j$ ordonné par valeurs croissantes. Cette nouvelle suite s'appelle **suite variationnelle**, et $y_j$ s'appelle **l'ordre statistique de rang $j$.** La valeur empirique de la médiane $x_{med}$ est l'élément central de la suite des $y_j$, c'est-à-dire :

$$\begin{cases} y_{\frac{n+1}{2}} & \text{si } n \text{ est impair,} \\ \left(y_{\frac{n}{2}} + y_{\frac{n}{2}+1}\right)/2 & \text{si } n \text{ est pair.} \end{cases}$$

À partir de l'échantillon initial des $x_j$, on va former une suite de deux caractères arbitraires, disons 0 et 1. Si la valeur de $x_k$ est inférieure à $x_{med}$ on place un 0, si elle est supérieure on place un 1. On abandonne les éléments qui sont égaux à $x_{med}$.

L'idée du test repose sur les remarques suivantes :

la suite de 0 et de 1 est constituée de sous-suites formées soit uniquement de 0 soit uniquement de 1. On désigne par $\mu$ le nombre total de sous-suites. Il apparaît comme une question de bon sens que $\mu$ doit avoir une valeur « pas trop faible comparée à $n$ », car les sous-suites ne peuvent pas être trop longues si le tirage est dû au hasard. Il s'ensuit que la longueur $\lambda$ de la plus grande sous-suite doit avoir une valeur « pas trop grande comparée à $n$ ». Il est clair que $\mu$ et $\lambda$ sont des grandeurs qui dépendent de $n$ et on les notera : $\mu(n)$ et $\lambda(n)$.

Si l'hypothèse d'indépendance statistique est vraie, on peut montrer que $\mu(n)$ obéit à une loi normale dont la moyenne et l'écart type sont donnés par les expressions :

$$M[\mu(n)] = \frac{n+2}{2}$$
$$D[\mu(n)] = \frac{n-1}{4}.$$

La distribution de $\lambda(n)$ est quelque peu plus compliquée, aussi préfère-t-on utiliser les résultats suivants (d'après S. Aïvazian) :

$$\mu(n) > PE\left[\frac{1}{2}(n+1) - 1{,}96\sqrt{(n-1)}\right]$$
$$\lambda(n) < PE\left[3{,}3\,(\log_{10}(n)+1)\right]$$

où $PE$ définit l'opération troncature (partie entière). On rejettera l'hypothèse d'indépendance stochastique si l'une des inégalités n'est pas vérifiée avec moins de 10 chances sur cent de rejeter la loi d'indépendance à tort.

***b – Application à une suite de nombres pseudo-aléatoires*** – Le tableau 25.2 donne la hauteur d'un échantillon de 100 personnes prélevées au hasard parmi des conscrits.

**Tableau 25.2.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1,725 | 1,69 | 1,73 | 1,663 | 1,727 | 1,705 | 1,717 | 1,72 | 1,716 | 1,692 |
| 1,713 | 1,741 | 1,706 | 1,693 | 1,688 | 1,74 | 1,705 | 1,683 | 1,776 | 1,723 |
| 1,706 | 1,715 | 1,746 | 1,714 | 1,727 | 1,678 | 1,724 | 1,708 | 1,701 | 1,711 |
| 1,702 | 1,685 | 1,698 | 1,734 | 1,729 | 1,719 | 1,701 | 1,713 | 1,702 | 1,708 |
| 1,716 | 1,733 | 1,71 | 1,703 | 1,719 | 1,711 | 1,693 | 1,721 | 1,728 | 1,68 |
| 1,711 | 1,704 | 1,696 | 1,744 | 1,694 | 1,723 | 1,748 | 1,704 | 1,781 | 1,676 |
| 1,696 | 1,722 | 1,728 | 1,68 | 1,746 | 1,722 | 1,723 | 1,721 | 1,665 | 1,69 |
| 1,735 | 1,724 | 1,7 | 1,686 | 1,722 | 1,723 | 1,682 | 1,736 | 1,708 | 1,692 |
| 1,721 | 1,698 | 1,706 | 1,692 | 1,712 | 1,715 | 1,709 | 1,724 | 1,765 | 1,742 |
| 1,663 | 1,736 | 1,719 | 1,701 | 1,716 | 1,679 | 1,707 | 1,726 | 1,687 | 1,709 |

La lecture des données s'effectue ligne par ligne. On trouvera sur le Web(*) le programme `teststat.c` qui réalise ces simulations et ces calculs. Il s'agit donc de s'assurer du bien-fondé de l'indépendance stochastique du tirage. Nous avons trouvé $x_{med} = 1{,}712$, le nombre de sous-chaînes est 56 et la sous-chaîne la plus longue a la taille 6. Par ailleurs, on calcule :

$$\mu(n) = 40 \quad \text{et} \quad \lambda(n) = 9.$$

On peut donc conclure qu'il n'y a aucune raison de repousser l'hypothèse d'un tirage au hasard.

***c – Test des suites ascendante et descendante*** – Ce test permet de déceler une tendance de déviation progressive de la moyenne de la distribution.

Comme dans le cas précédent, on forme une suite de 0 et de 1 à partir de la suite donnée par l'échantillon initial. On compare chaque élément au suivant immédiat, s'il est plus grand on place un 0 et s'il est plus petit on place un 1. Le cas d'égalité est écarté.

Ici encore la philosophie demeure la même et l'on comptera le nombre de sous-chaînes $\mu(n)$ ainsi que la longueur de la sous-chaîne la plus grande $\lambda(n)$. D'un point de vue pratique on devra

* `http://www.edpsciences.com/guilpin/`

vérifier les inégalités suivantes :

$$\mu(n) > PE\left[\frac{1}{3}(2n-1) - 1{,}96\sqrt{\frac{16n-29}{90}}\right]$$
$$\lambda(n) < \lambda_0(n),$$

$\lambda_0(n)$ étant donné par le tableau :

| $n$ | $n \leq 26$ | $26 < n \leq 153$ | $153 < n \leq 1\,170$ |
|---|---|---|---|
| $\lambda_0(n)$ | 5 | 6 | 7 |

Si l'une des deux inégalités n'est pas vérifiée, on rejettera l'hypothèse d'indépendance stochastique avec moins de 10 chances sur cent de rejeter la loi d'indépendance à tort.

***d – Application à une suite de nombres pseudo-aléatoires*** – On analyse le même échantillon (Tab. 25.2, page précédente) à la lueur de ce nouveau test. On obtient pour le nombre de sous-chaînes la valeur 67 et pour la sous-chaîne la plus longue la valeur 3. Le calcul donne $\mu(n) = 58$ et le tableau $\lambda(n) = 6$. Ici encore rien ne permet de repousser l'hypothèse d'un tirage au hasard.

***e – Test des carrés des différences successives*** – Ce test ne concerne que les données qui obéissent à une distribution normale. Toujours à partir de la suite initiale, on forme la demi-somme de la moyenne des carrés des différences entre chacun des éléments et son suivant immédiat, soit :

$$q^2(n) = \frac{1}{2(n-1)} \sum_{k=1}^{n-1} (x_{k+1} - x_k)^2$$

on calcule également la variance empirique de l'échantillon :

$$\sigma^2(n) = \frac{1}{(n-1)} \sum_{k=1}^{n} (x_k - x_0)^2$$

où $x_0$ est la moyenne de $x_j$. On mesure la « déviation éventuelle » au moyen de la grandeur :

$$\gamma(n) = \frac{q^2(n)}{\sigma^2(n)}$$

il s'agit en fait de la demi-moyenne des carrés des écarts ramenée à l'écart type.

Pour un échantillon de taille supérieure à 20, on rejettera l'hypothèse de l'indépendance stochastique des tirages si $\gamma(n) < \gamma_\alpha(n)$. $\gamma_\alpha(n)$ est directement relié à la loi normale et l'on montre que l'on a :

$$\gamma_\alpha(n) = 1 + \frac{u_\alpha}{\sqrt{n + 0{,}5(1 + u_\alpha^2)}}$$

où $u_\alpha$ est le quantile d'ordre $\alpha$ de la loi normale réduite. On rappelle la définition du quantile d'ordre $\alpha$ :

$$P(x < u_\alpha) = \alpha,$$

soit si $\alpha = 0{,}05$ , on trouve dans les tables :

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1{,}65} \exp\left(-\frac{x_2}{2}\right) \mathrm{d}x = 0{,}05$$

le quantile d'ordre 0,05 est $-1{,}65$.

***Remarque*** : Il ne faut pas confondre le quantile d'ordre $\alpha$ avec le point de pourcentage 100 $Q$. Le point de pourcentage 100 $Q$ de la variable aléatoire $\xi$ est la valeur $w_q$ telle que $P(x \geq w_q) = 1 - F(w_q) = Q$ où $F(x)$ est la répartition de la variable aléatoire $\xi$.

***f – Application à une suite de nombres pseudo-aléatoires*** – Toujours sur le même exemple, on examine ce que donne le test. On obtient $q^2(n) = 0{,}000\,59$ et $\sigma^2(n) = 0{,}000\,48$ puis :

$$\gamma(n) = 1{,}228 \quad \text{et} \quad \gamma_{0{,}05}(n) = 0{,}837.$$

Encore une fois il n'y a pas de raison d'abandonner l'hypothèse d'un tirage au hasard.

## 2.2. Homogénéité des variances conditionnelles

La variance conditionnelle $D[\eta|x]$ de la variable dépendante doit rester inchangée, au sens statistique, quand $x$ varie. Autrement dit, $D[\eta|x]$ doit être une constante :

$$D[\eta|x] = \sigma^2 = \text{cte}.$$

Si tel n'est pas le cas, elle doit être proportionnelle à une fonction connue de $x$, notée $h^2(x)$ (car $D$ est une forme quadratique), qui sera déterminée empiriquement :

$$D[\eta|x] = \sigma^2 h^2(x).$$

Le problème qui nous préoccupe maintenant, c'est de connaître un test qui nous permette de juger de l'homogénéité ou non de la série des variances.

On suppose que les variables ont été regroupées dans $k$ intervalles. Soit $x_{0i}$ le milieu de chacun des intervalles, la variance empirique de chaque intervalle est donnée par l'expression :

$$s_i^2 = \frac{1}{m_i - 1} \sum_{i=1}^{m_i} (y_{ij} - \bar{y}_i)^2$$

où $m_i$ est le nombre d'éléments tombant dans l'intervalle de regroupement $i$ et $\bar{y}_i$ la moyenne de $y$ dans le même intervalle.

Il s'agit de vérifier que les différences entre les $k$ valeurs obtenues pour les $s_i^2$ sont dues au hasard ou, dans le cas contraire, si l'on doit envisager une tendance systématique à varier en fonction de $x$. Quand $D[y|x]$ est une constante, alors :

$$D[y|x_{01}] = D[y|x_{02}] = \ldots = D[y|x_{0k}] = \sigma^2,$$

et la variable aléatoire :

$$\Lambda = -\frac{1}{c} \sum_{i=1}^{k} m_i \log_e \left(\frac{s_i^2}{s^2}\right)$$

suit approximativement pour $m_i > 2$ une loi du $\chi^2$ à $k-1$ degrés de liberté ; dans expression, on a noté :

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{m_i} - \frac{1}{n}\right]$$

$$\text{et} \quad s^2 = \frac{1}{n-k}\sum_{i=1}^{k}(m_i - 1)s_i^2,$$

$n$ étant la taille de l'échantillon.

On rejettera l'hypothèse de l'indépendance de la suite des variances si la valeur empirique de $\Lambda$ tombe dans le domaine des valeurs peu probables de la variable aléatoire. Il faudra alors trouver une fonction $h^2(x)$.

### *Exemples*

1. Le tableau 25.3 donne les résultats d'une première expérience qui a été réalisée 10 fois pour cinq valeurs de la variable $x$ qui sont : 5, 10, 17, 22 et 27. La onzième ligne donne la moyenne obtenue pour chaque valeur de $x$, la douzième la variance et la treizième les valeurs de $x$.

   Tous calculs faits, on trouve la valeur $\Lambda = 5{,}12$. Le nombre de degrés de liberté est $5-1=4$, et la probabilité de dépasser la valeur $\Lambda$ est 0,275. On peut affirmer qu'il n'y a aucune raison pour repousser l'hypothèse de l'homogénéité des variances conditionnelles.

**Tableau 25.3.**

| | $y_{1j}$ | $y_{2j}$ | $y_{3j}$ | $y_{4j}$ | $y_{5j}$ |
|---|---|---|---|---|---|
| 0 | 125,8 | 148,71 | 182,53 | 211,92 | 239,61 |
| 1 | 134,19 | 151,39 | 177,44 | 217,01 | 229,10 |
| 2 | 123,87 | 160,73 | 181,26 | 210,08 | 241,02 |
| 3 | 119,82 | 151,21 | 192,15 | 207,93 | 220,77 |
| 4 | 118,46 | 155,23 | 190,81 | 212,80 | 240,19 |
| 5 | 133,87 | 140,27 | 187,63 | 210,39 | 233,52 |
| 6 | 123,51 | 154,06 | 182,36 | 205,00 | 237,15 |
| 7 | 116,95 | 149,36 | 185,76 | 213,24 | 238,11 |
| 8 | 144,81 | 147,22 | 182,49 | 215,28 | 236,93 |
| 9 | 128,99 | 150,36 | 184,29 | 201,03 | 229,74 |
| Moyenne | 127,03 | 150,86 | 184,67 | 210,47 | 234,61 |
| Variance | 74,31 | 28,87 | 20,26 | 22,98 | 40,89 |
| Valeur de $X$ | 5 | 10 | 17 | 22 | 27 |

***Remarque*** : Les calculs ont été exécutés avec 16 chiffres significatifs, mais nous n'en avons reportés que 5, les nombres ayant été arrondis.

Sur le Web(*), on trouvera le programme `varian_1.c` qui réalise cette simulation.

2. Le tableau 25.4, page ci-contre, donne les résultats d'une deuxième expérience qui a été réalisée 10 fois pour cinq valeurs de la variable $x$ qui sont : 70, 80, 115, 176 et 212.

* `http://www.edpsciences.com/guilpin/`

**Tableau 25.4.**

| | $y_{1j}$ | $y_{2j}$ | $y_{3j}$ | $y_{4j}$ | $y_{5j}$ |
|---|---|---|---|---|---|
| 0 | 148,2 | 138,7 | 109,5 | 75,91 | 72,77 |
| 1 | 155,8 | 141,5 | 100,7 | 96,89 | 0,33 |
| 2 | 146,5 | 151,2 | 107,3 | 68,32 | 82,47 |
| 3 | 142,8 | 141,3 | 126,1 | 59,47 | −57,01 |
| 4 | 141,6 | 145,5 | 123,8 | 79,55 | 76,75 |
| 5 | 155,5 | 129,8 | 118,3 | 69,59 | 30,85 |
| 6 | 146,1 | 144,2 | 109,2 | 47,4 | 55,81 |
| 7 | 140,2 | 139,3 | 115,1 | 81,36 | 62,44 |
| 8 | 165,4 | 137,1 | 109,4 | 89,73 | 54,27 |
| 9 | 151,1 | 140,4 | 112,5 | 31,05 | 4,76 |
| Moyenne | 149,3 | 140,9 | 113,2 | 69,93 | 38,34 |
| Variance | 61,01 | 31,54 | 60,18 | 389,9 | 1 0941,0 |
| Valeur de $X$ | 70,0 | 80,0 | 115,0 | 176,0 | 212,0 |

Le calcul de $\Lambda$ donne 50,57. Il y a environ deux chances sur $10^{10}$ de pouvoir dépasser cette valeur. On renonce donc à l'hypothèse de l'homogénéité des variances conditionnelles et l'on va chercher une loi $h^2(x)$ qui rende les variances homogènes. On essaye la loi $h^2(x) = \exp(x/39{,}0)$ et l'on recommence les calculs. On trouve alors : $\Lambda = 4{,}516$ et la probabilité de dépasser cette valeur est $P = 0{,}211$. Rappelons que l'on a perdu un degré de liberté en choisissant $h^2(x) = \exp(x/39{,}0)$ et donc il convient de calculer une valeur du $\chi^2$ à $k-2$ degrés de liberté. À présent, il n'y a aucune raison de rejeter l'hypothèse d'une variation exponentielle des variances, les données expérimentales ne la contredisant pas.

Sur le Web(*), on donne le programme `varian_2.c` qui réalise cette simulation.

### 2.3. Les distributions sont gaussiennes

Il faut vérifier le caractère gaussien des $y_{ij}$ pour chaque valeur fixée de $i$. C'est le critère du $\chi^2$ qui sera le plus souvent utile, quoique l'aplatissement et la dissymétrie rendent également de grands services.

## 3. Conclusions

Nous avons évoqué les conditions idéales d'application de la méthode des moindres carrés que nous avons déjà présentée au chapitre 4 à propos de l'interpolation. En pratique les trois critères ne sont pas toujours vérifiés simultanément, mais ce n'est pas pour cela que l'on va renoncer aux méthodes d'analyse de corrélation-régression et en particulier à la méthode des moindres carrés.

Il a été montré notamment que l'estimation des coefficients de corrélation par la méthode des moindres carrés conserve toutes les bonnes propriétés à condition que $n$ soit grand devant 1,

* `http://www.edpsciences.com/guilpin/`

même si la variable dépendante n'est pas distribuée normalement à condition toutefois que les erreurs de mesure soient indépendantes. On dit alors que la méthode des moindres carrés est robuste.

## 4. Éléments de bibliographie

S. AÏVAZIAN (1970) *Étude statistique des dépendances*, Moscou.
H. CRAMER (1945) *Mathematical methods of statistic*, Princeton.
P. DAGNELIE (1998) *Statistique théorique et appliquée*, de Boeck.
R. RISSER et C.E. TRAYNARD (1969) *Les principes de la statistique mathématique*, Gauthier-Villars.
H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# 26 Analyse de corrélation et de régression

## 1. La corrélation

Dans le chapitre précédent, nous avons vu que deux variables aléatoires pouvaient avoir une certaine tendance à varier simultanément. La visualisation des données sur un graphique aide fortement à caractériser une éventuelle dépendance ; c'est la dépendance linéaire qui donne toute sa puissance à l'analyse de corrélation. Avant de s'attaquer à la détermination des paramètres de la liaison de corrélation, il convient de s'assurer de l'existence effective d'une telle dépendance, et c'est l'étude du coefficient de corrélation qui permet de répondre à la question.

### 1.1. Calcul du coefficient de corrélation dans le cas où les variances conditionnelles sont homogènes

Le calcul du **coefficient de corrélation** et son interprétation ne fonctionnent correctement que dans le cas d'une dépendance linéaire entre les variables aléatoires. Si une telle dépendance ne peut être envisagée, c'est au **rapport de corrélation** que nous aurons affaire. Pour l'instant, on suppose que l'on a établi le caractère linéaire de la dépendance entre les deux variables aléatoires $\eta$ et $\xi$. Ainsi le coefficient de corrélation est défini par l'expression :

$$r_{xy} = \frac{M\{(\xi - M[\xi]) \cdot (\eta - M[\eta])\}}{\sqrt{D[\xi]D[\eta]}}.$$

La valeur empirique ou estimateur est donnée par les expressions :

$$\bar{r}_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y},$$

soit encore :

$$\bar{r}_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{k} m_i x_i^0 \bar{y}_i - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

où :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{k} m_i x_i^0$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{k} m_i \bar{y}_i$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{k} m_i (x_i^0 - \bar{x})^2$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

expressions dans lesquelles $m_i$ est le nombre d'éléments dans l'intervalle $i$, et $x_i^0$ est le milieu (centre) de l'intervalle $i$.

Nous verrons en appendice de quelle manière ces expressions sont transformées lorsque les variances conditionnelles ne sont pas homogènes. Quoi qu'il en soit, $\bar{r}_{xy}$ est un coefficient calculable pour un système quelconque de deux variables aléatoires et sa valeur est comprise entre $-1$ et $+1$.

Dans le cas où **les distributions des variables $\eta$ et $\xi$ sont gaussiennes**, le coefficient de corrélation admet une signification concrète en tant que caractéristique de l'intensité de liaison entre les variables. Ainsi :

1. $|\bar{r}_{xy}| = 1$ confirme une **dépendance fonctionnelle linéaire**, il n'y a pas d'aspect statistique,
2. $\bar{r}_{xy} = 0$ montre la complète indépendance des variables.
3. Les deux moyennes, les deux variances et le coefficient de corrélation donnent une information exhaustive sur la dépendance stochastique des variables étudiées.

Dans le cas où **les distributions des variables $\eta$ et $\xi$ ne sont pas gaussiennes**, le coefficient de corrélation ne peut être adopté que comme l'une des caractéristiques possibles de l'intensité de la liaison. Il s'ensuit que l'interprétation n'est pas sûre, et notamment :

1. $\bar{r}_{xy} = 0$ ne signifie plus l'indépendance,
2. $|\bar{r}_{xy}| = 1$ ne signifie pas nécessairement la dépendance linéaire.

***Remarque*** : Une dépendance étroite entre deux variables aléatoires ne signifie en aucune sorte une interdépendance causale. Il y a une corrélation étroite entre l'arrêt des voitures et l'apparition d'un feu rouge. Si l'on se place du point de vue du physicien le feu rouge n'est pas une cause qui arrête véritablement les automobiles. Si l'on se place du point de vue du juriste la corrélation entre, non plus l'arrêt du véhicule mais le comportement du chauffeur, et l'apparition du feu rouge est une causalité à caractère juridique. Autrement dit la corrélation entre le véhicule et le feu rouge est factice, tandis qu'elle ne l'est pas entre le chauffeur et le feu rouge.

## 1.2. Distribution empirique du coefficient de corrélation

Le but de ce paragraphe est d'établir le bien-fondé d'une liaison de corrélation. Dans le cas où les distributions des variables sont normales et où la taille $n$ de l'échantillon est grande devant l'unité, on peut alors montrer que la distribution de $\bar{r}_{xy}$ est normale en première approximation, sa moyenne étant $r$ et sa variance donnée par l'expression :

$$\sigma_r^2 = \frac{\left(1 - r^2\right)^2}{n - 1}\,.$$

Il faut noter que cette approximation devient très grossière pour $n$ prenant des valeurs de quelques unités ou pour $r$ voisin de 1 en valeur absolue.

Cette approximation par une distribution normale nous permet de tester l'hypothèse d'une absence de liaison de corrélation, soit $r = 0$. Si $r = 0$, la grandeur

$$w = \frac{|\bar{r}_{xy}|\sqrt{n-2}}{\sqrt{1-\bar{r}_{xy}^2}}$$

suit une loi de Student à $(n-2)$ degrés de liberté.

On cherche dans la table de la loi de Student la valeur $t_\alpha(n-2)$ du point de pourcentage $\alpha$ qui est le seuil de signification que l'on choisit usuellement égal à 0,05.

Si $w < t_\alpha(n-2)$, on rejette alors **l'hypothèse d'une liaison de corrélation avec la probabilité $\alpha$ de la rejeter à tort.**

***Intervalle de confiance pour la vraie valeur du coefficient de corrélation*** – On utilise le fait que la distribution de $r$ est normale. On évalue les limites du domaine au moyen des relations :

$$r = \bar{r}_{xy} + \frac{\bar{r}_{xy}(1-\bar{r}_{xy}^2)}{2n} \pm u_{\alpha/2}\frac{1-\bar{r}_{xy}^2}{\sqrt{n-1}}$$

$u_{\alpha/2}$ est le point de pourcentage $100\alpha/2$ de la distribution normale réduite. Il en résulte que $r$ appartient à cet intervalle avec un niveau de confiance $(1-\alpha)$ à condition toutefois que $r$ ne soit pas trop près de 1 et que $n$ soit suffisamment grand (plusieurs dizaines).

## 1.3. Cas d'une dépendance non linéaire, rapport de corrélation dans le cas où les variances conditionnelles sont homogènes

Dans le cas où la dépendance s'écarte notablement de la forme linéaire, le coefficient de corrélation perd son sens en tant que caractéristique de l'intensité de liaison entre les deux variables aléatoires. On utilise alors le rapport de corrélation $\rho_{\eta/\xi}$ qui est beaucoup plus sûr parce que son interprétation ne dépend pas de la forme de la dépendance de la régression étudiée. Le rapport de corrélation est donné par l'expression :

$$\bar{\rho}_{\eta/\xi}^2 = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{k} m_i(\bar{y}_i - \bar{y})^2}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y})^2} = \frac{s_{\bar{y}}^2}{s_y^2},$$

avec

$$\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{k} m_i\bar{y}_i.$$

$s_{\bar{y}}^2$ exprime la dispersion des moyennes partielles $\bar{y}_i$ autour de la moyenne globale $\bar{y}$, et $s_y^2$ mesure la dispersion des résultats individuels autour de la moyenne globale $\bar{y}$.

Contrairement au coefficient de corrélation, les rapports de corrélation $\rho_{\eta/\xi}$ et $\rho_{\xi/\eta}$ ne sont pas symétriques, en revanche, les propriétés de $r$ et de $\rho$ sont très semblables :

1. $\bar{\rho}$ est compris entre $-1$ et $+1$.
2. Si $|\rho| = 1$, il existe une dépendance fonctionnelle univoque entre $\eta$ et $\xi$ et réciproquement.

3. L'absence d'une liaison de corrélation entre $\eta$ et $\xi$ signifie que $\rho_{\eta/\xi} = 0$. Réciproquement, si $\rho_{\eta/\xi} = 0$, alors $\bar{y}_i = \bar{y}$. Les moyennes $\bar{y}_i$ ne dépendent pas de $x$, autrement dit la droite correspondante de régression est parallèle à l'axe horizontal.

***Remarque*** : Il n'y a pas de dépendance simple entre $\rho_{\eta/\xi}$ et $\rho_{\xi/\eta}$, et la non-corrélation de $\eta$ et $\xi$ ne signifie pas nécessairement la non-corrélation de $\xi$ et $\eta$.

***Absence d'une liaison de corrélation*** – Tout comme le coefficient de corrélation, $\bar{\rho}^2_{\eta/\xi}$ suit une loi normale pourvu que $n$ soit au moins de l'ordre de quelques dizaines.

Ici encore, dans le cas où :

$$\frac{\bar{\rho}_{\eta/\xi}\sqrt{n-2}}{\sqrt{1-\bar{\rho}^2_{\eta/\xi}}} < t_\alpha(n-2),$$

on cherche dans la table de la loi de Student la valeur $t_\alpha(n-2)$ du point de pourcentage $\alpha$ qui est le seuil de signification.

Si l'on est amené à rejeter **l'hypothèse d'une liaison de corrélation, on le fait avec la probabilité $\alpha$ de la rejeter à tort**. Dans le cas contraire, cela veut dire que la liaison étudiée est statistiquement significative.

***Remarque*** : Il est possible de montrer que le rapport de corrélation ne peut être inférieur à la valeur absolue du coefficient de corrélation $r$. Dans le cas d'une dépendance linéaire, les deux caractéristiques coïncident et, par conséquent, la différence $\bar{\rho}^2_{\eta/\xi} - \bar{r}^2_{xy}$ est une mesure de l'écart de la dépendance de régression par rapport à la forme linéaire.

## 1.4. Cas où les variances conditionnelles ne sont pas homogènes

On suppose alors que l'on a trouvé les deux fonctions $h^2(x)$ et $g^2(y)$ qui rendent homogènes les variances conditionnelles :

$$D[\eta|\xi = x] = \sigma^2_\eta h^2(x) \quad \text{et} \quad D[\xi|\eta = y] = \sigma^2_\xi g^2(y).$$

On pose alors :

$$w_{fi} = \frac{1}{h^2(x_i)} \quad \text{et} \quad w_{gi} = \frac{1}{g^2(y_i)}$$
$$W_f = \sum_{i=1}^{n} w_{fi} \quad \text{et} \quad W_g = \sum_{i=1}^{n} w_{gi}$$

où $x_i$ et $y_i$ sont les centres des intervalles de regroupement respectivement pour $x$ et pour $y$. Dans ces conditions, on a une autre expression du coefficient de corrélation :

$$\bar{r}_{xy} = \frac{\left\{\left(\frac{1}{W_f}\sum_{i=1}^{n} w_{fi}x_iy_i - \bar{x}\bar{y}\right)\left(\frac{1}{W_g}\sum_{i=1}^{n} w_{gi}x_iy_i - \bar{\bar{x}}\bar{\bar{y}}\right)\right\}^{1/2}}{\{\bar{s}_x\bar{s}_y\bar{\bar{s}}_x\bar{\bar{s}}_y\}^{1/2}}$$

avec

$$\bar{x} = \frac{1}{W_f}\sum_{i=1}^{n} w_{fi}x_i \qquad \text{et} \quad \bar{y} = \frac{1}{W_f}\sum_{i=1}^{n} w_{fi}y_i$$

$$s_x^2 = \frac{1}{W_f}\sum_{i=1}^{n} w_{fi}(x_i - \bar{x})^2 \quad \text{et} \quad s_y^2 = \frac{1}{W_f}\sum_{i=1}^{n} w_{fi}(y_i - \bar{y})^2$$

$$\bar{\bar{x}} = \frac{1}{W_g}\sum_{i=1}^{n} w_{gi}x_i \qquad \text{et} \quad \bar{\bar{y}} = \frac{1}{W_g}\sum_{i=1}^{n} w_{gi}y_i$$

$$\bar{\bar{s}}_x^2 = \frac{1}{W_g}\sum_{i=1}^{n} w_{gi}(x_i - \bar{\bar{x}})^2 \quad \text{et} \quad \bar{\bar{s}}_y^2 = \frac{1}{W_g}\sum_{i=1}^{n} w_{gi}(y_i - \bar{\bar{y}})^2.$$

Les relations surmontées d'une barre sont obtenues avec les valeurs de $h^2(x_i)$ tandis que celles surmontées d'une double barre sont obtenues avec les valeurs de $g^2(y_i)$.

***Remarque*** : Le signe de $\bar{r}$ doit coïncider avec le signe des facteurs sous le radical (les signes des expressions situées entre les accolades sont les mêmes).

D'une façon semblable, le rapport de corrélation prend la forme suivante :

$$\bar{\rho}^2_{\eta/\xi} = \frac{\sum_{i=1}^{k} w_{fi}m_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k} w_{fi}\sum_{j=1}^{m_i}(y_{ij} - \bar{y})^2}.$$

## 2. Régression linéaire

L'analyse de corrélation nous a montré que deux variables aléatoires sont liées. Maintenant, il s'agit de déterminer la forme générale de la liaison étudiée, puis d'estimer les paramètres entrant dans l'expression analytique de la liaison. Les seuls renseignements dont nous disposons sont les données expérimentales.

Il n'y a pas de méthode standard pour obtenir la forme analytique de la liaison, cependant, la connaissance de la nature du problème est utile pour orienter les choix adéquats. Il convient d'ajouter que les fonctions à partir desquelles la liaison entre les variables est décrite doivent être linéaires par rapport aux paramètres à déterminer. Dans le cas où la dépendance n'est pas linéaire, il s'agira alors de décrire une dépendance curviligne et l'on utilisera une parabole de degré peu élevé.

Pour ce qui concerne l'estimation des paramètres, elle sera obtenue en règle générale par **la méthode de moindres carrés** à partir des données expérimentales. C'est pour cette raison que l'on demande la linéarité des fonctions par rapport aux paramètres à déterminer : la dérivation de la forme quadratique (définie positive) par rapport à chacun des paramètres à déterminer nous conduit à l'obtention d'un système linéaire que nous savons résoudre.

### 2.1. Calcul des paramètres par la méthode des moindres carrés

On suppose que nous avons vérifié sur les variables aléatoires $x$ et $y$ un certain nombre de propriétés qui sont les suivantes :

1. il existe entre les variables une liaison du type de celles que nous avons évoquées au chapitre précédent ;

2. les variables sont gaussiennes et les résultats d'observation sont bien dus au hasard ;
3. la variance de la variable dépendante satisfait soit à l'homogénéité des variances conditionnelles $D[\eta|\xi] = \sigma_0^2$, soit à une certaine loi $h^2(x)$ qui rend les variances conditionnelles homogènes $D[\eta|x] = \sigma^2 h^2(x)$ ;
4. la liaison étudiée est statistiquement significative, c'est-à-dire que la valeur du coefficient de corrélation ou du rapport de corrélation est suffisamment éloignée de zéro pour que l'hypothèse d'une absence de liaison de corrélation ne puisse pas raisonnablement s'expliquer par des fluctuations dues au hasard ;
5. la forme de la régression est linéaire du type $Y = aX + b'$. Si cela n'est pas le cas, on transforme la variable de telle sorte qu'on trouve une forme linéaire.

En définitive, il convient d'obtenir des estimateurs de $a$ et $b'$. Il faut bien comprendre qu'il s'agit effectivement d'estimateurs car ces valeurs dépendent de l'échantillon.

## 2.2. Méthode des moindres carrés

Pour déterminer la droite empirique de régression, on minimise la somme des carrés des résidus c'est-à-dire la variance empirique $s^2$ des données expérimentales par rapport à cette droite, soit :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{Y}(x_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{a}x_i - \bar{b}']^2$$

$$\text{où} \quad \bar{s}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m_i} w_i[y_{ij} - \bar{a}x_i^0 - \bar{b}']^2$$

$$\text{et} \quad w_i = \frac{1}{h^2(x_i)}\,.$$

***Calcul de la droite de régression*** – Notons d'abord qu'il est plus habile d'effectuer les calculs de la droite par rapport au barycentre des abscisses qui n'est rien d'autre que la moyenne de $x$. On écrira donc la loi sous la forme :

$$\bar{Y} = \bar{a}(x - \bar{x}) + \bar{b},$$

$$\text{avec} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Cette remarque n'est pas une coquetterie, non seulement les calculs se retrouvent beaucoup plus aisément par ce procédé mais $\bar{a}$ et $\bar{b}$ s'avèrent être des variables (estimateurs) stochastiquement indépendantes à la différence des variables (estimateurs) $\bar{a}$ et $\bar{b}'$.

Si on dérive l'expression de $s^2$ par rapport à $\bar{a}$ et $\bar{b}$ et qu'on annule les dérivées, on obtient alors :

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} = \frac{1}{n}\sum_{i=1}^{k} m_i \bar{y}_i$$

et

$$\bar{a} = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{n s_x^2} = \frac{\sum_{i=1}^{k} m_i \bar{y}_i x_i^0 - n\bar{x}\bar{y}}{n s_x^2} = \bar{r}\frac{s_y}{s_x},$$

avec

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2.$$

Dans le cas où $D[\eta|\xi] = \sigma^2 h^2(x)$, on obtient un autre jeu d'expressions :

$$\bar{b} = \frac{1}{W}\sum_{i=1}^{n} w_i y_i = \bar{y} = \frac{1}{\sum_{i=1}^{n} w_i m_i}\sum_{i=1}^{k} w_i m_i \bar{y}_i$$

avec

$$W = \sum_{i=1}^{n} w_i,$$

et

$$\bar{a} = \frac{\sum_{i=1}^{n} w_i y_i x_i - \bar{x}\bar{y}W}{\sum_{i=1}^{n} w_i [x_i - \bar{x}]^2} = \frac{\sum_{i=1}^{k} w_i m_i \bar{y}_i x_i^0 - \bar{x}\bar{y}\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} w_i m_i [x_i^0 - \bar{x}]^2}.$$

On remarquera que $\bar{b}$ n'est plus simplement relié au coefficient de corrélation comme dans le cas d'une variance constante.

## 2.3. Estimation de la précision

***a – Sur les paramètres $\bar{a}$ et $\bar{b}$*** – Il s'agit de déterminer les intervalles de confiance pour la ligne de régression inconnue à partir de l'expression de la droite estimée :

$$\bar{Y} = \bar{a}(x - \bar{x}) + \bar{b},$$

alors que la droite $Y = a(x - \bar{x}) + b$ est inconnue.

La distribution de Student à $(n - 2)$ degrés de liberté permet de résoudre le problème qui consiste donc à déterminer la précision sur les coefficients $\bar{a}$ et $\bar{b}$. On calcule pour ce faire les quantités suivantes :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2 \quad \text{et} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \bar{Y}(x_i)\right]^2$$

à la suite de quoi, nous pouvons affirmer que :

1. $\bar{b} - t_{\alpha/2}(n-2)\frac{s}{\sqrt{n}} < b < \bar{b} + t_{\alpha/2}(n-2)\frac{s}{\sqrt{n}}$,
2. $\bar{a} - t_{\alpha/2}(n-2)\frac{s}{s_x\sqrt{n}} < a < \bar{a} + t_{\alpha/2}(n-2)\frac{s}{s_x\sqrt{n}}$, avec la probabilité $(1-\alpha)$ de tomber dans chacun de ces intervalles.

Sur le Web(*), on trouvera le programme `regres.c` qui calcule l'incertitude sur la pente de la droite de régression.

* `http://www.edpsciences.com/guilpin/`

***b – Pour une valeur donnée $x_0$ de la variable x*** – On peut montrer que la variable :

$$z(x) = \frac{|\bar{Y}(x) - Y(x)|\sqrt{n}}{s\sqrt{1 + \dfrac{(x - \bar{x})^2}{s_x^2}}}$$

suit pour toute valeur fixée de $x$ la distribution de Student à $(n-2)$ degrés de liberté.

La différence entre la valeur moyenne empirique $\bar{Y}(x_0)$ et la valeur théorique $Y(x_0)$ ne dépasse pas en valeur absolue la grandeur :

$$t_{\alpha/2}(n-2)\frac{s}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \bar{x})^2}{s_x^2}}\,.$$

***c – Cas des variances conditionnelles non homogènes*** – Il est bien entendu que toutes les formules de ce paragraphe doivent être modifiées au cas où les variances conditionnelles ne sont pas constantes. Les expressions prennent alors la forme suivante :

1. $$\bar{b} - t_{\alpha/2}(n-2)\frac{\bar{s}}{\sqrt{\sum_{i=1}^{k} w_i m_i}} < b < \bar{b} - t_{\alpha/2}(n-2)\frac{\bar{s}}{\sqrt{\sum_{i=1}^{k} w_i m_i}}$$

2. $$a \in \left( \bar{a} \pm t_{\alpha/2}(n-2)\frac{\bar{s}}{\sqrt{\sum_{i=1}^{k} w_i m_i (x_i^0 - \bar{x})^2}} \right)$$

3. $$Y(x) \in \left( \bar{Y}(x_0) \pm t_{\alpha/2}(n-2)\bar{s}\sqrt{\frac{1}{W} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2}} \right)$$

avec

$$\bar{s}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m_i} w_i [y_{ij} - \bar{a}x_i^0 - \bar{b}]^2$$

$$\text{et} \quad \bar{x} = \frac{1}{W}\sum_{i=1}^{n} w_i x_i.$$

## 2.4. Régression multilinéaire, estimation de la précision

À la fin du chapitre concernant l'interpolation, nous avons étudié les systèmes linéaires surdéterminés qui s'écrivent :

$$\sum_{k=0}^{m} a_{lk} x_k + \alpha_k = 0,$$

pour $l = 0, 1, 2, \ldots, n$. Soit en notation matricielle :

$$\mathbf{A}X = b, \tag{26.1}$$

$\mathbf{A}$ étant une matrice de $n$ lignes et de $m$ colonnes (avec $m \leq n$). Le système des équations normales associé au système surdéterminé s'écrit alors :

$$[\mathbf{A}^T\mathbf{A}]X = [\mathbf{A}^T]b, \tag{26.2}$$

expression dans laquelle $\mathbf{A}^T$ est la matrice transposée de la matrice $\mathbf{A}$. Il est intéressant de pouvoir estimer les incertitudes qui entachent les résultats obtenus. La résolution du système (26.2) fournit une solution que l'on note $\Xi(\xi_0, \xi_1, \xi_2, \ldots, \xi_m)$ et l'on désire obtenir les valeurs de la dispersion $\sigma_k$ sur chacune des valeurs $\xi_k$. Pour cela on calcule d'abord la dispersion $\sigma$ sur les valeurs des seconds membres, on l'obtient en calculant d'abord la somme des carrés des résidus :

$$R^2 = \sum_{l=0}^{n} \left[ \sum_{k=0}^{m} a_{lk}\xi_k + \alpha_k \right]^2,$$

$\sigma$ est donné par l'expression suivante :

$$\sigma = \frac{R}{\sqrt{m-n}}.$$

Désignons par $d_{lk}$ les éléments de la matrice $\mathbf{D} = [\mathbf{A}^T\mathbf{A}]^{-1}$ (inverse de la matrice des équations normales). Alors la dispersion $\sigma_k$ sur chacune des valeurs $\xi_k$ est donnée par l'expression :

$$\sigma_k = \sigma\sqrt{d_{kk}}.$$

Il est aisé d'obtenir la matrice de corrélation $r_{qp}$ entre les variables $x_q$ et $x_p$ :

$$r_{qp} = \frac{d_{qp}}{\sqrt{d_{qq}d_{pp}}}.$$

Le programme `surdeter.c` mentionné à la fin du chapitre 6 sur l'interpolation s'occupe de calculer la matrice de corrélation et les incertitudes sur les valeurs calculées des inconnues ; le détail des calculs se trouve dans l'ouvrage de H. Mineur cité en bibliographie.

***Remarque*** : Si le coefficient de corrélation entre deux variables est proche de l'unité (hormis les éléments de la diagonale), cela signifie qu'il faut supprimer dans le modèle l'une des deux variables.

## 3. Éléments de bibliographie

S. AÏVAZIAN (1970) *Étude statistique des dépendances*, Moscou.
H. CRAMER (1945) *Mathematical methods of statistic*, Princeton.
H. MINEUR (1938) *Technique de la méthode des moindres carrés*, Gauthier-Villars.
R. RISSER et C.E. TRAYNARD (1969) *Les principes de la statistique mathématique*, Gauthier-Villars.
H. VENTSEL (1973) *Théorie des probabilités*, Éditions MIR, Moscou.

# ANNEXES

# A Les suites de Sturm. Application à la détermination du nombre de racines réelles d'un polynôme

En 1829, le mathématicien Charles Sturm (1803–1855) fait paraître un mémoire, célèbre alors, consacré à la détermination du nombre de racines réelles d'une équation $f(x) = 0$. L'intérêt est avant tout théorique, et, comme on s'en apercevra au cours des applications pratiques, les suites de Sturm ne sont pas d'un emploi aussi aisé qu'une première impression pourrait le laisser supposer dans la mesure où leur usage impose une très grande précision dans les calculs numériques. L'introduction des erreurs d'arrondi à chaque étape des calculs conduit d'une façon quasi inexorable à des résultats aberrants. Ceci explique la limitation des suites de Sturm au cas des polynômes de degré assez faible (quelques unités) à coefficients entiers.

## 1. Notion de variations d'une suite numérique

Considérons une suite numérique quelconque finie ou infinie notée $a_0, a_1, a_2, \ldots, a_k, \ldots$ On dit qu'il existe une variation dans cette suite lorsque deux termes consécutifs possèdent des signes opposés. Autrement dit, il existe une variation lorsque le produit de deux termes consécutifs est négatif. Par exemple, considérons la suite finie suivante : $15, 12, -8, -7, 18, -15, -12, -4, 7, 9$.

Cette suite possède quatre variations ou quatre changements de signe. C'est sur cette notion de variations dans une suite que repose le théorème de Sturm.

## 2. Suite de Sturm générée à partir d'un polynôme

Soit un polynôme de degré $n$, à coefficients réels, que l'on note $P_n(x)$. À partir de $P_n(x)$, on génère une suite de polynômes $Q_j(x)$ définie ainsi : on pose $Q_0(x) = P_n(x)$ et $Q_1(x) = P'_n(x)$, $P'_n(x)$ étant le polynôme dérivé de $P_n(x)$ par rapport à $x$. Ensuite, on divise $Q_0(x)$ par $Q_1(x)$. On obtient alors :

$$Q_0(x) = (a_0x + b_0)Q_1(x) + R_{n-2}(x),$$

expression dans laquelle on a désigné par $R_{n-2}(x)$ le reste de la division, c'est au plus un polynôme de degré $(n-2)$. On pose alors :

$$Q_2(x) = -R_{n-2}(x).$$

On opère de la même façon en divisant $Q_1(x)$ par $Q_2(x)$, ce qui s'écrit :

$$Q_1(x) = (a_1x + b_1)Q_2(x) + R_{n-3}(x),$$

le reste de la division $R_{n-3}(x)$ étant un polynôme de degré est au plus $(n-3)$. À nouveau, on pose $Q_3(x) = -R_{n-3}(x)$. Nous allons poursuivre les mêmes opérations jusqu'à ce que nous parvenions au terme ultime de la suite :

$$Q_n(x) = -R_0(x),$$

où $R_0(x)$ est alors une constante. **La suite des $Q_j(x)$ s'appelle suite de Sturm.** Pour des raisons de concision d'écriture, on peut désigner par $m_j$ la quantité $(a_j x + b_j)$, et l'on obtient la relation de récurrence entre trois polynômes consécutifs :

$$Q_k(x) = m_k Q_{k+1}(x) - Q_{k+2}(x),$$

avec les deux premiers polynômes de la suite : $Q_0(x)$ et $Q_1(x)$.

## 3. Quelques propriétés des suites de Sturm

D'abord, nous allons supposer que le polynôme $P_n(x)$ ne possède aucune racine multiple, autrement dit, $Q_0(x)$ et $Q_1(x)$ n'ont aucune racine commune. Sous cette hypothèse, on déduit les propriétés suivantes :

**a.** $Q_n(x)$ est une constante différente de zéro. En effet, $Q_n(x)$ est le plus grand commun diviseur de $Q_0(x)$ et de $Q_1(x)$, et si $Q_n(x)$ est égal à zéro cela signifie que $Q_0(x)$ et $Q_1(x)$ ont au moins une racine commune ce qui est contraire à l'hypothèse.

**b.** Deux polynômes consécutifs $Q_k(x)$ et $Q_{k+1}(x)$ ne possèdent aucune racine commune. En effet, la relation de récurrence montre que, s'il n'en était pas ainsi, $Q_{k+2}(x)$ s'annulerait aussi et ainsi de suite jusqu'à $Q_n(x)$ qui serait nul ; mais nous venons d'établir au paragraphe précédent que cela n'était pas possible.

**c.** Si, pour une valeur $x = \eta$ de la variable, le polynôme $Q_k(x)$ s'annule, la relation de récurrence nous donne :

$$Q_k(\eta) = -Q_{k+2}(\eta) \qquad \text{ou encore} \qquad Q_k(\eta) \cdot Q_{k+2}(\eta) < 0.$$

**d.** Lorsque $Q_0(x)$ s'annule pour la valeur $x = \xi$, le rapport $Q_0(x)/Q_1(x)$ passe du négatif au positif quand $x$ passe par $\xi$ par valeurs croissantes. Ceci n'est vrai qu'au voisinage immédiat de la racine de $Q_0(x)$ dans un domaine où $Q_1(x)$ garde un signe constant — on rappelle que $Q_0(x)$ et $Q_1(x)$ n'ont pas de racine commune.

Comme par hypothèse $\xi$ est une racine simple, nous pouvons écrire :

$$Q_0(x) = (x - \xi)(\alpha_{n-1} + \alpha_{n-2}(x - \xi) + \cdots).$$

À ce propos, signalons qu'on peut toujours écrire la deuxième parenthèse du deuxième membre sous forme d'un polynôme $\Pi_{n-1}(x)$ de degré $(n-1)$ :

$$\Pi_{n-1}(x) = c_n + c_{n-1}x + c_{n-2}x^2 + \cdots + c_1 x^{n-1}.$$

Le changement de variable $x$ en $x - \xi$ permet d'écrire :

$$\Pi_{n-1}(x) = \alpha_{n-1} + \sum_{k=1}^{n-1} \alpha_{n-k}(x - \xi)^k.$$

D'un autre côté, on peut écrire $Q_1(x)$ sous la forme :

$$Q_1(x) = \alpha_{n-1} + 2\alpha_{n-2}(x-\xi) + 3\alpha_{n-3}(x-\xi)^2 + \cdots + (n-1)\alpha_1(x-\xi)^{n-1}.$$

Dans le voisinage immédiat de la racine $\xi$, c'est-à-dire dans le domaine $(\xi - \varepsilon^2, \xi - \varepsilon^2)$ où $Q_1(x)$ n'a pas de racine, on peut écrire :

$$\frac{Q_0(x)}{Q_1(x)} = (x-\xi)\frac{\alpha_{n-1}}{\alpha_{n-1}} = (x-\xi)$$

les signes sont donnés ci-dessous :

| $x$ | $\xi - \varepsilon^2$ | $\xi$ | $x - \varepsilon^2$ |
|---|---|---|---|
| $\frac{Q_0(x)}{Q_1(x)}$ | $-$ | $0$ | $+$ |

Il s'ensuit que $Q_0(x)$ et $Q_1(x)$ ont leurs zéros entrelacés. À présent, il nous faut envisager le cas des racines multiples. À partir d'un certain rang, noté $(p+1)$, qui ne dépend que de l'ordre de multiplicité des racines, la suite des polynômes devient identiquement nulle. Le polynôme $Q_p(x)$ est alors le PGCD de $Q_0(x)$ et $Q_1(x)$. Si l'on pose :

$$Q_0^* = \frac{Q_k(x)}{Q_p(x)}$$

alors la nouvelle suite $Q_0^*(x), Q_1^*, Q_2^*, \ldots Q_p^*$ est une suite de Sturm, et le polynôme $Q_0^*$ n'a plus que des racines simples. Ainsi s'est-on ramené au cas précédent.

## 4. Le théorème de Sturm (1829)

Soient $a$ et $b$ deux nombres réels tels que $a < b$. Le nombre $N$ de racines réelles du polynôme $P_n(x)$ à coefficients réels et de degré $n$, qui sont comprises entre $a$ et $b$, est égal à la différence des variations prises par la suite de Sturm $\{Q_j(x)\}$ aux points $a$ et $b$ : $N = N(a) - N(b)$, où $N(y)$ désigne le nombre de variations de la suite de Sturm $\{Q_j(x)\}$.

### Démonstration

Pour parvenir à nos fins, nous allons faire croître la variable $x$ depuis la valeur $a$ jusqu'à la valeur $b$. Chaque fois que l'on passe par une racine $\xi$ de $Q_k(x)$, les signes des deux polynômes voisins $Q_{k-1}(x)$ et $Q_{k+1}(x)$ sont conservés et demeurent, bien sûr, opposés. Que le signe de $Q_{k-1}(x)$ change ou non, le nombre de variations de la suite partielle $Q_{k-1}(x)$, $Q_k(x)$ et $Q_{k+1}(x)$ ne change pas. Donc, d'une façon générale, le fait de passer par une racine d'un polynôme $Q_k(x)$ ne change pas le nombre de variations de la suite **excepté si $k = 0$.** En effet, dans ce dernier cas, au passage de chaque racine de $Q_0(x)$, le rapport $Q_0(x)/Q_1(x)$ passe du négatif au positif pour les $x$ croissants et l'on perd alors une variation. Il est évident qu'une racine multiple n'est comptée que comme racine unique, et il est indispensable d'examiner en détail la suite des $Q_j(x)$. D'un point de vue pratique, le nombre total de racines réelles est donné par :

$$N_1 = N(-\infty) - N(+\infty).$$

Dans ce cas, seul le coefficient du degré le plus élevé de chaque polynôme $Q_j(x)$ est utile pour calculer $N_1$.

Le nombre de racines positives est donné par :

$$N_2 = N(0) - N(+\infty).$$

Il faut, pour calculer $N_2$, connaître les coefficients des termes de degré zéro dans chacun des polynômes (il s'agit de la constante). Nous pouvons donc en déduire que, pour qu'un polynôme de degré $n$ possède $n$ racines réelles, il faut que les coefficients de degré le plus élevé soient positifs pour chacun des polynômes de la suite de Sturm. Par ailleurs, pour qu'un polynôme n'ait que des racines positives, il faut en plus que les termes de degré zéro aient des signes alternés. Insistons sur le fait que le théorème de Sturm peut se révéler très efficace pour localiser une racine ou pour vérifier qu'il existe ou non une racine dans un domaine choisi à l'avance (fini ou infini).

## 5. Disposition des calculs, schéma de Routh (1831–1907)

Le but de cet algorithme (1905) est avant tout d'éviter les nombres fractionnaires résultant de la division de deux polynômes consécutifs. À cette fin, il suffit de multiplier tous les coefficients du dividende et tous les coefficients du diviseur par deux nombres tels que le reste et le quotient n'aient pas de nombres fractionnaires comme coefficients. On dispose sur deux rangées les coefficients de $Q_0(x)$ et de $Q_1(x)$ selon les puissances décroissantes.

$$\begin{array}{cccccccc} a_0 & a_1 & a_2 & a_3 & a_4 & \dots & a_{n-1} & a_n \\ b_0 & b_1 & b_2 & b_3 & b_4 & \dots & b_{n-1} & b_n. \end{array}$$

On calcule une troisième rangée de termes appelés $c_j$ possédant $(n-1)$ termes et que l'on exprime facilement à partir des $a_j$ et des $b_j$ au moyen de la relation :

$$c_{j-1} = b_0 a_j - a_0 b_j \qquad \text{avec} \qquad j = 1 \dots n.$$

Cette ligne correspond au reste partiel de la division de $Q_0(x)$ et $Q_1(x)$, on obtient le reste définitif en calculant une quatrième rangée de termes $d_j$ exprimée au moyen d'une combinaison linéaire des $b_j$ et des $c_j$ :

$$d_{j-1} = c_0 b_j - b_0 c_j.$$

La rangée des $d_k$ est donc l'ensemble des coefficients, disposés selon les puissances décroissantes, du polynôme $Q_2(x)$. On recommence rigoureusement les mêmes opérations en utilisant les polynômes $Q_1(x)$ et $Q_2(x)$ ; on poursuit jusqu'à parvenir à $Q_n(x)$.

## 6. Quelques exemples de suites de Sturm

Chaque suite de polynômes orthogonaux constitue une suite de Sturm. L'étude des zéros de ces polynômes peut être envisagée par ce moyen.

## 7. Mise en œuvre du théorème de Sturm

Sur le Web(*), nous proposons le programme `sturm.c` qui dénombre les racines réelles d'un polynôme quelconque à coefficients réels. Afin de limiter la taille très rapidement croissante des nombres calculés par la méthode de Routh, à chaque calcul de la suite des $d_j$, nous effectuons la recherche du PGCD de tous les nombres $d_j$, puis, le cas échéant, nous procédons à la simplification, c'est-à-dire à la division par le PGCD.

* `http://www.edpsciences.com/guilpin/`

### 7.1. Exemple 1

Voici les coefficients dans l'ordre des puissances décroissantes d'un polynôme de degré 4.

$$a_0 = 1 \qquad a_1 = -5 \qquad a_2 = 10 \qquad a_3 = -10 \qquad a_4 = 4.$$

Ce polynôme admet deux racines réelles qui ont pour valeur 1 et 2, et deux racines complexes conjuguées $1 \pm i$. Le programme donne :

nombre de racines réelles multiples = 0 et nombre de racines réelles distinctes = 2.

### 7.2. Exemple 2

Voici les coefficients dans l'ordre des puissances décroissantes d'un polynôme de degré 5.

$$a_0 = 1 \qquad a_1 = -2 \qquad a_2 = -1 \qquad a_3 = 4 \qquad a_4 = -2 \qquad a_5 = -4.$$

Ce polynôme admet une racine réelle double qui a pour valeur $-1$, deux racines complexes conjuguées $1 \pm i$, et une racine réelle qui a la valeur 2. Le programme donne :

nombre de racines réelles multiples = 1 et nombre de racines réelles distinctes = 2.

### 7.3. Exemple 3

Voici les coefficients d'un polynôme de degré 6 :

$$a_0 = 1 \quad a_1 = -21 \quad a_2 = 175 \quad a_3 = -735 \quad a_4 = 1\,624 \quad a_5 = -1\,764 \quad a_6 = 720.$$

Ce polynôme admet pour racines 1, 2, 3, 4, 5, 6. Le programme donne 6 racines réelles distinctes et aucune racine multiple.

### 7.4. Exemple 4

Voici un polynôme de degré 10 qui admet comme racines : 4 fois la valeur 1, 3 fois la valeur 2, 2 fois la valeur 3, 1 fois la valeur 4. Ses coefficients sont :

$$a_0 = 1 \qquad a_1 = -20 \qquad a_2 = 175 \qquad a_3 = -882 \qquad a_4 = 2\,835 \qquad a_5 = -6\,072$$
$$a_6 = 8\,777 \qquad a_7 = -8\,458 \qquad a_8 = 5\,204 \qquad a_9 = -1\,848 \qquad a_{10} = 288.$$

Le programme indique qu'il y a 6 racines multiples, et 4 racines réelles distinctes. Cet exemple montre qu'il ne faut pas espérer trop de cette méthode dès que le degré du polynôme atteint la dizaine, la perte de signification des nombres calculés devient la cause principale de cette mise en échec. Nous avons rencontré un problème semblable lors du calcul des racines d'un polynôme : si l'on dispose d'une machine utilisant 16 chiffres significatifs, on ne peut guère espérer calculer des racines correctes pour les polynômes de degré égal ou supérieur à 13.

## 8. Éléments de bibliographie

E. DURAND (1971) *Solution numérique des équations algébriques*, Tome I, Éditions Masson, Paris.

H. MINEUR (1966) *Techniques de calcul numérique*, Éditions Dunod.

# B Polynômes orthogonaux relativement à une fonction poids. Généralisation de la méthode de Gauss

Compte tenu de la grande simplicité d'emploi agrémentée d'une grande précision, on peut se demander dans quelle mesure la méthode de Gauss est susceptible d'être généralisée aux cas d'intégrales, ayant bien sûr un sens, mais présentant des singularités. Le plus souvent, ces singularités existent aux bornes du domaine d'intégration lequel peut être fini ou infini.

Il n'est pas question de chercher à passer en revue le plus grand nombre de singularités, mais d'étudier, dans la mesure du possible, les cas les plus fréquemment rencontrés dans les applications. La méthode consiste à conserver les principes généraux qui ont permis d'obtenir la formule d'approximation de Gauss, et bien entendu, notre attention devra tout particulièrement porter sur l'utilisation des polynômes orthogonaux classiques.

Nous montrerons l'existence de racines de ces polynômes exclusivement réelles et distinctes, ainsi que quelques autres propriétés générales. En résumé, ce chapitre présente les définitions et théorèmes utiles pour l'établissement de méthodes d'intégration généralisant celle de Gauss.

## 1. Généralisation de la notion de polynômes orthogonaux

On appelle suite de polynômes orthogonaux sur l'intervalle fini ou infini $(a, b)$ relativement à la fonction poids $\omega(x)$ définie positive ou nulle sur $(a, b)$ une suite de polynômes :

$$W_0(x), W_1(x), W_2(x), \ldots, W_n(x), \ldots$$

de degré respectivement $0, 1, 2, \ldots, n, \ldots$ et qui vérifient les relations :

$$\int_a^b \omega(x) W_k(x) W_j(x)\, \mathrm{d}x = 0 \quad \text{si } k \neq j. \tag{B.1}$$

Cela implique que :

$$\int_a^b x^n \omega(x)\, \mathrm{d}x$$

doit converger quel que soit $n$.

Remarquons que la suite n'est définie qu'à une constante multiplicative près. On peut choisir cette constante, comme nous l'avons déjà écrit, selon divers critères que nous rappelons :

1. $W_n(1) = 1$ pour tout $n$ ;
2. orthonormalité : $\int_a^b \omega(x) W_n^2(x)\, dx = 1$, pour tout $n$ positif ou nul ;
3. le coefficient de la puissance la plus élevée du polynôme a une valeur fixée à l'avance ; lorsque cette valeur est égale à 1, on dit que le polynôme est à coefficient principal réduit.

Il nous reste à justifier l'existence d'une telle suite. Considérons la suite de fonction linéairement indépendantes :

$$\sqrt{\omega(x)}, x\sqrt{\omega(x)}, x^2\sqrt{\omega(x)}, \ldots, x^n\sqrt{\omega(x)}, \ldots$$

Le procédé d'orthogonalisation de Schmidt (1876–1959) nous permet d'obtenir un système de fonctions orthogonales répondant à notre question. On obtient alors :

$$W_0(x)\sqrt{\omega(x)}, W_1(x)\sqrt{\omega(x)}, \ldots, W_n(x)\sqrt{\omega(x)}, \ldots$$

où $W_0(x), W_1(x), \ldots, W_1(x), \ldots$ sont respectivement des polynômes de degré $0, 1, \ldots, n, \ldots$

## 2. Décomposition d'une fonction f(x) sur la base des polynômes $W_k(x)$ orthogonaux sur l'intervalle (a, b)

### Théorème liminaire

Tout polynôme $Q_n(x)$ de degré $n$ est représentable par une somme finie de polynômes $W_j(x)$ de degré inférieur et égal à $n$.

Pour démontrer cette proposition, il suffit de reprendre la technique analogue à celle que nous avons mise en œuvre lors de l'étude des polynômes de Lagrange (chapitre 6). Développons $Q_n(x)$ selon les puissances décroissantes de $x$ :

$$Q_n(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n = \sum_{k=0}^{n} a_k x^{n-k}.$$

Il nous suffit de remarquer, alors, que :

1. 1 est proportionnel à $W_0(x)$ :

$$1 = a_{00} W_0(x) ;$$

2. $x$ est une fonction linéaire de $W_0(x)$ et $W_1(x)$ :

$$x = a_{01} W_0(x) + a_{11} a_{00} W_0(x) ;$$

3. $x^2$ est une fonction linéaire de $W_0(x)$, $W_1(x)$ et $W_2(x)$ :

$$x^2 = a_{02} W_0(x) + a_{12} W_1(x) + a_{22} W_2(x),$$

et ainsi de suite.

Ces différentes expressions reportées dans $Q_n(x)$ donnent directement pour $Q_n(x)$ une forme linéaire en $W_k(x)$ :

$$Q_n(x) = b_0 W_0(x) + b_1 W_1(x) + \cdots + b_n W_n(x).$$

On en déduit la relation suivante :

$$\int_a^b \omega(x) W_{n+1}(x) Q_n(x)\, \mathrm{d}x = 0.$$

À présent, intéressons-nous à la fonction $f(x)$ pouvant être légitimement approchée par le polynôme $R_n(x)$ sur l'intervalle $(a, b)$ dans les conditions du théorème de Weierstrass ; nous pouvons écrire :

$$f(x) = R_n(x) + e(x) \quad \text{où } e(x) \text{ est le résidu ou le reste.}$$

$R_n(x)$ peut se décomposer sur la base des polynômes $W_j(x)$ :

$$R_n(x) = c_0 W_0(x) + c_1 W_1(x) + \cdots c_n W_n(x).$$

Pour obtenir les coefficients $c_k$, il suffit de multiplier les deux membres de l'équation précédente par $\omega(x) W_k(x)$, puis d'effectuer l'intégration sur l'intervalle $(a, b)$ dans la mesure où celle-ci a un sens.

$$\int_a^b \omega(x) R_n(x) W_k(x)\, \mathrm{d}x = c_k \int_a^b \omega(x) W_k^2(x)\, \mathrm{d}x$$

en vertu des relations d'orthogonalité. En posant :

$$I_k = \int_a^b \omega(x) W_k^2(x)\, \mathrm{d}x$$

on peut alors écrire :

$$c_k = \frac{1}{I_k} \int_a^b \omega(x) R_n(x) W_k(x)\, \mathrm{d}x.$$

Autrement dit, on choisit comme approximation de $f(x)$ un polynôme $R_n(x)$ de degré $n$ qui vérifie les relations :

$$\int_a^b \omega(x) R_n(x) W_k(x)\, \mathrm{d}x = \int_a^b \omega(x) f(x) W_k(x)\, \mathrm{d}x$$

avec $k = 1, 2, \ldots, n$.

Du point de vue strictement mathématique, la décomposition est d'autant plus précise que la valeur de $n$ est élevée, et cela en vertu du théorème de Weierstrass.

## 3. Racines des polynômes orthogonaux

### Théorème

Les zéros d'un polynôme de degré $n$ appartenant à une suite de polynômes orthogonaux relativement à la fonction poids $\omega(x)$ sur l'intervalle $(a, b)$ sont réels, distincts et compris dans l'intervalle $(a, b)$.

Considérons l'expression suivante :

$$\int_a^b \omega(x) W_n(x)\, \mathrm{d}x = 0,$$

elle signifie que $W_0(x)$ et $W_n(x)$ sont orthogonaux. Comme $\omega(x)$ n'est jamais une fonction négative dans l'intervalle $(a, b)$, il s'ensuit nécessairement que $W_n(x)$ s'annule au moins en un point dans l'intervalle $(a, b)$. Désignons par $x_1, x_2, \ldots, x_k$ tous les points où $W_n(x)$ s'annule. Le polynôme :

$$\Pi(x) = W_n(x)(x - x_1)(x - x_2)\ldots(x - x_k) = W_n(x) X_k(x)$$

conserve un signe constant dans l'intervalle $(a, b)$. $X_k(x)$ est un polynôme de degré $k$. Faisons l'hypothèse que $k$ soit plus petit que $n$. Il vient :

$$\int_a^b \omega(x) W_n(x) X_k(x)\, \mathrm{d}x \neq 0$$

puisque $W_n(x)X_k(x)$ garde un signe constant sur l'intervalle $(a, b)$. Par ailleurs, $X_k(x)$ est décomposable en polynômes $W_j(x)$ de degré égal et inférieur à $k$, soit :

$$X_k(x) = \sum_{j=0}^{k} \mathrm{d}_j W_j(x).$$

Il s'ensuit que :

$$\int_a^b \omega(x) W_n(x) X_k(x)\, \mathrm{d}x = 0$$

en vertu des relations d'orthogonalité puisque $n \neq k$ par hypothèse. On en déduit que $k$ doit être égal à $n$ puisque l'intégrale précédente est différente de zéro pour $k = n$. Par conséquent, ceci nous conduit à affirmer que le polynôme $W_n(x)$ de degré $n$ possède $n$ racines réelles dans l'intervalle $(a, b)$.

Il nous reste à montrer qu'il ne peut y avoir de racines doubles. Pour cela, supposons que la $h^{\mathrm{e}}$ racine soit double, dans ce cas la forme :

$$\Pi(x) = W_n(x)(x - x_1)(x - x_2)\ldots(x - x_h)^2 \ldots (x - x_{n-1})$$

est définie positive sur $(a, b)$, mais il en est de même pour le polynôme :

$$\Pi(x) = W_n(x)(x - x_1)(x - x_2)\ldots(x - x_{h-1})(x - x_{h+1})\ldots(x - x_{n-1})$$

et l'on retombe exactement sur le raisonnement précédent. On en conclut donc que toutes les racines sont distinctes (ou racines simples), car par le même processus on montrerait qu'il ne peut y avoir de racines triples, quadruples et ainsi de suite.

## 4. Relation de récurrence entre trois polynômes orthogonaux consécutifs

### Théorème

Entre trois polynômes orthogonaux consécutifs, il existe toujours une relation de récurrence de la forme :

$$W_n(x) - (A_n x + B_n)W_{n-1}(x) + C_n W_{n-2}(x) = 0 \tag{B.2}$$

où $A_n$, $B_n$ et $C_n$ sont trois constantes qui ne dépendent que de $n$.

Pour obtenir ce théorème, on procède de la manière suivante : d'abord on ajuste le coefficient $A_n$ en identifiant les termes de degré $n$, ce qui revient à écrire que :

$$\Pi(x) = W_n(x) - A_n W_{n-1}(x)x$$

est un polynôme de degré $(n-1)$ au plus.

Développons alors $\Pi(x)$ sur la base de polynômes $W_j(x)$ :

$$\Pi(x) = a_0 W_0(x) + a_1 W_1(x) + \cdots + a_{n-1} W_{n-1}(x).$$

En écrivant les relations d'orthogonalité pour $k < (n-2)$, on obtient :

$$\int_a^b \omega(x)\Pi(x)W_k(x)\,\mathrm{d}x = A_n \int_a^b \omega(x)xW_k(x)W_{n-1}(x)\,\mathrm{d}x = 0$$

car $xW_k(x)$ est au plus un polynôme de degré $(n-2)$. On en conclut que :

$$a_0 = a_1 = a_2 = a_3 = \ldots = a_{n-3} = 0.$$

Il suffit de poser : $a_{n-2} = -C_n$ et $a_{n-1} = B_n$ pour obtenir la relation de récurrence proposée.

L'intérêt de ce théorème repose sur le calcul effectif des coefficients des polynômes orthogonaux ; il suffit de connaître la relation de récurrence ainsi que les coefficients de deux polynômes consécutifs (généralement les deux premiers $W_0(x)$ et $W_1(x)$) pour obtenir tous les coefficients des autres polynômes.

## 5. Généralisation de la méthode de Gauss

La méthode de Gauss généralisée s'applique au calcul d'intégrales de la forme :

$$I = \int_a^b \omega(x)f(x)\,\mathrm{d}x$$

où $f(x)$ est une fonction régulière sur $(a,b)$ représentable par un polynôme sur cet intervalle et où $\omega(x)$ est une fonction définie positive ou nulle sur $(a,b)$ pouvant présenter des singularités sans pour autant altérer la convergence de l'intégrale $I$. Du reste, dans le cas le plus général, les singularités se présentent aux bornes $a$ ou $b$ du domaine, lequel, rappelons-le, peut être fini ou infini selon les cas. On conserve la relation (7.12) du chapitre 7 pour calculer une valeur approchée de l'intégrale $I$, soit $J$ cette approximation :

$$J = \sum_{j=0}^{n} H_j f(x_j) \tag{B.3}$$

et l'on conserve également le critère précédemment retenu pour obtenir les $H_j$ et les $x_j$ : il faut que $I = J$ si $f(x)$ est un polynôme arbitraire dont le degré est le plus élevé possible ; on désigne par $Q_n(x)$ ce polynôme.

### 5.1. Calcul des $x_j$

Reprenant la relation (7.12) du chapitre 7, on écrit l'expression :

$$0 = \sum_{j=0}^{n} H_j Q_n(x) W_{n+1}(x_j). \tag{B.4}$$

Les $x_j$ sont les racines du polynôme $W_{n+1}(x)$ appartenant à la suite de polynômes orthogonaux sur l'intervalle fini ou infini $(a, b)$ par rapport à la fonction de base $\omega(x)$ non négative sur $(a, b)$.

### 5.2. Calcul des $H_j$

Comme nous avons vu que chaque polynôme $W_k(x)$ possédait $k$ racines distinctes, on peut encore écrire :

$$W_{n+1}(x) = a \prod_{k=0}^{n} (x - x_k)$$

expression dans laquelle $a$ est le coefficient du terme de degré le plus élevé. On poursuit les calculs de la même manière que celle abordée lors de l'étude des polynômes de Legendre. Du reste il suffit de s'apercevoir que le calcul des $H_j$ ne fait appel qu'à la seule orthogonalité pour obtenir la nouvelle expression des $H_j$, soit :

$$H_j = \frac{1}{W'_{n+1}(x_j)} \int_a^b \omega(x) \frac{W_{n+1}(x)}{x - x_j} \, \mathrm{d}x$$

ou bien

$$H_j = \frac{K}{W'_{n+1}(x_j)} W_n(x_j) \tag{B.5}$$

où $K$ dépend de la suite particulière de polynômes envisagée. Son expression est donnée par :

$$K = I_n \frac{a_{0,n+1}}{a_{0,n}}$$

où $a_{0,n}$ et $a_{0,n+1}$ sont respectivement le coefficient du terme en $x^n$ du polynôme $W_n(x)$ et le coefficient du terme en $x^{n+1}$ du polynôme $W_{n+1}(x)$ ; ce rapport est calculé à partir de la relation de récurrence entre trois polynômes consécutifs en identifiant les coefficients des termes de degré $(n + 1)$. Par ailleurs, $I_n$ est la norme de $W_n(x)$ :

$$I_n = \int_a^b \omega(x) W_n^2(x) \, \mathrm{d}x.$$

En revanche, pour obtenir une expression analogue à (7.17) du chapitre 7, il sera nécessaire d'examiner chaque suite particulière de polynômes orthogonaux.

## 6. Expression de l'erreur en remplaçant I par J

L'expression (7.20) du chapitre 7 est quelque peu modifiée bien que le calcul demeure le même dans son principe. On obtient alors :

$$E = \frac{I_{n+1}}{(a_{n+1})^2} \sup \frac{\left| f^{(2n+1)}(x) \right|}{(2n+2)!}$$

où $I_{n+1}$ est la norme du polynôme $W_{n+1}(x)$ et $a_{n+1}$ le coefficient principal (coefficient de $x^{n+1}$).

## 7. Éléments de bibliographie

G. HACQUES (1971) *Mathématiques pour l'informatique*, Armand Colin.
F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Mc Graw-Hill.
A. LICHNEROWICZ (1960) *Algèbre et analyse linéaire*, Masson.
H. MINEUR (1966) *Techniques de calcul numérique*, Éditions Dunod.
H. STAHL et V. TOTIK (1992) *General orthogonal polynomials*, Cambridge University Press.

# C | Les fractions continues

Aujourd'hui, l'enseignement traditionnel effectué en France ne fait plus guère état de l'existence des fractions continues, pourtant ces êtres mathématiques ne sont pas tombés complètement en désuétude car ils demeurent un remarquable procédé de calcul des fonctions usuelles qui sont installées dans les calculateurs. De plus, l'étude des approximants de Padé (1863–1953) ainsi que les fondements théoriques de l'epsilon-algorithme sont indissociables des propriétés essentielles des fractions continues. C'est la raison pour laquelle nous avons pensé qu'il était utile d'en présenter les éléments.

## 1. Un exemple de fraction continue

L'exemple qui va nous servir d'introduction à la présentation des fractions continues n'est rien d'autre qu'un aspect du problème de l'interpolation. On se propose de déterminer une certaine fraction rationnelle :

$$y = \frac{P_n(x)}{Q_n(x)}, \tag{C.1}$$

où $P_n(x)$ et $Q_n(x)$ sont chacun des polynômes de degré $n$ de telle sorte que $y$ prenne la valeur $b_j$ quand $x$ prend la valeur $a_j$. On voit que $2n+1$ couples de valeurs sont nécessaires à la détermination de la solution ; on les numérote de 0 à $2n$. On peut écrire (C.1) de la façon suivante :

$$y = b_0 + \frac{x - a_0}{\frac{Q_n(x)}{P_{n-1}(x)}} \tag{C.2}$$

cette forme montre bien que $y = b_0$ quand $x = a_0$. Écrivons que l'expression (C.2) prend la valeur $b_j$ quand $x = a_j$ $(j \neq 0)$ :

$$\frac{Q_n(a_j)}{P_{n-1}(a_j)} = \frac{a_j - a_0}{b_j - b_0} \tag{C.3}$$

le second membre n'étant rien d'autre que l'inverse de la **différence première divisée** ; notons-la $\delta(a_0, a_j)$, $j$ prenant les valeurs $1, 2, \ldots 2n$. Rappelons que les **différences divisées** sont les différences premières, deuxièmes etc. (rencontrées lors de l'étude des polynômes d'interpolation) divisées chaque fois par un terme du type $(a_j - a_k)$ $j \neq k$.

Rien ne nous empêche de traiter l'équation (C.3) comme nous avons traité l'équation (C.1) en écrivant :

$$\frac{Q_n(x)}{P_{n-1}(x)} = \delta(a_0, a_1) + \frac{x - a_1}{\frac{P_{n-1}(x)}{Q_{n-1}(x)}} \cdot \tag{C.4}$$

Écrivons que l'expression (C.4) prend la valeur $b_j$ quand $x = a_j$ $(j \neq 0, 1)$ :

$$\frac{P_{n-1}(a_j)}{Q_{n-1}(a_j)} = \frac{a_j - a_0}{\delta(a_0, a_j) - \delta(a_0, a_1)} \tag{C.5}$$

le second membre n'étant rien d'autre que l'inverse de la différence deuxième divisée ; notons-la $\delta(a_0, a_1, a_j)$, $j$ prenant les valeurs $2, 3 \ldots 2n$.

En poursuivant la procédure jusqu'à $j = 2n$, et en notant $\delta(a_0, a_1, a_2, \ldots a_{2n})$ l'inverse de la différence $2n^{\text{e}}$ divisée, on aboutit à l'expression :

$$y = b_0 + \cfrac{x - a_0}{\delta(a_0, a_1) + \cfrac{x - a_1}{\delta(a_0, a_1, a_2) + \cfrac{x - a_2}{\delta(a_0, a_1, a_2, a_3) + \cdots + \cfrac{x - a_{2n-1}}{\delta(a_0, a_1, a_2, \ldots, a_{2n})}}}} \cdot \tag{C.6}$$

La fonction $y$ est donnée sous forme appelée **fraction continue**, et l'on verra un peu plus loin comment calculer numériquement une telle expression au moyen de relations de récurrence qui permettent d'obtenir la quantité définie par la relation (C.1).

## 2. Les fractions continues finies

Par définition, on appelle fraction continue $R_n$ d'ordre $n$ une expression telle que :

$$R_n = b_0 + \cfrac{a_0}{b_1 + \cfrac{a_1}{b_2 + \cfrac{a_2}{+b_3 + \cfrac{a_3}{+b_4 + \cdots \cfrac{a_{n-1}}{b_n}}}}} \cdot$$

Comme cette notation n'utilise pas une écriture linéaire, on préfère de façon très usuelle les formes conventionnelles suivantes :

$$R_n = b_0 + a_0/b_1 + a_1/b_2 + \cdots + a_{n-1}/b_n,$$

ou encore :

$$R_n = b_0 + \frac{a_0|}{|b_1} + \frac{a_1|}{|b_2} + \cdots + \frac{a_{n-1}|}{|b_n},$$

**$a_k$ s'appelle le numérateur partiel** tandis que **$b_k$ s'appelle le dénominateur partiel.** Si l'on développe une fraction continue finie, comme on l'a vu à propos de l'exemple précédent, elle prend la forme d'une fraction rationnelle que l'on note :

$$R_n = \frac{A_n}{B_n} \cdot$$

Toujours par définition, on appelle **réduite d'ordre $p$** (avec $q \leq n$) d'une fraction continue d'ordre $n$, la fraction continue tronquée à l'ordre $q$, notée :

$$R_q = \frac{A_q}{B_q} \cdot$$

## 2.1. Relations de récurrence entre les termes des réduites

En écrivant les réduites successives on trouve :

$$\begin{aligned}
\frac{A_0}{B_0} &= b_0, \\
\frac{A_1}{B_1} &= b_0 + \frac{a_1}{b_1} = \frac{b_0 b_1 + a_1}{b_1}, \\
\frac{A_2}{B_2} &= b_0 + \frac{a_1}{b_1 + a_2/b_2} = \frac{b_0 b_1 b_2 + a_1 b_2 + a_1 b_0}{b_1 b_2 + a_2},
\end{aligned}$$

soit encore :

$$\frac{A_2}{B_2} = \frac{b_2 A_1 + a_2 A_0}{b_2 B_1 + a_2 B_0} \cdot$$

Cette dernière expression se généralise de la manière suivante :

$$\frac{A_p}{B_p} = \frac{b_p A_{p-1} + a_p A_{p-2}}{b_p B_{p-1} + a_p B_{p-2}} \cdot \tag{C.7}$$

Signalons que cette façon d'écrire contient deux égalités : celle des numérateurs et celle des dénominateurs. Nous allons démontrer la relation (C.7) par récurrence en faisant la supposition qu'elle est vraie pour la valeur $p$ ; examinons alors la réduite d'ordre $(p+1)$ :

$$R_{p+1} = \frac{A_{p+1}}{B_{p+1}} \cdot$$

Pour réaliser les calculs, il suffit de remplacer $b_p$ par $b_p + a_{p+1}/b_{p+1}$ dans la réduite d'ordre $p$. Nous obtenons :

$$R_{p+1} = \frac{\left(b_p + \dfrac{a_{p+1}}{b_{p+1}}\right) A_{p-1} + a_p A_{p-2}}{\left(b_p + \dfrac{a_{p+1}}{b_{p+1}}\right) B_{p-1} + a_p B_{p-2}},$$

soit encore :

$$R_{p+1} = \frac{(b_p b_{p+1} + a_{p+1}) A_{p-1} + a_p A_{p-2} b_{p+1}}{(b_p b_{p+1} + a_{p+1}) B_{p-1} + a_p B_{p-2} b_{p+1}},$$

ce qui peut encore s'écrire :

$$R_{p+1} = \frac{b_{p+1}\,(b_p A_{p-1} + a_p A_{p-2}) + a_{p+1} A_{p-1}}{b_{p+1}\,(b_p B_{p-1} + a_p B_{p-2}) + a_{p+1} B_{p-1}},$$

$$\begin{aligned}
\text{et puisque} \quad & A_p = b_p A_{p-1} + a_p A_{p-2} \\
\text{et} \quad & B_p = b_p B_{p-1} + a_p B_{p-2}
\end{aligned}$$

nous obtenons :

$$R_{p+1} = \frac{A_{p+1}}{B_{p+1}} = \frac{b_{p+1}A_p + a_{p+1}A_{p-1}}{b_{p+1}B_p + a_{p+1}B_{p-1}}.$$

Ces deux dernières relations sont celles que l'on obtient en remplaçant $p$ par $(p+1)$ dans les équations (C.7). Comme les expressions (C.7) sont vraies pour $p = 2$, on en déduit par récurrence qu'elles sont générales. Par ailleurs, si l'on désire pouvoir également les utiliser pour $n = 1$, il suffit de choisir les conventions $A_{-1} = 1$ et $B_{-1} = 0$.

### 2.2. Une autre relation intéressante

Multiplions $A_{p+1}$ par $B_p$ et $B_{p+1}$ par $A_p$, puis retranchons membre à membre les relations de récurrence. Nous obtenons :

$$A_{p+1}B_p - B_{p+1}A_p = -a_{p+1}(A_pB_{p-1} - A_{p-1}B_p).$$

En posant $U_{p+1} = A_{p+1}B_p - B_{p+1}A_p$, on obtient la relation :

$$U_{p+1} = -a_{p+1}U_p.$$

Comme $U_1 = A_1B_0 - B_1A_0 = a_1$, on trouve en définitive :

$$U_{p+1} = A_{p+1}B_p - B_{p+1}A_p = (-1)^p \prod_{k=1}^{p+1} a_k, \tag{C.8}$$

ou encore, en divisant (C.2) par le produit $B_pB_{p+1}$, on établit :

$$\frac{U_{p+1}}{B_pB_{p+1}} = \frac{A_{p+1}}{B_{p+1}} - \frac{A_p}{B_p} = (-1)^p \frac{1}{B_pB_{p+1}} \prod_{k=1}^{p+1} a_k. \tag{C.9}$$

### 2.3. Propriétés essentielles

On admettra sans démonstration les deux propositions suivantes :

1. les polynômes $A_p$ et $B_p$ des variables $a_0, a_1, a_2, \ldots, a_p, b_0, b_1, b_2, \ldots, b_p$ sont irréductibles entre eux, ou encore premiers entre eux, si les variables $a_k$ et les variables $b_k$ sont indépendantes ;
2. aucune réduite ne se présente sous la forme $0/0$.

## 3. Les fractions continues infinies

Dans ce cas la fraction continue devient illimitée et la notion de convergence s'introduit tout naturellement lors de cette étude.

Considérons la réduite $R_n$ d'ordre $n$. Si $R_n$ tend vers une limite $R$ lorsque $n$ croît indéfiniment, on dira que la fraction continue est convergente et a pour limite $R$. Dans le cas contraire, nous dirons qu'elle est divergente.

***Remarque*** : Supposons que $a_k = 0$. Les relations de récurrence (C.8) et (C.9) nous donnent :

$$A_kB_{k-1} - A_{k-1}B_n = 0,$$

$$\text{soit :} \quad \frac{A_k}{B_k} = \frac{A_{k-1}}{B_{k-1}} = 0$$

$$\text{donc :} \quad R_{k-1} = R_k = 0,$$

et la fraction continue est alors finie. On conclut que, si une fraction continue est infinie, tous ses $a_k$ sont différents de zéro.

### Calcul des $a_k$ et des $b_k$ à partir des réduites

Du point de vue numérique, il va de soi que le calcul d'une fraction continue se ramène toujours à une approximation réalisée au moyen d'une réduite. Cependant, il est indispensable d'examiner sous l'angle de la légitimité les opérations susceptibles d'être réalisées. Il est possible de calculer les coefficients $a_k$ et $b_k$ au moyen des réduites $R_q$, $R_{q-1}$ et $R_{q-2}$. En effet, à partir des relations de récurrence établies au paragraphe 2, nous obtenons un système linéaire dont la solution s'écrit :

$$a_q = \frac{A_q B_{q-1} - A_{q-1} B_q}{A_{q-1} B_{q-2} - A_{q-2} B_{q-1}} \quad \text{et} \quad b_q = \frac{A_q B_{q-2} - A_{q-2} B_q}{A_{q-1} B_{q-2} - A_{q-2} B_{q-1}} .$$

On remarquera que les relations (C.7) sont définies à une constante multiplicative près ; on obtient alors des fractions continues dites équivalentes, tout simplement parce que toutes les réduites sont égales. Il s'ensuit que les réduites dépendent de paramètres arbitraires non nuls. Choisissons comme contrainte linéaire :

$$B_q = c_q B_{q-1}.$$

Nous pouvons écrire :

$$a_q = c_q c_{q-1} \frac{R_{q-1} - R_q}{R_{q-1} - R_{q-2}} \quad \text{et} \quad b_q = c_q \frac{R_q - R_{q-2}}{R_{q-1} - R_{q-2}} .$$

Il est souvent commode de choisir arbitrairement les $c_q$ égaux à 1, on obtient alors des expressions très simples :

$$a_q = \frac{R_{q-1} - R_q}{R_{q-1} - R_{q-2}} \quad \text{et} \quad b_q = \frac{R_q - R_{q-2}}{R_{q-1} - R_{q-2}} .$$

## 4. Développement en fraction continue à partir d'un développement en série entière

On considère le développement en série suivant :

$$S = \sum_{i=0}^{\infty} u_i,$$

et l'on souhaite que les sommes partielles $S_p$ soient égales aux réduites $R_p$ de la fraction continue $R$ :

$$R = b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} + \cdots$$

Compte tenu des résultats obtenus au paragraphe précédent, on peut écrire :

$$a_q = \frac{S_{q-1} - S_q}{S_{q-1} - S_{q-2}} \quad \text{et} \quad b_q = \frac{S_q - S_{q-2}}{S_{q-1} - S_{q-2}} .$$

Si l'on tient compte du fait que $S_q - S_{q-1} = u_q$, nous pouvons écrire :

$$a_q = -\frac{u_q}{u_{q-1}} \quad \text{et} \quad b_q = \frac{u_q + u_{q-1}}{u_{q-1}} = 1 + \frac{u_q}{u_{q-1}}$$

ce qui permet d'obtenir le développement suivant :

$$R = u_0 + \frac{u_1|}{|1} - \frac{u_2/u_1|}{|1 + u_2/u_1} - \cdots - \frac{u_p/u_{p-1}|}{|1 + u_p/u_{p-1}} - \cdots$$

***Remarque*** : Si nous connaissons les réduites d'une fraction continue, réciproquement, il lui correspond une série dont le terme général $u_q$ s'écrit :

$$u_k = R_k - R_{k-1}.$$

## 5. Développement en fractions continues de séries usuelles

Considérons le développement général d'une fonction en série de MacLaurin :

$$S = \sum_{k=0}^{\infty} \alpha_k x^k.$$

Nous obtenons, par application directe du résultat établi au paragraphe 4, le développement en fraction continue suivant :

$$R = \alpha_0 + \frac{\alpha_1 x|}{|1} - \frac{\alpha_2/\alpha_1 x|}{|1 + \alpha_2/\alpha_1 x} - \cdots - \frac{\alpha_p/\alpha_{p-1} x|}{|1 + \alpha_p/\alpha_{p-1} x} - \cdots$$

D'un point de vue pratique, c'est-à-dire dès que les applications sont en vue, on a tout intérêt à procéder à la transformation de la suite des $\alpha_p$ de la manière suivante : on forme la suite des $\beta_p$ obtenue en posant :

$$\begin{aligned} \beta_0 &= \alpha_0 \\ \beta_1 &= \alpha_1 \\ \beta_q &= \alpha_q/\alpha_{q-1} \quad \text{avec } q \geq 2, \end{aligned}$$

ceci a pour but une économie évidente de calcul. Par application de ce procédé, on obtient les développements suivants :

$$\log_e(1-x) = \frac{x|}{|1} + \frac{1^2 x|}{|2-x} + \frac{2^2 x|}{|3-2x} + \cdots + \frac{(p-1)^2 x|}{|p-(p-1)x} + \cdots \quad \text{avec } |x| < 1.$$

$$(1+x)^m = 1 + \frac{mx|}{|1} - \frac{x(m-1)/2|}{|1+x(m-1)/2} - \cdots - \frac{x(m-p-1)/p|}{|1+x(m-p-1)/p} - \cdots \quad \text{avec } |x| < 1.$$

$$\arctan(x) = \frac{x|}{|1} + \frac{1^2 x^2|}{|3-x^2} + \frac{3^2 x^2|}{|5-3x^2} + \cdots + \frac{(2p-1)^2 x^2|}{|(2p+1)-(2p-1)x^2} + \cdots \quad \text{avec } |x| \leq 1.$$

$$\exp(x) = 1 + \frac{x|}{|1} - \frac{x|}{|2+x} - \frac{2x|}{|3+x} + \cdots + \frac{(p-1)x|}{|p+x} + \cdots$$

On notera que la convergence de la fraction continue est assurée dans le même domaine que celui qui assure la convergence de la série entière génératrice.

## 6. Développement en fraction continue à partir d'un produit infini

Bien que les produits infinis ne jouent pas un rôle aussi important que celui tenu par les développements en série entière, il est tout de même intéressant de voir comment de tels produits permettent de générer un développement en fraction continue. Précisons que nous limitons notre étude aux produits infinis de monômes que nous écrirons donc sous la forme :

$$P = \prod_{i=1}^{\infty}(1 + v_i).$$

Pour étudier la convergence de $P$, il suffit de passer aux logarithmes, ce qui permet de se ramener à l'étude d'une suite classique dont le terme général se met sous la forme :

$$w_n = \log_e(1 + u_n).$$

Comme les suites $|u_n|$ et $|w_n|$ convergent ou divergent simultanément, il suffit de s'assurer de la convergence de $|u_n|$ pour obtenir la convergence absolue du produit, lequel, alors, admet une limite non nulle. Donc, en définitive, on souhaite former la fraction continue $R$ dont les réduites $R_q$ sont égales aux produits partiels $P_q$. On entend par produit partiel d'ordre $q$ le produit fini des $q$ premiers termes du produit infini. Puisque :

$$R_p = P_p = \prod_{i=1}^{p}(1 + v_i),$$

nous pouvons écrire :

$$a_q = \frac{P_{q-1}P_q}{P_{q-1} - P_{q-2}} = -\frac{P_{q-1}v_q}{v_{q-1} - P_{q-2}} = \frac{v_q}{v_{q-1}}(1 + v_{q-1}),$$

$$\text{et} \quad b_q = \frac{P_q - P_{q-2}}{P_{q-1} - P_{q-2}} = \frac{-1 + (1 + v_q)(1 + v_{q-1})}{v_{q-1}} = 1 + \frac{v_q(1 + v_{q-1})}{v_{q-1}},$$

ce qui permet de conclure que :

$$b_q = 1 - a_q.$$

En adoptant comme notation $P_0 = 1$, cela nous permet d'écrire $b_0 = 1$, et puisque $P_1 = 1 + v_1 = R_1 = b_0 + a_1/b_1$, cela entraîne que :

$$a_1 = v_1 \qquad \text{et} \qquad b_1 = 1,$$

ainsi nous obtenons la fraction continue

$$R = 1 + \frac{a_1|}{|1} + \frac{a_2|}{|1 - a_2} + \frac{a_3|}{|1 - a_3} + \cdots + \frac{a_q|}{|1 - a_q} + \cdots$$

### Application à quelques exemples

$$x^{-1}\sin(x) = \prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right) \qquad a_q = -\frac{(q-1)^2\pi^2 - x^2}{q^2\pi^2}$$

$$x^{-1}\sinh(x) = \prod_{k=1}^{\infty}\left(1+\frac{x^2}{k^2\pi^2}\right) \qquad a_q = -\frac{(q-1)^2\pi^2 + x^2}{q^2\pi^2}$$

$$\cos(x) = \prod_{k=1}^{\infty}\left(1-\frac{4x^2}{(2k-1)^2\pi^2}\right) \qquad a_q = -\frac{(2q-3)^2\pi^2 - 4x^2}{(2q-1)^2\pi^2}$$

$$\cosh(x) = \prod_{k=1}^{\infty}\left(1+\frac{4x^2}{(2k-1)^2\pi^2}\right) \qquad a_q = -\frac{(2q-3)^2\pi^2 + 4x^2}{(2q-1)^2\pi^2}$$

## 7. Éléments de bibliographie


C. BREZINSKI (1978) *Algorithmes d'accélération de la convergence*, Éditions Technip, Paris.

É. DURAND (1971) *Solutions numériques des équations algébriques*, Tome I pp. 20 et 91, Éditions Masson, Paris.

I.S. GRADSHTEYN et I.M. RYSHIK (1980) *Table of Integrals, Series and Products*, Academic Press.

F. HILDEBRAND (1956) *Introduction to the numerical analysis*, Édition Mc Graw-Hill.

C. JORDAN (1959) *Cours d'analyse de l'École Polytechnique*, Tome I, Gauthier-Villars, Paris.

A-M. LEGENDRE (1955) *Théorie des nombres*, A. Blanchard, Paris.

P. MONTEL (1957) *Leçons sur les récurrences et leurs applications*, ch. VI, VII, VIII et IX, Gauthier-Villars, Paris.

H. PADÉ (1893) *Sur la représentation approchée d'une fonction par des fractions rationnelles*, Tome IX, p. 1, 93, Annales Scientifiques de l'École Normale Supérieure.

(1899) *Mémoire sur les développements en fractions continues de la fonction exponentielle pouvant servir d'introduction à la théorie des fractions continues algébriques*, Tome XVI, p. 395, 426 Annales Scientifiques de l'École Normale Supérieure.

J. TRIGNAN (1994) *Introduction aux problèmes d'approximation : fractions continues, différences finies*, Édition du Choix.

G. VALIRON (1955) *Théorie des fonctions*, Masson, Paris.

# D | Les approximants de Padé et de Maehly

Le cadre de ce chapitre a pour toile de fond l'approximation des fonctions, qu'elles soient usuelles ou spéciales, que l'on retrouve très fréquemment en calcul. En filigrane, se pose le problème de connaître les moyens les plus économiques qui permettent de calculer les « fonctions de bibliothèque ».

C'est à H. Padé (1863–1953) que revient le mérite de s'être interrogé sur ces problèmes et d'avoir recherché la méthode la plus efficace possible pour réaliser des calculs numériques de fonctions (voir la bibliographie). La méthode a été ensuite reprise par Maehly comme on le verra. Aujourd'hui encore bien des théoriciens continuent de poursuivre de tels travaux qui demeurent un vaste champ de recherche. Rappelons toutefois que les approximants de Padé sont intimement liés aux fractions continues de Lagrange (1736–1813) et que les algorithmes d'accélération de la convergence sont indissociables de ces études (*cf.* C. Brézinski p. 72 et suivantes). Du reste, les publications originales de Padé en 1892 et 1899 portent respectivement pour titre : « Sur la représentation approchée d'une fonction par des fractions rationnelles » et « Sur les développements en fractions continues de la fonction exponentielle ».

## 1. Le théorème fondamental de Padé

Par définition, un approximant est une fraction continue (*cf.* annexe C) qui a été tronquée à un certain ordre. Il en résulte qu'un approximant se présente toujours sous la forme d'une fraction rationnelle au sens le plus large.

Soit $f(x)$ une fonction développable en série entière au voisinage de $x = 0$ qui s'écrit :

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots \qquad \text{avec } a_0 \neq 0.$$

Considérons le polynôme de degré $q$ :

$$P_q(x) = l_0 + l_1 x + l_2 x^2 + \cdots + l_q x^q = \sum_{i=0}^{q} l_i x^i,$$

et formons le produit :

$$f(x) \cdot P_q(x) = \left( \sum_{i=0}^{\infty} a_i x^i \right) \cdot \left( \sum_{k=0}^{q} l_k x^k \right) = \sum_{k=0}^{\infty} (a_k l_0 + a_{k-1} l_1 + a_{k-2} l_2 + \cdots + a_{k-q} l_q) x^k,$$

expression dans laquelle on conviendra que les coefficients affectés d'un indice négatif sont nuls. Égalons à zéro les coefficients de degré $p+1, p+2, \ldots, p+q$. On obtient un système linéaire constitué de $q$ équations à $q+1$ inconnues que l'on écrit :

$$\begin{aligned} a_{p+1}l_0 + a_p l_1 + a_{p-1}l_2 + \cdots + a_{p-q+1}l_q &= 0 \\ a_{p+2}l_0 + a_{p+1}l_1 + a_p l_2 + \cdots + a_{p-q+2}l_q &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{p+q}l_0 + a_{p+q-1}l_1 + a_{p+q-2}l_2 + \cdots + a_p l_q &= 0 \end{aligned}$$

que l'on peut encore écrire sous forme matricielle :

$$\begin{vmatrix} a_{p+1} & a_p & a_{p-1} & \cdots & a_{p-q+1} \\ a_{p+2} & a_{p+1} & a_p & \cdots & a_{p-q+2} \\ \cdots & \cdots & \cdots & & \\ a_{p+q} & a_{p+q-1} & a_{p+q-2} & \cdots & a_p \end{vmatrix} \cdot \begin{vmatrix} l_0 \\ l_1 \\ \\ l_q \end{vmatrix} = 0 \qquad \text{(D.1)}$$

Comme le nombre d'inconnues est supérieur d'une unité au nombre d'équations, il est toujours possible de satisfaire l'ensemble de ces équations par des valeurs des inconnues telles qu'elles ne soient pas toutes nulles ensemble. Si l'un des déterminants de la matrice (D.1), obtenu par suppression d'une colonne, est différent de zéro, alors toutes les inconnues sont déterminées à un facteur multiplicatif près. En revanche, si tous les déterminants sont nuls, alors plusieurs valeurs peuvent être choisies arbitrairement. De préférence, on s'intéresse à la suite $l_0, l_1, l_2, \ldots, l_q$ qui commence par le plus grand nombre de zéros possible. Le polynôme $P_q(x)$ est alors divisible par une puissance de $x$ dont l'exposant est noté $\omega_{pq}$, exposant au moins égal à celui que l'on obtiendrait en adoptant toute autre solution. Cette valeur de l'exposant est nulle lorsque $l_0$ est différent de zéro. On note par $R_p(x)$ le polynôme dont le degré est au plus égal à $p$, soit :

$$R_p(x) = \sum_{i=0}^{p} (a_i l_0 + a_{i-1}l_1 + a_{i-2}l_2 + \cdots + a_{i-q}l_q)x^i = \sum_{k=0}^{p} b_k x^k. \qquad \text{(D.2)}$$

Alors on peut écrire :

$$f(x) \cdot P_q(x) = R_p(x) + \varepsilon(x^{p+q+1}),$$

expression dans laquelle $\varepsilon(x^{p+q+1})$ est un infiniment petit dont l'ordre est au moins égal à $p+q+1$. Il s'ensuit que l'on peut écrire :

$$f(x) \approx \frac{R_p(x)}{P_q(x)},$$

$$\text{ou encore :} \quad f(x) \approx \frac{R_p^*(x)}{P_p^*(x)},$$

dans le cas où les polynômes $R_p(x)$ et $P_q(x)$ ne sont pas premiers entre eux ($R_p^*(x)$ et $P_p^*(x)$ sont premiers entre eux). Quoi qu'il en soit, nous avons établi que la fonction $f(x)$ pouvait être représentée ou approchée par une fraction rationnelle au voisinage de zéro. Notons que sous la dernière forme on aurait :

$$P_q^*(x) f(x) = R_p^*(x) + \varepsilon(x^{p+q+1}).$$

***Remarque*** : Comme dans l'expression de $P_q(x)$ ou de $P_q^*(x)$, il y a nécessairement un terme constant différent de zéro, alors $R_p(x)/P_q(x)$ diffère de $f(x)$ d'un infiniment petit dont l'ordre est $(p+q+1-\omega_{pq})$ qui est plus grand que la somme des degrés des polynômes de la fraction.

### Théorème de Padé

Parmi toutes les fractions rationnelles irréductibles dont les termes ont des degrés égaux au plus à $p$ pour le numérateur et au plus à $q$ pour le dénominateur, $p$ et $q$ étant des entiers positifs ou nuls égaux ou inégaux, il y a une fraction $R_p(x)/P_q(x)$ qui fournit une approximation dont l'ordre est supérieur à celui de l'approximation fournie par une quelconque des autres fractions.

Nous admettrons ce théorème sans démonstration (*cf.* les mémoires de Padé).

***Remarque 1*** : Dans l'état actuel des choses, à notre connaissance, on ne sait pas dire quelle est cette fraction rationnelle. Cependant, l'expérience montre que ce sont les termes de la diagonale pour laquelle $p$ et $q$ sont égaux ou les termes de la première parallèle supérieure à la diagonale (première sur-diagonale) telle que $p = q + 1$, qui fournissent les meilleures approximations.

***Remarque 2*** : On peut montrer que le nombre $\omega_{pq}$ est au plus égal à $q$, mais qu'il est aussi au plus égal à $p$.

***Théorème*** – Les polynômes $R_p(x)$ et $P_q(x)$ ont l'un et l'autre un terme constant différent de zéro ; $\omega_{pq}$ désignant zéro ou un entier positif au plus égal au plus petit des deux nombres $p$ et $q$, les degrés des termes $R_p(x)$ et $P_p(x)$ ont respectivement pour limite supérieure $p - \omega_{pq}$, $q - \omega_{pq}$ ; l'ordre de l'approximation a pour limite inférieure $p + q + 1 - \omega_{pq}$.

Ici encore, nous ne démontrerons pas ce théorème (*cf.* les mémoires de Padé).

## 2. Sur le calcul effectif des coefficients

Au départ, on connaît les coefficients du développement en série entière que l'on désigne par $a_i$. Ensuite on calcule les $l_k$ en faisant dans le système (D.1) $l_0 = 1$, ce qui permet de calculer le second membre du système linéaire. Comme les $l_k$ sont définis à une constante multiplicative près, on peut rechercher à exprimer les $l_k$ sous forme entière, mais cela demeure un problème secondaire. Pour terminer, on calcule les $b_k$ à partir de l'expression (D.2) qui se développe selon les égalités suivantes :

$$\begin{aligned} b_0 &= a_0 l_0 \\ b_1 &= a_1 l_0 + a_0 l_1 \\ b_2 &= a_2 l_0 + a_1 l_1 + a_0 l_2 \\ b_3 &= a_3 l_0 + a_2 l_1 + a_1 l_2 + a_0 l_3 \\ b_4 &= a_4 l_0 + a_3 l_1 + a_2 l_2 + a_1 l_3 + a_0 l_4 \end{aligned}$$

et ainsi de suite.

## 3. Estimation de l'erreur commise

L'erreur strictement mathématique $E[f(x)]$ commise en remplaçant $f(x)$ par une fraction rationnelle s'écrit :

$$E[f(x)] = f(x) - \frac{R_p(x)}{P_q(x)} \cdot$$

Pour évaluer cette erreur il est commode de revenir aux équations de départ. On peut alors écrire :

$$f(x) \cdot P_q(x) = R_q(x) + S_{p+q}(x)$$

où $S_{p+q}(x) = \varepsilon(x^{p+q+1})$. Il est pratique alors d'écrire $S_{p+q}(x)$ sous une forme plus exploitable, soit :

$$S_{p+q}(x) = x^{p+q+1} \sum_{k=0}^{\infty} c_k x^k.$$

Les coefficients $c_k$ décroissent en général extrêmement rapidement si bien que le premier terme $c_0$ est de loin prépondérant. Aussi cette remarque permet-elle d'estimer l'erreur qui est donc de l'ordre de :

$$E[f(x)] = c_0 \frac{x^{p+q+1}}{P_q(x)} \cdot$$

Reste à calculer $c_0$. Pour cela, il suffit de reprendre le développement de $f(x) \cdot P_q(x)$ et d'identifier les coefficients des termes en $x^{p+q+1}$. On obtient alors :

$$c_0 = a_{p+q+1} l_0 + a_{p+q} l_1 + \cdots + a_{p+1} l_q = \sum_{k=0}^{q} l_k a_{p+q+1-k}.$$

# 4. Développements de quelques fonctions en approximants de Padé

## 4.1. Développements de exp(x) (exemple donné par Padé)

On part du développement en série entière :

$$\exp(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots = \sum_{k=0}^{\infty} a_k x^k.$$

**Nota -** Pour des raisons de commodité d'écriture, on notera l'inverse d'un nombre au moyen du nombre souligné d'une barre, ainsi $\underline{x} = 1/x$.

***a - Approximation par la fraction rationnelle* p = 1, q = 3 -** En choisissant $l_0 = 1$, on obtient le système linéaire suivant :

$$\begin{vmatrix} 1 & 1 & 0 & -\underline{2} \\ \underline{2} & 1 & 1 & -\underline{6} \\ \underline{6} & \underline{2} & 1 & -\underline{24} \end{vmatrix}$$

La résolution donne :

$$\begin{aligned} l_1 &= -0{,}75 \\ l_2 &= \phantom{-}0{,}25 \\ l_3 &= -4{,}166\,666\,10^{-2} \end{aligned}$$

ou encore en multipliant par un facteur convenable :

$$\begin{aligned} l_0 &= \phantom{-}24 \\ l_1 &= -18 \\ l_2 &= \phantom{-}6 \\ l_3 &= -1. \end{aligned}$$

À partir de l'équation (D.2), on tire les valeurs des $b$ :

$$\begin{aligned} b_0 &= 24 \\ b_1 &= 6. \end{aligned}$$

De là, on exprime l'approximant :

$$\exp(x) = \frac{24 + 6x}{24 - 18x + 6x^2 - x^3}\,.$$

***b* – *Approximation par la fraction rationnelle* p = 2, q = 3 –** On résout le système linéaire :

$$\begin{vmatrix} \underline{2} & 1 & 1 & \underline{-6} \\ \underline{6} & \underline{2} & 1 & \underline{-24} \\ \underline{24} & 6 & \underline{2} & \underline{-120}. \end{vmatrix}$$

D'où l'on tire, après multiplication par 60 :

$$\begin{aligned} l_0 &= \phantom{-}60 \\ l_1 &= -36 \\ l_2 &= \phantom{-}9 \\ l_3 &= -1. \end{aligned}$$

En reportant ces résultats dans les équations (D.2), on accède aux $b_j$ :

$$\begin{aligned} b_0 &= 60 \\ b_1 &= 24 \\ b_2 &= 3. \end{aligned}$$

En définitive, on obtient l'approximant :

$$\exp(x) = \frac{60 + 24x + 3x^2}{60 - 36x + 9x^2 - x^3}\,.$$

***c* – *Approximation par la fraction rationnelle* p = q = 3 –** Les équations (D.2) donnent le système linéaire :

$$\begin{vmatrix} \underline{6} & \underline{2} & 1 & \underline{-24} \\ \underline{24} & \underline{6} & \underline{2} & \underline{-120} \\ \underline{120} & \underline{24} & \underline{6} & \underline{-720} \end{vmatrix}.$$

Après multiplication par 120, nous trouvons :

$$\begin{aligned} l_0 &= \phantom{-}120 \\ l_1 &= -60 \\ l_2 &= \phantom{-}12 \\ l_3 &= -1 \end{aligned}$$

puis :

$$b_0 = 120$$
$$b_1 = 60$$
$$b_2 = 12$$
$$b_3 = 1$$

d'où l'approximant recherché :

$$\exp(x) = \frac{120 + 60x + 12x^2 + x^3}{120 - 60x + 12x^2 - x^3}.$$

***d – Approximation par la fraction rationnelle*** $\mathbf{p = 4, q = 3}$ – Tous calculs faits, on trouve :

$$\exp(x) = \frac{840 + 480x + 120x^2 + 16x^3 + x^4}{840 - 360x + 60x^2 - 4x^3}.$$

Dans le cas présent, nous avons multiplié les coefficients $l_k$ par 4 pour obtenir les $b_j$ sous forme entière.

Avec une calculette, il est intéressant d'expérimenter ces différents approximants et d'en apprécier la précision.

## 4.2. Développement de la fonction de Bessel d'ordre zéro $J_0(x)$

$J_0(x)$ s'exprime à l'aide d'un développement en série entière donné par l'expression :

$$J_0(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(k!)^2} \left(\frac{x}{2}\right)^{2k}$$

soit encore :

$$J_0\left(\frac{z^2}{4}\right) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{k!} z^k.$$

***a – Approximation par la fraction rationnelle*** $\mathbf{p = 3, q = 4}$ – Il faut résoudre le système linéaire suivant :

$$\begin{vmatrix} -\underline{6}^2 & \underline{4} & -1 & 1 & -\underline{24} \\ \underline{24}^2 & -\underline{6}^2 & \underline{4} & -1 & \underline{120}^2 \\ -\underline{120}^2 & \underline{24}^2 & -\underline{6}^2 & \underline{4} & -\underline{720}^2 \\ \underline{720}^2 & -\underline{120}^2 & \underline{24}^2 & -\underline{6}^2 & \underline{5\,040}^2 \end{vmatrix}$$

$$l_0 = 1$$
$$l_1 = 0{,}113\,937\,367\,4$$
$$l_2 = 0{,}671\,266\,793 \times 10^{-2}$$
$$l_3 = 0{,}255\,002\,275\,8 \times 10^{-3}$$
$$l_4 = 0{,}565\,105\,603\,8 \times 10^{-5}.$$

On en déduit les $b_k$ :

$$\begin{aligned} b_0 &= 1 \\ b_1 &= -0{,}886\,062\,632\,6 \\ b_2 &= 0{,}142\,775\,300\,5 \\ b_3 &= 0{,}600\,355\,363 \times 10^{-2}. \end{aligned}$$

***b* – *Approximation par la fraction rationnelle* p = q = 4** – On donne les résultats du calcul :

$$\begin{aligned} & l_0 = 1 \\ & l_1 = 0{,}113\,937\,367\,4 \\ & l_2 = 3{,}315\,091\,014 \times 10^{-2} \\ & l_3 = 8{,}160\,869\,521 \times 10^{-3} \\ & l_4 = 1{,}084\,070\,461 \times 10^{-5} \\ \text{puis,} \quad & b_0 = 1 \\ & b_1 = -0{,}918\,085\,771\,3 \\ & b_2 = 0{,}171\,400\,862\,3 \\ & b_3 = -0{,}105\,327\,029\,2 \times 10^{-1} \\ & b_4 = 0{,}208\,963\,998\,1 \times 10^{-2}. \end{aligned}$$

***Remarque sur l'utilisation des approximants*** – Lorsque le degré des polynômes dépasse deux, on a alors intérêt, pour réaliser les calculs, à utiliser le schéma de Horner (1786–1837) qui fait gagner un nombre intéressant d'opérations. Soit dans le cas présent :

$$J_0\left(\frac{x^2}{4}\right) = \frac{\{[(b_4x + b_3)x + b_2]x + b_1\}x + b_0}{\{[(l_4x + l_3)x + l_2]x + l_1\}x + l_0}.$$

Pour se faire une idée plus tangible des approximants, on se propose de tabuler quelques valeurs de $J_0(x)$.

Cette tabulation amène plusieurs remarques. D'une part, les approximants de Padé étant extraits de développements en série de MacLaurin (ou de séries entières), donc valables au voisinage de zéro, sont également eux-mêmes valables au voisinage de zéro. Le tableau D.1, page suivante, montre à l'évidence que la précision de l'approximation se dégrade avec l'éloignement de l'origine. Cependant il est intéressant de noter que pour la valeur $x = 6$ l'approximation obtenue avec la fraction rationnelle $p = q = 4$ donne une précision de l'ordre de cinq pour mille ce qui reste tout à fait raisonnable.

## 4.3. Développement de $x^{-1}\arctan(x)$ (exemple donné par Padé)

On part du développement suivant :

$$x^{-1}\arctan(x) = 1 - \frac{x^2}{3} + \frac{x^4}{3} - \frac{x^6}{3} + \frac{x^8}{3} + \cdots$$

***Approximation par la fraction rationnelle* p = 3, q = 2** – On résout le système linéaire :

$$\begin{vmatrix} \underline{5} & \underline{-3} & \underline{7} \\ -7 & \underline{5} & \underline{-9} \end{vmatrix}$$

**Tableau D.1.** *Résultats numériques obtenus.*

| $x$ | $J_0(x)$ tables | $J_0(x)$ Padé $p=4 \quad q=4$ | $J_0(x)$ Padé $p=3 \quad q=4$ |
|---|---|---|---|
| 0,1 | 0,999 75 | 0,997 501 | 0,997 501 |
| 0,5 | 0,938 5 | 0,938 469 | 0,938 467 |
| 1,0 | 0,765 2 | 0,765 197 | 0,765 009 |
| 2,0 | 0,223 9 | 0,223 890 | 0,222 080 |
| 2,5 | −0,048 4 | −0,048 383 | −0,047 9 |
| 5,5 | −0,006 8 | −0,006 697 | −0,022 6 |
| 6,0 | 0,150 6 | 0,151 214 | |
| 7,0 | 0,300 1 | 0,305 75 | |
| 8,0 | 0,171 7 | 0,208 | |
| 9,0 | −0,090 3 | 23,38 | |
| 10,0 | −0,245 9 | 0,327 | |

ce qui donne :

$$\begin{aligned} & & l_0 &= 1 \\ & & l_1 &= 1{,}111\,111\,111 \\ & & l_2 &= 2{,}380\,952\,355 \\ \text{d'où l'on déduit :} & & b_0 &= 1 \\ & & b_1 &= 0{,}777\,777\,777 \\ & & b_2 &= 2{,}210\,581\,985 \\ & & b_3 &= -0{,}714\,285\,705\,7 \end{aligned}$$

## 4.4. Calcul de sinus et cosinus

On se propose de calculer la fraction rationnelle adaptée à chacune des fonctions sinus et cosinus en se fixant pour $x$ l'intervalle $(0, \pi/4)$ car ce sont ces développements que nous allons utiliser pour fabriquer les fonctions de bibliothèque ultérieurement. Nous partons donc des développements en série de MacLaurin :

$$\sin(x) = x\left(1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \cdots + (-1)^n \frac{x^n}{(2n+1)!} + \cdots\right)$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n \frac{x^n}{2n!} + \cdots$$

On se propose de déterminer les fractions rationnelles $p = q$ de telle sorte que l'erreur sur chacune des valeurs calculées (appartenant à l'intervalle précité) soit inférieure à $10^{-8}$. Il s'agit avant tout d'une erreur purement mathématique liée à la troncature des séries. Cela du reste ne préjuge en rien des autres erreurs liées à l'exécution des calculs qui peuvent devenir prépondérantes notamment lorsque les calculs font apparaître des différences. Comme nous

l'avons vu précédemment, il faut d'abord connaître les approximants avant de procéder aux évaluations des erreurs. Nous allons donc calculer les approximants pour $p = q = 2$ puis $p = q = 3$ (*cf.* Tab. D.2), cela nous permettra d'effectuer un choix par la suite.

**Tableau D.2.**

| sinus | cosinus |
|---|---|
| **Cas $p = q = 2$** | |
| $l_0 = 1$ | $l_0 = 1$ |
| $l_1 = 3{,}282\,828\,283 \times 10^{-2}$ | $l_1 = 4{,}365\,079\,365 \times 10^{-2}$ |
| $l_2 = 4{,}509\,379\,51 \times 10^{-4}$ | $l_2 = 8{,}597\,883\,6 \times 10^{-4}$ |
| $b_0 = 1$ | $b_0 = 1$ |
| $b_1 = -1{,}338\,383\,838 \times 10^{-1}$ | $b_1 = -4{,}563\,492\,064 \times 10^{-1}$ |
| $b_2 = 3{,}312\,890\,813 \times 10^{-3}$ | $b_2 = 2{,}070\,105\,82 \times 10^{-2}$ |
| **Cas $p = q = 3$** | |
| $l_0 = 1$ | $l_0 = 1$ |
| $l_1 = 2{,}414\,321\,216 \times 10^{-2}$ | $l_1 = 2{,}940\,421\,157 \times 10^{-2}$ |
| $l_2 = 2{,}760\,512\,714 \times 10^{-4}$ | $l_2 = 4{,}237\,288\,124 \times 10^{-4}$ |
| $l_3 = 1{,}595\,035\,872 \times 10^{-6}$ | $l_3 = 3{,}235\,543\,474 \times 10^{-6}$ |
| $b_0 = 1$ | $b_0 = 1$ |
| $b_1 = -1{,}425\,234\,545 \times 10^{-1}$ | $b_1 = -4{,}705\,957\,884 \times 10^{-1}$ |
| $b_2 = 4{,}585\,515\,911 \times 10^{-3}$ | $b_2 = 2{,}738\,828\,969 \times 10^{-2}$ |
| $b_3 = -4{,}163\,277\,311 \times 10^{-5}$ | $b_3 = -3{,}723\,422\,695 \times 10^{-4}$ |

L'erreur sera maximum en valeur absolue pour $x = \pi/4$. Donc, nous obtenons les résultats présentés dans le tableau D.3.

**Tableau D.3.**

| sinus | cosinus |
|---|---|
| **Cas $p = q = 2$** | |
| $c = 2{,}4 \times 10^{-8}$ | $c = 4{,}2 \times 10^{-7}$ |
| $E = 2{,}1 \times 10^{-9}$ | $E = 3{,}6 \times 10^{-8}$ |
| **Cas $p = q = 3$** | |
| $c = 6 \times 10^{-13}$ | $c = 1{,}34 \times 10^{-11}$ |
| $E = 2 \times 10^{-14}$ | $E = 4{,}5 \times 10^{-13}$ |

En conclusion, pour obtenir huit chiffres significatifs exacts, on pourra prendre la fraction rationnelle $p = q = 2$ pour le sinus et $p = q = 3$ pour le cosinus.

## 5. Généralisation des approximants de Padé, méthode de Maehly

Comme nous avons eu l'occasion de nous en rendre compte les approximants de Padé donnent une excellente précision au voisinage de zéro ; rien d'étonnant à cela puisqu'ils ont été construits pour remplir cet office. Malheureusement quand on s'éloigne de zéro, cette précision décroît notablement. Si l'on souhaite pouvoir obtenir une bonne précision sur tout l'intervalle fini $(a, b)$ que l'on réduira grâce à une transformation linéaire à l'intervalle canonique $(-1, +1)$, il nous faudra développer non plus les fonctions en série de MacLaurin mais, bien sûr, en une série de polynômes de Tchebycheff (1821–1894). En effet, on sait que, parmi tous les polynômes de degré $n$ à coefficient principal réduit, c'est le polynôme de Tchebycheff de degré $n$ qui s'écarte le moins de l'axe des $x$ sur l'intervalle $(-1, +1)$ au sens de la norme du sup. Donc, si l'on développe une fonction en série de polynômes de Tchebycheff, la précision de l'approximation sera homogène sur tout le domaine, l'erreur oscillant entre deux extremums.

On est en droit de penser que les erreurs auront un comportement identique pour les approximations par fractions rationnelles que l'on déduira non plus à partir de la série de MacLaurin mais à partir du développement en série de Tchebycheff. De plus on peut ainsi espérer construire de meilleurs algorithmes dans le cas de séries de puissances lentement convergentes et qui donneront de meilleurs résultats que les approximants de Padé dans le cas où l'on s'éloigne de l'origine. En revanche, ils seront moins précis dans le voisinage de zéro.

Nous avons étudié les polynômes de Tchebycheff quand nous avons évoqué la méthode d'intégration de Gauss-Tchebycheff (*cf.* chapitre 8). Aussi serons-nous bref pour ce qui concerne leurs propriétés.

Nous avons défini le polynôme de Tchebycheff de degré $n$ au moyen de la relation (*cf.* chapitre 8) :

$$T_n(x) = \cos[n \arccos(x)],$$

$$\text{on a donc : } \quad T_0(x) = 1 \quad \text{et} \quad T_1(x) = x.$$

Les polynômes de Tchebycheff obéissent aux relations de récurrence :

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0$$

$$T_{n+m}(x) + T_{|n-m|}(x) = 2T_n(x) \cdot T_m(x),$$

ainsi qu'à la propriété fondamentale d'orthogonalité (relativement à la fonction poids $1/\sqrt{1-x^2}$) :

$$\int_{-1}^{+1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{\pi}{2}\delta_{mn}.$$

À présent nous allons rappeler comment décomposer une fonction $f(x)$ en série de polynômes de Tchebycheff sur l'intervalle canonique $(-1, +1)$ sans préjudice de la généralité puisqu'il est toujours possible de ramener tout intervalle fini $I$ à cet intervalle. $f(x)$ admet un développement unique en série de Tchebycheff donné par l'expression :

$$f(x) = \sum_{n=0}^{\infty} a_n T_n(x),$$

avec

$$a_n = \frac{2}{\pi} \int_{-1}^{+1} \frac{f(x) T_n(x)}{\sqrt{1-x^2}} \, \mathrm{d}x$$

$$\text{et} \quad a_0 = \frac{1}{\pi} \int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}} \, \mathrm{d}x.$$

***Remarque*** : Il n'est pas souvent simple d'obtenir les coefficients $a_k$ sous forme littérale. En revanche on peut les obtenir assez facilement par voie numérique au moyen de la méthode de Gauss-Tchebycheff qui fournit une remarquable précision.

Cela dit, nous obtiendrons la méthode de Maehly par transposition directe de la méthode de Padé. On écrira :

$$f(x) = \frac{R_p(x)}{P_q(x)} = \frac{\sum_{i=0}^{p} b_i T_i(x)}{\sum_{j=0}^{q} l_j T_j(x)},$$

où ici $R_p(x)$ et $P_q(x)$ représentent des développements en série de Tchebycheff. On peut montrer que l'erreur est de l'ordre de $T_{p+q+1}(x)$. Comme dans le cas des approximants de Padé, les meilleures approximations sont obtenues pour $p = q$ et $p = q + 1$.

À partir de la connaissance des coefficients $a_k$, nous allons effectuer des calculs tout à fait semblables à ceux de Padé. Soit un polynôme de degré $q$ écrit en termes de polynômes de Tchebycheff :

$$P_q(x) = \sum_{j=0}^{q} l_j T_j(x).$$

On écrit alors :

$$f(x) P_q(x) = \sum_{k=0}^{\infty} b_k T_k(x) \sum_{i=0}^{q} l_i T_i(x).$$

Il nous faut à présent développer ce produit de sommes, et pour cela il faut utiliser la seconde relation de récurrence que nous venons de rappeler. On obtient successivement :

$$\begin{aligned} f(x) P_q(x) &= \sum_{k=0}^{\infty} \cdot \sum_{i=0}^{q} b_k l_i T_i(x) \cdot T_k(x) = \sum_{i=0}^{q} \sum_{k=0}^{\infty} b_k l_i T_i(x) \cdot T_k(x) \\ &= \sum_{k=0}^{\infty} \sum_{i=0}^{q} \frac{b_k l_i}{2} \left[ T_{i+k}(x) + T_{|i-k|}(x) \right] = \frac{1}{2} \sum_{k=0}^{\infty} \left[ \sum_{i=0}^{q} b_k l_i T_{i+k}(x) + \sum_{i=0}^{q} b_k l_i T_{|i-k|}(x) \right]. \end{aligned}$$

Rappelons pour mémoire que les indices négatifs correspondent à des coefficients nuls. On peut également développer les deux sommes, puis regrouper les termes, ce qui donne en définitive :

$$\begin{aligned}2f(x)P_q(x) = {} & T_0(x)(b_0l_0 + b_1l_1 + \cdots + b_ql_q + b_0l_0) + T_1(x)(b_0l_1 + b_1l_0 + b_0l_1 \\ & + b_1l_2 + b_2l_3 + \cdots + b_{q-1}l_q + b_1l_0 + b_2l_1 + b_3l_2 + \cdots + b_{q+1}l_q) \\ & + T_2(x)(b_0l_2 + b_1l_1 + b_2l_0 + b_0l_2 + b_1l_3 + b_2l_4 + \cdots + b_{q-2}l_q + b_2l_0 + b_3l_1 \\ & + b_4l_2 + \cdots + b_{q+2}l_q) + T_3(x)(b_0l_3 + b_1l_2 + b_2l_1 + b_3l_0 + b_0l_3 \\ & + b_1l_4 + b_2l_5 + \cdots + b_{q-3}l_q + b_3l_0 + b_4l_1 + b_5l_2 + \cdots + b_{q+3}l_q) + \cdots\end{aligned}$$

La technique de calcul demeure la même que celle utilisée lors de l'étude des approximants de Padé. On égale à zéro les coefficients des polynômes de degré $p+1, p+2, \ldots, p+q$, et l'on considère le cas classique où $p > q$. Dans ce cas, la somme

$$b_0l_k + b_1l_{k+1} + \cdots + b_{q-k}l_q$$

disparaît. En apparence on retrouve la matrice obtenue lors de l'établissement de la méthode de Padé, mais il convient de faire attention, car à chaque ligne de la matrice, il correspond une valeur de $p$ que l'on ne retrouve évidemment pas sur l'autre ligne. Il faut, par conséquent, différencier attentivement les éléments selon la ligne à laquelle ils appartiennent. Pour mieux situer le problème considérons les matrices dans le cas $p = q = 2$. Avec la méthode de Padé on écrira :

$$\begin{vmatrix} a_2 & a_1 & -a_3 \\ a_3 & a_2 & -a_4 \end{vmatrix}$$

et en se servant de la méthode de Maehly on aura :

$$\begin{vmatrix} \alpha_{32} & \alpha_{32} & -\alpha_{33} \\ \alpha_{43} & \alpha_{42} & -\alpha_{44} \end{vmatrix}$$

et l'on n'aura pas l'égalité $\alpha_{32} = \alpha_{42}$, ni l'égalité$\alpha_{43} = \alpha_{33}$.

En revanche, on devra écrire :

$$\begin{aligned}2\alpha_{33} &= 2\beta_3 \\ 2\alpha_{32} &= \beta_2 + \beta_4 \\ 2\alpha_{31} &= \beta_1 + \beta_5 \\ 2\alpha_{44} &= 2\beta_4 \\ 2\alpha_{42} &= \beta_2 + \beta_6 \\ 2\alpha_{43} &= \beta_3 + \beta_5\end{aligned}$$

expressions dans lesquelles les $\beta_k$ sont les coefficients du développement en série de Tchebycheff que l'on souhaite distinguer du développement en série entière.

## Un exemple : étude de la fonction $\log_e(1 - 2ty + t^2)$ sur $(-1, +1)$

Nous allons procéder au calcul de deux approximants de Maehly sur cette fonction tout simplement parce que le calcul des coefficients de la série de Tchebycheff est réalisable formellement.

En effet, on trouve dans les tables :

$$\int_0^{\pi} \log_e[1 - 2t\cos(x) + t^2]\cos(nx)\,\mathrm{d}x = -\frac{\pi}{n}t^n, \quad \text{si } t^2 < 1,$$

$$\text{et} \int_0^{n\pi} \log_e[1 - 2t\cos(x) + t^2]\cos(x)\,\mathrm{d}x = -\frac{\pi}{n}t^n, \quad \text{si } t^2 < 1.$$

Le changement de variable $x = \arccos(y)$ nous permet de calculer les coefficients du développement en série de Tchebycheff. Pour fixer les idées, choisissons tout à fait arbitrairement le coefficient $t$ égal à 0,5 ; reste donc à obtenir les coefficients du développement en polynômes de Tchebycheff de la fonction : $\log_e(1 - x + 0{,}25)$.

Il n'y a pas de difficulté à calculer les coefficients $a_k$ (*cf.* Tab. D.4).

**Tableau D.4.**

| | |
|---|---|
| $a_0 = 0$ | $a_1 = -\pi/2$ |
| $a_2 = -\pi/8$ | $a_3 = -\pi/24$ |
| $a_4 = -\pi/64$ | $a_5 = -\pi/160$ |
| $a_6 = -\pi/384$ | $a_7 = -\pi/896$ |
| $a_8 = -\pi/2\,048$ | $a_9 = -\pi/4\,608$ |
| $a_{10} = -\pi/10\,240$ | $a_{11} = -\pi/45\,056$ |
| $a_{12} = -\pi/98\,304$ | |

***a* – *Étude du cas* p = q = 3 –** Ce cas est présenté dans le tableau D.5. Pour des raisons évidentes de commodité de calcul, il sera judicieux de développer les polynômes de Tchebycheff pour obtenir $f(x)$ sous la forme :

$$f(x) = \frac{\sum_{k=0}^{p} b'_k x^k}{\sum_{j=0}^{q} l'_j x^j}.$$

On obtient alors les valeurs du tableau D.6, page suivante.

**Tableau D.5.** *Étude du cas* $p = q = 3$

| | |
|---|---|
| $l_0 = 1$ | $b_0 = 0{,}516\,155\,562$ |
| $l_1 = -1{,}078\,971\,778$ | $b_1 = -0{,}966\,519\,671\,6$ |
| $l_2 = 1{,}890\,890\,323 \times 10^{-1}$ | $b_2 = 0{,}335\,223\,066\,2$ |
| $l_3 = -7{,}339\,252\,452 \times 10^{-3}$ | $b_3 = -0{,}273\,101\,370\,2 \times 10^{-1}$ |

**Tableau D.6.** *Étude du cas* $p = q = 3$

| | |
|---|---|
| $l'_0 = \ \ 0{,}810\,910\,968$ | $b'_0 = \ \ 0{,}180\,932\,496$ |
| $l'_1 = -1{,}056\,954\,021$ | $b'_1 = -0{,}884\,589\,260$ |
| $l'_2 = \ \ 0{,}378\,178\,065$ | $b'_2 = \ \ 0{,}670\,446\,132$ |
| $l'_3 = -0{,}029\,357\,010$ | $b'_3 = -0{,}109\,240\,548$ |

***b – Étude du cas* p = q = 4** – Les résultats sont présentés dans le tableau D.7. De là on tire les valeurs du tableau D.8.

**Tableau D.7.** *Étude du cas* $p = q = 4$

| | |
|---|---|
| $l_0 = +1$ | $b_0 = -0{,}085\,339\,731$ |
| $l_1 = +0{,}439\,993\,085\,1$ | $b_1 = -0{,}460\,239\,216$ |
| $l_2 = -1{,}171\,670\,047$ | $b_2 = -0{,}614\,231\,378$ |
| $l_3 = +0{,}288\,160\,179$ | $b_3 = +0{,}453\,766\,143$ |
| $l_4 = -0{,}013\,102\,673$ | $b_4 = -0{,}047\,218\,321$ |

**Tableau D.8.** *Étude du cas* $p = q = 4$

| | |
|---|---|
| $l'_0 = +2{,}158\,567\,374$ | $b'_0 = +0{,}481\,673\,325$ |
| $l'_1 = -0{,}424\,487\,452$ | $b'_1 = -1{,}821\,537\,645$ |
| $l'_2 = -2{,}238\,518\,707$ | $b'_2 = -0{,}850\,716\,186$ |
| $l'_3 = +1{,}152\,640\,716$ | $b'_3 = +1{,}815\,064\,572$ |
| $l'_4 = -0{,}104\,821\,387$ | $b'_4 = -0{,}377\,746\,570$ |

Pour mieux apprécier les différentes méthodes, nous avons tabulé la fonction sous diverses formes dont le développement classique (*cf.* Tab. D.9, page ci-contre) :

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + (-1)^n \frac{x^n}{n} + \cdots$$

## 6. Erreur liée à l'usage des approximants de Maehly

On peut se faire une opinion sur la précision des approximations proposées. En effet, on peut écrire :

$$D(x) = f(x) \cdot P_q(x) - \sum_{k=0}^{p} a_k T_k(x) = \omega_{pq} T_{p+q+1}(x).$$

L'erreur est donc de l'ordre de : $\frac{D(x)}{P_q(x)}$. Comme on peut considérer que $P(x) = l_0 = 1$, on en déduit que :

$$E(x) = D(x) = \omega_{pq} T_{p+q+1}(x),$$

$\omega_{pq}$ ayant été défini au paragraphe 1.

**Tableau D.9.**

| $x$ | $\log(1,25-x)$ | $\sum_{k=0}^{6} a_k T_k(x)$ | $\sum_{n=1}^{6} \frac{(-1)^n x^n}{n}$ | Maehly $p=q=4$ |
|---|---|---|---|---|
| −1,2 | 0,896 088 024 | 0,847 958 | 0,042 759 | 0,896 046 |
| −0,5 | 0,559 615 787 | 0,558 333 | 0,548 071 | 0,559 616 |
| −0,1 | 0,300 104 592 | 0,301 856 | 0,300 034 | 0,300 104 |
| −0,01 | 0,231 111 721 | 0,232 049 | 0,231 102 | 0,231 113 |
| 0,0 | 0,223 143 551 | 0,223 958 | 0,223 139 | 0,223 145 |
| 0,01 | 0,215 111 379 | 0,215 798 | 0,215 106 | 0,215 113 |
| 0,1 | 0,139 761 942 | 0,139 185 | 0,139 761 | 0,139 765 |
| 0,5 | −0,287 682 072 | −0,287 500 | −0,287 671 | −0,287 688 |
| 0,7 | −0,597 837 000 | −0,595 837 | −0,596 951 | −0,597 832 |

## 7. Difficultés liées à la recherche d'une généralisation

Il est naturel d'envisager l'usage d'autres polynômes orthogonaux, voire de fonctions orthogonales, offrant des propriétés intéressantes afin de former d'autres approximants. La difficulté essentielle repose sur le fait qu'il n'est pas aisé d'obtenir une relation simple entre les produits de polynômes orthogonaux telle, par exemple, la relation de récurrence utilisée avec les polynômes de Tchebycheff. On pourrait penser à des approximants formés à partir des séries de Fourier, mais il n'existera pas de différences fondamentales avec les approximants de Maehly, car on passe d'une représentation à l'autre par un changement de variable du type $y = \arccos(x)$. Quoiqu'il en soit, les approximants de Padé et de Maehly sont à la base du calcul de toutes les fonctions de bibliothèque existant dans les calculateurs arithmétiques quels qu'ils soient. Nous verrons dans la prochaine annexe E comment sont effectivement calculées la plupart des fonctions de bibliothèque à partir des approximants que nous venons d'étudier.

## 8. Éléments de bibliographie

G.A. BAKER et P.R. GRAVES-MORRIS (1978) *Padé Approximants*, Cambridge University Press.

C. BREZINSKI (1978) *Algorithmes d'accélération de la convergence*, Éditions Technip, Paris.

H. PADÉ (1892) *Sur la représentation approchée d'une fonction par des fractions rationnelles*, Annales Scientifiques de l'École Normale Supérieure.

H. PADÉ (1899) *Mémoire sur les développements en fractions continues de la fonction exponentielle*, Annales Scientifiques de l'École Normale Supérieure.

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

# E Calcul des fonctions de bibliothèque élémentaires

Le calcul des fonctions de bibliothèque repose sur la décomposition préalable de la fonction en deux parties qui seront le cas échéant une somme ou un produit. Cette décomposition est effectuée de telle sorte qu'une des parties soit calculée exactement et que l'autre soit une fonction dont l'argument est généralement petit devant l'unité. Très souvent, on évalue cette dernière fonction au moyen des approximants de Padé, de Maehly ou des fractions continues. Malheureusement, cette technique qui donne des résultats remarquables n'est pas systématiquement applicable. Considérons par exemple le cas des fonctions de Bessel, il n'est pas possible dans l'état actuel de nos connaissances de procéder à leur calcul par cette voie. En effet, on ne sait pas réduire le calcul d'une fonction de Bessel à un intervalle canonique au voisinage de zéro ou sur l'intervalle $(-1, +1)$ qui nous permettrait ensuite d'utiliser les approximants déjà évoqués.

Avant d'entreprendre à proprement parler cette étude, insistons sur le fait qu'on n'utilise jamais directement un développement en série pour calculer les fonctions de bibliothèque. On peut s'en persuader facilement en considérant le développement de $\exp(-x)$ qui nous servira de technique de calcul pour $x = 10$. Nous obtenons une série alternée dont le terme général va commencer seulement à décroître lorsque :

$$10^n < n!$$

ce qui donne approximativement :

$$n \approx 10.$$

On conçoit sans difficulté que les sommes partielles vont être relativement importantes avant de décroître jusqu'à une valeur petite. Cela imposerait des mots-mémoire de taille importante et de fort longs calculs.

Dans ce chapitre, nous nous attacherons à l'esprit de la technique du calcul d'un certain nombre de fonctions élémentaires.

## 1. Calcul de exp(x) pour x appartenant à $(-\infty, +\infty)$

L'idée de base pour calculer $\exp(x)$ (sur une machine qui fonctionne avec des nombres en représentation binaire) repose sur la décomposition de cette fonction en un produit du type :

$$\exp(x) = 2^N \exp(Q) \tag{E.1}$$

de telle sorte que $N$ soit un entier et $Q$ un terme compris entre $-1$ et $+1$. Passant aux logarithmes en base 2, nous pouvons écrire alors :

$$x \log_2(e) = N + Q \log_2(e) = N + n. \tag{E.2}$$

En remarquant au passage que $\log_2(e) \log_e(2) = 1$, on obtient sans grande difficulté $N$ et $n$. $N$ est la partie entière de $x/\log_e(2)$ et $n$ la partie fractionnaire du même rapport. Le calcul de $\log_e(2)$ s'effectue au moyen d'un développement en série et l'on obtient pour $\log_e(2)$ et donc pour $\log_2(e)$ autant de chiffres significatifs qu'on le souhaite. Il s'agit là de constantes calculées une fois pour toute : $\log_e(2) = 0{,}693\,147\,180$ et $\log_2(e) = 1{,}442\,695\,04$. Il reste à calculer $N$ et $n$, puis $Q = n \log_e(2)$.

Du simple fait que la machine calcule en binaire, il n'y a pas d'erreur en calculant $2^N$ puisque $N$ est entier. On obtient donc $n$ sans problème et l'on note que $Q$ est compris dans l'intervalle $-0{,}693\,147 < Q < +0{,}693\,147$.

Il reste à calculer la fonction $\exp(Q)$ ce que l'on réalisera par les approximants de Maehly par exemple. Nous aurons :

$$\exp(Q) = \sum_{k=0}^{\infty} a_k T_k(x)$$

avec

$$a_0 = \frac{1}{\pi} \int_{-1}^{+1} \frac{\exp(x)}{\sqrt{1-x^2}}\, \mathrm{d}x$$

$$a_k = \frac{2}{\pi} \int_{-1}^{+1} \frac{\exp(x) T_k(x)}{\sqrt{1-x^2}}\, \mathrm{d}x.$$

**Tableau E.1.**

| | |
|---|---|
| $b'_0 = 1{,}905\,369\,087$ | $l'_0 = +1{,}905\,369\,014$ |
| $b'_1 = 0{,}954\,415\,684$ | $l'_1 = -0{,}950\,951\,382$ |
| $b'_2 = 0{,}190\,992\,815$ | $l'_2 = +0{,}189\,261\,972$ |
| $b'_3 = 0{,}015\,822\,193$ | $l'_3 = -0{,}015\,540\,073$ |

Le calcul des coefficients $a_k$ peut se faire par la méthode de Gauss-Tchebycheff, mais il faudra calculer les valeurs de $\exp(x_k)$, $x_k$ étant un zéro du polynôme de Tchebycheff de degré $p$, soit par un développement en série, soit par un approximant de Padé déterminé dans un précédent chapitre. Les coefficients que nous avons trouvés pour $p = q = 3$, en ayant développé les polynômes de Tchebycheff, sont présentés dans le tableau E.1. On aura donc :

$$\exp(x) = \frac{[(b'_3 x + b'_2)x + b'_1]x + b'_0}{[(l'_3 x + l'_2)x + l'_1]x + l'_0}.$$

Il faut bien noter que les degrés $p$ et $q$ des polynômes formant la fraction rationnelle sont fonction de la précision recherchée. Il en va de même du nombre de chiffres significatifs qu'il convient de conserver. Ce problème doit être étudié pour chaque cas particulier comme il a été dit à l'occasion de l'étude des approximants de Padé et de Maehly.

***Un exemple numérique*** – On se propose de calculer exp(35,257). pour cela on réalise la décomposition :

$$x \log_2(e) = N + n$$

qui donne : $N = 50$ et $n = 0{,}865\,099\,03$ d'où $Q = 0{,}599\,640\,954$.

Voici les résultats trouvés par différents procédés, il convient de les multiplier par $10^{15}$ :

| machine $X$ | $\sum_{k=0}^{9} a_k T_k(x)$ | Padé $p = q = 3$ | Maehly $p = q = 3$. |
|---|---|---|---|
| 2,050 786 965 | 2,050 786 931 | 2,050 787 5 | 2,050 787 2 |

## 2. Calcul de sin(x) et cos(x) pour x appartenant à $(-\infty, +\infty)$

Compte tenu de la parité des fonctions, on se ramène à l'intervalle $(0, \pi)$, puis ensuite, on réduit le champ à $(0, \pi/2)$ et enfin à l'intervalle $(0, \pi/4)$ puisque l'on dispose de relations du genre :

$$\sin\left(\frac{\pi}{2} - x\right) = \cos(x) \quad \text{et} \quad \cos\left(\frac{\pi}{2} - x\right) = \sin(x)$$

qui permettent de calculer sinus et cosinus sur l'intervalle $(0, \pi/4)$. D'abord, on détermine les nombres $n$ et $N$ tels que :

$$x = \pi(N + n) \quad \text{avec} \quad 0 \leq n < 1,$$

on en déduit que :

$$\sin(x) = (-1)^N \sin(n\pi) \quad \text{et} \quad \cos(x) = (-1)^N \cos(n\pi).$$

Donc, pour l'instant, nous nous sommes ramené à l'intervalle $(0, \pi)$. À présent, voyons comment se ramener à l'intervalle $(0, \pi/2)$. Posons :

$$s_1 = \text{signe} \left(\frac{1}{2} - n\right)$$
$$\text{puis :} \quad n_1 = \frac{1}{2} - s_1 \left(\frac{1}{2} - n\right).$$

Si $s_1$ est positif, on écrit : $\sin(x) = (-1)^N \sin(n_1\pi)$ et $\cos(x) = (-1)^N \cos(n_1\pi)$.
Si $s_1$ est négatif, on écrit : $\sin(x) = (-1)^N \cos(n_1\pi)$ et $\cos(x) = (-1)^N \sin(n_1\pi)$.

La dernière transformation consiste à obtenir l'argument de la fonction sinus ou cosinus dans l'intervalle $(0, \pi/4)$. Il suffit de reprendre la technique déjà utilisée pour parvenir à ces fins. Posons :

$$s_2 = \text{signe} \left(\frac{1}{4} - n_1\right)$$
$$\text{puis :} \quad n_2 = \frac{1}{4} - s_2 \left(\frac{1}{4} - n_1\right).$$

Si $s_2$ est positif, on aura : $\sin(n_1\pi) = \sin(n_2\pi)$ et $\cos(n_1\pi) = \cos(n_2\pi)$.
Si $s_2$ est négatif, on écrit : $\sin(n_1\pi) = \cos(n_2\pi)$ et $\cos(n_1\pi) = \sin(n_2\pi)$.

On calcule les fonctions cosinus et sinus des réductions au moyen des approximants de Padé obtenus respectivement pour $p = q = 3$ et $p = q = 2$ qui assurent une précision d'au moins huit chiffres significatifs. Il est possible de faire mieux encore, mais il nous a semblé préférable de nous attacher surtout à la philosophie générale. Cependant, lorsqu'il s'agit d'implanter une fonction dans un système, il n'y a plus aucun détail qui puisse être délaissé. À précision égale, l'algorithme qui contient le moins d'opérations logiques et arithmétiques — qui sera par conséquent le plus rapide — sera toujours préféré, économie de temps oblige. Ainsi, pour le calcul des fonctions cosinus et sinus, on obtiendra de meilleures performances si l'on réduit l'intervalle à $(0, \pi/6)$ sur lequel on adoptera les approximants de Maehly ; toutes choses égales par ailleurs, on conservera les mêmes principes que ceux qui viennent d'être développés, mais on ajoutera l'analyse des cas particuliers usuels tels que $\pi/2$ et $\pi/4$ ou encore $\pi/3$ et $\pi/6$.

## Exemples d'application

***a*** – On se propose de calculer $\sin(69,674\,361)$. Nous pouvons écrire :

$$N = 22 \quad \text{et} \quad n = 0{,}178\,037\,92.$$

Puis nous obtenons :

$$s_1 = \text{signe}\ (0{,}5 - n) = +1 \quad \text{et} \quad n_1 = 0{,}5 - (0{,}5 - n) = n.$$

Il nous reste donc à calculer $\sin(n_1\pi)$. L'usage de l'approximant de Padé pour $p = q = 2$ établi dans un chapitre précédent permet d'obtenir le résultat suivant :

$$\sin(69{,}674\,361\pi) = 0{,}530\,612\,163\,8$$

le dixième chiffre « significatif » étant sujet à caution. Maintenant, on se propose de calculer $\cos(58{,}125\,769\,8)$.On obtient sans problème :

$$N = 18 \quad \text{et} \quad n = 0{,}502\,007\,17.$$

À partir de là, nous déduisons :

$$\begin{aligned} & s_1 = -1 \quad \text{et} \quad n = 0{,}5 + (0{,}5 - n) = 0{,}497\,992\,83, \\ \text{puis :}\ & s_2 = -1 \quad \text{et} \quad n = 0{,}25 + (0{,}25 - n) = 0{,}002\,007\,17. \end{aligned}$$

Il nous faut donc calculer $\sin(0{,}002\,007\,17\pi)$, on trouve alors :

$$\cos(58{,}125\,769\,8) = -0{,}630\,566\,873\,9 \times 10^{-2},$$

les dix chiffres significatifs mentionnés sont exacts.

***b*** – Un autre exemple. On se propose de calculer $\sin(7{,}587\,214\,583)$. On établit que :

$$N = 2 \quad \text{et} \quad n = 0{,}415\,085\,41$$

ce qui donne :

$$s_2 = -1 \quad \text{et} \quad n_2 = 0{,}084\,914\,59.$$

En adoptant l'approximant de Padé $p = q = 3$ pour évaluer le cosinus, nous obtenons :

$$\sin(7{,}587\,214\,583) = 0{,}964\,628\,186\,9.$$

L'erreur absolue sur ce résultat est inférieure à $5\,10^{-10}$.

## 3. Calcul de $\log_e(x)$ pour x appartenant à $(0, +\infty)$

Écrivons le nombre $x$ dans le système binaire. Nous obtenons en choisissant la représentation flottante :

$$x = 0{,}b_0b_1b_2b_3b_4b_5b_6b_7 \ldots 2^{L+1} \quad \text{avec} \quad b_0 \neq 0.$$

Posons $N = 2^L$, de telle sorte que l'on puisse écrire :

$$\log_e(x) = L\log_e(2) + \log_e(Q).$$

Autrement dit, on a $x = 2^L Q$ avec $1 \leq Q < 2$. On reconnaît une technique tout à fait semblable à celle utilisée pour la décomposition de $\exp(x)$.

Il reste à calculer $\log_e(Q)$ avec $1 \leq Q < 2$, opération que l'on réalisera soit avec les approximants de Padé soit avec les approximants de Maehly. Encore une fois insistons sur le fait que cette idée ne constitue que le principe du calcul. À titre d'exemple, on peut très bien utiliser les approximants de la fonction $\log(1-2tx+t^2)$ en choisissant $t = 0{,}5$, soit $\log(1{,}25-x)$. On obtient alors $x = 0{,}25$ et $x = -0{,}75$ comme limites de variation pour $x$.

Proposons-nous de calculer log(35 784,351 74). En appliquant les règles précédentes, on obtient $L = 15$ et $Q = 1{,}092\,051\,75$. Il faut faire :

$$x = 1{,}25 - 1{,}092\,051\,75 = 0{,}157\,948\,25$$

et l'on trouve :

$$\log(Q) = 0{,}088\,061\,275$$

en adoptant l'approximant de Maehly pour $p = q = 4$. D'où le résultat :

$$\log(35\,784{,}351\,74) = 10{,}485\,269.$$

Ici le dernier chiffre est entaché d'erreur.

## 4. Calcul de tangente et cotangente pour x appartenant à $(-\infty, +\infty)$

Les raisons de parité permettent de travailler sur l'intervalle $(0, \infty)$, tandis que la division par $\pi$ ramène l'intervalle à $(0, \pi)$. En appliquant la technique déjà évoquée pour les fonctions sinus et cosinus, on se ramène à l'intervalle $(0, \pi/4)$. Bien sûr, on se souviendra que le produit de tangente et de cotangente du même argument est égal à l'unité.

Ensuite, il y a plusieurs façon d'évaluer ces fonctions dans un intervalle réduit. On peut utiliser la fraction continue suivante :

$$\frac{1}{2}\cot\left(\frac{x}{2}\right) = 1 + \sum_{n=1}^{\infty} \frac{-(x/2)^2|}{|2n+1},$$

mais on peut aussi obtenir des approximants de Padé ou de Maehly à partir du développement banal de $\cot(x)$ :

$$x\cot(x) = 1 - \sum_{k=1}^{\infty} \frac{2^{2k}|B_{2k}|x^{2k}}{(2k)!}$$

où les $B_{2k}$ sont les nombres de Bernoulli (*cf.* p. 178). Cette série converge pour $x^2 < \pi^2$.

## 5. Calcul de argtanh(x) pour x appartenant à (0, 1)

Comme nous avons la relation :

$$\operatorname{argtanh}(x) = \log\left(\frac{1+x}{1-x}\right),$$

on peut adopter soit la fraction continue :

$$\log_e\left(\frac{1+x}{1-x}\right) = \frac{2x|}{|1} + \sum_{m=1}^{\infty} \frac{-m^2x^2|}{|4m^2-1}$$

ou encore le développement de Tchebycheff :

$$\log_e\left(\frac{1+2ax+a}{1-2ax+a}\right) = 4\sum_{n=0}^{\infty} \frac{a^{2n+1}}{2n+1},$$

où $|x| \leq 1$ dans chacun des deux développements.

## 6. Calcul de arctan(x) pour x appartenant à (0, +∞)

On pose $x = \tan(Q)$ ce qui est toujours possible. Ensuite, on divise l'intervalle $(0, \pi/2)$ selon des angles en progression arithmétique donnés par la relation :

$$Q_k = (k - 0{,}5)\frac{\pi}{q} \quad \text{avec toutefois } Q_0 = 0.$$

Si $x$ est compris entre $Q_k$ et $Q_{k+1}$, alors on peut écrire :

$$\arctan(x) = \frac{k\pi}{q} + \arctan\left(a_k - \frac{b_k}{x + a_k}\right)$$

expression dans laquelle $a_k = \cot(k\pi/q)$ et $b_k = 1 + a_{k^2}$.

En règle générale, $q$ dépasse rarement la valeur 12. Ensuite, on calcule la valeur de arctan du second membre au moyen de la fraction continue de Gauss :

$$\arctan(t) = \frac{t|}{|1} + \sum_{m=1}^{\infty} \frac{m^2t^2|}{|2m+1}.$$

## 7. Calcul de arcsin(x) et arccos(x) pour x appartenant à (0, 1)

On peut se ramener sans grand problème au calcul de arctan en remarquant que :

$$\arcsin(x) = \arctan\left(x/\sqrt{1-x^2}\right)$$

$$\text{et} \qquad \arccos(x) = \frac{\pi}{2} - \arcsin(x).$$

Cette remarque étant faite, il faut préciser que ce procédé est « long » en temps-machine, aussi préfère-t-on implanter une fonction arcsin(x) fondée sur les approximants ou les fractions rationnelles.

## 8. Calcul de la racine carrée pour x appartenant à (0, ∞)

Il s'agit probablement de l'un des plus vieux algorithmes toujours en vigueur, dont la paternité reviendrait à Héron d'Alexandrie (I^er siècle). *Grosso modo*, la technique consiste à obtenir un ordre de grandeur « convenable » de la racine carrée puis à appliquer la méthode de Newton dont la convergence est, rappelons-le, quadratique. En calcul binaire, on pourra poser :

$$x = 2^{2m}n \quad \text{avec} \quad \frac{1}{4} < n < 1 \quad \text{donc} \quad 4^{m-1} < x < 4^m.$$

On obtient alors $\sqrt{x} = 2^m\sqrt{n}$.

À présent, il nous reste à calculer $\sqrt{n}$. Cette opération est réalisée au moyen de la fraction continue :

$$\sqrt{n} = \frac{25}{7} - \frac{5\,10^3\,7^{-3}|}{|n + 235\,7^{-2}} - \frac{4\,10^2\,7^{-4}|}{|n + 15\,7^{-2}}\,.$$

En introduisant cette valeur dans la formule de Héron-Newton :

$$y_{k+1} = (y_k + n/y_k)/2$$

après trois tours d'itération, on obtient alors une précision relative sur $y_2$ qui sera meilleure que $4{,}5\,10^{-9}$.

### Exemple

On se propose de calculer $\sqrt{17\,584{,}348\,33}$. On établit que :

$$4^7 < 17\,584{,}348\,33 < 4^8$$

et donc que : $\sqrt{17\,584{,}348\,33} = 256\sqrt{0{,}268\,315\,862}$.

L'application de la fraction continue donne : $\sqrt{n} = 0{,}518\,098\,408\,4$, laquelle, introduite dans la formule de Newton-Héron, permet d'obtenir : $\sqrt{n} = 0{,}517\,992\,155\,5$ à la première itération, et $\sqrt{n} = 0{,}517\,992\,144\,7$ à la seconde.

En définitive, on trouve : $\sqrt{17\,584{,}348\,33} = 132{,}605\,989$, tous les chiffres significatifs sont exacts.

## 9. Éléments de bibliographie

A. RALSTON et H.S. WILF (1965) *Méthodes mathématiques pour calculateurs arithmétiques*, Dunod.

# F Calcul numérique des fonctions de Bessel

Les fonctions de Bessel occupent une place très importante dans les solutions de problèmes offrant la symétrie cylindrique et l'on peut dire que, dans la hiérarchie des fonctions qu'il est nécessaire de tabuler, elles viennent juste après les fonctions trigonométriques, la fonction logarithme et la fonction exponentielle, mais on a coutume de les faire rentrer dans la classe des fonctions dites spéciales. Il existe beaucoup d'ouvrages qui traitent des propriétés remarquables de ces fonctions, on en trouvera quelques uns cités en bibliographie. Notre but est de montrer comment on peut obtenir aisément ces tables numériques au moyen des algorithmes que nous avons établis.

## 1. L'équation différentielle des fonctions de Bessel (1784–1846)

L'équation de Laplace (ou de Poisson) à laquelle obéit une grandeur $V(\rho, \theta, z)$ s'écrit en coordonnées cylindriques $(\rho, \theta, z)$ :

$$\frac{\partial^2 V}{\partial \rho^2} + \frac{1}{\rho} \cdot \frac{\partial V}{\partial \rho} + \frac{1}{\rho^2} \cdot \frac{\partial^2 V}{\partial \theta^2} + \frac{\partial^2 V}{\partial z^2} = 0.$$

On recherche les solutions sous forme de produits de Laplace :

$$V(\rho, \theta, z) = R(\rho)\Theta(\theta)Z(z),$$

ce qui nous autorise ensuite à passer en dérivées droites :

$$\frac{\mathrm{d}^2 R}{\mathrm{d}\rho^2} + \frac{1}{\rho} \cdot \frac{\mathrm{d}R}{\mathrm{d}\rho} + \left(1 - \frac{\nu^2}{\rho^2}\right) = 0, \tag{F.1}$$

c'est cette équation différentielle qui est appelée équation de Bessel. Réécrivons-la plus conventionnellement en terme de $y(x)$ :

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + \frac{1}{x} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} + \left(1 - \frac{\nu^2}{x^2}\right) = 0.$$

Cherchons une solution sous la forme d'un développement en série entière que l'on écrit :

$$y = x^{\mu} \sum_{j=0}^{\infty} a_j x^j \tag{F.2}$$

On peut voir sans grande difficulté que tous les coefficients $a_j$ dépendent du coefficient $a_0$. Par définition, les fonctions de Bessel de première espèce sont celles pour lesquelles le coefficient arbitraire $a_0$ est égal à :

$$a_0 = \frac{1}{2^\mu \Gamma(1+\mu)}$$

où $\Gamma(x)$ est la fonction gamma ; en définitive, elles s'écrivent sous la forme :

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(\nu+m+1)} \left(\frac{x}{2}\right)^{2m}. \tag{F.3}$$

Sans entrer dans les détails, il existe une deuxième solution, indépendante de la première, appelée solution de Weber (1842–1913), qui donne naissance aux fonctions de Bessel de deuxième espèce ; elles s'écrivent (si $\nu$ est différent d'un nombre entier) :

$$Y_\nu(x) = \frac{J_\nu(x)\cos(\pi\nu) - J_{-\nu}(x)}{\sin(\pi\nu)}. \tag{F.4}$$

Si $\nu$ est entier égal à $n$, cette dernière expression est une forme indéterminée ; en appliquant la règle de l'Hospital, nous obtenons la solution :

$$\begin{aligned} Y_n(x) = {} & \frac{2}{\pi} J_n(x) \log_e\left(\frac{x}{2}\right) - \frac{1}{\pi} \sum_{m=0}^{n-1} \frac{(n-m-1)!}{m!} \left(\frac{x}{2}\right)^{2m-n} \\ & - \frac{1}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m}{m!(n+m)!} \left(\frac{x}{2}\right)^{2m+n} [\Psi(m) + \Psi(m+n)], \end{aligned} \tag{F.5}$$

où $\Psi(x)$ est la dérivée logarithmique de la fonction factorielle, à savoir :

$$\Psi(x) = \frac{\Gamma'(1+x)}{\Gamma(1+x)} \tag{F.6}$$

encore appelée fonction digamma. Si $x$ est un nombre entier appelé $n$, nous avons :

$$\Psi(n) = -\gamma + 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n},$$

avec $\Psi(0) = -\gamma = -0{,}577\,215\,664\,901\,532\,8$ (constante d'Euler).

Ce sont les expressions (F.3), (F.4) et (F.5) qui font l'objet de la programmation pour obtenir les tables désirées.

## 2. Relations de récurrence

Les deux espèces obéissent aux mêmes relations de récurrence, donc en désignant par $W_\nu(x)$ l'une et l'autre des deux fonctions nous pouvons écrire :

$$\begin{aligned} xW'_\nu(x) &= \nu W_\nu(x) - xW_{\nu+1}(x) \\ xW'_\nu(x) &= -\nu W_\nu(x) + xW_{\nu-1}(x) \\ 2W'_\nu(x) &= W_{\nu-1}(x) - W_{\nu+1}(x) \\ 2\frac{\nu}{x} W_\nu(x) &= W_{\nu-1}(x) + W_{\nu+1}(x). \end{aligned} \tag{F.7}$$

On peut fonder quelques espoirs sur cette dernière relation de récurrence si l'on cherche à calculer les fonctions de Bessel d'ordre entier comme c'est généralement le cas.

À partir de $W_0(x)$ et $W_1(x)$, on peut calculer $W_n(x)$ ; malheureusement, le jeu des erreurs dû aux soustractions nous conduira très rapidement à des résultats peu précis et l'on ne peut pas espérer dépasser pour $n$ une dizaine d'unités, mais tout dépend du nombre de chiffres significatifs retenus pour effectuer les calculs. Du reste une application sera présentée pour le calcul de $J_{-\nu}(x)$ servant au calcul de $Y_\nu(x)$ pour $\nu$ non entier.

## 3. Représentation de $J_\nu(x)$ par une intégrale définie

On admettra sans démonstration les résultats suivants :

$$J_\nu(x) = \frac{2}{\sqrt{\pi}\Gamma(\nu+1/2)}\left(\frac{x}{2}\right)^\nu \int_0^{\pi/2} \sin^{2\nu}(t)\cos[x\sin(t)]\,\mathrm{d}t$$

$$= \frac{1}{\sqrt{\pi}\Gamma(\nu+1/2)}\left(\frac{x}{2}\right)^\nu \int_0^{\pi} \sin^{2\nu}(t)\cos[x\sin(t)]\,\mathrm{d}t \qquad \text{(F.8)}$$

$$J_\nu(x) = \frac{2}{\sqrt{\pi}\Gamma(\nu+1/2)}\left(\frac{x}{2}\right)^\nu \int_0^{1} (1-u^2)^{\nu-1/2}\cos(xu)\,\mathrm{d}u \qquad \text{(F.9)}$$

avec cette restriction toutefois que $\nu > -1/2$.

Cette dernière relation est une excellente expression qui s'intègre élégamment par la méthode de Gauss-Tchebycheff et fournit des résultats d'une remarquable précision.

## 4. Technique de calcul

***a*** – Le calcul de $J_n(x)$ avec $n$ entier s'effectue au moyen de l'intégration numérique de la relation (F.9), et pour plus de précision nous calculerons séparément $J_0(x)$ et $J_1(x)$ au moyen des relations :

$$J_0(x) = \frac{1}{\pi}\int_{-1}^{1} \frac{\cos(xt)}{\sqrt{1-t^2}}\,\mathrm{d}t$$

$$\text{et} \quad J_1(x) = \frac{x}{\pi}\int_{-1}^{1} \cos(xt)\sqrt{1-t^2}\,\mathrm{d}t.$$

***b*** – Le calcul de $J_\nu(x)$ avec $\nu$ non entier s'effectue au moyen de la relation (F.9). Bien qu'il n'y ait aucune différence apparente dans le calcul de $J$ que l'ordre soit entier ou non entier, nous avons tenu à la séparation des deux types, les fonctions de Bessel d'ordre entier étant de très loin les plus utilisées.

En ce qui concerne le calcul de la fonction gamma qui figure dans les formules, on utilise l'expression intégrale suivante :

$$\Gamma(x) = \int_0^{\infty} \exp(-t)t^{x-1}\,\mathrm{d}t,$$

qui se calcule aisément au moyen de la méthode de Gauss-Laguerre.

***c*** – Le calcul de $Y_\nu(x)$ avec $\nu$ non entier est réalisé au moyen de la formule (F.4) dans laquelle on devra calculer $J_{-\nu}(x)$ que ne permet pas de calculer la relation (F.9). On l'obtiendra au moyen de la relation de récurrence (F.7).

***d*** – Le calcul de $Y_n(x)$ avec $n$ entier est fondé sur l'exploitation de l'expression (F.5) qui utilise le calcul de$J_n(x)$.

Sur le Web(*), on trouvera un ensemble de programmes `bessel1n.c`, `bessel1f.c`, `bessel2n.c`, `bessel2f.c`, `besseljf.h`, `besseljn.h` et `besselyn.h` qui réalisent ces différents calculs. La précision obtenue est donnée dans les commentaires en tête des programmes.

## 5. Calcul de l'erreur sur $J_0(x)$

Nous partons de l'expression intégrale de $J_0(x)$

$$J_0(x) = \frac{1}{\pi} \int_{-1}^{1} \frac{\cos(xt)}{\sqrt{1-t^2}}\, \mathrm{d}t$$

évaluée par la méthode de Gauss-Tchebycheff pour laquelle pèse une erreur de l'ordre de :

$$E = \frac{M_{2n+2}}{(2n+2)!} \cdot \frac{(2n+1)!!}{2^{n+1}(n+1)!} = \frac{M_{2n+2}}{2n+2} \cdot \frac{1}{2^{2n+2}[(n+1)!]^2}\,.$$

$M_{2n+2}$ est aisé à obtenir et sa valeur est :

$$|x|^{2n+2}.$$

Comme nous avons choisi $n = 100$, nous pouvons faire usage de la formule de Stirling pour s'affranchir de la fonction factorielle, et nous obtenons :

$$\log_e E(x) = (2n+2)\log_e\left(\frac{|x|e}{2n+2}\right).$$

Pour fixer les idées, cherchons les valeurs de $x$ qui permettent d'obtenir une précision absolue de $10^{-20}$ sur $J_0(x)$, en choisissant toujours $n = 20$ :

$$|x| = 54.$$

D'un point de vue pratique, cela signifie que tous les chiffres significatifs de la tabulation sont exacts, la seule source d'erreur étant la propagation de l'erreur à travers l'opération de troncature des opérations arithmétiques effectuées...

Pour ce qui concerne les autres fonctions de Bessel de première espèce, il est difficile d'exploiter une relation générale qui donnerait directement une majoration de $M_{2n+2}$ pour $J_k(x)$, toutefois, pour les premières valeurs de $k$, le calcul direct en faisant usage de la formule de dérivation de Leibnitz permet d'aboutir.

## 6. Éléments de bibliographie

A. ANGOT (1972) *Compléments de mathématiques*, Éditions Masson.
G. GOUDET (1962) *Les fonctions de Bessel*, Éditions Masson.
C. MULLER (1998) *Analysis of spherical symetries in euclidian spaces*, Springer.
E.T. WHITTAKER et G.N. WATSON (1992) *Modern Analysis*, Cambridge University Press.

* `http://www.edpsciences.com/guilpin/`

# G Éléments succincts sur le traitement du signal

Dans ce chapitre, nous allons évoquer quelques aspects des principales fonctions mathématiques utilisées en théorie du traitement du signal. Il s'agit plus spécifiquement des notions de convolution et de corrélation.

Ce chapitre ne fait pas partie à proprement parler du calcul numérique mais en est une application directe, c'est dans ce sens que nous présentons un certain nombre de notions élémentaires qui font aujourd'hui l'objet d'un traitement direct sur ordinateur. En vérité, il n'est pas à dissocier des chapitres consacrés à l'étude statistique des résultats d'expériences en général.

Par définition, un signal est le résultat de la mesure d'une grandeur physique qui varie dans le temps. Le signal peut se trouver à la sortie d'un capteur qui délivre soit un signal continu (analogique), soit un signal numérique (échantillonnage). C'est généralement une tension qui est recueillie dont l'amplitude et la phase dépendront du temps $t$. Nous désignerons par $a(t)$ et $b(t)$ les parties réelle et imaginaire associées au signal $x(t)$ :

$$x(t) = a(t) + jb(t). \tag{G.1}$$

## 1. Puissance et énergie d'un signal

***a*** – Par définition, la puissance instantanée du signal $x(t)$ est donnée par l'expression :

$$p(t) = [a(t)]^2 + [b(t)]^2 = [a(t) + jb(t)][a(t) - jb(t)] = x(t)x^*(t), \tag{G.2}$$

où $x^*(t)$ désigne la quantité complexe conjuguée.

***b*** – La puissance moyenne sur un intervalle $T$, à partir du temps $t_0$, est donnée par :

$$P(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} x(t)x^*(t)\, \mathrm{d}t, \tag{G.3}$$

s'il s'agit de signaux physiques, $x(t)$ est réel, et l'on écrit :

$$P(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} x(t)^2\, \mathrm{d}t, \tag{G.4}$$

**c** – La puissance instantanée d'interaction de deux signaux $x(t)$ et $y(t)$ s'écrit :

$$\begin{aligned} p_{xy} &= x(t)y^*(t) \\ \text{et} \quad p_{yx} &= y(t)x^*(t) \\ \text{d'où la relation :} \quad p_{xy} &= p_{yx}^* \\ \text{et dans le cas des signaux réels :} \quad p_{xy} &= p_{yx} = x(t)y(t). \end{aligned}$$

**d** – La puissance moyenne d'interaction de deux signaux sur un intervalle $T$ s'exprime au moyen de la relation :

$$P_{xy}(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} x(t)y^*(t)\, \mathrm{d}t, \tag{G.5}$$

$$P_{yx}(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} y(t)x^*(t)\, \mathrm{d}t, \tag{G.6}$$

ce qui permet d'écrire :

$$P_{xy}(t_0, T) = P_{yx}^*(t_0, T),$$

et dans le cas de signaux réels :

$$P_{xy}(t_0, T) = P_{yx}(t_0, T) = \frac{1}{T} \int_{t_0}^{t_0+T} x(t)y(t)\, \mathrm{d}t.$$

**e** – On apporte quelques modifications à ces définitions si l'on a affaire à des signaux transitoires, lesquels sont, par définition, à support borné. (Ils sont nuls hors d'un intervalle fini bien déterminé.) On préfère alors définir l'énergie d'une autre façon au moyen de la relation :

$$E_{xy} = \int_{t_0}^{t_1} x(t)y^*(t)\, \mathrm{d}t = \int_{-\infty}^{+\infty} x(t)y^*(t)\, \mathrm{d}t. \tag{G.7}$$

***f* – *Rappel du théorème de Parseval*** – Il s'agit, pour le physicien, de l'expression du principe de conservation de l'énergie (ou de l'information) quel que soit l'espace de représentation :

$$\int_{-\infty}^{+\infty} x(t)y^*(t)\, \mathrm{d}t = \int_{-\infty}^{+\infty} X(\omega)Y^*(\omega)\, \mathrm{d}\omega,$$

où $X(\omega)$ et $Y(\omega)$ sont respectivement les transformées de Fourier de $x(t)$ et $y(t)$. Dans le cas où $x(t) = y(t)$, on obtient :

$$\int_{-\infty}^{+\infty} |x(t)|^2(t)\, \mathrm{d}t = \int_{-\infty}^{+\infty} |X(\omega)|^2\, \mathrm{d}\omega.$$

***Densité de puissance ou densité spectrale énergétique*** – C'est la fonction $S_{xx}(\omega_0)$ égale à la dérivée de la puissance $P(\omega)$ par rapport à la fréquence $\omega$ (il s'agit de la puissance définie au paragraphe 1 exprimée en fonction de la fréquence) :

$$S_{xx}(\omega_0) = \frac{\mathrm{d}P(\omega)}{\mathrm{d}\omega} \quad \text{pour } \omega = \omega_0.$$

***Types de signaux*** – Les signaux rencontrés au cours des expériences réelles sont essentiellement de trois types :

**a.** Les signaux périodiques,
**b.** les signaux « transitoires » qui ne sont pas périodiques mais qui sont localisés dans l'espace direct,
**c.** le bruit.

Aux signaux évoqués en a et b se superpose, dans le cas général, un bruit. Celui-ci est bien représenté par une fonction aléatoire ayant la propriété de stationnarité (la moyenne, l'écart quadratique etc. sont indépendants du temps). De plus il n'est ni périodique ni localisé dans l'espace direct.

## 2. La corrélation et ses propriétés

Sans nuire à la généralité, nous limiterons notre propos aux phénomènes dépendant du temps. Plus précisément, il s'agira le plus souvent de considérer un ensemble de fonctions représentant le même phénomène physique donc de fonctions réelles.

Voici quelques exemples : les enregistrements de tensions délivrées par des capteurs pour réaliser un encéphalogramme, un électrocardiogramme, l'enregistrement de la température en un lieu donné, l'enregistrement du taux de $CO_2$ dans une ville donnée, l'enregistrement du spectre d'émission d'un astre, etc.

La question alors qui se pose est de savoir si le processus $P_x$ qui donne naissance aux fonctions $x(t)$ et le processus physique $P_y$ qui donne naissance aux fonctions $y(t)$ sont en relation. La réponse se trouve dans la fonction de corrélation qui consiste à comparer $x(t)$ et $y(t)$ à différents instants. L'idée repose sur la mesure de la surface commune au graphe de $x(t)$ et au graphe de $y(t)$ décalés entre eux d'une quantité $\tau$. Cette surface est une fonction du décalage $\tau$. Plus le signal $y(t)$ aura tendance à reproduire l'allure du signal $x(t)$ après un décalage $\tau_0$, plus la surface commune sera « grande » et plus les processus physiques seront corrélés. Si les signaux $x(t)$ et $y(t)$ sont différents, la fonction de corrélation s'appelle **intercorrélation**, tandis que si les signaux sont les mêmes on parle de fonction d'**autocorrélation**. L'expression de la fonction de corrélation s'écrit au temps $t_0$ :

$$C_{xy}(t_0, \tau) = \lim_{T \to \infty} \frac{1}{T} \int_{t_0}^{t_0+T} x(t) y(t-\tau)\, \mathrm{d}t, \tag{G.8}$$

$$C_{xx}(t_0, \tau) = \lim_{T \to \infty} \frac{1}{T} \int_{t_0}^{t_0+T} x(t) x(t-\tau)\, \mathrm{d}t. \tag{G.9}$$

Un exemple physique de fonction de corrélation : les interférences.

On peut imaginer qu'il s'agit d'interférences lumineuses. On désigne alors par $x(t)$ l'amplitude de la radiation émise par une source lumineuse. Après la différence de marche $\tau$, les interférences seront données par la fonction $\Psi(t, \tau)$ :

$$\Psi(t, \tau) = x(t) + x(t - \tau).$$

Comme l'intensité $I_M$ au point $M$ est la moyenne du carré de l'amplitude, nous écrivons :

$$I_M = \langle \Psi^2 \rangle = \langle x(t)^2 \rangle + \langle x(t - \tau)^2 \rangle + 2\langle x(t)x(t - \tau) \rangle.$$

S'il s'agit d'une source lumineuse non cohérente :

$$I_M = 2[\sigma_x^2 + C_{xx}(\tau)],$$

où :

$$\sigma_x^2 = \langle x(t)^2 \rangle = \frac{1}{T} \int_0^T x^2(t)\, \mathrm{d}t.$$

$C_{xx}(\tau)$ détermine la variation (spatiale) de l'énergie moyenne en fonction de la différence de marche $\tau$. Ici la fonction de corrélation donne la forme des franges d'interférences.

## 2.1. Cas de la fonction d'autocorrélation

Il est intéressant de donner les principales propriétés de la fonction d'autocorrélation concernant les trois types de signaux. Quel que soit le signal :

1. La fonction d'autocorrélation $C_{xx}(\tau)$ est paire, soit :

$$C_{xx}(\tau) = C_{xx}(-\tau)$$

   et sa transformée de Fourier est réelle et paire.
2. Le maximum de la fonction de corrélation est atteint pour la valeur $\tau = 0$. $C_{xx}(0)$ représente la valeur quadratique moyenne du signal.
3. Sauf pour un signal périodique, $\lim_{\tau \to \infty} C_{xx}(\tau) = 0$.
4. **Théorème de Wiener-Kinchine** (sans démonstration). La densité spectrale d'énergie et la fonction d'autocorrélation sont transformées de Fourier l'une de l'autre.

$$C_{xx}(t) \underset{TF^{-1}}{\overset{TF}{\longleftrightarrow}} S_{xx}(\omega).$$

Dans le cas où un signal n'a pas de composante continue $S_{xx}(0) = 0$, alors à partir de :

$$S_{xx}(\omega) = \int_{-\infty}^{+\infty} C_{xx}(t) \exp(2\pi j \omega t)\, \mathrm{d}t,$$

on déduit que :

$$S_{xx}(\omega) = 0 = \int_{-\infty}^{+\infty} C_{xx}(t)\, \mathrm{d}t,$$

ce qui signifie que la fonction d'autocorrélation a une surface algébrique nulle.

Dans le cas d'un signal périodique $e(t)$ de période $T$, la fonction d'autocorrélation est également périodique de période $T$. Dans ce cas, le signal est décomposable en série de Fourier, soit :

$$e(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[ a_k \cos \frac{2\pi kt}{T} + b_k \sin \frac{2\pi kt}{T} \right],$$

on en déduit la fonction d'autocorrélation (définie sur une période) :

$$C_{xx}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} e(t)e(t-\tau)\, \mathrm{d}t = \frac{1}{T} \int_{-T/2}^{T/2} \sum_{k=1}^{\infty} \left[ a_k \cos \frac{2\pi kt}{T} + b_k \sin \frac{2\pi kt}{T} \right]$$
$$\times \sum_{k=1}^{\infty} \left[ a_k \cos \frac{2\pi k(t-\tau)}{T} + b_k \sin \frac{2\pi k(t-\tau)}{T} \right] \mathrm{d}t = \frac{a_0^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} \left[ a_k^2 + b_k^2 \right] \cos \frac{2\pi k\tau}{T}.$$

### 2.2. Cas de la fonction d'intercorrélation

Dans ce cas, il n'y a plus de propriétés extrémales ni de parité.

***a*** – $C_{xy}(t) = C_{yx}(-t)$.

***b*** – Le théorème de Wiener (1894–1952) - Kinchine (1894–1959) demeure :

$$C_{xy}(t) \underset{TF^{-1}}{\overset{TF}{\longleftrightarrow}} S_{xy}(\omega).$$

$S_{xy}(\omega)$ est la densité énergétique d'interaction, elle est liée à l'information échangée entre les deux signaux.

Dans le cas de deux signaux périodiques de même période $T$, la fonction d'autocorrélation est également périodique de période $T$.

## 3. Applications de la corrélation

La corrélation trouve des applications fondamentales dans les domaines suivants :

### 3.1. Détection et extraction d'un signal périodique noyé dans du bruit

On considère un signal $e(t)$ périodique, de période $T_e$ auquel s'est superposé un bruit $\beta(t)$. Le signal délivré est alors :

$$s(t) = e(t) + \beta(t).$$

On désigne par $C_{ss}(\tau)$ la fonction d'autocorrélation :

$$C_{ss}(\tau) = \frac{1}{T} \int_0^T [e(t) + \beta(t)][e(t-\tau) + \beta(t-\tau)]\, \mathrm{d}t,$$

qui s'écrit encore :

$$C_{ss}(\tau) = C_{ee}(\tau) + C_{\beta\beta}(\tau) + C_{\beta e}(\tau) + C_{e\beta}(\tau).$$

Puisque le bruit est indépendant du signal $e(t)$, on en déduit que :

$$C_{\beta e}(\tau) = C_{e\beta}(\tau) = 0.$$

Par ailleurs, nous avons dit que $C_{\beta\beta}(\tau)$ tendait vers zéro quand la période d'intégration $T_i$ tendait vers l'infini. Bien que la période $T_i$ ne tende pas vers l'infini mais est « suffisamment grande », on pourra admettre que :

$$C_{ss}(\tau) = C_{ee}(\tau).$$

Ainsi, les extremums successifs donnent la période $T_e$ du signal.

Comment extraire ce signal périodique?

C'est la fonction $q(t) = \text{sha}(t)$ (peigne de Dirac) qui nous apporte la solution. Comme c'est une fonction paire, le résultat de la corrélation et le résultat de la convolution sont identiques à un facteur multiplicatif près (si l'on ne divise pas par $T_i$ la corrélation). Or la convolution d'une fonction périodique $e(t)$ avec une fonction peigne de même période laisse la fonction $e(t)$ inchangée. Ce qui revient au même de dire que l'intercorrélation de la fonction périodique $e(t)$ avec la fonction peigne de même période fournit la fonction $e(t)$ elle-même. Donc :

$$C_{sq}(\tau) = C_{eq}(\tau) + C_{\beta q}(\tau)$$

$C_{\beta q}(\tau)$ est nulle et comme $C_{eq}(\tau) = e(\tau)$, on obtient en définitive :

$$C_{sq}(\tau) = e(\tau).$$

***Remarque*** : Lorsque le signal est susceptible d'être reproduit aux erreurs près dues au bruit, la somme des signaux successifs permet d'éliminer statistiquement le bruit. On a enregistré $n$ fois le signal $s_i(t) = e_i(t) + \beta_i(t)$. Nous écrivons alors :

$$s(t) = \frac{1}{n}\sum_{k=1}^{n} s_i(t) = \frac{1}{n}\sum_{k=1}^{n} e_i(t) + \frac{1}{n}\sum_{k=1}^{n} \beta_i(t) = \frac{1}{n}\sum_{k=1}^{n} e_i(t)$$

car la moyenne du bruit est nulle $\frac{1}{n}\sum_{k=1}^{n} \beta_i(t) = 0$.

## 3.2. Calcul de la densité spectrale d'énergie

La densité spectrale énergétique $S_{xy}(\omega)$ est la transformée de Fourier de la fonction de corrélation $C_{xy}(t)$ (théorème de Wiener-Kinchine). En général, c'est la fonction de corrélation que l'on sait calculer ; on en effectue la transformée de Fourier, et l'on obtient la densité spectrale d'énergie.

## 3.3. Recherche du maximum de la fonction d'intercorrélation

La recherche du maximum de la fonction d'intercorrélation fournit une mesure du temps de décalage entre une excitation $e(t)$ et la réponse $r(t)$ et par conséquent une mesure de la vitesse de propagation du phénomène. L'étude de la dispersion de la vitesse en fonction de la fréquence peut être abordée au moyen de la fonction d'intercorrélation puis du théorème de Wiener-Kinchine.

## 4. La convolution

Tout instrument de mesure modifie le signal d'entrée $e(t)$ pour délivrer un signal de sortie $s(t)$. Cela est lié au fait qu'un signal d'entrée infiniment bref (impulsion de Dirac) a une réponse finie. Cette réponse impulsionnelle $a(t)$ du système de mesure (supposé linéaire) est appelée fonction d'appareil.

La fonction d'appareil d'un dispositif électronique peut s'obtenir en attaquant l'entrée par une très brève impulsion. Bien qu'il ne s'agisse pas d'un signal temporel, signalons toutefois la fonction d'appareil d'un spectrographe qui peut être obtenue, dans le voisinage de $\lambda_0$, par un laser de longueur d'onde $\lambda_0$.

On peut voir maintenant de quelle manière un signal d'entrée $e(t)$ se trouve perturbé à la sortie $s(t)$ par une fonction d'appareil $a(t)$.

Pour cela il suffit de décomposer le signal d'entrée $e(t)$ en une suite «d'impulsions » de largeur $\Delta t$ et dont l'amplitude sera $e(k\Delta t)$ avec $k = 0, 1, 2, \ldots, n$. L'impulsion $e(k\Delta t)$ fournira à la sortie le signal :

$$s_k(t) = e(k\Delta t)a(t - k\Delta t)\Delta t.$$

Comme **l'instrument de mesure est supposé linéaire**, il y a additivité des $s_k(t)$ pour reconstituer le signal à la sortie, soit :

$$s(t) = \sum_{k=0}^{n} e(k\Delta t)a(t - k\Delta t)\Delta t.$$

Par ailleurs la fonction $a(t)$ étant nulle pour $t < 0$, cela revient au même d'écrire que $a(t-k\Delta t)$ est nulle pour $t < k\Delta t$. Donc $a(t)$ est à présent définie sur $(-\infty, +\infty)$ et l'on a la relation :

$$s(t) = \sum_{-\infty}^{+\infty} e(k\Delta t)a(t - k\Delta t)\Delta t.$$

Un classique passage à la limite obtenu en faisant tendre $\Delta t$ vers zéro donne le résultat :

$$s(t) = \int_{-\infty}^{+\infty} e(z)a(t - z)\,\mathrm{d}z,$$

l'écriture symbolique de cette équation étant $s(t) = e(t) * a(t)$.

***Remarque 1*** : En général, $s(t)$ et $a(t)$ sont les deux fonctions connues et c'est $e(t)$ qui est la fonction inconnue. L'équation de convolution, qui est pourtant une équation fonctionnelle linéaire, offre bien des difficultés quant à sa résolution. Sur le plan numérique, les fonctions sont toujours à support borné et l'intégrale est finie. Cette équation devient donc une intégrale de Fredholm de première espèce qui est typiquement **un problème mal posé** (*cf.* chapitre 17). Sa résolution nécessite l'usage de méthodes appropriées telles que la méthode des régularisants que nous avons évoqué dans un paragraphe spécial consacré à l'équation de convolution.

***Remarque 2*** : Les opérateurs de convolution et de corrélation se ressemblent beaucoup ; notamment si une des fonctions est symétrique ils ne différeront que d'une constante multiplicative. Dans les autres cas, la convolution consiste à renverser une des fonctions avant de procéder au même calcul que celui défini pour la corrélation.

### Fonctions propres de l'opérateur de convolution

En d'autres termes, il s'agit de rechercher les fonctions qui, convoluées par elles-mêmes, restent inchangées. Désignons par $\phi(t)$ une de ces fonctions propres ; elle doit obéir à l'équation :

$$\Phi(t)^*\Phi(t) = k\Phi(t),$$
$$\text{qui deviendra :} \quad \Psi(\omega)\Psi(\omega) = k\Psi(\omega)$$

dans l'espace de Fourier où $\Psi(\omega)$ est la transformée de Fourier de $\phi(t)$.

La solution au sens classique est $W(\omega) = \alpha \neq 0$ sur un compact $|\omega| < \omega_0$, et 0 ailleurs. C'est la fonction « porte » usuelle dont la transformée de Fourier inverse est :

$$2\alpha\omega_0 \frac{\sin(2\pi\omega_0 t)}{2\pi\omega_0 t}.$$

Il existe aussi une solution au sens des distributions ; il s'agit de $W(\omega) = \beta$ sur l'intervalle $(-\infty, +\infty)$ dont la transformée de Fourier inverse est $\beta\delta(t)$.

## 5. Notions sur le filtrage

Par définition, un filtre temporel est une opération d'atténuation — voire d'interruption — sur un signal. Désignons par $z(t)$ cette opération d'atténuation, par $s(t)$ le signal à filtrer et par $x(t)$ le résultat ; nous pouvons écrire :

$$x(t) = z(t) \cdot s(t).$$

Le spectre de fréquence est modifié par cette opération et en passant dans l'espace de Fourier, nous obtenons :

$$x(t) = z(t) \cdot s(t) \quad \begin{matrix} TF \\ \longleftrightarrow \\ TF^{-1} \end{matrix} \quad X(\omega) = Z(\omega)^* S(\omega).$$

On voit bien que tout filtrage temporel modifie le spectre du signal filtré. À titre d'exemple, on peut évoquer l'usage d'un potentiomètre (ou d'un interrupteur), et le simple fait de « monter » ou « baisser » le son sur un amplificateur d'audiofréquences modifie le spectre entendu. Pour s'en convaincre, il suffit de réaliser plusieurs fois ces opérations plus ou moins rapidement sur le premier poste radiophonique venu.

### 5.1. Exemple

La fonction « porte » laisse passer le signal sur l'intervalle $(-T, T)$ et sa transformée de Fourier $Z(\omega)$ est :

$$Z(\omega) = 2T \frac{\sin(2\pi T\omega)}{2\pi T\omega}.$$

$S(\omega)$ sera convolué par $Z(\omega)$ et plus $T$ sera « grand », moins l'effet du filtrage sera sensible.

***Remarque 1*** : Le fait de découper un événement temporel dans le temps afin de l'étudier a comme conséquence la modification de son spectre.

***Remarque 2*** : On peut appliquer l'opération de filtrage à la représentation fréquentielle (dans l'espace de Fourier). L'opération de convolution se fera dans l'espace direct.

### 5.2. Les filtres physiques

Aux instants $t < 0$, un filtre physique ne peut avoir qu'une réponse nulle (principe de causalité), et à l'impulsion $\delta(t)$ on observe la réponse $a(t)$ qui est nulle pour $t < 0$ et définie pour $t > 0$. Soit $A(\omega)$ sa transformée de Fourier que nous écrivons :

$$A(\omega) = U(\omega) + jV(\omega) \quad U \text{ et } V \text{ sont réels.}$$

$V(\omega)$ n'est jamais nulle car pour cela il faudrait que $a(t)$ soit symétrique par rapport à $t = 0$ ce qui n'est pas possible. On peut adopter une autre présentation :

$$A(\omega) = |A(\omega)| \exp[j\phi(\omega)]$$

où $|A(\omega)|$ est le module et $\phi(\omega)$ la phase. De là, on conclut que tout filtre physique déphase le signal d'entrée.

### 5.3. Cas du monochromateur

Il s'agit de concevoir un filtre qui ne laisse passer qu'une seule fréquence. Cela n'est pas réalisable. En effet, dans l'espace de Fourier, on doit obtenir deux raies infiniment étroites (*cf.* la fonction de Dirac) symétriques par rapport à $\omega = 0$ ou symétriques par rapport à l'axe $\omega = 0$. Dans l'espace direct, nous avons vu qu'il s'agissait soit d'une sinusoïde pure, soit d'une cosinusoïde pure lesquelles s'étendent toutes les deux sur $(-\infty\ ; +\infty)$.

Dès que l'on aborde les applications pratiques, on ne peut que tronquer ces fonctions et les considérer uniquement sur un intervalle fini. Cette opération de troncature est évidemment un filtrage temporel qui a comme conséquence immédiate l'élargissement des «raies » dans l'espace de Fourier. Transcrivons mathématiquement ces opérations :

Soient $e(t)$ le signal à filtrer et $E(\omega)$ sa transformée de Fourier. Il s'agit donc de convoluer $e(t)$ et $a(t) = \cos(2\pi\omega_0 t)$, mais on voit bien qu'on ne peut pas réaliser la convolution sur l'espace des $t$ tout entier, et il nous faut tronquer par la fonction porte $p(t)$ dont la transformée de Fourier est $P(\omega)$.

$$\begin{array}{lcl} & \overset{TF}{\longleftrightarrow} & \\ e(t)^* a(t) & \longleftrightarrow & E(\omega)A(\omega) \\ e(t)^*[a(t)p(t)] & \longleftrightarrow & E(\omega)[A(\omega)^* P(\omega)] \\ & \underset{TF^{-1}}{\longleftrightarrow} & \end{array}$$

## 6. Notion de bruit

Il y a diverses approches du bruit qui sont plus ou moins fécondes et riches de développements fructueux :

### 6.1. Le bruit blanc

On dit qu'un bruit est blanc lorsque son spectre de puissance est constant dans toute l'étendue des fréquences. Autrement dit, sa fonction d'autocorrélation est une impulsion de Dirac $\delta(t)$. Il s'agit là d'une analogie avec la lumière blanche.

D'un point de vue pratique, un tel bruit n'existe pas ; cependant, on dira que l'on a affaire à un bruit blanc lorsque la fonction est constante en moyenne dans la bande de fréquence utilisée par les systèmes considérés. Bien sûr la fonction d'autocorrélation calculée ne sera pas un pic de Dirac mais une courbe très étroite dans le temps, l'unité de temps étant les temps de corrélation considérés. Il s'ensuit que les hautes fréquences seront filtrées (le bleu s'il s'agit de la lumière) et il restera les basses fréquences (le rouge s'il s'agit de la lumière), d'où le nom de bruit rose.

### 6.2. Le bruit gaussien

Soit un bruit dont la moyenne est $m$ et l'écart quadratique moyen $\sigma$. On parle de bruit gaussien lorsque celui-ci suit une loi de probabilité $p(t)$ gaussienne à savoir :

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-m)^2}{2\sigma^2}\right].$$

Le grand intérêt de ce type de bruit repose sur la simplicité de sa formulation analytique qui ne fait appel qu'aux deux premiers moments. De plus, dans bien des cas, la somme de processus aléatoires non gaussiens tend vers un processus gaussien, c'est ce que l'on rencontre dans la théorie des erreurs (voir le théorème central limite). Attention, cela n'est pas systématique, une somme de processus de Cauchy reste un processus de Cauchy.

## 7. Éléments de bibliographie

A. ANGOT (1972) *Compléments de mathématiques*, Éditions Masson.
R. BOITE et H. LEICH (1980) *Les filtres numériques*, Éditions Masson.
J. MAX (1980) *Méthodes et techniques de traitement du signal et applications aux mesures physiques*, Éditions Masson.
A. PAPOULIS (1977) *Signal analysis*, McGraw-Hill.
J.C. RADIX (1970) *Introduction au filtrage numérique*, Eyrolles.
E. ROUBINE (1970) *Introduction à la théorie de la communication*, Tomes I, II et III, Masson.

# H | Problèmes et exercices

L'annexe I contient les corrigés de la plupart des exercices énoncés ici.

## 1. Généralités sur le calcul numérique

### 1.1. Calcul du nombre d'or

Soit la suite définie par les relations suivantes :

$$U_0 = 0 \qquad U_1 = 1 \qquad \ldots \qquad U_{n+1} = U_n + U_{n-1} \qquad \ldots$$

Cette suite s'appelle suite de Fibonacci (1180?–1250). Montrer que la suite de terme général $k_n = U_n/U_{n-1}$ tend vers une limite lorsque $n$ tend vers l'infini.

Effectuer un programme permettant de calculer les termes de cette suite $k_n$. Quel critère retiendra-t-on pour obtenir une approximation de la limite de la suite? Estimer la précision obtenue sur ce nombre.

On peut calculer directement cette limite $k$. On montrera que $k$ est la solution positive de l'équation du second degré : $\gamma^2 - \gamma - 1 = 0$

**N.B.** - La suite de Fibonacci est utilisée en informatique dans les problèmes de tris externes.

### 1.2. Tabulation des polynômes de Tchebycheff

Les polynômes de Tchebycheff sont définis de la manière suivante :

$$T_n(x) = \cos[n \arccos(x)] \quad \text{avec} \quad -1 \leq x \leq +1.$$

Établir la relation de récurrence suivante entre trois polynômes consécutifs :

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

et définir les deux premiers polynômes $T_0$ et $T_1$.

Effectuer un programme qui permette de calculer la valeur des polynômes quels que soient le nombre $n$ appartenant à l'ensemble des entiers positifs et la valeur de l'argument $x$ à l'intervalle $(-1, +1)$.

Quelles limitations viennent affecter cet algorithme? On comparera les résultats donnés par le calcul direct utilisant les fonctions inverses et l'utilisation de la relation de récurrence.

**N.B.** - Les polynômes de Tchebycheff jouent un rôle essentiel dans le problème de l'optimisation des approximations de fonctions (chapitre 8).

## 1.3. Calcul de la fréquence des notes de deux gammes

Calculer les fréquences des notes de la gamme naturelle et de la gamme tempérée dans l'intervalle compris entre le $la_3$ (440 Hz) et le $la_4$ (880 Hz).

## 1.4. Calcul de la constante d'Euler

La constante d'Euler que l'on rencontre lors du calcul des fonctions de Bessel de deuxième espèce est donnée par l'expression :

$$\gamma = \lim \left[ 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m} - \log_e(m) \right]$$

lorsque $m$ tend vers l'infini.

Réaliser un programme qui calcule $\gamma$ selon cet algorithme, et effectuer une estimation de la précision obtenue.

Comme la vitesse de convergence est loin d'être satisfaisante, on préfère utiliser une autre expression pour calculer $\gamma$ :

$$\sum_{k=1}^{m-1} \frac{1}{k} = \log_e(m) + \gamma - \frac{1}{2m} - \frac{1}{12m^2} + \frac{1}{120m^4} - \frac{1}{252m^6} + \cdots + (-1)^k \frac{B_k}{2km^{2k}} + \cdots$$

expression dans laquelle les $B_k$ sont les nombres de Bernoulli (*cf.* chapitre 12). On donne les premiers nombres :

$$B_1 = \frac{1}{6} \qquad B_2 = \frac{1}{30} \qquad B_3 = \frac{1}{42}$$
$$B_4 = \frac{1}{30} \qquad B_5 = \frac{5}{66} \qquad B_6 = \frac{691}{2\,730}$$
$$B_7 = \frac{7}{6} \qquad B_8 = \frac{3\,617}{510} \qquad B_9 = \frac{43\,867}{798}$$
$$B_{10} = \frac{174\,611}{330} \qquad B_{11} = \frac{854\,513}{138} \qquad B_{12} = \frac{236\,364\,091}{2\,730} \cdots$$

Les nombres de Bernoulli sont les coefficients du développement de :

$$\frac{t}{\exp(t) - 1} = 1 - \frac{t}{2} + B_1 \frac{t^2}{2!} - B_2 \frac{t^4}{4!} + \cdots + (-1)^{n+1} B_n \frac{t^{2n}}{(2n)!} + \cdots$$

et par ailleurs, ils sont reliés à la fonction $\zeta$ de Riemann laquelle permet d'obtenir les relations suivantes :

$$B_n = \frac{2S_{2n}(2n)!}{4^n \pi^{2n}}$$
$$\text{avec} \quad S_{2n} = 1 + \frac{1}{2^{2n}} + \frac{1}{3^{2n}} + \frac{1}{4^{2n}} + \cdots + \frac{1}{k^{2n}} \cdots$$

Effectuer un programme qui réalise le calcul de $\gamma$ selon cette procédure.

En machine (précision relative de $10^{-s}$), déterminer expérimentalement combien on doit utiliser de nombres de Bernoulli au minimum pour obtenir la meilleure précision avec la machine utilisée (une fois $m$ fixé). Ou bien calculer $m$ sachant que l'on arrête le développement à l'ordre $j$ compris, c'est-à-dire après $B_j$. Pour mener à bien les calculs, on utilisera le fait que l'erreur sur $\gamma$ est inférieure à deux fois le module du premier terme de la série abandonné (*cf.* G. Valiron, Théorie des fonctions, Masson, p. 218 et 219).

La convergence est très rapide, cependant, la série infinie est-elle convergente?

## 1.5. Calcul numérique des dérivées d'un polynôme à coefficients réels pour la valeur x = r

On considère un polynôme $P_n(x)$ à coefficients réels que l'on écrit :

$$P_n(x) \sum_{k=0}^{n} a_k x^k \quad \text{avec} \quad a_0 \neq 0.$$

***a*** – Rappeler comment calculer effectivement la valeur numérique du polynôme $P_n(x)$ pour la valeur $x = r$ au moyen du schéma de Horner. On écrira les étapes du calcul au moyen d'une relation de récurrence que l'on explicitera donnant une suite que l'on appellera $v_k$ par exemple.

***b*** – On divise le polynôme $P_n(x)$ par $(x - r)$, et l'on écrit :

$$P_n(x) = (x - r)P_{n-1}(x) + R_n. \tag{H.1}$$

On poursuit la division des polynômes $P_{n-q}(x)$ par $(x - r)$. Écrire la relation de récurrence qui donne le polynôme $P_{n-q}(x)$, et montrer que : $P_1(x) = (x - r)P_0(x) + R_1$, par la suite on posera $P_0(x) = R_0$ pour des raisons de commodité d'écriture et d'homogénéité.

Donner la relation de récurrence qui permet de calculer les coefficients du polynôme $P_{k-1}(x)$ en fonction des coefficients du polynôme $P_k(x)$.

***c*** – Réécrire l'expression (H.1) en éliminant tous les $P_k(x)$ : $P_n(x)$ s'exprime en fonction de $x, r$ et des $R_k$.

***d*** – Donner l'expression du développement du polynôme $P_n(x)$ en série de Taylor au voisinage de $x = r + h$ en fonction de $x, r$ et des dérivées de $P_n(x)$. En déduire la valeur de $R_k$.

***e*** – Montrer comment calculer effectivement tous les $R_k$ à l'aide du schéma de Horner.

## 1.6. Calcul numérique de la fonction ζ de Riemann

***Rappel du théorème de Cauchy*** – Soit $f(x)$ une fonction positive et décroissante pour $x \geq 1$, alors l'intégrale $I = \int_1^\infty f(x)\ \mathrm{d}x$ et la série $f(k)$ (où $k$ est un entier strictement positif) convergent ou divergent ensemble.

Soit la fonction de Riemann

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \cdots + \frac{1}{n^s} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n^s}\,.$$

***a*** – Associer l'intégrale $I = \int_1^\infty f(x)\ \mathrm{d}x$ à la fonction $\zeta(s)$ en explicitant $f(x)$, puis donner les différentes expressions de $I$ (obtenues par quadratures simples sous le signe somme) en fonction de $s$. En déduire la condition de convergence de la série ; au préalable, on aura représenté sur un graphique les termes $1, \frac{1}{2^s}, \frac{1}{3^s}, \ldots$ ainsi que la fonction associée $f(x)$.

***b*** – On désigne par $S_k(s)$ la série de Riemann tronquée à l'ordre $k$ :

$$S_k(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \cdots + \frac{1}{k^s}\,.$$

En approchant $\zeta(s)$ par $S_k(s)$ on effectue une erreur de troncature strictement mathématique (indépendante de la troncature numérique des nombres) :

$$e_k = \frac{1}{(k+1)^s} + \frac{1}{(k+2)^s} + \frac{1}{(k+3)^s} + \cdots$$

Au moyen de l'intégrale associée, proposer un encadrement de $e_k$, puis en déduire une méthode de calcul de $\zeta(s)$ lorsque celle-ci converge.

### 1.7. Recherche d'une tangente commune à deux courbes

On considère deux fonctions $y = f(x)$ et $y = g(x)$. À l'intérieur d'un domaine $D$ du plan réel $(Ox, Oy)$ elles admettent chacune, hormis les valeurs prises à la frontière du domaine, un extremum et un seul qui est un minimum. On suppose que, dans $D$, les fonctions $f(x)$ et $g(x)$ sont continues, dérivables au moins deux fois et qu'elles ne possèdent pas de points d'inflexion. Dans la suite, on s'intéressera aux courbes représentatives, $C_1$ pour $y = f(x)$ et $C_2$ pour $y = g(x)$, qui admettent une tangente commune dans $D$. On se propose de déterminer numériquement cette tangente commune.

***a*** – Soit $(x_0, y_0)$ un point $A$ du plan par lequel il est possible de tracer une tangente quelconque à chacune des courbes $C_1$ et $C_2$. Écrire les coordonnées du point $(x_f, y_f)$ tangent à $y = f(x)$, puis les coordonnées du point $(x_g, y_g)$ tangent à $y = g(x)$, la tangente étant issue de $A$. Indiquer une méthode de calcul numérique de ces coordonnées.

***b*** – S'il est possible de tracer, à partir de $A$, une tangente à une courbe alors il est possible, en général, d'en tracer une seconde. Laquelle conviendra-t-il de retenir pour la solution qui nous intéresse?

***c*** – Pour l'instant, il n'y a aucune raison pour que ces deux tangentes constituent une tangente commune, et nous allons déplacer le point $A$ sur la droite $x = x_0$ pour obtenir cette tangente commune. Donner un algorithme qui permette d'obtenir numériquement cette tangente commune. Comment devra-t-on choisir $x_0$ pour obtenir la précision optimum?

***d*** – Les courbes représentatives $C_1$ et $C_2$ ne sont plus données par des fonctions explicites mais par des fonctions implicites qui s'écrivent :

$$\Phi(x, y) = 0 \quad \text{et} \quad \Psi(x, y) = 0.$$

Ici encore, on suppose que les fonctions ont les bonnes propriétés usuelles de régularité (continuité et dérivabilité à l'ordre deux, pas de points d'inflexion dans $D$). Donner une procédure qui permette de calculer numériquement les coordonnées $(x_g, y_g)$ et $(x_f, y_f)$ des points tangents aux courbes, les deux tangentes étant issues de $A$. Puis, proposer une procédure pour obtenir numériquement la tangente commune aux deux courbes.

## 2. Algorithmes accélérateurs de la convergence des suites

### 2.1. L'epsilon-algorithme scalaire

***a*** – En réaliser la programmation sous forme de sous-programme. Application à une suite lentement convergente :

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + (-1)^n \frac{1}{2n+1} + \cdots$$

***b*** – Appliquer l'epsilon-algorithme au calcul de

$$\gamma = \lim \left( 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m} - \log_e(m) \right).$$

À quel ordre $m$ obtient-on le meilleur résultat avec une machine donnée?

### 2.2. L'epsilon-algorithme scalaire

En réaliser la programmation sous forme de sous-programme.

## 3. Les développements asymptotiques

### 3.1. Développement asymptotique

***a*** – Soit la fonction : $f(x) = \int_x^\infty t^{-1}\exp(x-t)\,\mathrm{d}t$ avec $x > 0$ encore appelée exponentielle intégrale. En intégrant successivement par parties montrer que $f(x)$ admet un développement asymptotique dont on explicitera les coefficients. Montrer que la série est bien divergente quand $n$ tend vers l'infini. Calculer l'erreur $\varepsilon_n(x)$ commise lorsque l'on remplace $f(x)$ par son développement $S_n(x)$ et montrer que $\varepsilon_n(x)$ passe par un minimum qui dépend de $n$ pour une valeur fixée de $x$. Montrer qu'en machine ce minimum est obtenu dans le voisinage de $\varepsilon_n(x) \approx \varepsilon_{n+1}(x)$ ou $\varepsilon_{n-1}(x) \approx \varepsilon_n(x)$. On désigne par $N$ la valeur de $n$ pour laquelle on obtient ce minimum ($N$ est une fonction de $x$).

Effectuer un programme qui, pour une valeur quelconque de $x$, détermine $N$ puis calcule la somme $S_N(x)$ et enfin l'erreur mathématique $\varepsilon_N(x)$.

Appliquer l'epsilon-algorithme scalaire (*cf.* chapitre 2) aux sommes partielles $S_n(x)$ pour des valeurs de $n$ quelconques plus petites et plus grandes que $N$, pour des valeurs quelconques de $x$ (mêmes voisines de l'unité).

***b*** – On propose la même étude sur les fonctions suivantes :

$$\Theta(x) = \frac{2}{\pi}\int_0^x \exp(-t^2)\,\mathrm{d}t \qquad \text{et} \qquad \operatorname{cerf}(x) = 1 - \Theta(x).$$

On montrera que

$$\operatorname{cerf}(x) = \frac{-x^2}{x\pi}\left\{1 - \frac{1}{2x^2} + \frac{1\cdot 3}{2^2x^4} - \frac{1\cdot 3\cdot 5}{2^3x^6} + \cdots + (-1)^n\frac{1\cdot 3\cdot 5\cdots(2n-3)}{2^{n-1}x^{2n-2}} + \cdots\right\}.$$

## 4. Résolution des équations numériques

### 4.1. Racines d'une équation f(x) = 0

***a*** – Réaliser un programme qui donne les racines d'une équation $f(x) = 0$ par dichotomie. Compte tenu de la précision relative de la machine utilisée qui est de $10^{-n}$, donner le nombre de tours de dichotomie suffisant pour obtenir la meilleure précision sur la racine. On aura soin d'injecter la valeur de la racine trouvée dans l'équation $f(x) = 0$...

***b*** – Réaliser un programme qui calcule les racines selon la méthode itérative du premier ordre.

***c*** – Réaliser un programme qui calcule les racines selon la méthode de Newton (méthode itérative du deuxième ordre).

***d*** – ***Application*** – (Équation rencontrée lors de l'étude du corps noir)
Chercher la racine non nulle de l'équation transcendante :

$$y + 5\exp(-y) - 5 = 0.$$

Comparer les différentes méthodes, estimer la précision obtenue sur chacune des valeurs.

## 4.2. Méthode de Bairstow

Dans le cas particulier des polynômes, plutôt que de faire usage des méthodes du genre de celles programmées au paragraphe 1, on préfère utiliser des programmes spécifiques dont la méthode de Bairstow est un exemple typique. Programmation de la méthode de Bairstow.

***Applications*** – Rechercher les racines du polynôme de Legendre d'ordre $n$ que l'on note $P_n(x)$. Les polynômes sont donnés par les relations de récurrence :

$$P_0(x) = 1$$
$$P_1(x) = x$$
$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0.$$

À partir de quel ordre $n$ commence-t-on à avoir de sérieux ennuis de calculs avec une machine dont les nombres ont une précision relative de $10^{-s}$ ?

**N.B.** - Les zéros des polynômes de Legendre jouent un rôle fondamental dans la méthode d'intégration numérique de Gauss-Legendre.

Même problème pour les polynômes de Laguerre et les polynômes d'Hermite qui sont donnés respectivement par les relations de récurrence suivantes :

$$L_0(x) = 1$$
$$L_1(x) = 1 - x$$
$$L_{n+1}(x) + (x - 2n - 1)L_n(x) + n^2 L_{n-1}(x) = 0$$

et

$$H_0(x) = 1$$
$$H_1(x) = x$$
$$H_{n+1}(x) - xH_n(x) + nH_{n-1}(x) = 0.$$

## 4.3. Calcul de la fonction arctan(x)

Connaissant la fonction tangente, $a = \tan(x)$, on se propose de calculer $x = \arctan(a)$. Montrer que l'on a l'identité :

$$x = \arctan(a) - \arctan[\tan(x_0)] + x_0,$$

puis en déduire que :

$$x = x_0 + \arctan\left(\frac{a - \tan(x_0)}{1 + a\tan(x_0)}\right) = x_0 + \arctan[F(x_0)].$$

Donner le développement de $x$ en fonction de $X = F(x_0)$. À quelle condition la série converge-t-elle ?

Calculer la valeur approchée de $\arctan\left(\frac{1}{\sqrt{3}}\right)$ en choisissant comme première approximation $x_0 = 0{,}8$ et en se limitant à l'ordre 3. Quelle est la précision obtenue sur ce résultat ?

## 4.4. Racine d'une équation f(x) = 0

On considère une fonction $f(x)$ continue dans un domaine $\Delta$. À l'intérieur de ce domaine $\Delta$, $f(x)$ admet **une seule racine** $X$ telle que $f(X) = 0$. On connaît une première approximation de $X$ que l'on note $x_0$. On se propose d'obtenir une meilleure approximation $x_1$ de la manière suivante :

En $A_0$ de coordonnées $[x_0, f(x_0)]$, on trace la tangente à la courbe $y = f(x)$, puis on effectue une projection orthogonale du point $B_0$ de coordonnées $(x_0, 0)$ sur cette tangente. On désigne par $C_0$ cette projection dont les coordonnées sont $(x_1, y_1)$.

***a*** – Donner l'expression formelle de $x_1$ en fonction de $x_0$, de la fonction $f$ et éventuellement de ses dérivées. Proposer un processus itératif du type :

$$x_{n+1} = \Phi(x_n),$$

on donnera explicitement $\Phi(x)$.

***b*** – On désigne par $e_n$ l'écart à la solution $X$ après le $n^e$ tour d'itération : $e_n = X - x_n$.

Montrer qu'au premier ordre on peut écrire : $e_{n+1} = ke_n$. On explicitera $k$ en fonction de $\Phi$ et de ses dérivées (ou de $f$ et de ses dérivées).

Donner les conditions pour que la suite des approximations soit toujours convergente vers $X$.

***c*** – En s'aidant de quelques figures simples, retrouver les conditions sur $f$ (ou ses dérivées) pour que l'approximation $x_1$ soit effectivement meilleure que $x_0$. Représenter au moins un cas pour lequel la méthode ne converge pas vers $X$.

***d*** – Comparer cette méthode à celle de Newton, c'est-à-dire que l'on donnera les avantages et les inconvénients que ces deux méthodes présentent comparées l'une à l'autre.

### 4.5. Racines d'un polynôme à coefficients complexes

On considère un polynôme de la variable complexe à coefficients complexes :

$$P_n(z) = (\alpha_0 + j\beta_0)z^n + (\alpha_1 + j\beta_1)z^{n-1} + (\alpha_2 + j\beta_2)z^{n-2} + \cdots + (\alpha_n + j\beta_n)$$
$$\text{avec } \alpha_0 + j\beta_0 \neq 0.$$

On désire obtenir une racine complexe de ce polynôme que l'on a localisée dans le domaine complexe $\Delta$. Pour ce faire, on divise $P_n(z)$ par le monôme $D_1(z) = z - (\rho + j\sigma)$.

***a*** – Montrer que le quotient, noté $Q_{n-1}(z)$, est un polynôme de degré $n-1$ et le reste est une constante complexe notée $R = \Phi + j\Psi$. On écrit le polynôme $Q_{n-1}(z)$ sous la forme :

$$Q_{n-1}(z) = (\gamma_0 + j\delta_0)z^{n-1} + (\gamma_1 + j\delta_1)z^{n-2} + (\gamma_2 + j\delta_2)z^{n-3} + \cdots + (\gamma_{n-1} + j\delta_{n-1}).$$

***b*** – Expliciter les coefficients $\gamma_k$ et $\delta_k$ en fournissant les relations de récurrence auxquelles ils obéissent. On précisera convenablement le début et la fin des relations de récurrence. Donner les expressions de $\Phi$ et de $\Psi$. Montrer que, pour que $(\rho + j\sigma)$ soit une racine, il faut et il suffit que les fonctions $\Phi$ et $\Psi$ soient nulles.

***c*** – $\Phi$ et $\Psi$ sont des fonctions implicites des deux variables $\rho$ et $\sigma$, rappeler la méthode de Newton permettant de les annuler simultanément. On écrira les relations en fonction $\Phi$ et $\Psi$ et de leurs dérivées partielles par rapport à $\rho$ et $\sigma$ : $\dfrac{\partial\Phi}{\partial\rho}$, $\dfrac{\partial\Phi}{\partial\sigma}$, $\dfrac{\partial\Psi}{\partial\rho}$ et $\dfrac{\partial\Psi}{\partial\sigma}$.

***d*** – Calculer effectivement $\partial\Phi/\partial\rho$ et $\partial\Psi/\partial\rho$ en dérivant les relations de récurrence des $\gamma_k$ et des $\delta_k$ par rapport à $\rho$ puis en posant $\dfrac{\partial\gamma_k}{\partial\rho} = t_{k-1}$ et $\dfrac{\partial\delta_k}{\partial\rho} = u_{k-1}$. Ici encore, on explicitera bien le début et la fin des relations de récurrence.

***e*** – Opérer de la même manière pour calculer $\partial\Phi/\partial\sigma$ et $\partial\Psi/\partial\sigma$. On dérivera les relations de récurrence des $\gamma_k$ et des $\delta_k$ par rapport à $\sigma$ et l'on posera $\dfrac{\partial\gamma_k}{\partial\sigma} = v_{k-1}$ et $\dfrac{\partial\delta_k}{\partial\sigma} = w_{k-1}$.

***f*** – Donner l'algorithme permettant de calculer la racine appartenant au domaine $\Delta$. En supposant que toutes les racines aient été localisées chacune dans un domaine $\Delta_k$, proposer une méthode générale permettant de calculer toutes les racines complexes du polynôme $P_n(z)$ à coefficients complexes.

## 4.6. Résolution d'un système non linéaire

On se propose de résoudre numériquement un système non linéaire de $n$ équations à $n$ inconnues qui se note :

$$F_q(x_1, x_2, x_3, \ldots, x_n) = 0 \quad \text{avec} \quad q = 1, 2, 3, \ldots, n.$$

***a*** – Montrer que l'on peut toujours écrire ce système sous la forme :

$$x_p = \Phi_p(x_1, x_2, x_3, \ldots, x_n) \quad \text{avec} \quad p = 1, 2, 3, \ldots, n.$$

***b*** – Partant de valeurs approchées $\left\{x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right\}$ de la racine recherchée, on forme les suites récurrentes :

$$x_p^{(k)} = \Phi_p\left(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)}\right) \quad \text{avec} \quad p = 1, 2, \ldots, n.$$

On se propose d'étudier la convergence d'un tel processus itératif sur l'ensemble des valeurs $x_p^{(k+1)}$. Pour cela on désigne par $x_j^*$ la solution exacte du système, et l'on note par $\xi_j$ l'écart entre la solution et sa valeur approchée au $k^{\text{e}}$ tour :

$$x_j^* = x_j^{(k)} + \xi_j.$$

Après avoir développé la fonction $\Phi p$ en série de Taylor au premier ordre au voisinage de la valeur obtenue au $k^{\text{e}}$ tour d'itération, donner l'expression de $x_p^{(k)}$ en fonction de $\Phi p$ et de ses dérivées partielles.

***c*** – L'erreur après la $k^{\text{e}}$ itération est définie au moyen de l'expression :

$$e_j^{(k)} = x_j^* - x_j^{(k)}.$$

Donner l'expression de $\xi_j$ au $k^{\text{e}}$ tour d'itération, puis, donner l'expression de $e_p^{(k+1)}$ en fonction des $e_q^{(k)}$.

On peut remarquer que les quantités $e_j^{(k+1)}$ sont les composantes d'un vecteur $E^{(k+1)}$, alors montrer que $E^{(k+1)}$ peut s'écrire sous forme matricielle :

$$E^{(k+1)} = ME^{(k)}.$$

On explicitera la matrice $M$.

***d*** – Soit $E^{(0)}$ le vecteur des erreurs sur les valeurs initiales. Donner l'expression de $E^{(k+1)}$ en fonction de $E^{(0)}$.

Quelle condition doit remplir la matrice $M$ pour que le processus soit convergent ? (Application contractante).

***e*** – Dans le cas où la convergence est assurée, proposer une amélioration qui exploite mieux les résultats des calculs et qui, par conséquent, augmente la vitesse de convergence de la méthode.

## 4.7. Résolution d'un système non linéaire

On considère un système non linéaire de deux équations à deux inconnues :

$$f(x,y) = 0$$
$$g(x,y) = 0.$$

On suppose qu'il existe un domaine $\Delta$ à l'intérieur duquel le système admet une solution et une seule ; en outre, on suppose que les fonctions possèdent des dérivées par rapport à $x$ et à $y$ jusqu'à l'ordre $n$ inclus, et que les fonctions et leurs dérivées sont continues dans $\Delta$ par rapport à $x$ et à $y$.

On se propose de calculer la solution $(X,Y)$ en minimisant la forme quadratique suivante :

$$\Phi^2(x,y) = f^2(x,y) + g^2(x,y).$$

***a*** – Montrer que ce second problème est équivalent au premier.

***b*** – On choisit comme première approximation de $(X,Y)$ un point $M_0(x_0,y_0)$ appartenant à $\Delta$. Dans un premier temps on considère que $y_0$ ne change pas et l'on se propose de chercher une meilleure approximation seulement de $x_0$ appelée $x_1$ telle que :

$$x_1 = x_0 + \xi.$$

Écrire le développement de $\Phi^2(x_0+\xi, y_0)$ au deuxième ordre en $\xi$ au voisinage de $x_0$ ; puis calculer $x$ pour que $\Phi^2(x,y)$ soit minimum.

***c*** – Maintenant, on maintient $x_1$ constant et l'on va calculer par le même procédé une meilleure approximation de $y_0$ appelée $y_1$ telle que :

$$y_1 = y_0 + \eta.$$

Donner la valeur de $\eta$.

***d*** – Comme les développements successifs sont limités au deuxième ordre, les valeurs de $\xi$ et de $\eta$ ne sont que des approximations. Proposer un algorithme qui permet d'atteindre $X$ et $Y$. Comment vérifier que les valeurs calculées sont bien une solution du système proposé?

## 4.8. Ordre d'un processus itératif

Soit $f(x)$ une application de $R$ sur $R$ continue et indéfiniment dérivable. À partir d'une valeur $x_0$ on génère une suite $x_{k+1} = f(x_k)$.

Si dans le domaine $D$ l'application est contractante, après un nombre fini d'itérations on obtiendra la relation $X = f(X)$ à la précision de la machine utilisée.

Soit $\Phi(x) = 0$ une fonction dont on désire calculer la racine unique $X$ localisée dans le domaine $\Delta$. Montrer que les recherches de la racine par approximations successives et par la méthode de Newton obéissent au schéma fonctionnel que nous avons défini.

La recherche d'une racine $X$ de l'équation $\Phi(x) = 0$ dans le domaine $\Delta$ consiste à associer à cette équation une autre équation équivalente $x = f(x)$ où $f(x)$ est une application contractante dans $\Delta$ de telle sorte que l'on ait $X = f(X)$. On appelle $e_k$ l'erreur après la $k^e$ itération et l'on pose :

$$e_k = X - x_k,$$

donner l'expression de $e_{n+1}$ en fonction de $e_n$, de $f(x)$ ainsi que des dérivées de $f(x)$. En déduire que $e_{n+1}$ est un polynôme (ou série entière) en $e_n$ que l'on écrit :

$$e_{n+1} = \alpha e_n + \beta e_n^2 + \gamma e_n^3 + \cdots$$

On explicitera les coefficients $\alpha$, $\beta$ et $\gamma$ en fonction des dérivées de $f(x)$.

Dans l'hypothèse où l'application est contractante $\lim_{n \longrightarrow \infty} e_n = 0$, cependant, en calcul numérique, on cesse les itérations à l'ordre $N$ ($N$ tours d'itération) et l'erreur $e_N$ est finie. Par définition, on dira que l'on a affaire à un processus itératif du premier ordre si $\alpha \neq 0$ et à un processus du deuxième ordre si $\alpha = 0$ et $\beta \neq 0$ et ainsi de suite. On comprend que la convergence sera d'autant plus rapide que le processus itératif contractant sera d'ordre plus élevé.

***1. Application à la méthode de Newton*** – On cherche la racine $X$ de l'équation $\Phi(x) = 0$ dans le domaine $\Delta$ où l'on suppose que la méthode de Newton est convergente.

***a*** – Rappeler rapidement la méthode de Newton, puis exprimer $f(x)$ en fonction de $\Phi(x)$, et calculer enfin les coefficients $\alpha$ et $\beta$ en fonction de $\Phi(x)$ et de ses dérivées.

***b*** – À quelle condition l'application est-elle contractante?

***2.*** – Associée à l'équation $\Phi(x) = 0$, on considère à présent la formule d'itération suivante :

$$f(x) = x - \frac{g(x)\Phi(x)}{g(x)\Phi'(x) + h(x)\Phi(x)},$$

où $h(x)$ et $g(x)$ sont des fonctions arbitraires dont on exige seulement la continuité et la dérivabilité à un ordre suffisant.

Ici encore on suppose que $\Phi(x)$ a une racine unique dans le domaine $\Delta$, domaine où l'application est contractante. Calculer les coefficients $\alpha$ et $\beta$.

Quelle est la condition qui permet d'annuler le coefficient $\beta$? Quel sera alors l'ordre du processus?

## 4.9. Résolution d'un système de deux équations à deux inconnues. Méthode de Kacmarz (1937)

Soit :

$$f(x, y) = 0$$
$$g(x, y) = 0$$

un système de deux équations à deux inconnues. On suppose que les deux fonctions sont continues et dérivables (à l'ordre nécessaire) et que dans un domaine $D$ ce système admet une solution unique $(X, Y)$.

***1. Cas où le système est linéaire*** – On écrit alors :

$$a_1 x + b_1 y + c_1 = 0 \quad \text{(droite } D_1\text{)}$$
$$a_2 x + b_2 y + c_2 = 0 \quad \text{(droite } D_2\text{)}.$$

Partant d'une approximation $A_0(x_0, y_0)$ appartenant au domaine $\Delta$, on projette perpendiculairement ce point sur la droite $D_1$, on obtient alors le point $A_1$ de coordonnées $(x_1, y_1)$. Ensuite on projette le point $A_1$ sur la droite $D_2$ ce qui nous donnera le point $A_2$ de coordonnées $(x_2, y_2)$. On poursuit la génération des points $A_k$ jusqu'à ce que l'on ait obtenu la précision souhaitée.

***a*** – Donner l'expression des coordonnées $x_1$ et $y_1$ en fonction de $x_0$, $y_0$, $a_1$, $b_1$ et $R_1$ cette dernière quantité étant ainsi définie : $R_1 = a_1x_0 + b_1y_0 + c_1$.

***b*** – Donner l'expression des coordonnées $x_2$ et $y_2$ en fonction de $x_1$, $y_1$, $a_2$, $b_2$ et $R_2$ cette dernière quantité étant ainsi définie : $R_2 = a_2x_1 + b_2y_1 + c_2$.

***2. Cas général*** – Linéariser le système

$$f(x, y) = 0$$
$$g(x, y) = 0.$$

Par ce procédé nous sommes ramené au problème précédent, il suffit de remplacer les coefficients $a_1, a_2, b_1, b_2$ et $R_1$, $R_2$ par les nouvelles valeurs fournies par la linéarisation ; donner les expressions de $a_1, a_2, b_1, b_2$ et $R_1$, $R_2$ en fonction de $f(x, y)$, $g(x, y)$ et de leurs dérivées partielles.

Donner alors les expressions de $x_k$, $y_k$ en fonction de $x_{k-1}$, $y_{k-1}$ puis $x_{k+1}$, $y_{k+1}$ en fonction de $x_k, y_k$.

Donner les expressions des différents coefficients calculés sur l'exemple suivant :

$$f(x, y) = x^2 + 4y^2 - 36 = 0$$
$$g(x, y) = x^2 + y^2 - 12x + 27 = 0$$

# 5. Éléments de calcul matriciel

## 5.1. Résolution d'un système linéaire

Soit un système de $n$ équations à $n$ inconnues. Réaliser un programme permettant la résolution de ce système linéaire par la méthode des pivots, puis le transformer en sous-programme.

Dans le cas où la matrice est mal conditionnée ou encore lorsque son déterminant est voisin de zéro, on rencontre des difficultés pour obtenir un résultat raisonnable. Voici une méthode permettant éventuellement de se tirer d'affaire.

On peut soupçonner de tels problèmes quand la permutation de lignes du système linéaire fournit des résultats différents les uns des autres et que les troncatures des nombres ne peuvent pas directement expliquer.

***a*** – On calcule une première valeur $|\Delta_0|$ déterminant par le produit des pivots. Pour obtenir une erreur minimum à chaque étape du calcul il faut amener non pas le pivot le plus grand (car cela conduit à amener en fin de calcul les pivots les plus petits qui seront entachés d'une importante erreur puisque le produit des pivots est une constante...) mais le pivot dont le module est le plus proche de $\sqrt[n]{|\Delta_0|}$ ce qui conduit à une erreur minimum sur l'ensemble des calculs.

***b*** – On calcule une seconde valeur $\Delta_1$ du déterminant en amenant chaque fois le pivot dont le module est le plus proche de $\sqrt[n]{|\Delta_0|}$. On répète l'opération jusqu'à ce que l'on obtienne des valeurs stables $\Delta_n$ et de $\sqrt[n]{|\Delta_n|}$.

***c*** – On inverse la matrice (ou l'on résout le système linéaire) en amenant chaque fois le pivot dont le module est le plus proche de $\sqrt[n]{|\Delta_n|}$.

***d*** – Réaliser un tel programme.

## 5.2. Système linéaire surdéterminé

On envisage l'étude d'un système linéaire de $n$ équations à $m$ inconnues avec $n \geq m$. Le système appelé $S_1$ se présente sous la forme :

$$\sum_{j=1}^{m} a_{lk} x_k + a_l = 0 \quad \text{avec} \quad k = 1, 2, \ldots, n$$

que l'on écrira en notation matricielle :

$$AX = B.$$

Au moyen de la méthode des moindres carrés, il s'agit de réduire ce système $S_1$ au système $S_2$ appelé système des équations normales, de $m$ équations à $m$ inconnues.

Montrer que le système des équations normales s'écrit :

$$A^T AX = A^T B$$

où $A^T$ est la matrice transposée de $A$.

Réaliser un programme qui établit le système des équations normales puis qui le résout numériquement.

Étudier les incertitudes qui entachent les résultats du calcul selon le procédé donné page 378.

## 5.3. Résolution d'un système linéaire volumineux

On considère un système linéaire inversible $AX = B$ dans lequel $A$ est une matrice carrée d'ordre $N$. Malheureusement la matrice $A$ est beaucoup trop volumineuse pour tenir dans la mémoire centrale du micro-ordinateur dont on dispose. On se propose de fractionner la matrice $A$ et le vecteur $B$ respectivement en quatre sous-matrices et deux sous-vecteurs de la manière suivante :

$$\begin{vmatrix} A_1 & A_2 \\ A_3 & A_4 \end{vmatrix} \cdot \begin{vmatrix} X_1 \\ X_2 \end{vmatrix} = \begin{vmatrix} B_1 \\ B_2 \end{vmatrix}$$

Montrer que le problème peut se ramener à la résolution des deux systèmes suivants :

$$A_1 X_1 + A_2 X_2 = B_1 \tag{H.2}$$

$$A_3 X_1 + A_4 X_2 = B_2 \tag{H.3}$$

et donner les conditions sur l'ordre des matrices pour que de telles équations puissent avoir un sens ; pour cela on désignera par $l_1$ et $k_1$ respectivement le nombre de lignes et le nombre de colonnes de la matrice $A_1$, $l_1$ et $k_1$ positifs non nuls strictement inférieurs à $N$.

On se propose d'éliminer le vecteur $X_1$ entre les deux équations (H.2) et (H.3). Donner l'expression de $X_1$ tirée de (H.2). En reportant cette expression dans (H.3), donner l'expression de $X_2$.

Expliquer la méthode à suivre pour calculer effectivement $X_2$. Pour cela on détaillera chacune des opérations à réaliser.

Détailler ensuite la méthode pour obtenir $X_1$.

## 5.4. Résolution d'un système linéaire par la méthode itérative de Jacobi

On considère un système linéaire de $n$ équations à $n$ inconnues que l'on écrit sous la forme matricielle usuelle :

$$A'X = B'$$

où $A'$ est une matrice d'ordre $n$ sur laquelle on fait l'hypothèse qu'elle est non singulière, $X$ le vecteur des inconnues à $n$ composantes et $B'$ le vecteur second membre à $n$ composantes.

***a*** – On désire se ramener à un système identique noté :

$$AX = B \tag{H.4}$$

de telle sorte que l'on puisse décomposer la matrice $A$ en une différence de deux matrices :

$$A = I - M \tag{H.5}$$

où $I$ est la matrice unité d'ordre $n$ et où $M$ est une matrice d'ordre $n$ n'ayant que des zéros sur la diagonale principale.

Les éléments de $A'$ sont notés $a'_{lk}$, ceux de $A$ notés $a_{lk}$, et ceux de $M$ sont notés $m_{lk}$.

Montrer que toujours la décomposition proposée est possible dans les conditions de l'énoncé. Exprimer les éléments de $A$ en fonction de ceux de $A'$ puis ceux de $M$. Donner l'expression des éléments $b_k$ du vecteur $B$ en fonction des éléments $b'_k$ du vecteur $B'$ et des éléments de $A'$.

***b*** – En remplaçant dans l'équation (H.4) la matrice $A$ par son expression donnée en (H.5), montrer que l'on peut générer un processus itératif de valeurs $X_k$ de $X$ qui permet effectivement de calculer la solution $X$. On écrira explicitement l'équation matricielle à laquelle ce processus obéit.

Partant du vecteur $X_0$ (dont toutes les composantes sont nulles par exemple), expliciter les vecteurs $X_1, X_2, X_3, \ldots, X_k$, en fonction de $I$, $M$ et $B$.

***c*** – Quelle est la limite $Z$ de $X_k$ quand $k$ tend vers l'infini? Montrer que $Z$ est bien solution du problème.

***d*** – On se propose d'étudier la convergence de la procédure. Montrer que la suite des $X_k$ converge à condition que $M^p$ tende vers zéro quand $p$ tend vers l'infini. Quelle propriété doit vérifier la matrice $M$ pour qu'il en soit ainsi?

***e*** – Quand cette dernière condition n'est pas vérifiée, le processus est divergent. Proposer néanmoins une procédure qui permette quand même l'obtention de la solution $X$ en exploitant les données fournies par l'algorithme divergent.

### 5.5. Résolution d'un système linéaire dépendant d'une matrice symétrique (méthode de Choleski)

***1.*** – On considère une matrice carrée $A$ d'ordre $n$ dont les éléments sont notés $a_{lk}$. On suppose que cette matrice est régulière (ou non singulière : son déterminant est différent de zéro). On décompose cette matrice en un produit de deux matrices triangulaires $L$ et $S$, $L$ étant triangulaire inférieure avec des éléments quelconques sur la diagonale, $S$ étant triangulaire supérieure avec des 1 sur la diagonale, de telle sorte que $A = LS$. On désigne par $l_{lk}$ les éléments de $L$ et par $s_{lk}$ les éléments de $S$, mais on ne demande pas de les calculer.

***a*** – On se propose de résoudre le système linéaire $AX = B$ où $B$ est un vecteur colonne donné dont les éléments sont notés $b_l$. Montrer qu'il est équivalent de résoudre le système $SX = L^{-1}B = Y$.

***b*** – Supposons que l'on connaisse les éléments $y_l$ du vecteur colonne $Y$. Montrer comment calculer directement la solution du système $SX = Y$.

***c*** – Maintenant, il s'agit de calculer $Y$. Écrire explicitement le système $B = LY$, puis les relations entre les éléments $y_k$, $l_{pq}$ et $b_r$. En déduire la valeur des $y_k$ en fonction des $l_{pq}$, $b_r$ et des $y_s$ déjà calculés $(s = 1, 2, \ldots, (k-1))$.

***d*** – Montrer que la matrice $L$ peut être rendue triangulaire avec des 1 sur la diagonale en la multipliant à droite par une matrice diagonale $D^{-1}$ convenablement choisie. On écrira $L = \Lambda D$. Donner l'expression des éléments de $\lambda_{lk}$ la matrice $\Lambda$ en fonction des éléments $l_{ij}$ de $L$. Montrer que $A = LS = (\Lambda D)S$ peut aussi s'écrire (seulement dans ce cas précis) $A = \Lambda(DS)$.

***2.*** – À partir de maintenant on suppose que la matrice $A$ est symétrique : $A = A^T$ ($A^T$ est la matrice transposée de $A$).

***a*** – Donner l'expression de $A^T$ en fonction de $D$, $S^T$ et $\Lambda^T$, en déduire deux relations entre $\Lambda$, $S$, $S^T$ et $\Lambda^T$.

***b*** – Déduire que $A = \Lambda D \Lambda^T$ (ou encore que $A = S^T D S$). Comme $D$ est une matrice diagonale, on peut écrire : $A = \Lambda\sqrt{D}\sqrt{D}\Lambda^T = MM^T$. Donner les éléments $\delta_{ll}$ de $\sqrt{D}$ en fonction des éléments $d_{pp}$ de $D$.

***c*** – Écrire directement les relations donnant les éléments $m_{lk}$ de la matrice triangulaire inférieure $M$ en fonction des éléments $a_{ij}$ de $A$.

***d*** – Indiquer une méthode de résolution des systèmes linéaires dépendant d'une matrice symétrique non singulière.

## 5.6. Résolution d'un système linéaire par la méthode du gradient conjugué

On se propose de résoudre un système linéaire de $N$ équations à $N$ inconnues :

$$AX = B$$

où $A$ est une **matrice définie positive et symétrique** d'ordre $N$ d'élément $a_{lk}$, $X$ le vecteur inconnu, $B$ le vecteur second membre d'éléments $\{b_l\}$.

***1.*** – Soient $p$ et $q$ deux vecteurs de dimension $N$. On note par $(p, q)$ le produit scalaire de $p$ et $q$ : $(p, q) = p^T q$, $p^T$ étant le vecteur transposé de $p$ en écriture matricielle.
Montrer que $(p, Aq) = (Ap, q) = (q, Ap) = (Aq, p)$.

***2.*** – Par définition, on dit qu'une suite de vecteurs $\{p_i\}$ est A-orthogonale ($i = 0, 1, \ldots, N-1$) quand les produits scalaires $(Ap_i, p_j)$ sont nuls pour $i \neq j$.

***a*** – Montrer que $(Ap_i, p_i) > 0$.

***b*** – Montrer que les $\{p_i\}$ forment une base de l'espace à $N$ dimensions.

***c*** – On désigne par $h$ le vecteur qui est solution du système $AX = B$, montrer que $h$ peut s'écrire sous la forme :

$$h = \sum_{j=0}^{N-1} c_j p_j, \quad \text{avec } c_j = \frac{(B, p_j)}{(Ap_j, p_j)} \cdot \tag{H.6}$$

Montrer que $h$ peut s'obtenir par le schéma itératif suivant :

$$\begin{aligned} x_0 &= c_0 p_0 \\ c_j &= \frac{(B, p_j)}{(Ap_j, p_j)} \\ x_{j+1} &= x_j + c_{j+1} p_{j+1} \\ x_M &= h \quad \text{avec } M \leq N. \end{aligned}$$

**3.** – On se propose maintenant de construire une base $\{p_k\}$ de vecteurs A-orthogonaux à partir d'une base $\{v_k\}$ de $N$ vecteurs linéairement indépendants (orthogonalisation de Gram-Schmidt) pour $k = 0, 1, 2, \ldots, N-1$.

***a*** – On pose $p_0 = v_0$, puis $p_1 = v_1 + \alpha_{10} p_1$.
Montrer que

$$\alpha_{10} = -\frac{(Av_1, p_0)}{(Ap_0, p_0)}$$

puis que le vecteur $p_{k+1}$ s'exprime en fonction des $k$ précédents vecteurs $p$ de la façon suivante :

$$p_{k+1} = v_{k+1} + \alpha_{k+1,0} p_0 + \alpha_{k+1,1} p_1 + \cdots + \alpha_{k+1,k} p_k = v_{k+1} + \sum_{j=0}^{k} \alpha_{k+1,j} p_j$$

avec $k < N-2$. (H.7)

Donner l'expression des $\alpha_{k+1,j}$ en fonction des produits scalaires de vecteurs (connus au fur et à mesure). Montrer que

$$(v_k, Ap_j) = 0 \quad \text{pour} \quad j > k.$$

***b*** – Si $A$ est la matrice unité ($A = I$), la suite des vecteurs A-orthogonaux devient une suite de vecteurs orthogonaux. Dans ce cas particulier, montrer que la suite des $\{p'_k\}$ dont la définition (légèrement modifiée par rapport à (H.7)) est la suivante :

$$p'_{k+1} = p'_k + \sum_{j=0}^{k-1} \gamma_{k+1,j} p'_j + \gamma_{k+1,k} v_{k+1}$$

est une suite orthogonale que l'on peut écrire sous la forme :

$$p'_{k+1} = p'_k + \frac{(p'_k, p'_k)}{(v_{k+1}, p'_k)} \left\{ \sum_{j=0}^{k-1} \frac{(v_{k+1}, p'_j)}{(p'_j, p'_j)} p'_j - v_{k+1} \right\}. \quad \text{(H.8)}$$

**4.** – On se propose d'utiliser les résultats précédents pour résoudre le système $AX = B$. Pour cela, il nous suffit de connaître une base quelconque de l'espace de dimension $N$, à partir de celle-ci nous obtiendrons une base $\{p_k\}$ A-orthogonale qui nous permettra de calculer le vecteur solution $h$.

Cette base $\{p_k\}$ A-orthogonale sera déterminée à partir d'une base $\{r_k\}$ orthogonale. On propose de calculer simultanément — ou en parallèle — les deux bases en procédant de la manière suivante :

Soit $x_0$ une approximation arbitraire du vecteur $h$. On pose :

$$\begin{aligned}
p_0 &= r_0 = B - Ax_0 \\
\alpha_i &= \frac{r_i^2}{(p_i, Ap_i)} \quad \text{où} \quad r_i^2 = (r_i, r_i) \\
x_{i+1} &= x_i + \alpha_i p_i \\
r_{i+1} &= x_i - \alpha_i A p_i \\
\beta_i &= -\frac{(r_{i+1}, Ap_i)}{(p_i, Ap_i)} \\
p_{i+1} &= r_{i+1} + \beta_i p_i.
\end{aligned}$$

**N.B.** - À chaque étape du processus itératif, $r_i$ donne l'écart à la solution exacte et s'appelle vecteur résidu ou résidu.

***a*** – Montrer que la suite des $\{r_i\}$ est orthogonale en utilisant la relation (H.8). On pourra poser :

$$v_{k+1} = Ap_k \quad \text{et} \quad p'_k = r_k.$$

***b*** – Montrer que la suite des $\{p_j\}$ est A-orthogonale en utilisant la relation (H.7).

***c*** – Quel est le nombre maximum de tours d'itération qui permettra d'obtenir la solution exacte, abstraction faite des erreurs dues aux troncatures des calculs effectifs?

***5.*** – Cette méthode de résolution d'un système linéaire, dépendant d'une matrice $A$ définie positive et symétrique, s'appelle méthode du gradient conjugué.

Les formules

$$x_0 = c_0 p_0 \quad \text{avec} \quad c_j = \frac{(B, p_j)}{(Ap_j, p_j)}$$
$$x_j = x_{j-1} + c_j p_j$$

montrent que l'approximation $x_j$ s'obtient à partir de l'approximation précédente en ajoutant une quantité algébrique parallèle au vecteur $p_j$. Nous allons voir que, précisément, la quantité $c_j p_j$ est celle qui minimise le carré du résidu dans la direction $p_j$. On note $E^2(x)$ le carré du résidu.

***a*** – Montrer que $E^2(x)$ est donné par l'expression :

$$E^2(x) = [A(h - x), h - x].$$

Il s'agit de minimiser $E^2(x)$ dans la direction $p$ sur la droite $x = x_j + l_j p_j$ et de déterminer $l_j$ correspondant à ce minimum.

***b*** – Montrer que $E^2(x_j + l_j p_j) = E^2(x_j) + l_j^2 (Ap_j, p_j) - 2l_j (r_j, p_j)$ expression dans laquelle on a posé : $r_j = B - Ax_j \ (= B - x_j)$.

Déduire que $l_j = \dfrac{(r_j, p_j)}{(Ap_j, p_j)}$.

***c*** – En remarquant que $(r_j, p_j) = (B - Ax_j, p_j) = (B, p_j) - (x_j, Ap_j)$ et que

$$x_j = \sum_{k=0}^{j-1} c_k p_k$$

déduire que $l_j = c_j$.

***6.*** – Soit $A$ une matrice quelconque. Quelles transformations doit-on appliquer au système linéaire pour pouvoir utiliser l'algorithme étudié?

## 5.7. Calcul direct des coefficients du polynôme caractéristique

Soit $A$ une matrice carrée d'ordre $n$ dont les éléments sont réels. On se propose de calculer directement les coefficients $\alpha_k$ de son polynôme caractéristique que l'on peut écrire sous les formes suivantes :

$$P_n(\lambda) = [\lambda I_n - A], \tag{H.9}$$

$$P_n(\lambda) = \sum_{j=0}^{n} \alpha_j \lambda^{n-j}, \tag{H.10}$$

où $I_n$ est la matrice unité d'ordre $n$, les crochets signifiant qu'il s'agit d'un déterminant.

***a*** – Identifier l'expression donnée par (H.10) avec le développement en série de MacLaurin de $P_n(\lambda)$ afin d'expliciter les coefficients $\alpha_k$. Donner alors l'expression de $\alpha_k$ en fonction de $P_n(0)$ et des dérivées en zéro qui sont notées $P_n^{(q)}(0)$, $q = 1, 2, \ldots, n$.

Pour obtenir les dérivées successives de $P_n(\lambda)$, on se propose de dériver l'expression (H.9) par rapport à $\lambda$, puis on fait $\lambda = 0$ pour obtenir les $P_n^{(q)}(0)$.

Comme $A$ ne dépend pas de $\lambda$, la règle de dérivation du déterminant donnant $P_n'(\lambda)$ **dans le cas particulier qui nous intéresse** est donnée par les expressions suivantes :

$$P'(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}[\lambda I_n - A] = \sum_{j=1}^{n} [\lambda I_{n-1} - A_{jj}],$$

où $I_{n-1}$ est la matrice unité d'ordre $(n-1)$ et où $A_{jj}$ désigne la matrice d'ordre $(n-1)$ obtenue en supprimant la $j^{\mathrm{e}}$ ligne et la $j^{\mathrm{e}}$ colonne de la matrice $A$.

Donner l'expression de $P'(0)$.

***b*** – En utilisant la règle de dérivation précédemment fournie, donner l'expression de $P''(\lambda)$, puis en déduire l'expression de $P_n''(0)$ en fonction des matrices $A_{jj,kk}$ qui désignent les matrices d'ordre $(n-2)$ obtenues en supprimant dans la matrice $A$ les lignes $j$ et $k$ d'une part et les colonnes $j$ et $k$ d'autre part (on précisera soigneusement les indices).

***c*** – Montrer que l'expression de $P_n^{(q)}(0)$ est donnée par :

$$P_n^{(q)}(0) = (-1)^{n-q} \sum_{j,k,\ldots,q} [A_{jj,kk,\ldots,qq}] = q!(-1)^{n-q} \sum_{j<k,\ldots,<q} [A_{jj,kk,\ldots,qq}].$$

***d*** – Montrer par un examen direct de (H.9) que $P_n^n(0) = n!$

***e*** – Dire comment exploiter cet algorithme pour obtenir les valeurs propres de la matrice $A$.

## 5.8. Calcul des valeurs propres d'une matrice réelle et symétrique par la méthode de Jacobi

***a*** – Soit $A$ une matrice carrée d'ordre $n$ à éléments $\{a_{lk}\}$ réels. On considère une autre matrice $T$ d'ordre $n$ à éléments réels $\{t_{lk}\}$ qui possède une matrice inverse $T^{-1}$. Les valeurs propres $\{\lambda_i\}$ et les vecteurs propres $\{X_i\}$ sont donnés par l'équation : $AX = \lambda X$. Montrer que $B = T^{-1}AT$ admet les mêmes valeurs propres que $A$. Pour cela on posera $X = TZ$, expression dans laquelle $Z$ est vecteur propre de $B$.

***b*** – Dans toute la suite du problème on considère que **$A$ est une matrice symétrique** $a_{rs} = a_{sr}$. À quelle propriété importante obéissent les valeurs propres de $A$?

***c*** – À présent, on cherche à diagonaliser la matrice $A$ au moyen d'une succession de transformations $T^{-1}AT$.

***Choix de la matrice $T$*** – $T$ est une matrice unité d'ordre $n$ à l'exception de quatre éléments se situant à l'intersection des lignes $p$ et $q$ d'une part et des colonnes $p$ et $q$ d'autre part :

$$t_{pp} = \cos(\varphi) \qquad t_{pq} = \sin(\varphi) \qquad t_{qp} = \sin(\varphi) \qquad t_{qq} = -\cos(\varphi).$$

Montrer que $T^{-1} = T$.

***d*** – Écrire les éléments de la matrice $(AT)$. Puis montrer comment ces éléments sont modifiés lorsque l'on multiplie $(AT)$ par $T^{-1}$. Écrire les éléments $b_{lk}$ de $B = T^{-1}(AT)$ et en particulier $b_{pp}, b_{qq}, b_{pq}$ et $b_{qp}$.

***e*** – En déduire que $B = T^{-1}AT$ est une matrice symétrique. (Attention, le produit de deux matrices symétriques n'est pas, en général, une matrice symétrique.)

***f*** – $\varphi$ est un paramètre que l'on choisit de telle sorte qu'on annule le coefficient $b_{pq}$ et par conséquent $b_{qp}$, donner alors l'expression de $\tan(2\varphi)$ en fonction des $a_{lk}$. On pose $\tan(2\varphi) = \tau$. Donner l'intervalle de variation de $\varphi$. Donner les expressions de $\cos(\varphi)$ et de $\sin(\varphi)$ en fonction de $\tau$ en prenant garde aux éventuels problèmes de signe, au préalable, il peut être utile de calculer $\sin(2\varphi)$ ou $\cos(2\varphi)$, et l'on désignera par $\varepsilon$ le signe de $\tau$.

***g*** – Définir un algorithme qui amène progressivement des zéros partout hormis sur la diagonale principale dans la suite de matrices $B_i$ que l'on précisera. Dans le cas où $\tau$ devient infini le procédé ne tombe pas pour autant en défaut. Donner alors les valeurs numériques de $\cos(\varphi)$ et de $\sin(\varphi)$ qui permettent de poursuivre la transformation.

***Convergence de la méthode*** – Montrer que pour une matrice symétrique telle que $A$, la somme des carrés des valeurs propres de $A$ notée $S^2$ est égale à la somme des carrés de tous les éléments de $A$. En déduire que cette valeur est invariante si la matrice subit une série de transformations générales du type de celles définies ci-dessus $T^{-1}AT$.

***h*** – Calculer $b_{pp} + b_{qq}$ et $b_{pp} - b_{qq}$ en fonction des $a_{lk}$ et de $\varphi$. De ces deux équations auxquelles on joint l'expression de $b_{pq}$, démontrer que :

$$b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2.$$

***i*** – En déduire que la méthode étudiée est inconditionnellement convergente.

***Quelques expressions utiles...***

$$\cos^2(\varphi) - \sin^2(\varphi) = \cos(2\varphi),$$
$$2\sin(\varphi)\cos(\varphi) = \sin(2\varphi).$$

### 5.9. Calcul des valeurs propres d'une matrice par la méthode de Souriau (1948)

Soit $\mathbf{A}$ une matrice carrée d'ordre $n$ dont on se propose de calculer les valeurs propres $\lambda j$, lesquelles sont les racines du polynôme caractéristique :

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n = 0 \quad \text{avec} \quad a_0 = 1.$$

Nous nous proposons de calculer directement les coefficients $a_k$ puis de calculer les racines du polynôme. Pour ce faire, considérons une suite de matrices $B_q$ définie de la façon suivante :

$$\mathbf{B}_q = \mathbf{A}^q + a_1 \mathbf{A}^{q-1} + a_2 \mathbf{A}^{q-2} + \cdots + a_q \mathbf{I},$$

où $\mathbf{I}$ est la matrice unité d'ordre $n$.

***a*** – Donner la relation de récurrence qui existe entre deux matrices consécutives $B_k$ et $B_{k+1}$. Quelle valeur convient-il de donner à $\mathbf{B}_0$ ?

***b*** – Démontrer les relations de Newton exprimant les coefficients $a_k$ en fonction des sommes des racines élevées à la même puissance $m$. On notera :

$$S^m = \sum_{j=1}^{n} \lambda_j{}^m = \mathrm{Tr}(\mathbf{A}^m).$$

En utilisant l'expression de $a_k$ donnée par la relation de Newton, montrer que

$$a_k = -\mathrm{Tr}(\mathbf{A}\mathbf{B}_{k-1})/k,$$

et déduire une procédure générale de calcul des valeurs propres d'une matrice.

## 6. L'interpolation

### 6.1. Le polynôme de Lagrange

On dispose d'un fichier de données constitué d'une suite de $N$ couples d'abscisses et d'ordonnées. On désire réaliser un programme qui permette d'effectuer une interpolation voire une extrapolation par le polynôme de Lagrange.

***a*** – Programmer directement la méthode de Lagrange.

***b*** – Calculer les coefficients du polynôme de Lagrange par inversion de la matrice fournie par l'écriture du système linéaire.

***c*** – Comparer les performances des deux approches du même problème.

### 6.2. Les fonctions « spline »

Réaliser un programme qui effectue les interpolations (des mêmes données) au moyen de fonctions «spline » de degré trois. La mise au point terminée, le transformer en sous-programme.

***Applications des § 6.1 et 6.2*** – Déduire des études précédentes un ou plusieurs programmes qui réalisent la différentiation et l'intégration.

Quelles sont les difficultés liées à ces méthodes ?

La différentiation est-elle une opération sûre ?

Laquelle des deux techniques d'interpolation donne les meilleurs résultats ?

Pourquoi ?

# 7. Intégration des équations différentielles dans le champ réel

## 7.1. Étude d'un pendule de longueur variable

On considère un pendule dont la longueur $l$ varie linéairement avec le temps selon la loi : $l = l_0 + vt$ où $v$ est une constante.

***a*** – Écrire l'équation différentielle à laquelle obéit l'angle $\theta$ que fait le pendule avec la verticale, sachant que ce pendule est soumis à l'action de la pesanteur dont l'intensité est $g$.

***b*** – Écrire l'équation dans le cas où $g$ est petit. Réaliser un programme qui effectue la tabulation de la solution de cette dernière équation, sachant que le pendule qui nous préoccupe est la benne d'une grue suspendue au bout d'un câble de longueur $l_0 = 10$ m. Au temps $t = 0$, $\theta = \pi/200$ radian, $\mathrm{d}\theta/\,\mathrm{d}t = 0$ radian/s et $v = 1$ m/s.

***c*** – La benne s'arrête quand elle a parcouru 7 m verticalement ($l_1 = 3$ m). Quelle est alors l'amplitude des oscillations? Peut-on considérer que $\theta$ est resté petit?

***d*** – Mêmes questions qu'au § c avec toutefois des conditions initiales différentes : $l_0 = 3$ m, $l_1 = 10$ m, $v = -1$ m/s, $\theta$ et $\theta'$ gardant les mêmes valeurs.

## 7.2. Système électromécanique dépendant d'une équation de Mathieu

On considère un circuit électrique fermé constitué d'une self $L$, d'une résistance $R$ et d'une capacité $C$, tous les trois montés en série. $L$ et $R$ sont fixes, en revanche, la capacité $C$ varie en fonction du temps selon la loi :

$$\frac{1}{C(t)} = \frac{1 + \varepsilon \sin \omega_1 t}{C_0}\,.$$

La réalisation de ce condensateur est simple : il est constitué de deux plaques parallèles séparées par une distance $e(t)$ qui varie au cours du temps à la fréquence $\omega_1/2\pi$ selon l'expression : $e(t) = e_0(1 + \varepsilon \sin \omega_1 t)$ avec la valeur absolue de $\varepsilon$ plus petite que l'unité. Ce mouvement des plaques est évidemment créé par un moteur qui peut donc fournir de l'énergie au circuit électrique.

Écrire l'équation différentielle à laquelle obéit la charge électrique $q(t)$. Intégrer numériquement cette équation avec les valeurs suivantes : $L = 1$ H, $R = 1\ \Omega$, $C_0 = 10^{-4}$ F, $\varepsilon = 0{,}57$ et $\omega_1 = 225\ \mathrm{s}^{-1}$.

Au temps $t = 0$, $q_0 = 10^{-4}$ C et $i = \mathrm{d}q/\,\mathrm{d}t = 0$ A.

## 7.3. L'équation de Van der Pol

Soit l'équation différentielle suivante :

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} - 2\varepsilon(1 - y^2)\frac{\mathrm{d}y}{\mathrm{d}t} + y = 0.$$

On donne $\varepsilon = 1{,}0$. Intégrer numériquement cette équation en choisissant des conditions initiales très différentes (ce sont des données qui sont communiquées en conversationnel).

Représenter les résultats de l'intégration dans l'espace des phases, c'est-à-dire dans l'espace où $y$ est porté en abscisse et $\mathrm{d}y/\,\mathrm{d}t$ en ordonnée. On essayera les valeurs $\varepsilon = 1{,}0$ et $\varepsilon = 10{,}0$. Qu'observez-vous? Quel intérêt y a-t-il à fréquenter de telles équations?

## 7.4. Intégration d'une équation différentielle du premier ordre par une méthode itérative (prédiction-correction)

Soit à intégrer numériquement l'équation différentielle suivante :

$$\mathrm{d}y/\,\mathrm{d}x = f(x,y) \quad \text{pour } x \text{ appartenant à } (a,b).$$

De plus, on désire que la solution prenne la valeur $y_0$ pour la valeur $x_0 = a$ de la variable.

***a*** – Rappeler dans quelles conditions il existe une solution unique $y(x)$ à l'équation différentielle.

***b*** – Par la suite, on supposera que les conditions énoncées au § a sont satisfaites, et l'on désignera par $h$ le pas d'intégration de la variable $x$ et par $y_m$ la valeur de la solution au point $x_m = a + mh$.

Soit $D$ la droite tangente en $(x_m, y_m)$ à la courbe $y(x)$. On procède à une interpolation parabolique de $y(x)$ sur l'intervalle $(x_{m-1},\ x_{m+1})$, la sécante passant par $(x_{m-1}, y_{m-1})$, et $(x_{m+1}, y_{m+1})$ est donc parallèle à la tangente en $(x_m, y_m)$, donner alors l'expression de $y_{m+1}$.

***c*** – Cette relation appelée prédiction montre que le calcul de $y_m$ exige la connaissance des valeurs des deux points précédents, et l'on ne peut pas calculer la valeur de $y_1$ par ce procédé. Dans ce cas, on utilisera la formule de Runge et Kutta pour calculer $y_1$. Donner l'expression de $y_1$.

***d*** – Hormis $y_1$, nous allons améliorer la précision de chaque $y_{m+1}$ en utilisant une formule itérative appelée correction. On désigne alors par $y^1_{m+1}$ la valeur calculée au § b et par $y^k_{m+1}$ la $k^{\mathrm{e}}$ valeur itérée. Pour calculer la valeur $y^2_{m+1}$, on considère qu'elle est déterminée par l'intersection de la droite $Q$ et de la droite $x = x_{m+1}$ ; la droite $Q$ est construite de la manière suivante : elle passe par le point $(x_m, y_m)$ et sa pente est la moyenne arithmétique de la pente de $y(x)$ calculée en $(x_m, y_m)$ et de la pente calculée en $(x_{m+1}, y^1_{m+1})$. Donner l'expression de $y^k_{m+1}$.

***e*** – Quel critère pourra-t-on retenir pour limiter les itérations?

***f*** – La méthode proposée converge lorsque $h < 2/M$, où $M = \sup|\partial f/\partial x|$ dans le voisinage du point où l'on effectue les calculs. Comment s'affranchir de cette contrainte?

***g*** – Après avoir transformé l'équation différentielle sous une forme intégrale sur les intervalles définis pour la formule de prédiction et pour la formule de correction, donner les expressions de l'erreur de la prédiction et de la correction. On pourra effectuer un développement analogue à celui qui a été réalisé lors de l'étude des méthodes des trapèzes et de Simpson.

## 7.5. Étude d'un phénomène transitoire obéissant à une équation différentielle du premier ordre

Si la constante de temps est faible devant l'unité, alors bien avant l'unité, la solution présentera de faibles variations donnant l'illusion d'un phénomène régit par une grande constante de temps. L'intégration numérique peut alors poser quelques problèmes.

***1. Position du problème*** – Soit l'équation différentielle suivante :

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -50y + 5 \quad \text{avec } y(0) = 0{,}20. \tag{H.11}$$

***a*** – Montrer que la solution formelle peut se mettre sous la forme :

$$y(t) = 0{,}1[1 + \exp(-50t)]. \tag{H.12}$$

Déterminer l'intervalle de variation de $y(t)$ quand $t$ appartient à $(0\ ;\ \infty)$, puis représenter graphiquement cette solution.

***b*** – Calculer par la méthode d'Euler les valeurs numériques $y(1)$ et $y(2)$ (respectivement pour les valeurs $t_1 = 1$ et $t_2 = 2$) avec un pas d'intégration $h = 0{,}05$. Comparer avec le calcul direct de l'expression (H.12).

Que suggérez-vous pour obtenir des valeurs numériques plus conformes à celles obtenues à l'aide de la relation (H.12)?

***2. Étude d'une méthode de résolution spécifique*** – On considère une équation un peu plus générale, soit :

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -Dy + s(t) \quad \text{avec} \quad D \gg 1. \tag{H.13}$$

où $s(t)$ représente une fonction continue à variations faibles (lentes) sur $(0, \infty)$.

***a*** – En multipliant les deux membres de (H.13) par $\exp(-D \cdot t)$ puis en intégrant entre $t_1$ et $t_2$, montrer que l'équation (H.13) peut se mettre sous forme d'une équation intégrale qui est équivalente :

$$y(t_2) = y(t_1)\exp[-D(t_2 - t_1)] + \int_{t_1}^{t_2} \exp[D(t - t_2)s(t)]\,\mathrm{d}t \tag{H.14}$$

***b*** – L'intégrale figurant dans (H.14), désignée par $I$, peut être calculée d'une manière approchée par le procédé suivant : si $s(t)$ est développable en série entière, on obtient l'approximation en tronquant le développement à l'ordre $N$ ;

$$s(t) = \sum_{k=0}^{N} t^k b_k\,;$$

qu'il sera plus habile d'écrire (l'intérêt apparaîtra plus loin...) :

$$s(t) = \sum_{k=0}^{N} \frac{(t - t_1)^k}{k!\mu^k} a_k\,;$$

en posant $\mu = t_2 - t_1$.

Montrer que $I$ peut être approchée par l'expression :

$$I = \sum_{k=0}^{N} C_k a_k.$$

Donner l'expression de $C_0$, puis la relation de récurrence entre deux valeurs consécutives $C_k$ et $C_{k+1}$. (Intégration de $C_k$ par parties.)

***Choix des points de calcul*** – Il est commode de choisir les points en progression arithmétique :

$$T_{N-k} = t_2 - k\mu \quad \text{avec} \quad k = 0, 1, \ldots, N.$$

***Étude du cas* N = 1** – Calculer les points $T_0$ et $T_1$, puis les valeurs $s_k = s(T_k)$ en fonction des $a_j$. Exprimer les $a_l$ en fonction des $s_k$. Donner alors l'expression de $y(T_1)$ en fonction des $C_i$ et $s_j$.

***Application numérique*** – En conservant le pas $h = 0{,}05$, calculer $y(1)$ et $y(2)$ de l'équation différentielle (H.11).

***Étude du cas* N = 2** – Calculer les points $T_0$, $T_1$ et $T_2$, puis les valeurs $s_k = s(T_k)$ en fonction des $a_j$. Inverser ces relations pour obtenir les $a_l$ en fonction des $s_j$. Donner l'expression de $y(T_2)$ en fonction des $C_i$ et $s_j$.

***3. Généralisation de l'équation (H.13)*** – On considère à présent que $s(t)$ dépend de $y(t)$ et l'on écrira :

$$s(t) = s[y(t), t].$$

L'équation (H.13) n'est évidemment plus linéaire, mais on supposera encore que $s[y(t); t]$ est une fonction à variations lentes et faibles. Rien du développement formel précédemment réalisé n'est modifié, seul le calcul de l'intégrale demandera le calcul explicite de $s[y(t); t]$, car les coefficients $a_k$ dépendent des valeurs obtenues pour les $y(T_j)$.

Écrire explicitement l'expression donnant $y(T_1)$ à partir de la formule obtenue pour $N = 1$. Indiquer un procédé pour obtenir effectivement $y(T_1)$ en fonction de $y(T_0) = y(0)$.

Connaissant à présent $y(T_0)$ et $y(T_1)$, donner la formule exprimant $y(T_2)$ à partir de l'expression obtenue pour $N = 2$. Puis indiquer un procédé permettant de calculer effectivement $y(T_2)$.

Décrire une procédure complète d'intégration de la dernière équation différentielle proposée.

***Application numérique*** – Soit l'équation : $\frac{\mathrm{d}y}{\mathrm{d}x} = -50y + \frac{5}{1+y}$ avec $y(0) = 0{,}2$. Calculer $y(h)$ et $y(2h)$ avec $h = 0{,}05$.

### 7.6. Problème de la poursuite

Un jardinier tourne à la vitesse constante de 6 km/h autour d'un parterre dont la forme est une ellipse de grand axe $2b = 20$ m et de petit axe $2a = 12$ m. Lorsqu'il est en $A$, son chien est en $B$ et court à sa rencontre de telle sorte que le vecteur vitesse du chien passe à chaque instant par le point où se trouve le jardinier. Le chien court à la vitesse constante de 20 km/h. Sachant que la distance $OB = 30$ m et que le chien a le droit de piétiner le parterre, trouver la trajectoire du chien et représenter la graphiquement (*cf.* Fig. H.1, page suivante).

Trouver, au besoin expérimentalement, des conditions initiales de ce problème pour lesquelles le chien ne rattrape pas le jardinier. Tracer dans ce cas le cycle limite de la trajectoire suivie par le chien.

## 8. Intégration des équations aux dérivées partielles

### 8.1. L'équation de Laplace, les isothermes d'un réfrigérateur

Nous nous proposons de déterminer dans le plan les isothermes d'un réfrigérateur à l'équilibre thermique, c'est-à-dire quand $\partial T/\partial t = 0$, expression dans laquelle $T$ représente la température et $t$ le temps (*cf.* Fig. H.2, page suivante).

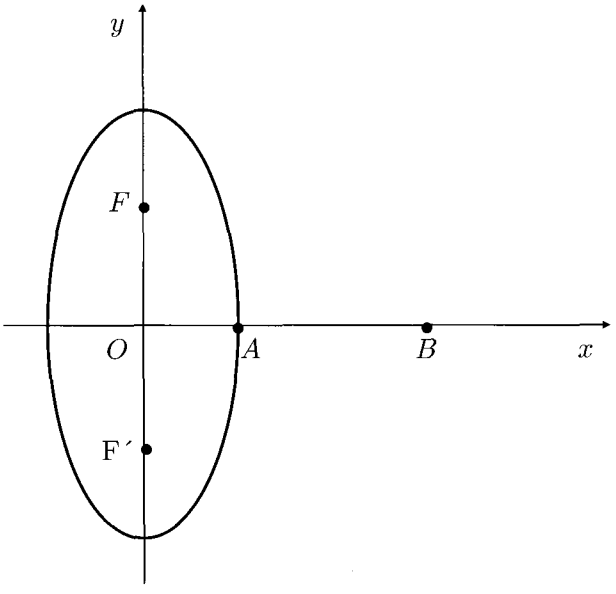


**Figure H.1.** *Problème de poursuite.*

La géométrie du réfrigérateur est donnée sur la figure H.2 : le congélateur est représenté par un rectangle à la température de $-10\,°\mathrm{C}$, en revanche, la paroi du réfrigérateur est à la température ambiante de $20\,°\mathrm{C}$.

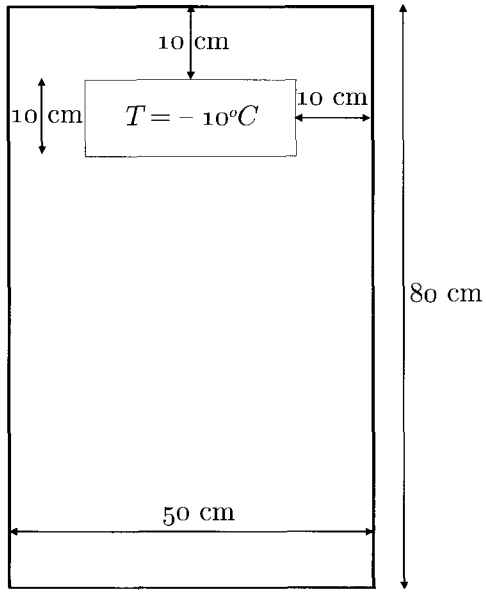


**Figure H.2.** *Isothermes d'un réfrigérateur.*

Procéder à un maillage convenable du domaine, calculer la température en chaque point du maillage et déterminer quelques isothermes de ce système thermique.

### 8.2. Refroidissement d'une sphère homogène

Une sphère homogène de cuivre de rayon $r = 5$ cm est portée à la température uniforme $\theta_0 = 90$ °C. Au temps $t = 0$, elle est plongée dans un grand réservoir d'eau à la température uniforme $\theta_1 = 10$ °C. On considère que ce réservoir est un excellent thermostat. On demande de calculer la température dans la sphère à différents instants. On dit que la sphère est complètement refroidie lorsque son centre est à une température inférieure à 10,1 °C. Déterminer la durée du refroidissement. On donne le coefficient de diffusion dans le cuivre $D = 1{,}12$ cm$^2$/s.
**N.B.** - Attention à l'usage des coordonnées sphériques que nous emploierons de toute façon, mais en $r = 0$ les équations sont discontinues. Comment pallier cet inconvénient?

### 8.3. Cas où tout le volume est le siège d'un dégagement de chaleur

Si tout le volume est le siège d'un dégagement de chaleur, montrer que l'équation de la chaleur s'écrit :

$$\frac{\lambda}{\rho}\Delta T + \frac{P}{\rho} = \frac{\partial T}{\partial t},$$

où $P$ est la puissance dissipée par unité de volume, $\rho$ est la capacité calorifique par unité de volume et $\lambda$ est la conductibilité thermique du milieu.

Un cylindre de cuivre de longueur indéfinie, de rayon $R = 1$ cm, est parcouru par un courant $I$. La résistivité du cuivre est $r(T) = r_0(1 + \alpha T)$, où $r_0 = 1{,}72\,10^{-8}$ Ω m, $T$ est exprimé en °C et $\alpha = 4{,}1(\text{°C})^{-1}$. On donne $\rho = 3{,}40$ J cm$^{-3}$ et $D = \frac{\rho}{\lambda} = 1{,}12$ cm$^{-2}$s$^{-1}$. On plonge ce fil dans un bain (thermostat) à la température de 10 °C, on attend que le fil soit en équilibre thermique, ensuite on fait passer un courant continu réparti uniformément d'intensité $I = 200$ A.

Après avoir montré que $P = r(T)j^2$ avec $j = \frac{I}{\pi R^2}$, déterminer le profil de température de seconde en seconde dans une section du fil.

## 9. Les transformées de Fourier

### 9.1. L'algorithme de Cooley-Tuckey

Réaliser le programme de la transformée de Fourier directe et de la transformée de Fourier inverse.

On commencera par réaliser des programmes de mise au point qui sont :

**a.** la table des adresses permettant de mettre les données dans le bon ordre ;
**b.** la table des sinus (et cosinus) au moyen d'un algorithme récurrent ;
**c.** la transformée de Fourier par décimation.

On pourra tester les deux transformées de Fourier sur une gaussienne tronquée (il faut réfléchir sur la troncature...). La transformée inverse doit permettre d'obtenir les valeurs de départ, éventuellement entachées d'une petite erreur, déphasées de $\pi$ selon le choix retenu pour représenter les données.

Transformée de Fourier de la fonction porte.

Transformée de Fourier d'un réseau (taches de diffraction).

## 9.2. Étude d'un vélocimètre

On se propose d'étudier un dispositif permettant de mesurer la vitesse d'écoulement laminaire d'un fluide, ou plus exactement la vitesse de très fines particules en suspension dans un liquide.

Le liquide s'écoule dans un tube cylindrique transparent au sein duquel on projette l'image d'une plaque opaque dans laquelle il y a $N$ trous de diamètre $a$ disposés sur une droite selon une progression arithmétique de raison $b$. L'image de l'axe portant les trous est parallèle à l'axe du cylindre contenant le fluide (*cf.* Fig. H.3).

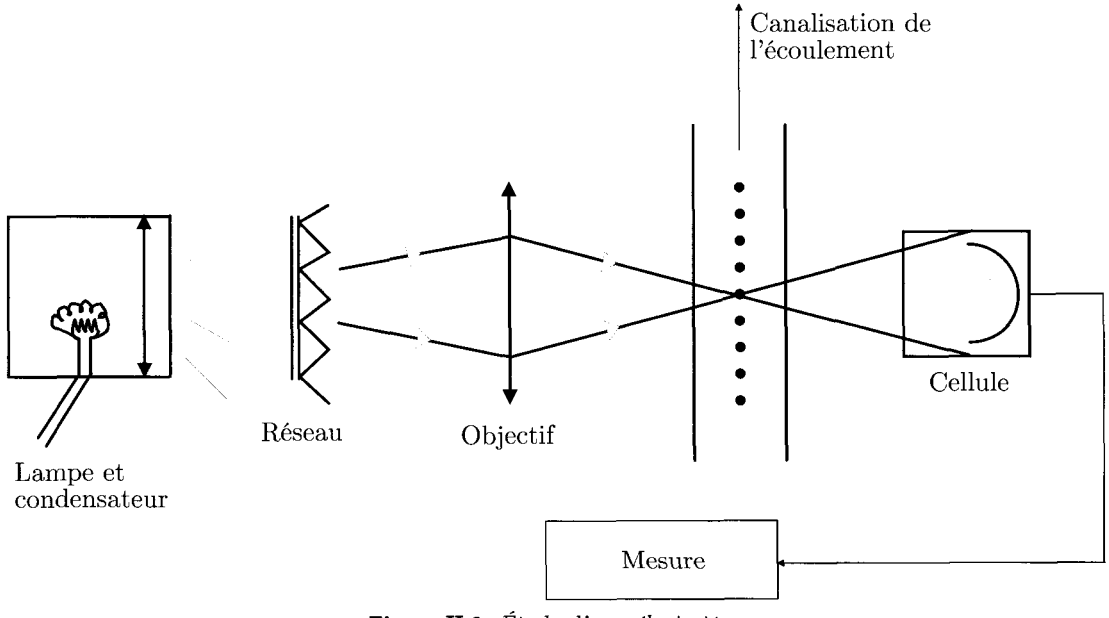


**Figure H.3.** *Étude d'un vélocimètre.*

Pour des raisons de commodité, on supposera que l'image a le grandissement 1 en valeur absolue. Notons que la plaque trouée peut être avantageusement remplacée par une diapositive sombre dans laquelle des points blancs jouent le rôle des trous. On éclaire le système avec une lampe alimentée en courant continu afin d'éviter les modulations parasites dues au courant sinusoïdal.

En l'absence de suspension, l'intensité reçue par la cellule photoélectrique est constante et égale à $I_0$. Le passage d'une particule $G$ devant les trous va provoquer l'absorption de lumière. On appelle $v$ la vitesse de la particule dont on supposera que la taille est négligeable devant la taille des trous.

Montrer que l'intensité absorbée $I_a(t)$ peut être représentée par la figure H.4, page ci-contre.

On donnera les valeurs de $\alpha$ et $\beta$ portés sur la figure H.4, page ci-contre, en fonction des paramètres précédemment définis. Ensuite, on calculera formellement $J(\nu)$ la transformée de Fourier de $I_a(t)$. Représenter alors $J(\nu)$, soit en tabulant la fonction $J(\nu)$, soit en calculant numériquement la $TF$ de $I_a(t)$ pour un nombre de trous donnés, de dimension et d'espacement donnés. L'allure de la courbe est donnée figure H.5, page ci-contre.

***a*** – Donner les expressions des quantités $A$, $B$ et $C$ portées par cette figure.

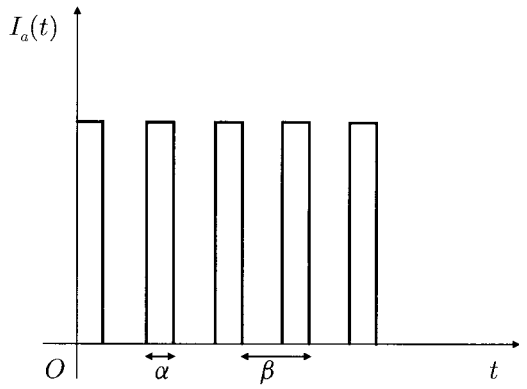


**Figure H.4.**

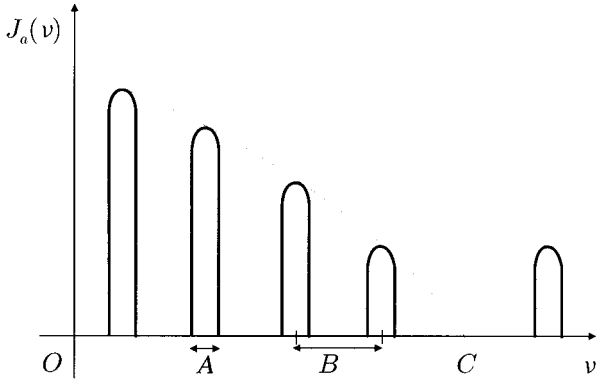


**Figure H.5.**

***b*** – Comment concevez-vous l'appareillage du cadre « mesure » chargée du traitement du signal?

Le parallélisme entre l'axe des trous et l'axe de déplacement des particules $G$ doit être rigoureux. Comment réaliser cette opération à l'aide du dispositif et de l'appareillage utilisés?

Dans la réalité, le nombre de grains $G$ est très élevé, et l'on recueillera sur la cellule photoélectrique un signal formé de la superposition de signaux tels que celui présenté figure H.4, mais déphasé d'une façon aléatoire les uns par rapport aux autres. Dire pourquoi l'allure générale de $J(\nu)$ ne sera pas modifiée.

## 9.3. Calcul numérique des transformées de Fourier pour un nombre de données $N = b^s$

Soit $f(x)$ une fonction appartenant à $L^2$, et l'on désigne par $\varphi(t)$ sa transformée de Fourier. On ne considère par la suite que les fonctions échantillonnées normalisées. Pour chacune des fonctions, on dispose de $N$ échantillons complexes (uniformément répartis) notés $f_m$ et $\varphi_k$ avec $(k, m) = 0, 1, 2, \ldots, N-1$. Au besoin, on peut écrire $f_m = f_m^r + jf_m^i$ et $\varphi_k = \varphi_k^r + j\varphi_k^i$.

La transformée de Fourier discrète s'écrit :

$$\varphi_k = \frac{1}{N} \sum_{m=0}^{N-1} f_m W_N^{km} \quad \text{avec} \quad W_N = \exp\left(j\frac{2\pi}{N}\right).$$

On suppose que $N$ est une puissance de trois : $N = 3^s$. Montrer que $\varphi_k$ est une combinaison linéaire de trois transformées de Fourier comprenant chacune $N/3$ échantillons ; soit :

$$\varphi_k = \frac{1}{3}[A_k + B_k W_N^k + C_k W_N^{2k}],$$

on explicitera $A_k$ $B_k$ et $C_k$.

Montrer que les transformées de Fourier $A_k$, $B_k$ et $C_k$ peuvent être calculées par le même procédé et ainsi de suite.

On suppose $s = 2$. Donner l'ordre dans lequel doivent être disposées les données pour mener à bien le calcul utilisant cet algorithme.

Dans le cas où $N = 2^s$ nous savons que la donnée n° $k$ doit être déplacée au rang $j$, valeur qui est donnée par :

$$j = R_N(k) = R_N(k - 2^q) + \frac{N}{2^{q+1}} \quad \text{avec } 2^q \leq k < 2^{q+1} \quad \text{et } R_N(0) = 0.$$

Montrer comment doit être modifiée cette relation pour l'adapter au cas $N = 3^s$. Vérifier sur l'exemple $N = 9$ la validité de la relation proposée.

Montrer comment on peut généraliser ce procédé de calcul des transformées de Fourier au cas où $N = b^s$ où $b$ est un entier positif supérieur à 1.

# 10. Introduction aux méthodes de Monte-Carlo

## Générateur de nombres pseudo-aléatoires

Réaliser un générateur de nombres pseudo-aléatoires à distribution rectangulaire (ou uniforme) selon le procédé de Lehmer. Tester ce générateur.

Réaliser un générateur de nombres pseudo-aléatoires à distribution gaussienne en usant du théorème central limite.

Cette dernière façon de procéder peut être médiocre si les queues de distribution jouent un rôle important. On préférera :

$$\eta_1 = \sqrt{(-2\log_e \xi_1)}\cos(2\pi\xi_2)$$
$$\eta_2 = \sqrt{(-2\log_e \xi_1)}\sin(2\pi\xi_2)$$

$\xi_1$ et $\xi_2$ étant des nombres pseudo-aléatoires indépendants, à distribution rectangulaire et appartenant à $(0, 1)$. Tester ce procédé.

Comment obtenir des nombres pseudo-aléatoires indépendants et gaussiens de moyenne donnée $m$ et d'écart type $\sigma$ ?

***a*** – Calculer $\pi$ par une méthode de Monte-Carlo.

***b*** – Calculer quelques intégrales définies par une méthode de Monte-Carlo.

***c*** – Résoudre localement l'équation de Laplace par une méthode de Monte-Carlo. On reprendra le problème concernant le calcul des isothermes d'un réfrigérateur et l'on comparera les différents résultats (exercice du chapitre 14).

# 11. Éléments de calcul des probabilités

## 11.1. L'ordre statistique

On considère un vecteur $V$ dont les $n$ composantes $\xi_i$ sont des nombres aléatoires indépendants obéissant tous à la même répartition continue $F(x)$ (distribution $f(x)$).

On effectue un classement de ces composantes selon les valeurs croissantes que l'on note :

$$x_1^{(n)} < x_2^{(n)} < \ldots < x_n^{(n)},$$

$x_1^{(n)}$ est la plus petite des valeurs $\xi_i$ et $x_n^{(n)}$ la plus grande. On admet que les inégalités sont strictes en probabilité car $F(x)$ est continue.

On dispose de $n$ urnes numérotées de 1 à $n$, et l'on place chacune des valeurs $x_k^{(n)}$ dans l'urne $k$ correspondant au rang $k$ attribué par le classement. $x_k^{(n)}$ est une variable aléatoire.

On effectue le tirage des composantes d'un nombre de vecteurs $V$ aussi grand que l'on veut, toutes les composantes de tous les vecteurs étant des variables aléatoires indépendantes de même répartition $F(x)$. Pour chaque vecteur $V$ on effectue le classement puis le remplissage des $n$ urnes. On se propose d'étudier la statistique des éléments contenus dans chacune des urnes, et plus précisément leur distribution.

***a*** – On désigne par $N_n(x)$ le nombre de composantes $\xi_i$ du vecteur $V$ qui sont inférieures à un nombre donné $x$ qui appartient à $(-\infty, +\infty)$. Montrer que $S_n(x) = N_n(x)/n$ est la fonction de répartition des variables $x_k^{(n)}$, pour cela on explicite les intervalles de définition de $x$ pour les diverses valeurs discrètes susceptibles d'être prises par $S_n(x)$. On note que, pour $x$ fixé, $\eta = S_n(x)$ est une variable aléatoire car si les valeurs prises par $S_n(x)$ sont connues, on ne sait pas pour quelles valeurs de $x$ ces valeurs sont prises. $S_n(x)$ s'appelle répartition empirique des $x_k^{(n)}$.

***b*** – Donner, en fonction de $F(x)$, l'expression de la probabilité $P(\xi_k < x)$, c'est-à-dire la probabilité que la composante $\xi_k$ soit inférieure au nombre $x$ évoqué au § a. Puis montrer que $P(\xi_k < x) = \text{constante} = p$, quelle que soit la composante d'indice $k$ ($k = 1, 2, \ldots, n$).

***c*** – On se propose de calculer la probabilité que $S_n(x)$ prenne l'une des valeurs possibles, soit $m/n$, donner son expression en fonction de $F(x)$ (ou $p$) ; pour cela, on remarque qu'il doit y avoir $m$ composantes du vecteur $V$ qui sont inférieures à $x$ (éventuellement une égale à $x$) et $(n - m)$ qui sont supérieures à $x$. On justifiera l'usage de la loi binomiale.

***d*** – À présent, on s'intéresse à la distribution de la variable aléatoire $x_j^{(n)}$ c'est-à-dire de tous les $\xi_i$ (de chaque vecteur $V$) tombés dans l'urne $j$. On désigne par $\Phi_j^{(n)}(x)$ la probabilité que $x_j^{(n)}$ soit inférieur à un nombre $x$, c'est-à-dire $P(x_j^{(n)} < x)$.

L'événement $E(x_j^{(n)} < x)$ a lieu quand au moins $j$ composantes de $V$ sont inférieures à $x$. Montrer que l'événement $E$ ($S_n(x)$ n'est pas plus petit que $\frac{j}{n}$) est équivalent à l'événement $E(x_j^{(n)} < x)$. En déduire que $\Phi_j^{(n)}(x) = \sum_{m=j}^{n} P[S_n(x) = \frac{m}{n}]$, puis donner l'expression de $\Phi_j^{(n)}(x)$ en fonction de $F(x)$.

***e*** – Montrer que :

$$\Phi_j^{(n)}(x) = \frac{n!}{(j-1)!(n-j)!} \int_0^{F(x)} t^{j-1}(1-t)^{n-j}\, \mathrm{d}t,$$

pour ce faire, on réalise une succession d'intégrations par parties de cette dernière expression.

En déduire la fonction de distribution $f_j^{(n)}(x)$ de la variable $x$ appartenant à l'urne $j$, puis donner l'expression de la valeur moyenne $\langle x_j^{(n)} \rangle$ des $\xi_i$ de chaque vecteur $V$ qui sont tombés dans l'urne $j$.

***Application numérique*** – La fonction $F(x)$ est la loi uniforme :

$$F(x) = \begin{cases} 0 & \text{pour } x \leq 0, \\ x & \text{pour } 0 < x < 1, \\ 1 & \text{pour } x \geq 1. \end{cases}$$

et l'on s'intéresse à la statistique d'ordre $n = 5$ (il y a 5 urnes). Donner l'expression de la densité de probabilité, puis calculer la valeur moyenne des $\xi_i$ tombés dans chacune des urnes. Vérifier que les résultats sont symétriques.

# 12. Lois (Binomiale, Poisson, Gauss-Laplace)

## 12.1. La loi Binomiale

Approximation de $n!$ au moyen de l'approximation de Stirling : $\log_e(n!) = n\log_e(n) - n$. Application numérique : calculer $n!$ pour $n = 50$ puis 60, 70, 80, 90 et 100. Quelle erreur est commise sur ces approximations?

Approximation de $n!$ au moyen de la formule de Stirling :

$$n! = n^n \exp(-n)\sqrt{2\pi n}(1 + \varepsilon_n)$$

avec $n\varepsilon_n \longrightarrow 1/12$ quand $n$ croît indéfiniment. Reprendre les calculs précédents.

## 12.2. La loi de Gauss-Laplace

Effectuer un sous-programme qui réalise le calcul des trois intégrales suivantes :

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right) \, \mathrm{d}t$$

$$g(x) = 1{,}0 - f(x)$$

$$h(x) = \sqrt{\frac{2}{\pi}} \int_0^x \exp\left(-\frac{t^2}{2}\right) \, \mathrm{d}t.$$

## 12.3. Loi du $\chi^2$

Effectuer un sous-programme qui réalise le calcul du $\chi^2$ à $m$ degrés de liberté :

$$L_m(x) = \frac{1}{\Gamma\left(\frac{m}{2}\right) 2^{m/2}} \int_0^x u^{(m/2)-1} \exp\left(-\frac{u}{2}\right) \, \mathrm{d}u$$

avec

$$\Gamma(z) = \int_0^\infty \exp(-t)t^{z-1}\,\mathrm{d}t$$
$$\Gamma(n+1) = n!$$
$$\Gamma\left(n+\frac{1}{2}\right) = \frac{1\cdot 3\cdot 5\ldots(2n-1)}{2^n}\sqrt{\pi}.$$

## 12.4. Loi de Student

Effectuer un sous-programme qui réalise le calcul de la loi de Student à $m$ degrés de liberté :

$$\Lambda_m(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}} \int_{-\infty}^{x} \left(1+\frac{u^2}{m}\right)^{-(m+1)/2} \mathrm{d}u.$$

## 12.5. Loi de Poisson

Effectuer un sous-programme qui réalise le calcul de la loi de Poisson :

$$P_k = \frac{\lambda^k}{k!}\exp(-\lambda).$$

On donne $\lambda = 30$, par quelle courbe de Gauss peut-on approcher cette loi? Donner le sup du module de la différence entre ces deux fonctions.

## 12.6. Loi de Poisson. Distribution de Cauchy

On considère la variable aléatoire $\xi$ obéissant à la loi de distribution

$$p(x) = k\exp -a|x| \quad \text{pour } x \text{ appartenant à } (-\infty;+\infty)$$

expression dans laquelle $a$ est un paramètre strictement positif et $k$ une constante de normalisation.

***a*** – Calculer la constante $k$.

***b*** – Rappeler la définition de la fonction caractéristique $\phi(t)$ de la variable aléatoire $\xi$ et donner l'intervalle de définition de cette fonction. Calculer effectivement $\phi(t)$.

***c*** – Calculer la moyenne $m$ ainsi que l'écart quadratique moyen $D$ de la variable aléatoire $\xi$.

***d*** – Rappeler l'expression de la distribution $l(m,x)$ de Student à $m$ degrés de liberté d'une variable aléatoire $\eta$ et donner son expression dans le cas où $m = 1$ ; on désignera par $q(x)$ cette distribution particulière.

***e*** – Donner l'expression $\Phi(t)$ de la fonction caractéristique de la variable aléatoire $\eta$.

***f*** – Soit $z = \frac{1}{n}\sum_{i=1}^{n}\eta_i$ la moyenne arithmétique de $n$ variables aléatoires $\eta_k$ indépendantes et distribuées chacune selon la loi $q(x)$. Donner l'expression de la fonction caractéristique $\Psi(t)$ de la variable $z$, puis en déduire sa fonction de distribution.

***g*** – Calculer l'espérance mathématique $M$ de la variable aléatoire $z$. Que concluez-vous?

***Quelques résultats utiles*** – La transformée de Fourier

$$\Phi(t) = \int_{-\infty}^{+\infty} f(x) \exp(2\pi j x t)\, \mathrm{d}x.$$

## 12.7. Loi binomiale négative. Loi de Poisson dont le paramètre suit une loi du $\chi^2$. Urne de Pólya (1887–1985)

Ce problème est constitué de trois parties notées 1, 2 et 3 dont les débuts sont indépendants les uns des autres, les symboles et notations utilisés sont également indépendants dans ces trois parties.

***1.*** – Une urne contient $P$ boules blanches et $Q$ boules noires ($P$ et $Q$ strictement positifs). On effectue dans cette urne des tirages non exhaustifs, c'est-à-dire que l'on procède à la remise dans l'urne de la boule tirée avant d'effectuer un autre tirage. On note $p$ la probabilité de tirer une boule blanche et $q$ celle de tirée une boule noire ($p + q = 1$). On appelle **succès** le tirage d'une boule blanche et **échec** le tirage d'une boule noire.

On s'intéresse aux suites de $n$ tirages notées $S_n^r$ telles qu'un $n^{\mathrm{e}}$ tirage fournisse le $r^{\mathrm{e}}$ succès qui clôt la suite, $r$ étant un nombre entier positif fixé à l'avance. Pour cela, on poursuit l'expérience aussi loin qu'il le faut, $n$ étant aussi grand que l'on veut. Les suites $S_n^r$ se terminent toutes par un succès qui est le $r^{\mathrm{e}}$. Comme $n \geq r$, il est préférable d'écrire $n = r + k$ avec $k = 0, 1, 2, \ldots$

***a*** – On note $f(k, r, p)$ la probabilité que le $r^{\mathrm{e}}$ succès ait lieu au tirage $r + k$. Au cours des $r + k$ tirages, montrer que la probabilité d'avoir tiré exactement $k$ échecs qui précèdent le tirage du $r^{\mathrm{e}}$ succès est aussi $f(k, r, p)$.

***b*** – On appelle événement $A$ la réalisation du $r^{\mathrm{e}}$ succès au tirage $r + k$, événement $B$ la réalisation de $k$ échecs au tirage $r + k - 1$, et enfin l'événement $C$ la réalisation d'un succès au tirage $r + k$. Écrire la relation entre les événements $A$, $B$ et $C$. Calculer la probabilité de la réalisation de l'événement $B$ en fonction de $r$, $k$, $p$ et $q$.

***c*** – Donner l'expression de $f(k, r, p)$. Montrer que $f(k, r, p)$ est le coefficient d'ordre $k$ du développement de la fonction : $p^r(1 - q)^{-r}$. En déduire que $\sum_{k=0}^{\infty} f(k, r, p) = 1$.

***d*** – Plus généralement, pour un réel $r > 0$ fixé et pour $0 < p < 1$, la séquence $\{f(k, r, p)\}$ avec $k = 0, 1, 2, \ldots$ est appelée **distribution binomiale négative**.

Montrer que $p^r(1 - qt)^{-r}$ est la fonction génératrice des moments non centrés. Calculer $m_1$ la moyenne de la variable aléatoire entière $\mu$ qui suit la distribution $f(k, r, p)$, puis $m_2$ le moment du deuxième ordre. En déduire la variance $\sigma^2$. Comparer ces résultats à ceux de la loi binomiale.

***e*** – Donner l'expression de la fonction caractéristique $\varphi(t)$ associée à la fonction $f$.

***2. a*** – On considère une variable aléatoire $\xi$ qui obéit à la loi de Poisson de paramètre $\lambda$. Rappeler ce que représente le paramètre $\lambda$, puis donner la probabilité $P_k$ pour que $k$ événements soient réalisés.

***b*** – En réalité, $\lambda$ est une variable aléatoire qui suit une loi du $\chi^2$ dont l'expression s'écrit :

$$g(x) = \frac{\alpha^\nu}{\Gamma(\nu)} x^{\nu-1} \exp(-\alpha x) \quad \text{avec} \quad x > 0.$$

Vérifier que $g(x)$ est bien une densité de probabilité. Calculer la probabilité $\pi(\xi = k)$ que la variable aléatoire de Poisson soit égale à une valeur entière positive donnée $k$. En déduire que la variable obéit à une loi binomiale négative.

***3.*** – On considère une urne qui contient $b$ boules noires et $r$ boules rouges. Après chaque tirage effectué, on remet dans l'urne la boule tirée plus $c$ boules de la même couleur. (Il s'agit d'un tirage contagieux dans une urne de Pólya.)

***a*** – Donner l'expression de la probabilité $\Pi_{n_1 n_2}$ de tirer d'abord $n_1$ boules noires puis $n_2$ boules rouges. Montrer que, quel que soit l'ordre fixé du tirage de $n_1$ boules noires et de $n_2$ boules rouges, la probabilité est toujours la même et vaut $\Pi_{n_1 n_2}$.

***b*** – Donner l'expression de la probabilité $P(n_1, n)$ de tirer $n_1$ boules noires parmi $n$ tirages.

***c*** – On pose :

$$p = \frac{b}{b+r}, \quad q = \frac{r}{b+r} \quad \text{et} \quad \gamma = \frac{c}{b+r}.$$

Montrer que

$$P(n_1, n) = \frac{C^{n_1}_{-p/\gamma} C^{n_2}_{-q/\gamma}}{C^{n}_{-1/\gamma}}.$$

***d*** – Lorsque $n \longrightarrow \infty$ et que $p$ et $\gamma$ tendent vers zéro de telle sorte que :

$$np \longrightarrow \lambda \quad \text{et} \quad n\gamma \longrightarrow \frac{1}{\rho},$$

montrer que la distribution asymptotique étudiée tend vers la loi binomiale négative :

$$P(n_1, n) \longrightarrow C^{n_1}_{-\lambda\rho} \left(\frac{\rho}{1+\rho}\right)^{\rho\lambda} \left(\frac{1}{1+\rho}\right)^{n_1}.$$

Pour cela, on montrera que

$$C^{n_1}_{-p/\gamma} \longrightarrow C^{n_1}_{-\rho\lambda},$$

que

$$C^{n}_{-1/\gamma} \longrightarrow C^{n_2}_{-1/\gamma} (1+\rho)^{n_1},$$

et enfin que :

$$\frac{C^{n_2}_{-q/\gamma}}{C^{n_2}_{-1/\gamma}} \longrightarrow \left(\frac{\rho}{1+\rho}\right)^{\rho\lambda}.$$

***Quelques résultats utiles*** – Coefficient du binôme, $n$ est un entier positif :

$$(1+x)^n = 1 + nx + \cdots + \frac{n(n-1)\cdots(n-k+1)}{k!} + \cdots + x^n = \sum_{k=0}^{n} C_n^k x^k$$

$$C_n^p = C_n^{n-p} = \frac{n!}{p!(n-p)!)} = \binom{n}{p}.$$

Coefficient du binôme généralisé, $\alpha$ est un réel quelconque positif :

$$(1+x)^\alpha = 1 + \alpha x + \cdots + \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}x^k + \cdots = \sum_{k=0}^{\infty} C_\alpha^k x^k$$

$$= \sum_{k=0}^{\infty} C_\alpha^k x^k = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k.$$

$$(1-x)^{-\alpha} = 1 + \alpha x + \cdots + \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{k!}x^k + \cdots = \sum_{k=0}^{\infty} C_{-\alpha}^k (-x)^k$$

$$= \sum_{k=0}^{\infty} C_{k+\alpha-1}^k x^k = \sum_{k=0}^{\infty} \binom{-\alpha}{k} (-1)^k x^k.$$

La fonction factorielle :

$$\Gamma(x+1) = x! = \int_0^\infty u^x \exp(-u)\, du.$$

$$\sum_{k=0}^{n_2} \frac{1}{1+k\gamma} \longrightarrow \frac{1}{\gamma} \log_e(1+\gamma n_2)$$

quand $n_2$ est très grand devant l'unité.

## 12.8. La loi binomiale négative (reprise et applications)

Une urne contient $P$ boules blanches et $Q$ boules noires ($P$ et $Q$ strictement positifs). On effectue dans cette urne des tirages non exhaustifs, c'est-à-dire que l'on procède à la remise dans l'urne de la boule tirée avant d'effectuer un autre tirage. On note $p$ la probabilité de tirer une boule blanche et $q$ celle de tirer une boule noire ($p+q=1$). On appelle **succès** le tirage d'une boule blanche et **échec** le tirage d'une boule noire.

On s'intéresse aux suites de $n$ tirages notées $S_n^r$ telles qu'un $n^{\text{e}}$ tirage fournisse le $r^{\text{e}}$ succès qui clôt la suite, $r$ étant un nombre entier positif fixé à l'avance. Pour cela, on poursuit l'expérience aussi loin qu'il le faut, $n$ étant aussi grand que l'on veut. Les suites $S_n^r$ se terminent toutes par un succès qui est le $r^{\text{e}}$. Comme $n \geq r$, il est préférable d'écrire $n = r + k$ avec $k = 0, 1, 2, \ldots$

***a*** – On note $f(k,r,p)$ la probabilité que le $r^{\text{e}}$ succès ait lieu au tirage $r+k$. Au cours des $r+k$ tirages, montrer que la probabilité d'avoir tiré exactement $k$ échecs qui précèdent le tirage du $r^{\text{e}}$ succès est aussi $f(k,r,p)$.

***b*** – On appelle événement $A$ la réalisation du $r^{\text{e}}$ succès au tirage $r+k$, événement $B$ la réalisation de $k$ échecs au tirage $r+k-1$, et enfin l'événement $C$ la réalisation d'un succès au tirage $r+k$. Écrire la relation entre les événements $A$, $B$ et $C$. Calculer la probabilité de la réalisation de l'événement $B$ en fonction de $r, k, p$ et $q$.

**c** – Donner l'expression de $f(k, r, p)$.

**d** – Plus généralement, pour un réel $r > 0$ fixé et pour $0 < p < 1$, la séquence $\{f(k, r, p)\}$ avec $k = 0, 1, 2, \ldots$ est appelée **distribution binomiale négative**.

On montre que $p^r(1 - qt)^{-r}$ est la fonction génératrice des moments non centrés. À l'aide de cette fonction génératrice, calculer $m_1$ la moyenne de la variable aléatoire entière $\mu$ qui suit la distribution $f(k, r, p)$, puis $m_2$ le moment du deuxième ordre. En déduire la variance $\sigma^2$.

**e** – Sur six pommiers d'un verger tirés au hasard, on a prélevé sur chacun 25 feuilles toujours au hasard, puis, on a dénombré les mites rouges sur chacune des feuilles. Les résultats obtenus sont présentés dans le tableau H.1 d'après C.I. Bliss, *Fitting the negative binomial distribution to biological data*, Biometrics, juin 1953 p. 176.

**Tableau H.1.**

| Nombre de mites par feuille | Nombre de feuilles observées |
|---|---|
| 0 | 70 |
| 1 | 38 |
| 2 | 17 |
| 3 | 10 |
| 4 | 9 |
| 5 | 3 |
| 6 | 2 |
| 7 | 1 |
| 8 et plus | 0 |

Calculer la moyenne $m_e$ et l'écart type $\sigma_e^2$ de cette distribution expérimentale.

**f** – Pour rendre compte de cette distribution expérimentale, on envisage l'hypothèse d'une distribution de Poisson de paramètre $\lambda$. Donner la valeur de $\lambda$. Calculer les différents $P_k$ puis le $\chi^2$ correspondant. Peut-on retenir l'hypothèse d'une distribution de Poisson?

**g** – On fait l'hypothèse que la loi expérimentale est la loi binomiale négative. Calculer $r$, $p$ et $q$ précédemment définis. Puis calculer la probabilité théorique que le nombre de mites sur les feuilles soit $m = 0, 1, 2, \ldots, 7$.

**h** – Calculer le $\chi^2$ correspondant. Peut-on retenir cette dernière hypothèse?

***Quelques résultats utiles*** – Coefficient du binôme, $n$ est un entier positif :

$$(1+x)^n = 1 + nx + \cdots + \frac{n(n-1)\cdots(n-k+1)}{k!} + \cdots + x^n = \sum_{k=0}^{n} C_n^k x^k$$

$$C_n^p = C_n^{n-p} = \frac{n!}{p!(n-p)!)} = \binom{n}{p}.$$

Coefficient du binôme généralisé, $\alpha$ est un réel quelconque :

$$(1+x)^{\alpha} = 1 + \alpha x + \cdots + \frac{\alpha(\alpha-1)\ldots(\alpha-k+1)}{k!}x^k + \cdots = \sum_{k=0}^{\infty} C_{\alpha}^{k} x^k$$

$$= \sum_{k=0}^{\infty} C_{\alpha}^{k} x^k = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$$

$$(1-x)^{-\alpha} = 1 + \alpha x + \cdots + \frac{\alpha(\alpha+1)\ldots(\alpha+k-1)}{k!}x^k + \cdots = \sum_{k=0}^{\infty} C_{-\alpha}^{k}(-1)^k x^k$$

$$= \sum_{k=0}^{\infty} C_{k+\alpha-1}^{k} x^k = \sum_{k=0}^{\infty} \binom{-\alpha}{k}(-1)^k x^k.$$

## 12.9. Loi de Poisson. Durée de vie d'un système simple (fiabilité)

***1.*** – Chaque jour un cycliste parcourt la même distance. L'expérience lui a appris qu'il survient une crevaison de sa bicyclette en moyenne deux fois par mois.

***a*** – Calculer la probabilité qu'il survienne exactement une crevaison en une semaine (1 mois = 4 semaines).

***b*** – Calculer la probabilité qu'il survienne exactement une crevaison en un mois.

***c*** – Calculer la probabilité qu'il ne survienne aucune crevaison en un mois.

***d*** – Calculer la probabilité qu'il survienne trois crevaisons ou plus en un mois.

***e*** – Calculer la probabilité qu'il survienne exactement quatre crevaisons en deux mois.

***2.*** – On se propose d'écrire explicitement la loi de Poisson en fonction du temps $t$ et l'on désigne par $\lambda$ le nombre moyen d'événements par unité de temps. Donner l'expression de la loi de Poisson $P_k(t)$ donnant la probabilité pour que $k$ événements surviennent durant le temps $t$.

On désigne par $t_\nu$ la suite discrète de la date des événements qui arrivent séquentiellement au cours du temps. La suite $t_\nu$ s'appelle flux d'événements, et on rappelle que les temps d'apparition des événements sont indépendants les uns des autres.

On désire étudier la statistique de la variable aléatoire $\xi = t_{i+1} - t_i$ qui est le temps qui s'écoule entre deux événements consécutifs. Après avoir remarqué que $\xi$ est une variable aléatoire continue, calculer la probabilité $F(t) = P(\xi \leq t)$ que la variable aléatoire $\xi$ soit plus petite que $t$. Pour cela on calculera d'abord la probabilité pour que, dans un intervalle $t$ débutant à l'instant $t_i$, il n'y ait aucun événement. En déduire la densité de probabilité $f(t)$ à laquelle obéit la variable aléatoire $\xi$. Calculer la moyenne $a$ de $\xi$ ainsi que sa variance $\sigma^2$.

***3.*** – Une propriété remarquable de cette loi.

Si l'intervalle de temps $\xi$ est déjà entamé depuis un certain temps $\tau$, on se propose de démontrer que cela n'a aucune influence sur la partie restante de l'intervalle, et que sa répartition sera la même que celle qui gouverne $\xi$, c'est-à-dire $F(t)$.

Pour démontrer cette assertion, on écrit que l'événement $\tau$ a eu lieu et l'on étudie la loi de répartition conditionnelle de la partie restante $\eta = \xi - \tau$, soit :

$$\phi(t) = P(\xi - \tau < t | \xi > \tau).$$

On désigne par $A$ l'événement $\xi > \tau$, $B$ l'événement $\xi - \tau < t$.

À l'aide du théorème des probabilités composées, donner l'expression de $P(A \cdot B)$ ; en déduire l'expression de $\phi(t)$ en fonction de $P$, $A$ et $B$. Montrer que l'événement $A \cdot B$ peut aussi s'écrire :

$$\tau < \xi < t + \tau,$$

expression dont on calculera la répartition en fonction de $F$, $t$ et $\tau$.

Donner à présent l'expression de $\phi(t)$ en fonction de $F$, $t$ et $\tau$.

Remplacer $F$ par l'expression calculée au § b et montrer que $F(t) = \phi(t)$.

## 12.10. Durée de vie d'un système. Loi de Weibull (1887–1979). Fiabilité

La durée de vie d'un système physique est mesurée par la variable aléatoire $\xi$. On se propose d'étudier la répartition $F(t)$ de cette variable, soit : $F(t) = P(\xi < t)$.

On désigne par $\lambda(t)$ le coefficient des défaillances en fonction du temps $t$ que l'on définit de la façon suivante :

$$\lambda(t) = \frac{n(t) - n(t + \Delta t)}{\Delta t\, n(t)},$$

$n(t)$ étant le nombre de systèmes ayant vécu jusqu'à l'instant $t$, $\Delta t$ étant un intervalle de temps petit (infiniment petit du premier ordre).

***a*** – Montrer que $F(t)$ peut se mettre sous la forme :

$$F(t) = 1 - \frac{n(t)}{n} \quad \text{où} \quad n = n(0).$$

***b*** – Donner l'équation différentielle du premier ordre à laquelle obéit la répartition $F(t)$ en fonction de $\lambda(t)$. On exploitera le fait que l'on peut écrire :

$$\lambda(t) = \frac{n(t) - n(0) + n(0) - n(t + \Delta t)}{\Delta t\, n(t)},$$

puis que $\Delta t$ est un infiniment petit du premier ordre.

***c*** – En intégrant l'équation différentielle obtenue, donner l'expression de $F(t)$ en fonction de $\lambda(t)$.

***d*** – Une classe intéressante de coefficients de défaillance $\lambda(t)$ est donnée par l'expression :

$$\lambda(t) = \lambda_0 \alpha t^{\alpha - 1} \quad \text{avec} \quad \lambda_0 > 0 \quad \text{et} \quad \alpha > 0,$$

donner alors l'expression de $F(t)$ ainsi que celle de $f(t)$.

***e*** – Calculer la moyenne $\bar{a}$ de $\xi$ ainsi que sa variance $\sigma^2$.

***f*** – Cette loi est connue sous le nom de loi de Weibull et trouve son emploi en théorie de la fiabilité. La densité de probabilité $f(t)$ (ou la répartition) établie à la question b du précédent problème est un cas particulier de la loi de Weibull, donner alors la valeur de $\alpha$ qui correspond à ce cas particulier.

**N.B. -** On donne :

$$\Gamma(x + 1) = \int_0^\infty u^x \exp(-u)\, du.$$

### 12.11. L'inégalité de Kolmogorov (1903–1987)

Soit $X$ une variable aléatoire continue de moyenne nulle $E(X) = 0$, d'écart quadratique $\sigma^2$ et de fonction de distribution $f(x)$. On considère un nombre arbitraire $t$ positif.

***a*** – Rappeler la démonstration de l'inégalité de Bienaymé-Tchebycheff :

$$P(|X| > t\sigma) \leq \frac{1}{t^2}$$

où $P$ désigne la probabilité, et l'on précisera le domaine de validité de $t$.

***b*** – À présent, on suppose que $|X|$ a une borne supérieure désignée par $B$. Montrer que :

$$E(X^2) \leq \int_{|x|\leq a} a^2 f(x)\,\mathrm{d}x + \int_{|x|>a} x^2 f(x)\,\mathrm{d}x,$$

expression dans laquelle $a$ est compris entre 0 et $B$. En déduire que

$$P(|X| > a) \geq \frac{E(X^2) - a^2}{B^2}\,.$$

Cette expression s'appelle inégalité de Kolmogorov.

### 12.12. Loi quasi normale

***a*** – On considère une variable aléatoire $Y$ de moyenne nulle $E(Y) = 0$, d'écart quadratique moyen $E(Y^2) = \sigma^2$ et obéissant à une distribution gaussienne. Rappeler, au besoin en les démontrant, les expressions des moments centrés d'ordre trois et quatre.

***b*** – Par définition, nous dirons qu'une variable aléatoire $X$ obéit à une distribution quasi normale lorsque :

$$E(X^3) \neq 0 \quad \text{et} \quad E(X^4) = 3[E(X^2)]^2.$$

On se propose de calculer les paramètres $a$ et $b$ de la loi :

$$f(x) = \frac{1}{\sqrt{2\pi}}(ax^2 + bx - 3a)^2 \exp\left(-\frac{x^2}{2}\right),$$

de telle façon que $f(x)$ soit une loi quasi normale. On montrera que $f(x)$ est bien une densité de probabilité, puis on calculera $E(X)$, $E(X^2)$, $E(X^3)$ et $E(X^4)$. Enfin, on déterminera $a$ et $b$.

Une fois les conditions établies, donner l'écart quadratique moyen $\sigma^2$ de la loi normale dont les moments deux et quatre coïncident avec ceux de la loi quasi normale envisagée.

## 13. La fonction caractéristique

### 13.1. $\left(\frac{\sin x}{x}\right)^n$ tend vers une gaussienne quand n croît positivement

***a*** – Soient deux variables aléatoires indépendantes $\xi_1$ et $\xi_2$, qui sont toutes les deux normales (gaussiennes) centrées et réduites. Calculer la fonction caractéristique de chacune d'elle, puis la fonction caractéristique de la variable $\eta = \xi_1 + \xi_2$. En déduire la fonction de distribution de $\eta$.

***b*** – Les variables indépendantes sont toujours gaussiennes mais ne sont plus centrées réduites. On désigne par $m_1$ et $m_2$ les moyennes respectives et par $\sigma_1$ et $\sigma_2$ les écarts quadratiques moyens. Donner la fonction de distribution de $\eta = \xi_1 + \xi_2$.

***c*** – Généraliser le calcul du § b en considérant la variable

$$\eta = \sum_{k=1}^{n} \xi_k,$$

puis en déduire le théorème d'addition des moyennes et le théorème d'addition des variances (écart type élevé au carré).

***d*** – On considère une variable aléatoire $\xi$ à distribution rectangulaire, c'est-à-dire uniformément répartie sur l'intervalle $(-0{,}5\ ;\ 0{,}5)$. Donner la fonction caractéristique $\Phi(t)$ de cette variable aléatoire, puis la fonction caractéristique de la variable aléatoire $\xi/n$, et enfin la fonction caractéristique $\Psi(t)$ de la variable aléatoire :

$$\mu = \frac{1}{n} \sum_{k=1}^{n} \xi_k$$

où les $\xi_k$ sont des variables aléatoires indépendantes à distribution rectangulaire uniformément réparties sur l'intervalle $(-0{,}5\ ;\ 0{,}5)$.

Au moyen d'un développement limité au deuxième ordre de $\log_e \Psi(t)$, montrer que la fonction caractéristique tend vers une gaussienne quand $n$ tend vers l'infini. En déduire la loi de distribution de $\mu$ quand $n$ devient grand devant l'unité.

Déduire de cet exercice le comportement asymptotique de la fonction :

$$f(x) = \left(\frac{\sin x}{x}\right)^n.$$

Effectuer deux programmes qui permettent la vérification expérimentale de conclusion obtenue. L'un effectuera les auto-convolutions successives de la fonction de distribution, l'autre le calcul direct de la fonction caractéristique.

À partir de quelle valeur $n_0$ de $n$ peut-on admettre l'approximation asymptotique?

## 14. La loi du $\chi^2$ et la loi de Student

### 14.1. Le test de Pearson

On désire montrer que le paramètre de Pearson

$$\chi^2 = \sum_{q=1}^{r} \frac{(Y - np_q)^2}{np_q}$$

obéit à une loi du $\chi^2$ à $(r-1)$ degrés de liberté.

On considère une variable aléatoire $X$ dont on se donne un échantillon constitué de $n$ éléments notés $X_1, X_2, \ldots, X_n$.

On réalise une partition de l'axe réel en $r$ intervalles notés $I_1, I_2, \ldots, I_r$ et l'on désigne par $p_1, p_2, \ldots, p_r$ les **probabilités théoriques** que $X$ appartienne respectivement aux intervalles $I_1, I_2, \ldots, I_r$. (Réalisation d'un histogramme.)

Soit $x_1, x_2, \ldots, x_n$ une réalisation de l'échantillon. On note par $y_q$ le nombre de valeurs $x_1, x_2, \ldots, x_n$ qui appartiennent à l'intervalle $I_q$ avec $q = 1, 2, \ldots, r$.

$y_q$ est la valeur d'une variable aléatoire $Y_q$ obtenue pour ladite réalisation. (Ici, les majuscules $X_k, Y_k, \ldots$ sont des variables aléatoires tandis que les minuscules $x_k, y_k, \ldots$ sont les réalisations de ces variables aléatoires.)

Pour parvenir à nos fins, nous montrerons que $\chi^2$ est la somme de carrés de variables gaussiennes centrés, réduites (au sens large) et indépendantes obéissant cependant à une relation linéaire.

***a*** – On rappelle que la fonction caractéristique d'une variable aléatoire à $n$ dimensions (ou vecteur aléatoire) $X = (X_1, X_2, \ldots, X_n)$ est l'espérance mathématique de la variable complexe

$$\exp[j(t_1X_1 + t_2X_2 + \cdots + t_nX_n)],$$

soit :

$$\phi(t_1, t_2, \ldots, t_n) = M\{\exp[j(t_1X_1 + t_2X_2 + \cdots + t_nX_n)]\}$$

où $M\{\ \}$ est l'opérateur espérance mathématique portant sur chacune des variables.

Montrer que le moment du premier ordre de la variable $X$ est donné par :

$$m_{1q} = \frac{1}{j} \cdot \frac{\partial f}{\partial t_q} \quad \text{pour } t_1 = t_{2\ldots} = t_n = 0$$

et celui du deuxième ordre par :

$$m_{2q} = \frac{1}{j^2} \cdot \frac{\partial^2 f}{\partial t_q^2} \quad \text{pour } t_1 = t_{2\ldots} = t_n = 0.$$

***b*** – On désigne par $p_k$ la **probabilité théorique** que la variable $X_j$ appartienne à l'intervalle $I_k$. Montrer que les variables $Y_1, Y_2, \ldots, Y_r$ obéissent à une loi de probabilité multinomiale :

$$P(Y_1 = n_1, Y_2 = n_2, \ldots Y_r = n_r) = \frac{n!}{n_1!n_2!\ldots n_r!} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}.$$

***c*** – Nous allons obtenir, par voie directe, la fonction caractéristique de cette loi multinomiale. Pour cela, on considère une suite de vecteurs aléatoires indépendants et de même distribution notée $(Z_1^q, Z_2^q, \ldots, Z_r^q)$ avec $q = 1, 2, \ldots, n$.

Ces composantes prennent la valeur 0 ou 1, et l'on pose :

$$P(Z_j^q = 1) = p_j \quad \text{et} \quad P(Z_j^q = 0) = 1 - p_j.$$

Montrer que $(Y_1, Y_2, \ldots, Y_r) = \sum_{q=1}^{n} (Z_1^q, Z_2^q, \ldots, Z_r^q)$, puis donner l'expression de la fonction caractéristique : $\phi q(t_1, t_2, \ldots, t_n)$ de $(Z_1^q, Z_2^q, \ldots, Z_r^q)$. En déduire, au moyen du théorème concernant la fonction caractéristique d'une somme de variables indépendantes, la fonction caractéristique :

$$\phi_q(t_1, t_2, \ldots, t_n) \text{ de } (Y_1, Y_2, \ldots, Y_r).$$

***d*** – En utilisant les relations établies au § a, calculer le moment du premier ordre $m_{1q}$ et le moment du deuxième ordre $m_{2q}$ des composantes $Y_q$ du vecteur aléatoire $(Y_1, Y_2, \ldots, Y_r)$.

***e*** – On désire étudier le comportement asymptotique de la loi multinomiale quand $n$ tend vers l'infini. Pour cela, il faudra transformer les composantes $Y_q$ en composantes centrées réduites. Montrer que le changement de variables :

$$U_q = \frac{Y_q - np_q}{\sqrt{np_q}}$$

répond à la question car la nouvelle variable $U_q$ ne dépend pas de $n$. Montrer que la fonction caractéristique du vecteur $U = (U_1, U_2, \ldots, U_r)$ tend vers la fonction

$$\exp\left\{-\frac{1}{2}\left(\sum_{q=1}^{r} t_q^2 - \left[\sum_{q=1}^{r} t_q\sqrt{p_q}\right]^2\right)\right\}$$

qui est la fonction caractéristique d'une loi normale (dégénérée). Pour y parvenir, on effectuera des développements limités à l'ordre deux...

Montrer que $\sum_{q=1}^{r} U_q\sqrt{p_q} = 0$.

***f*** – Déduire des résultats précédents, que la variable aléatoire

$$\chi^2 = \sum_{q=1}^{r} \frac{(Y_q - np_q)^2}{np_q}$$

obéit à une loi du $\chi^2$ à $(r-1)$ degrés de liberté.

***Remarque*** : On rappelle un résultat élémentaire d'analyse : $\lim\left(1 - \frac{s}{n}\right)^n$ tend vers $\exp(-s)$ quand $n$ tend vers l'infini.

## 14.2. Le test de Student

On considère une population parente gaussienne dont la moyenne théorique est $m$ et l'écart type théorique $\sigma$. On extrait de cette population un échantillon de $n$ individus. On désigne par $x$ la variable caractérisant la mesure sur la population et par $x_i$ la variable associée au $i^{\text{e}}$ individu.

On se propose d'étudier l'écart entre la moyenne de l'échantillon $\langle x\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i$ et la moyenne théorique $m$, c'est-à-dire la probabilité $P(|\langle x\rangle - m| > I)$, $I$ étant un intervalle qu'on se fixe à l'avance — ou encore on peut se fixer la probabilité $P$ et l'on détermine l'intervalle de confiance $I$.

Pour se fixer les idées, il s'agit d'un échantillon de 20 individus dont on a mesuré la taille $x$. On a trouvé expérimentalement : $x = 170{,}86$ cm et $\sigma = 3{,}65$ cm.

On souhaite que $|\langle x\rangle - m|$ appartienne à un intervalle de confiance où il y ait 90 chances sur 100 de trouver la valeur de $\langle x\rangle$ donnée par l'échantillon.

***1re approche*** – Calculer $I$ en utilisant l'inégalité de Bienaymé-Tchebycheff. Réaliser l'application numérique.

***2e approche*** – On suppose que, comme c'est le cas dans cet exemple, le nombre d'individus n'est pas très grand, et l'on se propose d'étudier la répartition de la variable aléatoire :

$$t = \frac{\sqrt{n}(\langle x\rangle - m)}{\sigma}\,.$$

***a*** – Montrer que

$$\sigma^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - m)^2 - n(\langle x\rangle - m)^2\right].$$

***b*** – Sur les variables centrées $(x_j - m)$, on effectue un changement de coordonnées orthogonales au moyen de la matrice orthogonale $A$ d'éléments $(a_{ij})$ :

$$\begin{aligned}
x'_1 &= \sum_{j=1}^{n} a_{1j}(x_j - m)\\
x'_2 &= \sum_{j=1}^{n} a_{2j}(x_j - m)\\
&\cdots\cdots\\
x'_{n-1} &= \sum_{j=1}^{n} a_{n-1j}(x_j - m)\\
x'_n &= \tfrac{1}{\sqrt{n}}\textstyle\sum_{j=1}^{n}(x_j - m).
\end{aligned}$$

Montrer que

$$\sum_{j=1}^{n} x'^2_j = \sum_{j=1}^{n}(x_i - m)^2.$$

***c*** – Montrer que $\langle x\rangle - m = \frac{x'_n}{\sqrt{n}}$.

***d*** – Montrer que

$$\sigma^2 = \frac{x'^2_1 + x'^2_2 + \cdots + x'^2_{n-1}}{n-1}.$$

***e*** – Montrer que les variables $x'_j$ sont centrées, normales, indépendantes et de même variance.

***f*** – Montrer alors que $t$ est une variable de Student dont on précisera le nombre de degrés de liberté.

***g*** – On désigne par $P(|t| > t_0) = 1 - \alpha$ la probabilité que $|t|$ dépasse une certaine valeur positive $t_0$. Déduire la relation entre $|\langle x\rangle - m|$, $\sigma$, $t_0$ et $n$.

***h*** – Effectuer l'application numérique. On donne :

$$t_{0,05}(19) = 1{,}729 \qquad t_{0,05}(20) = 1{,}725 \qquad t_{0,05}(21) = 1{,}721.$$

***3e approche*** – On suppose que la variable $t$ définie au § 14.2 est gaussienne. Effectuer l'application numérique dans ces nouvelles conditions. On donne :

$$\frac{1}{\sqrt{2\pi}}\int_0^{1{,}645} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = 1 - 0{,}05.$$

Commenter les trois résultats obtenus, et en déduire la valeur de la répartition de Student $t_{0,05}(\infty)$.

# 15. Systèmes à plusieurs variables aléatoires

## Système linéaire surdéterminé. Matrice de corrélation

Dans un document à caractère historique de 1321 on trouve le tableau H.2 d'imposition.

**Tableau H.2.**

| N° de la famille | Nombre de personnes | Quantité détenue | Impôt |
|---|---|---|---|
| 1 | 3 | 4,5 | 1,5 |
| 2 | 4 | 3,8 | 2 |
| 3 | 3 | 3,1 | 2,1 |
| 4 | 5 | 0 | 1,1 |
| 5 | 7 | 7,7 | 4 |
| 6 | 3 | 4,1 | 2 |
| 7 | 4 | 3,2 | 2 |
| 8 | 2 | 5,0 | 3,6 |
| 9 | 6 | 1,5 | 1,6 |
| 10 | 3 | 6,0 | 2,5 |

On désigne par $x$ la variable aléatoire associée au nombre de personnes, par $y$ la variable aléatoire associée à la quantité détenue et enfin $z$ la variable aléatoire associée à l'impôt perçu. Ensuite, on désigne par $n$ le nombre de familles. Par ailleurs on notera $\langle q \rangle$ la moyenne de la variable aléatoire $q$ et par $\sigma_q^2$ sa variance.

***1.*** – Dans l'hypothèse raisonnable de dépendances linéaires, calculer les coefficients de corrélation suivants :

***a*** – $r_{xz}$ entre le nombre de personnes et l'impôt payé,

***b*** – $r_{xy}$ entre le nombre de personnes et la quantité détenue,

***c*** – $r_{yz}$ entre la quantité détenue et l'impôt payé.

***2.*** – Quelles sont les liaisons de corrélation qui sont fondées ou significatives ? On donne un extrait de la table de Student :

$$t_{0,05}(8) = 1,860 \qquad t_{0,05}(9) = 1,833 \qquad t_{0,05}(10) = 1,812.$$

# 16. Critères de conformité

## 16.1. Exercices du cours

***a*** – Faire un programme qui réalise un histogramme. Les données faisant l'objet du classement seront fournies par le générateur de nombres pseudo-aléatoires de Lehmer.

***b*** – Vérifier que les nombres générés obéissent bien à une loi de distribution uniforme.

***c*** – Soit $\{\xi_i\}$ une suite de nombres pseudo-aléatoires fournie par le générateur de Lehmer. Montrer que la suite de nombres pseudo-aléatoires :

$$\eta_k = \left[\sum_{i=1}^{12} \xi_i - 6\right] \sigma + x_0$$

obéit à une distribution gaussienne de moyenne $x_0$ et d'écart type $\sigma$.

***d*** – Cette façon d'obtenir des nombres aléatoires à distribution gaussienne présente un défaut : les queues de distribution ne sont pas bonnes. On propose d'étudier un autre générateur de nombres gaussiens qui respecte les queues de distribution :

$$\eta_k = [-2\log_e(\xi_i)]^{1/2} \cos(2\pi\xi_{i+1})$$
$$\eta_{k+1} = [-2\log_e(\xi_i)]^{1/2} \sin(2\pi\xi_{i+1}).$$

Vérifier que ce générateur fournit des nombres aléatoires à distribution gaussienne.

## 16.2. Étalonnage d'un appareil de mesure

L'étalonnage d'un appareil de mesure a porté sur 400 mesures dont nous nous proposons d'étudier les incertitudes ; celles-ci, évaluées en mètres, sont regroupées dans le tableau H.3, où $m_i$ est le nombre de mesures tombant dans un intervalle donné d'après H. Ventsel.

**Tableau H.3.**

| $e_i$ (m) | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| $m_i$ | 21 | 72 | 66 | 38 | 51 | 56 | 64 | 32 | |

On cherche à représenter cette distribution expérimentale par une distribution rectangulaire :

$$p(x) = 1/(b-a) \quad \text{pour } a < x < b$$
$$p(x) = 0 \quad \text{ailleurs.}$$

***a*** – Pour déterminer $a$ et $b$, on se propose de calculer la moyenne $m$ et l'écart quadratique moyen $\sigma^2$, ce qui fournira un système de deux équations à deux inconnues. Donner les expressions de $m$ et $\sigma^2$ en fonction de $a$ et $b$. Vérifier que $a$ et $b$ jouent un rôle symétrique.

***b*** – Calculer $a$ et $b$ en fonction de $m$ et $\sigma^2$, puis numériquement $m, \sigma^2$, $a$ et $b$. Calculer la probabilité théorique de tomber dans chacun des intervalles de classement figurant dans le tableau.

***c*** – Calculer le $\chi^2$. Quel est le nombre de degrés de liberté ? Quelle est approximativement la probabilité de dépasser la valeur du $\chi^2$. Peut-on retenir l'hypothèse d'une distribution rectangulaire ?

## 16.3. Tests d'hypothèse

Étude de la distribution d'une hauteur d'arbres. On a effectué l'étude d'une population parente d'une espèce déterminée d'arbres, et en particulier, on s'est intéressé à leur taille. On a trouvé les résultats présentés dans le tableau H.4, page ci-contre, d'après P. Dagnelie.

**Tableau H.4.**

| $h$ (m) | 18 | 20 | 22 | 24 | 26 | 8 | 30 | 32 | 34 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|
| fréquence | 2 | 3 | 9 | 12 | 27 | 16 | 7 | 2 | 2 | |

***a*** – Calculer les deux premiers moments $m$ et $D$ puis l'écart type $\sigma$.

***b*** – Doit-on regrouper un certain nombre d'intervalles? Le cas échéant, proposer une solution.

***c*** – L'hypothèse $H$ est la loi normale. Calculer les probabilités par intervalles de regroupement.

***d*** – Calculer le $\chi^2_{\text{obs}}$.

***e*** – Peut-on raisonnablement conserver l'hypothèse $H$? Justifier votre réponse.

***f*** – Le test de Kolmogorov donne-t-il une confirmation du résultat obtenu au § e? On détaillera tous les calculs.

# 17. Étude des dépendances dans le cas linéaire

## 17.1. Test d'indépendance stochastique des résultats d'observations

***a*** – Générer 1 000 nombres pseudo-aléatoires à distribution gaussienne. On se propose de vérifier les tests usuels d'indépendance stochastique sur cet échantillon.

***b*** – Effectuer le test de la médiane.

***c*** – Effectuer le test des suites ascendante et descendante.

***d*** – Effectuer le test de la somme des carrés des différences successives.

## 17.2. Homogénéité d'une suite de variance

Dans le cadre du schéma de régression, on désire vérifier l'homogénéité d'une suite de variances. À cet effet, on fabrique deux séries de données sur lesquelles on se propose de retrouver les caractéristiques. Pour 5 valeurs de l'abscisse $x$, on détermine 5 valeurs de l'ordonnée $y$ selon la loi $y = ax + b$. On affecte un indice aux points $x_i, y_{i0}$ avec $i = 1, 2, \ldots, 5$. On choisira tous les paramètres comme on l'entend...

***a*** – Pour chaque $y_{i0}$ on prend 10 valeurs obtenues à partir d'un générateur de nombres pseudo-aléatoires gaussiens d'écart type $\sigma$ ; on les note $y_{ij}$, elles sont centrées sur $y_{i0}$ et ont la variance $\sigma^2$. Il s'agit du même écart type pour les 50 nombres tirés.

***b*** – Pour chaque $x_i$ on choisit un écart type $\sigma$ différent tel que les $\sigma_i^2$ obéissent à une loi crédible telle que :

$$\sigma_i^2 = \sigma_0^2 h^2(x_i).$$

Effectuer le test de l'homogénéité des suites de variances sur les deux échantillons. Tenter de trouver une fonction qui soit différente de $h^2(x_i)$ mais qui en soit une bonne approche.

## 17.3. Test d'indépendance stochastique du tirage d'un échantillon

On considère le jeu de pile ou face durant lequel on note la succession des apparitions. On note $A$ l'apparition de face et $B$ l'apparition de pile, la probabilité de présenter l'une des deux faces est 0,5.

On appelle chaîne de longueur $n$, la suite des $A$ et $B$ obtenue durant l'expérience, dans l'ordre de leur apparition, au moyen de $n$ tirages consécutifs. On appelle sous-chaîne une suite constituée uniquement de $A$ (ou de $B$) limitée à droite et à gauche par un $B$ (resp. $A$). Nous nous proposons d'étudier quelques caractéristiques des chaînes et des sous-chaînes.

***Partie 1. a*** – Montrer que la probabilité $\pi$ de tirer une sous-chaîne constituée de $\mu$ éléments (sous-chaîne de longueur $\mu$ contenant $\mu A$ ou $\mu B$) est égale à $(1/2)^{\mu}$. On énoncera le théorème sur lequel se fonde le calcul.

***b*** – Calculer $m$ la longueur moyenne et $\sigma$ l'écart type des sous-chaînes.

***c*** – On effectue le tirage de $K$ sous-chaînes. Donner $\lambda_K$ la longueur moyenne des chaînes contenant $K$ sous-chaînes.

***d*** – Donner la probabilité $P$ d'obtenir au moins une sous-chaîne (parmi les $K$), dont la longueur $\nu_j$ soit supérieure ou égale à une valeur arbitraire positive $M_0$ (on pourra éventuellement calculer la probabilité contraire). On énoncera le théorème sur lequel se fonde le calcul.

Application numérique : $K = 10, M_0 = 8$, calculer $P$.

***e*** – On souhaite que la probabilité $P$ d'obtenir au moins une sous-chaîne de longueur supérieure ou égale à $M_0$ soit inférieure ou égale à 0,05. Donner alors la relation entre $M_0$ et $K$, puis donner une expression approchée de $M_0$ en fonction de la longueur $n$ de la chaîne (on prendra $n = 2K$).

***f*** – Ce résultat peut être utilisé comme test n° 1 servant à vérifier l'indépendance stochastique des résultats d'observation (test fondé sur la médiane de l'échantillon). Effectuer l'application numérique pour une chaîne de longueur $n = 100$, et comparer le résultat avec celui obtenu à partir de l'expression approchée donnée dans le cours.

***Partie 2*** – On suppose que le nombre $K$ de sous-chaînes est grand devant l'unité, vérifier qu'il en est de même de $\lambda_K$. On se propose d'étudier la variable aléatoire $\xi$ (désignée par $n$ précédemment) qui est la longueur d'une chaîne constituée de $K$ sous-chaînes de longueur $l_j$ :

$$\xi = \sum_{j=1}^{K} l_j.$$

***a*** – Calculer $\Sigma$ l'écart type de la variable $\xi$. Rappeler le théorème central limite. En déduire la distribution de la variable aléatoire $\xi$, puis donner l'expression de la probabilité $P(\xi < \nu_0)$ pour que $\xi$ soit plus petit que la valeur arbitraire $\nu_0$.

***b*** – En choisissant le seuil de signification $\alpha = 0{,}05$, donner l'inégalité à laquelle obéissent $\xi$ et $K$.

On donne :

$$0{,}05 = \frac{1}{\sqrt{2\pi}} \int_{1{,}65}^{\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t.$$

**c** – Si l'on se donne la taille $\xi$, de la chaîne, montrer que le nombre de sous-suites $K$ (fonction de $\xi$) doit être plus grand qu'une certaine limite $K_{\min}$ pour ne pas tomber dans la queue de distribution (1,96 ; $\infty$), cas des événements peu probables.

**d** – Montrer que ce résultat peut servir de test n° 2 des séries fondé sur la médiane de l'échantillon. Effectuer le calcul numérique pour une chaîne de longueur $\xi = 100$, et comparer ce résultat à celui obtenu avec l'expression approchée donnée dans le cours.

Montrer que la limite du module de la différence de ces deux derniers tests (test n° 2 et test du cours correspondant) est 1/2 quand la longueur de la chaîne tend vers l'infini.

**e** – Quelles critiques de principe peut-on former concernant l'application simultanée de ces deux tests numérotés 1 et 2? (Il s'agit du strict point de vue de la rigueur.) Montrer brièvement comment on pourrait pallier ces inconvénients.

***Quelques résultats utiles***

$$\sum_{k=0}^{\infty} kx^{k-1} = \sum_{k=0}^{\infty}(k+1)x^k = \frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} kx^k + \sum_{k=0}^{\infty} x^k$$

$$\sum_{k=0}^{\infty} kx^k = \frac{x}{(1-x)^2}$$

$$\frac{2}{(1-x)^3} = \sum_{k=0}^{\infty} k^2x^k + 2\sum_{k=0}^{\infty} kx^k + \sum_{k=0}^{\infty} x^k + \sum_{k=0}^{\infty} kx^{k-1}$$

$$\sum_{k=0}^{\infty} k^2x^k = \frac{x(1+x)}{(1-x)^3}$$

$$(1+\alpha)^n = 1 + n\alpha \quad \text{si} \quad |\alpha| < 1.$$

## 17.4. Loi F(m, l). Étude du rapport de deux variances : loi de Fisher (1872–1962) - Snedecor (1881–1974)

On considère une population parente sur laquelle on prélève un échantillon de $n$ individus. Ces $n$ individus font chacun l'objet de la mesure de deux grandeurs physiques dont les résultats sont notés $(x_i, y_i)$, $i = 1, 2, \ldots, m$.

Les mesures des grandeurs $x$ et $y$ sont en réalité des variables aléatoires $\xi$ et $\eta$ dont une étude préalable a montré qu'elles vérifiaient les critères d'applicabilité du schéma de régression si bien que l'on peut écrire :

$$Y(x) = a + b(x - x_0).$$

On appelle variance empirique $s^2$ la moyenne empirique des carrés des résidus, soit :

$$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}[Y(x_i) - y_i]^2,$$

$a$, $b$ et $s^2$ caractérisent entièrement la régression étudiée.

On considère un autre échantillon de taille $n$, différente ou non de $m$, qui fournit une autre droite de régression :

$$Y'(x) = a' + b'(x - x'_0),$$

à laquelle est associée la variance empirique $s'^2$.

Imaginons que les échantillons aient été prélevés à des époques différentes, mais que $a$ et $a'$ d'une part puis $b$ et $b'$ d'autre part soient grosso modo les mêmes (ce qui nécessiterait une étude en soi). On désire savoir si au cours du temps on a bien affaire à la même population parente ou non. Autrement dit, est-ce que la différence entre les deux droites peut s'expliquer par les fluctuations aléatoires des échantillons?

L'étude du rapport des variances empiriques $v^2 = s^2/s'^2$ qui obéit à une loi que nous noterons $F(m,n)$ permet de répondre oui à la question dans le cas où :

$$F_{1-\alpha/2}(m-1,n-1) < \frac{s^2}{s'^2} < F_{\alpha/2}(m-1,n-1)$$

$\alpha$ étant le seuil de signification usuel. On se propose d'établir la loi de distribution $F(m,n)$ appelée loi de Fisher-Snedecor.

***1.*** – On désigne par $\xi^2$ la variable :

$$\xi^2 = \frac{1}{m-1}\sum_{i=1}^{m}\xi_i^2,$$

la somme des carrés de $m$ variables gaussiennes indépendantes, de moyenne nulle et d'écart type $\sigma$. En opérant d'une façon tout à fait semblable à la méthode utilisée pour établir la loi du $\chi^2$, donner la loi de distribution de $\xi^2$.

***2.*** – On désigne par $\eta^2$ la variable :

$$\eta^2 = \frac{1}{n-1}\sum_{i=1}^{n}\eta_i^2,$$

la somme des carrés de $n$ variables gaussiennes indépendantes, (indépendantes entre elles et également des $\xi_j^2$) de moyenne nulle et d'écart type $\sigma$.

***a*** – Écrire la loi de distribution de $\eta^2$.

***b*** – Rappeler la démonstration du théorème concernant la distribution du rapport de deux variables aléatoires indépendantes $y = \tau/\phi, \tau$ et $\phi$ ayant respectivement pour fonctions de distribution $p_1(\tau)$ et $p_2(\phi)$.

***c*** – Déduire la distribution du rapport des deux variables $y = \dfrac{\xi^2}{\eta^2}$.

***d*** – Calculer la moyenne de $y$ notée $\langle y \rangle$.

***e*** – Calculer l'écart type de $y$ noté $\langle \sigma^2 \rangle$.

***f*** – Préciser les degrés de liberté $l$ et $k$ dans la loi $F(l,k)$ en fonction de $m$ et $n$. Justifier la réponse.

***g*** – Montrer que $F_{1-\alpha}(k,l) = \dfrac{1}{F_\alpha(l,k)}$.

***Quelques expressions utiles***

$$\Gamma(x) = \int_0^\infty \exp(-t)t^{x-1}\,\mathrm{d}t$$

$$\Gamma(1+x) = x\Gamma(x)$$

$$\int_0^\infty \frac{x^{\mu-1}}{(1+x)^\nu}\,\mathrm{d}x = \frac{\Gamma(\mu)\Gamma(\nu-\mu)}{\Gamma(\nu)}$$

$$\frac{1}{2^{m/2}\Gamma(m/2)}x^{m/2-1}\exp(-x/2) \quad \xrightarrow{TF} \quad (1-2jt)^{-m/2}.$$

# 18. Analyse de corrélation

## 18.1. Le coefficient de corrélation. Schéma de corrélation

Il nous faut ici encore fabriquer un lot de données que l'on se propose à la suite d'étudier. La variable $x$ est aléatoire à distribution gaussienne d'écart type $\sigma$. Dans un premier temps, on calcule une valeur $z = ax + b$ après avoir tiré $x = \xi$. On choisira les paramètres comme on l'entend... À $z$ on ajoute une variable gaussienne $\delta$ de **moyenne nulle** et d'écart type $\sigma'$. Il faut veiller à ce que $\delta$ et $\xi$ ne soient pas corrélées. Comment s'en assurer ou y parvenir?

Remplir un premier fichier de 50 à 100 couples de données selon ce procédé en choisissant $\sigma' = \sigma/10$.

Remplir un deuxième fichier de 50 à 100 couples de données en choisissant $\sigma' = \sigma$.

Remplir un troisième fichier de 50 à 100 couples de données en choisissant $\sigma' = 10\sigma$.

Calculer chacun des trois coefficients de corrélation $r$. Examiner éventuellement l'hypothèse $r = 0$.

Calculer l'intervalle de confiance associé à chacun des coefficients de corrélation pour un seuil de signification donné $\alpha$.

## 18.2. La distribution gaussienne à deux dimensions. Coefficient de corrélation

On considère une population parente sur laquelle on a prélevé un échantillon de taille $n$ très grand devant 1. Chaque individu est caractérisé par deux grandeurs $\xi$ et $\eta$ qui sont des variables aléatoires. On désigne par $m_\xi$ et $m_\eta$ les moyennes, par $\sigma_\xi^2$ et $\sigma_\eta^2$ les écarts quadratiques moyens respectivement de $\xi$ et $\eta$.

***a*** – On appelle $x$ et $y$ les variables centrées réduites associées à $x$ et $y$. Donner les expressions de $x$ et $y$.

Une analyse appropriée a montré que la fonction de distribution des variables $x$ et $y$ était de la forme :

$$\Phi(x,y) = A\exp\left(-\frac{x^2 - 2\alpha xy + y^2}{\beta^2}\right),$$

expression dans laquelle $A$, $\alpha$ et $\beta$ sont des paramètres.

***b*** – $\Phi(x,y)$ étant une densité de probabilité, exprimer $A$ en fonction des autres paramètres $\alpha$ et $\beta$. Pour effectuer le calcul, on écrira que $x^2 - 2\alpha xy + y^2 = (x - \alpha y)^2 + (1-\alpha^2)y^2$, puis on effectuera un changement de variables : attention au jacobien...

***c*** – Donner l'expression de la fonction de répartition des variables $x$ et $y$ notée $\Psi(x, y)$. On sait que la répartition de la variable $x$ est $\Phi_x(x) = \Psi(x, \infty)$. Calculer la fonction de distribution de la variable $x$.

***d*** – Quelle relation doivent satisfaire les coefficients $\alpha$ et $\beta$ pour que $\Phi_x(x)$ soit la loi normale réduite. Montrer que, dans ces conditions, la fonction de distribution de la variable $y$ notée $\Phi_y(y)$ suit aussi une loi normale réduite.

***e*** – Calculer $K_{xy}$ la covariance des variables $x$ et $y$, en déduire le coefficient de corrélation $r_{xy}$. Dans le cas où $\alpha = 0$, donner l'expression de $\Phi(x, y)$ en fonction de $\Phi_x(x)$ et $\Phi_y(y)$. Peut-on déduire que la non-corrélation entraîne l'indépendance des variables $x$ et $y$? Argumenter.

***f*** – Dans le cas où $\alpha \neq 0$, donner l'expression de la droite de régression de $y$ en $x$.

***g*** – Calculer la fonction caractéristique de $\Phi(x, y)$.

## 18.3. Analyse de régression. Distribution de Student

Reprendre les fichiers fabriqués au cours des exercices 17.1 et 17.2.

Calculer les coefficients de la droite de régression $y = ax + b$ dans chacun des différents cas.

Estimer la précision sur chacun des coefficients $a$ pour un seuil de signification donné $\alpha$.

Donner l'intervalle de confiance concernant $y$ pour une valeur donnée $x_0$ de la variable $x$ dans chacun des cas typiques.

## 18.4. Étude d'un changement de phase

L'étude d'un changement de phase permet d'effectuer la mesure d'une grandeur thermodynamique $Y$ en fonction de la température $X$. La représentation graphique donnée par la figure H.6 montre que les points expérimentaux se disposent selon deux segments de droite désignés par $S_1$ et $S_2$.

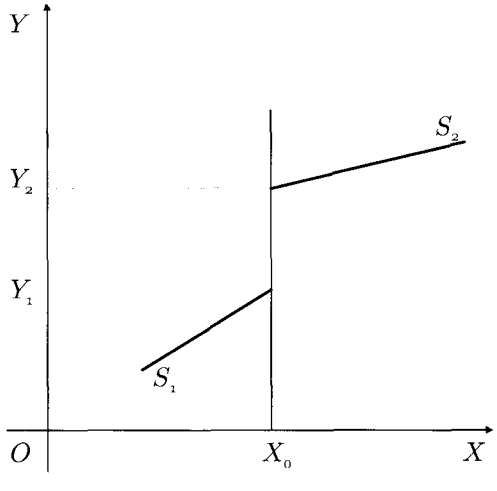


**Figure H.6.** *Étude d'un changement de phase.*

Expérimentalement, on sait que l'incertitude sur chacune des mesures effectuées $(Y_k)$, qu'il s'agisse des points du segment $S_1$ ou des points du segment $S_2$, correspond au même écart type $\sigma$. Les incertitudes constituent une variable aléatoire dont la moyenne est nulle. De plus, on

admettra qu'il n'y a aucune incertitude sur les valeurs $(X_k)$ du paramètre $X$. Une extrémité de $S_1$ et une extrémité de $S_2$ correspondent à la même abscisse $X_0$ qui est bien déterminée expérimentalement. On s'intéresse à la transition au point $X_0$, c'est-à-dire à la grandeur $H = Y_2 - Y_1$ où $Y_1$ et $Y_2$ sont représentés sur la figure.

***a*** – Sachant que le segment $S_1$ est déterminé par $N_1$ mesures et le segment $S_2$ par $N_2$ mesures, indiquer de quelle manière on pourra estimer la grandeur $H$. On justifiera l'option retenue.

***b*** – Estimer l'incertitude sur la détermination de $H$, il s'agit bien sûr d'une détermination statistique.

***c*** – Réaliser l'organigramme des opérations conduisant au traitement correct des données expérimentales.

## 18.5. Mesure d'une température par spectroscopie

L'enregistrement d'une bande électronique émise par une molécule diatomique permet, dans la mesure où l'équilibre thermodynamique local est atteint, de déterminer la température de ladite molécule.

Le principe est le suivant : la bande est constituée d'une suite de raies dont on peut mesurer pour chacune la surface, laquelle est proportionnelle à l'intensité relative $I_J$ (éventuellement après avoir tenu compte de la correction due à la fonction d'appareil), et auxquelles on peut affecter un nombre entier $J$ appelé nombre quantique de rotation.

Sans entrer dans les détails, pour de nombreuses molécules diatomiques, la théorie permet d'établir la relation suivante :

$$\log_e\left(\frac{I_J}{2J+1}\right) = A - B\frac{J(J+1)}{T}$$

où $I$ est l'intensité relative de la raie $J$ (donnée expérimentale), $A$ est une constante calculée (donnée considérée comme rigoureuse), $B$ est une constante calculée (donnée considérée comme rigoureuse), $T$ est la température absolue que l'on cherche à évaluer.

***a*** – Sachant que l'on dispose de $N$ données expérimentales $I_k$, comment déterminer $T$?

***b*** – Quelle est la précision statistique de cette détermination?

***c*** – Réaliser l'organigramme des opérations décrivant le traitement automatique des données expérimentales.

## 18.6. Solubilité du nitrate de sodium dans l'eau

Voici le tableau H.5 de résultats, obtenus par Mendéléev (1834–1907), donnant la solubilité $y$ dans l'eau du nitrate de sodium en fonction de la température $T$, soit $y = f(T)$, les grandeurs $y$ et $T$ étant exprimées en unités arbitraires.

**Tableau H.5.**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | 66,7 | 71 | 76,3 | 80,6 | 85,7 | 92,9 | 99,4 | 113,6 | 125,1 |
| $T$ | 0 | 4 | 10 | 15 | 21 | 29 | 36 | 51 | 68 |

Peut-on approcher la loi de solubilité par une relation linéaire? On justifiera la réponse puis on calculera les coefficients de la droite de régression de $y$ en $T$, coefficients obtenus par la méthode des moindres carrés.

On désire connaître la solubilité pour la valeur $T = 40$ laquelle ne figure pas dans le tableau. Donner alors la valeur correspondante de la solubilité ainsi que l'intervalle de confiance de cette détermination pour un niveau de confiance de 0,9.

On donne un extrait de la table de Student :

$$t_{0,05}(7) = 1{,}895 \qquad t_{0,05}(9) = 1{,}833 \qquad t_{0,1}(7) = 1{,}415 \qquad t_{0,1}(9) = 1{,}383.$$

# 19. Les fractions continues

## 19.1. Calcul d'une fonction donnée sous forme d'une fraction continue

On dit que la fonction $y(x)$ est donnée sous forme d'une fraction continue lorsqu'elle s'écrit sous la forme :

$$y(x) = A_0 + \cfrac{x - a_0}{A_1 + \cfrac{x - a_1}{A_2 + \cfrac{x - a_2}{+A_3 \cfrac{x - a_3}{+A_4 + \cdots \cfrac{x - a_{n-1}}{A_n + \cdots}}}}}.$$

On se propose de la calculer à l'ordre $n$. Pour y parvenir, on forme une suite d'approximations $\frac{P_1}{Q_1}, \frac{P_2}{Q_2}, \frac{P_3}{Q_3}, \cdots, \frac{P_n}{Q_n}$ qui sont respectivement à l'ordre 1, à l'ordre 2, ... à l'ordre $n$.

***a*** – Au moyen d'un raisonnement par récurrence, démontrer les relations suivantes :

$$P_{m+1}(x) = A_m P_m(x) + (x - a_{m-1}) P_{m-1}(x)$$
$$Q_{m+1}(x) = A_m Q_m(x) + (x - a_{m-1}) Q_{m-1}(x) \quad \text{avec} \quad P_0 = 1, P_1 = A_0, Q_0 = 0, Q_1 = 1.$$

***b*** – En supposant que $f(x)$ est une fonction continue indéfiniment dérivable sur un intervalle contenant $x$ et $x_0$, on se propose d'étudier le développement de la fonction $f(x)$ au voisinage de la valeur $x_0$ de la variable $x$ qui est de la forme suivante :

$$f(x) = A_0 + \cfrac{x - x_0}{A_1 + \cfrac{x - x_0}{A_2 + \cfrac{x - x_0}{+A_3 \cfrac{x - x_0}{+A_4 + \cdots \cfrac{x - x_0}{A_n + \cdots}}}}}.$$

Donner les valeurs de $A_0$ et de $A_1$ en fonction de $f(x)$ et de ses dérivées.

***c*** – On montre que les $A_k$ obéissent à la relation de récurrence $\frac{k+1}{A_{k+1}} = \frac{k-1}{A_{k-1}} + \frac{\mathrm{d}A_k}{\mathrm{d}x_0}$. Donner la valeur de $A_3$ en fonction des coefficients $c_k = \frac{1}{k!} f^{(k)}(x_0)$ avec $k = 0, 1, 2, \ldots$

***d*** – Donner le développement de $\log_e(1 + x)$ en fraction continue, puis calculer numériquement la valeur de $\log_e(2)$.

**e** – Rappeler le développement de $\log_e(1+x)$ en série de MacLaurin, puis calculer numériquement la valeur de $\log_e(2)$ en retenant les quatre premiers termes.

**f** – Estimer l'erreur entachant les résultats obtenus aux § d et e. Expliquer pourquoi le développement en fractions continues présente un avantage sur le développement en série de MacLaurin.

# 20. Éléments de traitement du signal

## 20.1. Détermination d'une constante de temps

On désire réaliser l'étude expérimentale de la décharge d'un condensateur $C$ à travers une résistance $R$. Pour cela on mesure la tension $V_i$ aux bornes du condensateur à des temps $t_i$. Les résultats obtenus sont présentés dans le tableau H.6.

**Tableau H.6.**

| $t$ (s) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $V$ (volt) | 100 | 75 | 55 | 40 | 30 | 20 | 15 | 10 | 10 | 5 | 5 |

On se propose d'étudier ces données selon deux approches différentes, mais avant de passer aux applications numériques, on établira des relations générales obtenues en considérant que :

1. les données sont numérotées de 0 à $N$,
2. les temps $t_i$ sont en progression arithmétique de raison $\Delta t$.

Rappeler l'expression analytique de la tension $V$ en fonction du temps $t$.

***Approche n° 1*** – Montrer qu'au moyen d'une transformation convenable, on peut se ramener à un problème linéaire dont on déterminera les constantes au moyen de la méthode des moindres carrés. On désigne par $a$ l'inverse de la constante de temps. Donner la valeur numérique de $a$ ainsi que l'intervalle de confiance $\pm\, da$ défini de telle sorte qu'il y ait 70 chances sur 100 que le résultat réel tombe à l'intérieur de cet intervalle.

***Approche n° 2*** – On recherche $V(t)$ sous la forme suivante :

$$V(t) = A + B\exp(b \cdot t),$$

et l'on se propose d'obtenir $A$, $B$ et $b$ par la méthode des moindres carrés. Pour se faire, on pose :

$$u = \exp(b \cdot \Delta t)$$
$$\text{et} \quad p_i = \exp(b \cdot t_i) = \exp[b(t_0 + i\Delta t)].$$

Donner l'expression de $V_j$ en fonction de $A$ et $p_j$, puis celles de $V_{j+1}$ , $V_{j+2}\ldots$ en fonction de $A$, $p_j$ et $u$.

Si l'on note $\Delta V_j = V_{j+1} - V_j$, donner alors les expressions de $\Delta V_j$, $\Delta V_{j+1}$ etc.

Éliminer $p_k$ entre deux $\Delta V_k$ consécutifs et donner l'équation qui les lie. Combien obtiendra-t-on de ces équations? On désignera par $K$ ce nombre. Résoudre ce système de $K$ équations à une

inconnue par la méthode des moindres carrés et en déduire la valeur de $b$. Donner l'expression de l'erreur quadratique correspondante $E_a^2$.

Il reste à déterminer $A$ et $B$. Montrer comment les obtenir par la méthode des moindres carrés, donner leurs expressions ainsi que l'écart quadratique correspondant $E_{ab}^2$.

Calculer numériquement les valeurs de $A$, $B$ et $b$.

Que concluez-vous quant à l'utilisation de ces deux méthodes?

Puisque les valeurs de $a$ et $b$ sont différentes, donner l'intervalle de confiance minimum $\Delta a$ permettant d'encadrer la valeur $b$. Donner alors la probabilité de tomber dans l'intervalle ainsi défini.

# I Corrigés des problèmes et exercices

## 1. Généralités sur le calcul numérique

### 1.1. Calcul du nombre d'or

La suite $U_n$ est formée de termes strictement positifs pour $n > 0$. La suite $k_n$ est évidemment minorée par 0, il suffit de montrer qu'elle admet une limite supérieure, pour cela, il suffit de remarquer que $U_n = U_{n-1} + U_{n-2}$ et d'écrire la relation de définition de $k_n$ :

$$k_n = \frac{U_n}{U_{n-1}} = \frac{U_{n-1}}{U_{n-1}} + \frac{U_{n-2}}{U_{n-1}} = 1 + \frac{U_{n-2}}{U_{n-1}} \cdot$$

Le dernier terme est plus petit que 2 car la suite $U_n$ est strictement croissante et positive ; on en conclut que $k_n$ est compris entre 0 et 2 quel que soit $n > 2$ et que la suite $k_n$ admet donc une limite.

Désignons par $\gamma$ cette limite, on écrit alors : $U_{n+1} - U_n - U_{n-1} = 0$ et l'on divise chaque terme par $U_{n-1}$ qui est différent de zéro pour $n > 2$. On obtient donc :

$$\frac{U_{n+1}}{U_{n-1}} - \frac{U_n}{U_{n-1}} - 1 = 0, \quad \text{soit encore :} \quad \frac{U_{n+1}}{U_n} \cdot \frac{U_n}{U_{n-1}} - \frac{U_n}{U_{n-1}} - 1 = 0.$$

Quand $n$ tend vers l'infini la dernière équation devient : $\gamma^2 - \gamma - 1 = 0$, dont la racine positive est :

$$\gamma = \frac{1 + \sqrt{5}}{2} = 1{,}618\,033\,989\ldots$$

Sur le plan du calcul numérique de $\gamma$ au moyen de la suite de Fibonacci, on remarque que $k_n$ est le rapport de deux entiers $U_n$ et $U_{n-1}$ lesquels admettent une représentation sans erreur (pourvu qu'ils possèdent moins de 16 chiffres significatifs en notation décimale si l'on envisage l'usage du langage C en double précision). La seule erreur sur $k_n$ est donc l'erreur liée à la division. Si l'on se ramène à l'erreur relative, il sera inutile de poursuivre les calculs après que deux valeurs consécutives de $k_n$ seront égales à la précision relative $\varepsilon_r$ donnée par la machine (et le langage utilisé) ; cela se traduit par l'égalité :

$$\frac{k_{n+1} - k_n}{k_n} = \varepsilon_r.$$

Cette condition donne l'erreur relative la plus petite sur le calcul de $\gamma$, mais le procédé est coûteux en temps de calcul. Le calcul direct de $\gamma$ donné par la résolution de l'équation du deuxième degré fournit une erreur relative plus grande : il faut tenir compte de l'erreur sur la racine carrée puis de l'erreur sur la soustraction suivie de l'erreur sur la division...

***Remarque*** : Sur le plan de la programmation, il ne faut pas réserver un tableau pour la variable indicée $k_n$, seules deux valeurs consécutives sont nécessaires et suffisantes, et l'on fera un décalage chaque fois que l'on calculera une nouvelle valeur. D'ailleurs, on ne connaît pas la taille de ce tableau.

## 1.2. Les polynômes de Tchebycheff

Posons $y = \arccos(x)$, puis écrivons la relation de définition pour $T_{n-1}(x)$ et $T_{n+1}(x)$ :

$$T_{n-1}(x) = \cos[(n-1)y] = \cos(ny)\cos(y) + \sin(ny)\sin(y)$$
$$T_{n+1}(x) = \cos[(n+1)y] = \cos(ny)\cos(y) - \sin(ny)\sin(y)$$

additionnons ces deux relations :

$$T_{n+1}(x) + T_{n-1}(x) = 2\cos(ny)\cos(y)$$

enfin retournons aux anciennes variables en remarquant que $x = \cos(y)$, soit :

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

C'est donc une relation de récurrence entre trois polynômes consécutifs. Pour l'utiliser, il suffit de connaître les deux premiers polynômes $T_0(x)$ et $T_1(x)$ à savoir :

$$T_0(x) = \cos(0) = 1,$$
$$\text{et} \quad T_1(x) = \cos[\arccos(x)] = x.$$

Ici encore, il ne faut pas utiliser de tableaux sauf si l'on désire effectuer une représentation graphique de ces polynômes.

Il existe une source de difficultés qui va affecter la précision : le second membre de la relation de récurrence est une **différence**, il s'ensuit que, si les deux termes sont sensiblement les mêmes, une erreur importante pourra surgir et qui se propagera au cours des calculs. Il est donc intéressant de comparer les résultats obtenus avec l'utilisation directe de la fonction $\arccos(x)$ en cherchant à mettre en défaut l'algorithme. Il s'agit de faire ses gammes...

## 1.3. Calcul des fréquences de la gamme naturelle et de la gamme tempérée

Une des façons de fonder la gamme naturelle consiste à calculer les fréquences au moyen des quintes qui se situent sept demi-tons majeurs au-dessus de la note précédente : le rapport des fréquences est trois sur deux. Ce processus permet de déterminer les dièses. Pour ce qui concerne les bémols, ils sont déterminés par les quintes qui se situent sept demi-tons majeurs au-dessous de la note précédente, les fréquences sont dans le rapport deux sur trois. Une manière simple d'opérer consiste à écrire les douze demi-tons de la gamme puis à les calculer successivement. Si la fréquence d'une note sort de la gamme à déterminer, on la divise par deux s'il s'agit des dièses ou on la multiplie par deux s'il s'agit des bémols. Commençons à partir du $la_3$ (440 Hz) par définir les dièses.

La quinte du *la* est la note *mi* dont la fréquence est $440 \times 3/2 = 660$ Hz. On note que 660 Hz se situe bien dans l'intervalle (440 Hz, 880 Hz). Calculons à présent la fréquence de la quinte de la note *mi*, il s'agit de la note *si* : $660 \times 3/2 = 990$ Hz, cette fréquence n'appartient pas au bon intervalle puisqu'elle est supérieure à 880 Hz, c'est l'octave de la note dont on cherche la fréquence. Il suffit de diviser par deux la valeur de la fréquence pour se ramener dans le bon intervalle, soit 495 Hz. En poursuivant ainsi, on trouve les notes et les fréquences présentées dans le tableau I.1.

**Tableau I.1.**

| | | | |
|---|---|---|---|
| *la* | 440 Hz | *ré#* | 626,48 Hz |
| *mi* | 660 Hz | *la#* | 469,86 Hz |
| *si* | 495 Hz | *fa* | 704,79 Hz |
| *fa#* | 742,5 Hz | *do* | 528,59 Hz |
| *do#* | 556,88 Hz | *sol* | 792,89 Hz |
| *sol#* | 835,31 Hz | *ré* | 594,67 Hz |

En poursuivant ainsi, on calculerait les doubles dièses. Maintenant, calculons les notes qui vont éventuellement être affectées d'un bémol (*cf.* Tab. I.2).

**Tableau I.2.**

| | | | |
|---|---|---|---|
| *la* | 440 Hz | $mi_b$ | 618,05 Hz |
| *ré* | 586,66 Hz | $la_b$ | 412,03 Hz |
| *sol* | 782,22 Hz | $ré_b$ | 549,38 Hz |
| *do* | 521,48 Hz | $sol_b$ | 732,51 Hz |
| *fa* | 695,31 Hz | *si* | 488,34 Hz |
| $si_b$ | 463,54 Hz | *mi* | 651,12 Hz |

En continuant de la sorte, on peut définir les doubles bémols. Quoi qu'il en soit on voit aisément que les différentes tonalités ne possèdent pas les mêmes notes, ou plus exactement, les mêmes notes ne possèdent pas les mêmes fréquences. Autrement dit, une fois l'instrument accordé dans une tonalité, il n'est pas possible d'en sortir (modulation) sans avoir le sentiment que l'instrument joue faux. La gamme naturelle propose un *do#* à 557 Hz et un $ré_b$ à 549 Hz. La gamme tempérée, due à J.S. Bach (1685–1750), propose d'accorder la même fréquence à ces deux notes. Autrement dit, l'intervalle d'octave sera divisé en douze demi-tons d'égale valeur dans une échelle logarithmique.

Donc le rapport du logarithme des fréquences de deux notes consécutives, séparées alors par un demi-ton, est une constante quelle que soit la fréquence initiale de la gamme retenue, elle vaut $(1/12)\log_e(2) = 0{,}057\,762$. On établit très facilement la table de la gamme tempérée (*cf.* Tab. I.3, page suivante).

On note que le *do#* et $ré_b$ ont la même fréquence de 554 Hz qui est intermédiaire entre les deux précédentes valeurs. Pour terminer, disons qu'il existe bien d'autres gammes et de façons de les concevoir et que l'accord d'un piano se rapproche de la gamme tempérée mais en diffère par de petites nuances qui donnent du relief, du volume, de la profondeur à l'instrument.

**Tableau I.3.**

| | | | |
|---|---|---|---|
| *la* | 440,00 Hz | *mi* | 659,25 Hz |
| *la#* ou $si_b$ | 466,16 Hz | *mi#* ou *fa* | 698,46 Hz |
| *si* ou $do_b$ | 493,88 Hz | *fa#* ou $sol_b$ | 739,99 Hz |
| *si#* ou *do* | 523,25 Hz | *sol* | 783,99 Hz |
| *do#* ou $ré_b$ | 554,37 Hz | *sol#* ou $la_b$ | 830,61 Hz |
| *ré* | 587,33 Hz | *la* | 880,00 Hz |
| *ré#* ou $mi_b$ | 622,25 Hz | | |

## 1.4. Calcul de la constante d'Euler (1707–1783)

Avec une machine qui effectue les calculs avec une précision relative de $10^{-q}$, on obtient la précision optimale lorsque deux valeurs consécutives de la suite des $\gamma_k$ seront égales à la précision de la représentation, soit :

$$\left|\frac{\gamma_{k+1}-\gamma_k}{\gamma_k}\right| = \left|\left\{\frac{1}{k+1} - \log_e\left(\frac{k+1}{k}\right)\right\}/\gamma_k\right| \approx 10^{-q}.$$

En admettant que $\gamma_k$ soit minoré par $1/2$, on obtient l'équation transcendante qui donne $k$ en fonction de $q$ c'est-à-dire :

$$\left|\frac{1}{k+1} - \log_e\left(\frac{k+1}{k}\right)\right| \approx 2\cdot 10^{-q}.$$

Sachant que $k$ est très grand devant l'unité, on peut se livrer aux approximations commodes :

$$\log_e\left(\frac{k+1}{k}\right) = \log_e\left(1+\frac{1}{k}\right) \approx \frac{1}{k},$$

il s'ensuit :

$$\left|\frac{1}{k+1} - \frac{1}{k}\right| \approx \frac{1}{k^2} \approx 2\cdot 10^{-q},$$

de là on tire une valeur approchée de $k \approx \sqrt{1/2}\,10^{q/2}$. En utilisant une représentation de 16 chiffres significatifs, il faut calculer environ dix millions de termes ce qui est évidemment prohibitif s'il existe d'autres méthodes. Estimons néanmoins l'erreur risquant d'affecter le résultat, dans le cas le plus pessimiste évidemment. Écrivons la relation de récurrence entre deux valeurs consécutives de $\gamma_k$ :

$$\gamma_{k+1} = \gamma_k + \frac{1}{k+1} - \log_e\left(\frac{k+1}{k}\right)$$

il faut réaliser à peu près $2k$ additions (soustractions) qui introduisent une erreur absolue de l'ordre de $2\cdot 10^{-q}$ dans le pire des cas. On ne pourra donc compter que sur 7 chiffres significatifs après tant de calculs...

Il n'y a pas de difficulté majeure à programmer la relation faisant intervenir les nombres de Bernoulli, il s'agit d'un développement asymptotique qui est divergent mais qui donne

d'excellents résultats si l'on tronque habilement la série. Lorsque l'on néglige les termes après le rang $\mu$, on commet sur $\gamma$ une erreur strictement mathématique inférieure à $2B_\mu/(2\mu m^{2\mu})$. Fixons $m$ à la valeur 100 par exemple et regardons quelle doit être la valeur de $\mu$ pour obtenir la meilleure précision mathématique. Raisonnons encore sur les erreurs absolues nous pouvons écrire :

$$2 \cdot 10^{-q} = 2\frac{B_\mu}{2\mu 100^{2\mu}} \cdot$$

Il n'y a pas d'autre solution que d'effectuer une tabulation pour $\mu = 1, 2, 3, \ldots$ Réalisons cette opération pour $\mu = 4$ et 5, on obtient respectivement pour le membre de droite :

$$\frac{1}{120\,100^8} \approx 10^{-18} \quad \text{et} \quad \frac{5}{660\,100^{10}} \approx 10^{-22}.$$

En conclusion, en choisissant $m = 4$, l'erreur de troncature strictement mathématique est inférieure à la précision usuelle de $10^{-16}$. Il convient d'évaluer les erreurs liées à l'exécution des calculs, elles seront majorées par $4m10^{-16}$ avec $m = 100$, autrement dit on aura mieux que 13 chiffres significatifs.

### 1.5. Calcul numérique des dérivées d'un polynôme à coefficients réels pour la valeur x = r

***a*** – Le schéma de Horner (1786–1837) est un algorithme qui permet de calculer la valeur d'un polynôme au moyen d'un nombre minimum d'opérations en utilisant un système de parenthèses :

$$P_n(x) = \sum_{k=0}^{n} a_k x^k = \Big\{\ldots\Big[\big\{[(a_0x + a_1)x + a_2]x + a_3\big\}x + a_4\Big]x + \cdots\Big\}x + a_n.$$

Pour $x = r$, on calcule donc la suite :

$$\begin{aligned}
v_0 &= a_0 \\
v_1 &= v_0 r + a_1 \\
v_2 &= v_1 r + a_2 \\
&\cdots\cdots \\
v_{k+1} &= v_k r + a_{k+1} \\
&\cdots\cdots \\
v_n &= v_{n-1} r + a_n = P_n(r).
\end{aligned}$$

***b*** – On écrit la suite des polynômes :

$$\begin{aligned}
P_n(x) &= (x - r)P_{n-1}(x) + R_n \\
P_{n-1}(x) &= (x - r)P_{n-2}(x) + R_{n-1} \\
&\cdots\cdots\cdots \\
P_{n-q}(x) &= (x - r)P_{n-q-1}(x) + R_{n-q} \\
&\cdots\cdots\cdots \\
P_1(x) &= (x - r)P_0(x) + R_1,
\end{aligned}$$

On écrit $P_0(x) = R_0$ car $P_0(x)$ est une simple constante.

***c*** – Le polynôme prend la forme :

$$P_n(x) = (x - r)^n R_0 + (x - r)^{n-1} R_1 + \cdots + (x - r)^{n-p} R_p + \cdots + (x - r)R_{n-1} + R_n.$$

***d*** – On écrit le développement en série de Taylor :

$$P_n(x) = P_n(r+h) = \sum_{q=0}^{n} \frac{(x-r)^q}{q!} P_n^{(q)}(r),$$

expression dans laquelle on a remplacé $h$ par $(x-r)$. Par identification avec l'expression du § c, on obtient :

$$R_{n-p} = \left[\frac{1}{p!}\frac{\mathrm{d}^p}{\mathrm{d}x^p} P_n(x)\right] \quad \text{pour } x = r.$$

***e*** – Pour calculer effectivement les $R_k$, il suffit d'appliquer le schéma de Horner aux coefficients de $P_n(x)$ puis à ceux de $P_{n-1}(x)$ et ainsi de suite jusqu'à celui de $P_0(x) = R_0$. Désignons par $b_j$ les coefficients de $P_{n-1}(x)$ ; on les calcule en fonction des $a_k$ de la façon suivante :

$$\begin{aligned}
P_{n-1}(x) &= \sum_{j=0}^{n-1} b_j x^j \\
b_0 &= a_0 \\
b_1 &= a_1 + b_0 r \\
&\cdots\cdots\cdots \\
b_k &= a_k + b_{k-1} r \\
&\cdots\cdots\cdots \\
R_n &= a_n + b_{n-1} r,
\end{aligned}$$

on poursuit les mêmes calculs pour le polynôme $P_{n-2}(x)$ en notant toutefois que l'on perd une rangée. Ensuite on opère de la même manière pour $P_{n-3}(x)$ et ainsi de suite jusqu'à l'obtention de $R_0$.

## 1.6. Calcul numérique de la fonction ζ de Riemann

***a*** – On a : $I = \int_1^\infty \frac{1}{x^s}\,\mathrm{d}x$ et deux cas se présentent selon que $s = 1$ ou $s \neq 1$ :

$$\begin{aligned}
&\text{si } s = 1 \text{ on obtient} \quad I = [\log_e(x)]_1^\infty \quad \text{qui diverge,} \\
&\text{et si } s \neq 1 \text{ on obtient} \quad I = [-\frac{1}{s-1} \cdot \frac{1}{x^{s-1}}]_1^\infty, \quad \text{cette expression est finie} \\
&\text{si } s > 1 \text{ et alors on a :} \quad I = \frac{1}{s-1}\cdot
\end{aligned}$$

***b*** – L'erreur $e_k$ est comprise entre $\int_k^\infty \frac{1}{x^s}\,\mathrm{d}x$ et $\int_{k+1}^\infty \frac{1}{x^s}\,\mathrm{d}x$, on en déduit alors les inégalités suivantes :

$$S_k(s) + \int_{k+1}^{\infty} \frac{1}{x^s}\,\mathrm{d}x < \zeta(s) < S_k(s) + \int_{k}^{\infty} \frac{1}{x^s}\,\mathrm{d}x$$

ce qui s'écrit encore :

$$S_k(s) + \frac{1}{s-1}\cdot\frac{1}{(k+1)^{s-1}} < \zeta(s) < S_k(s) + \frac{1}{s-1}\cdot\frac{1}{k^{s-1}}\cdot$$

Prenons un exemple : $\zeta(2) = \dfrac{\pi^2}{6} (= 1{,}644\,934)$, on calcule alors $S_{100}(2) = 1{,}634\,983\,903$ puis on déduit que :

$$1{,}634\,983 + \frac{1}{101} < \zeta(2) < 1{,}634\,983 + \frac{1}{100},$$
$$\text{soit encore :} \quad 1{,}644\,88 < \zeta(2) < 1{,}644\,98.$$

Il existe deux programmes `dzeta0.c` et `dzeta1.c` qui réalisent ces algorithmes.

## 1.7. Recherche d'une tangente commune à deux courbes

***a*** – On a $y_f = f(x_f)$ et $f(x_f) - y_0 - (x_f - x_0)f'(x_f) = 0$ pour la première courbe et pour la seconde : $y_g = g(x_g)$ et $g(x_g) - y_0 - (x_g - x_0)g'(x_g) = 0$.

En général il faut rechercher la racine d'équations transcendantes (voire implicites) qui donnent $x_f$ et $x_g$ puis on calcule $y_f$ et $y_g$. Pour cela, on utilise une des méthodes itératives proposées dans le cours.

***b*** – Passant par le point $A$, il existe en général deux tangentes à la courbe $C_1$ et à la courbe $C_2$. Il faut donc choisir la tangente la plus petite en module.

***c*** – On désigne par $A_1$ un point de la droite $x = x_0$ dont l'ordonnée est supérieure au plus grand des deux extremums de $C_1$ et $C_2$, et par $A_2$ un point de la droite $x = x_0$ dont l'ordonnée est inférieure au plus petit des deux extremums de $C_1$ et $C_2$.

On considère la fonction $\Phi = \dfrac{y_f - y_0}{x_f - x_0} - \dfrac{y_g - y_0}{x_g - x_0}$ qui est la différence des pentes des tangentes menées de $A$ aux deux courbes. Quand $A$ parcourt le segment $A_1A_2$ la fonction $\Phi$ ne s'annule qu'une seule fois, et c'est pour cette valeur précisément que l'on obtient la tangente commune aux deux courbes. La dichotomie, par exemple, peut nous permettre d'obtenir cette valeur de $y_0$. On examine le signe de cette fonction en $A_1$ puis en $(A_1 + A_2)/2$ et ainsi de suite. Si la longueur $A_1A_2$ est de l'ordre de quelques unités, alors en une cinquantaine de tours d'itération nous aurons obtenu la valeur de $y_0$ par où passe la tangente commune avec une précision de l'ordre de $10^{-15}$.

À condition que les fonctions $f(x)$ et $g(x)$ soit calculables de façon semblable, pour avoir des erreurs semblables sur les calculs de $x_f$ et $x_g$, et une bonne précision sur la fonction $\Phi$, il faut choisir $x_0$ approximativement au milieu de l'intervalle $(x_f, x_g)$.

***d*** – Dans le cas de fonctions implicites, on a $\Phi(x_f, y_f) = 0$. La tangente à $C_1$ s'écrit :

$$\frac{\mathrm{d}\Phi}{\mathrm{d}x} = \frac{\partial\Phi}{\partial x} + \frac{\partial\Phi}{\partial y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} = 0$$
$$\text{d'où :} \quad \frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{\Phi'_x}{\Phi'_y} \quad \text{pour} \quad x = x_f \quad \text{et} \quad y = y_f.$$

On écrit encore : $\dfrac{y_f - y_0}{x_f - x_0} = -\dfrac{\Phi'_x}{\Phi'_y}$ pour $x = x_f$ et $y = y_f$. Donc, pour connaître les valeurs $x_f$ et $y_f$, il faut résoudre le système non linéaire de deux équations à deux inconnues par une des méthodes itératives proposées dans le cours. On opère de la même façon pour l'autre tangente pour laquelle on écrit :

$$\Psi(x_g, y_g) = 0,$$

et :

$$\frac{y_g - y_0}{x_g - x_0} = -\frac{\Psi'_x}{\Psi'_y} \quad \text{pour } x = x_g \quad \text{et } y = y_g.$$

On règle le problème du point $A$ de la même manière qu'au paragraphe précédent.
On fournit le programme `tangent.c` concernant cette procédure.

## 2. Algorithmes accélérateurs

(Voir le cours chapitre 2)

## 3. Les développements asymptotiques

### 3.1. Développement asymptotique

Une première intégration par parties donne :

$$f(x) = \left[-\frac{1}{t}\exp(x-t)\right]_x^\infty - \int_x^\infty t^{-2}\exp(x-t)\,\mathrm{d}t = \frac{1}{x} - \int_x^\infty t^{-2}\exp(x-t)\,\mathrm{d}t.$$

En poursuivant ainsi les calculs on trouve :

$$f(x) = \frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \cdots + (-1)^n\frac{n!}{x^{n+1}} + (-1)^{n+1}(n+1)!\int_x^\infty t^{-n-2}\exp(x-t)\,\mathrm{d}t.$$

Cette série diverge quand $n$ tend vers l'infini puisque le terme général devient infini quel que soit $x$. Calcul de l'erreur $\varepsilon_n(x)$ lorsque l'on tronque la série après le terme d'ordre $n+1$ :

$$\varepsilon_n(x) = (n+1)!\int_x^\infty t^{-n-2}\exp(x-t)\,\mathrm{d}t$$

mais étant donné que $\exp(x-t) < 1$ car $x > 0$ et $t > x$ on peut écrire l'inégalité :

$$\varepsilon_n(x) < (n+1)!\int_x^\infty \frac{1}{|t^{n+2}|}\,\mathrm{d}t < \frac{n!}{|x^{n+1}|},$$

et pour $n$ fixé, on voit que l'erreur tend vers zéro quand $x$ croît indéfiniment.

La fonction $\varphi(n) = \dfrac{n!}{|x^{n+1}|}$ est une fonction strictement positive quels que soient $n$ et $x$ ; pour l'étudier, il est commode de lui associer la fonction $\psi(n) = \log_e[\varphi(n)]$ :

$$\psi(n) = n\log_e(n) - n - (n+1)\log_e(x)$$

dont la dérivée s'écrit ainsi :

$$\frac{\mathrm{d}\psi(n)}{\mathrm{d}n} = \log_e(n) - \log_e(x),$$

elle a un minimum pour $n = N = x$ ; cela signifie que la fonction $\varphi(n)$ a aussi un minimum. On va voir que c'est le même minimum. Pour $x$ fixé, l'erreur sera minimum pour une valeur de $n_0$ qui est approximativement donnée par l'égalité :

$$\varepsilon_{n_0}(x) = \varepsilon_{n_0+1}(x),$$

soit encore :

$$\frac{N!}{x^{N+1}} = \frac{(N+1)!}{x^{N+2}}$$

ce qui donne $n_0 = x - 1$. Ce résultat n'est pas en contradiction avec ce qui précède, en effet $x$ est une variable réelle tandis que $n$ est une variable entière. On aurait masqué cette remarque en écrivant l'égalité $\varepsilon_{n_0-1}(x) = \varepsilon_{n_0}(x)$ car on aurait trouvé $n_0 = x$. On aurait tout aussi bien pu écrire $\varepsilon_{n_0-1}(x) = \varepsilon_{n_0+1}(x)$ qui aurait donné encore un résultat légèrement différent : $x = \sqrt{n_0(n_0+1)}$ qui a une racine approchée de l'ordre de $n_0 = x$ ou $n_0 + 1 = x$.

***a – Application de l'epsilon-algorithme*** – Le développement asymptotique étant un prolongement analytique, on peut s'attendre à ce que la convergence se réalise même si la condition précédente n'est pas vérifiée. Dans la mesure de la « calculabilité effective », on doit trouver des résultats corrects aussi bien pour des petites valeurs de $x$ que pour de grandes valeurs de $n$. C'est ce que nous avons vérifié.

***b – Cas de la fonction erreur*** – Pour ce qui concerne la fonction :

$$\operatorname{cerf}(x) = 1 - \theta(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2)\, \mathrm{d}t,$$

on effectue une suite d'intégration par parties et l'on obtient :

$$\operatorname{cerf}(x) = 1 - \theta(x) = \frac{\exp(-x^2)}{x\sqrt{\pi}} \left\{ 1 - \frac{1}{2x^2} + \frac{1\cdot 3}{2^2 x^4} - \frac{1\cdot 3\cdot 5}{2^3 x^6} + \cdots + (-1)^{n-1} \frac{1\cdot 3\cdot 5 \ldots (2n-3)}{2^{n-1} x^{2n-2}} + \cdots \right\}.$$

## 4. Résolution des équations numériques

### 4.1. Voir le cours

### 4.2. Voir le cours

### 4.3. Calcul de la fonction arctan(x)

Comme $x_0 - \arctan(x_0) = 0$, on peut ajouter cette quantité à l'équation $x = \arctan(a)$, soit : $x = \arctan(a) + x_0 - \arctan(x_0)$ ou encore : $x - x_0 = \arctan(a) - \arctan(x_0)$ expression que l'on transforme ainsi :

$$\arctan[\tan(x - x_0)] = x - x_0 = \arctan(x) - \arctan(x_0)$$

puis en développant l'expression de $\tan(x - x_0)$, soit :

$$\tan(x - x_0) = \frac{\tan(x) - \tan(x_0)}{1 + \tan(x)\tan(x_0)}$$

d'où :

$$\arctan\left[\frac{a - \tan(x_0)}{1 + a\tan(x_0)}\right] = x - x_0$$

puisque l'on connaît $a = \tan(x)$.

Cette dernière expression peut s'écrire formellement :

$$x = x_0 + \arctan[F(x_0)] = x_0 + \arctan(F_0).$$

On développe la fonction arctan, on obtient :

$$x = x_0 + F_0 - \frac{F_0^3}{3} + \frac{F_0^5}{5} + \cdots + (-1)^n \frac{F_0^{2n+1}}{2n+1} + \cdots$$

le rayon de convergence est $|F_0| < 1$.

On se propose de calculer $\arctan\left(\frac{1}{\sqrt{3}}\right)$ avec $x_0 = 0{,}8$. Donc :

$$F_0 = \frac{\sqrt{3}/3 - \tan(0{,}8)}{1 + \sqrt{3}/3\tan(0{,}8)} = -0{,}283\,661\,987,$$

ce qui permet de calculer la valeur numérique de

$$x = x_0 + F_0 - \frac{F_0^3}{3} + \frac{F_0^5}{5} = 0{,}523\,598$$

entachée d'une erreur absolue de $2\,10^{-5}$ qui est le premier terme abandonné de la série alternée fournie par le développement.

## 4.4. Racine d'une équation f(x) = 0

***a*** – L'équation de la tangente en $A_0$ s'écrit : $y = f(x_0) + (x - x_0)f'(x_0)$. L'équation de la droite orthogonale à cette tangente et passant par $B_0$ s'écrit : $y = -(x - x_0)/f'(x_0)$, on en déduit la coordonnée du point d'intersection :

$$x_1 = x_0 - \frac{f(x_0)f'(x_0)}{1 + f'^2(x_0)}.$$

À partir de $x_1$ on peut calculer $x_2$ et ainsi de suite, d'une façon générale on aura :

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{1 + f'^2(x_k)} = \Phi(x_k).$$

***b*** – On a $e_n = X - x_n$, il s'ensuit que $x_{n+1} = \Phi(X - e_n)$ expression que l'on va développer en série de Taylor au deuxième ordre :

$$x_{n+1} = \Phi(X - e_n) = \Phi(X) - e_n \frac{\mathrm{d}\Phi}{\mathrm{d}x} + \frac{e_n^2}{2!} \cdot \frac{\mathrm{d}^2\Phi}{\mathrm{d}x^2} + \cdots,$$

et comme $\Phi(X) = X$, on obtient au premier ordre :

$$x_{n+1} = X - e_n \frac{\mathrm{d}\Phi}{\mathrm{d}x},$$

soit encore :

$$e_{n+1} = e_n \frac{\mathrm{d}\Phi}{\mathrm{d}x} = k e_n.$$

Pour que ce processus itératif fonctionne, il faut que les extremums relatifs dans $\Delta$ n'existent que sur la frontière de $\Delta$, et $f(x)$ doit donc être monotone dans $\Delta$. En outre, il faut que $f(x)$ soit continue dans $\Delta$ et qu'elle n'ait pas de tangente verticale (on dit qu'elle est lipschitzienne).

***c*** – Calculons $k = \dfrac{\mathrm{d}\Phi}{\mathrm{d}x}$. En rappelant que $f(X) = 0$ (ou $\Phi(X) = X$), après quelques calculs on trouve : $k = \dfrac{1}{1 + f'^2}$ ce qui montre que $k$ est toujours plus petit que 1 et donc que le processus est toujours convergent **à condition de ne pas sortir du domaine $\Delta$.**

***d*** – Ce processus est un processus du premier ordre, et, du point de vue de la vitesse de convergence, il est moins rapide que la méthode de Newton qui est un processus du deuxième ordre. En revanche, au voisinage d'une tangente horizontale, la méthode de Newton devient instable en rejetant très loin la valeur numérique de la suite des $x_n$. La méthode proposée ne présente pas ce défaut en offrant une grande stabilité, cependant si l'on utilise cette méthode en particulier, il faudra plus que jamais terminer le programme par le calcul de $f(x_n)$ et vérifier que cette valeur est suffisamment proche de zéro. En effet, en examinant la formule de récurrence que nous avons établie, on peut s'apercevoir que cette expression permet aussi de calculer les zéros de la dérivée $f'(x)$...

### 4.5. Racines d'un polynôme à coefficients complexes

Il s'agit là d'un exercice très proche de la méthode de Bairstow. On écrit :

$$\begin{aligned} P_n(z) &= (\alpha_0 + j\beta_0)z^n + (\alpha_1 + j\beta_1)z^{n-1} + (\alpha_2 + j\beta_2)z^{n-2} + \cdots + (\alpha_n + j\beta_n) \\ &= D_1(z)Q_{n-1}(z) + (\Phi + j\Psi) = [z - (\rho + j\sigma)](\gamma_0 + j\delta_0)z^{n-1} + (\gamma_1 + j\delta_1)z^{n-2} \\ &\quad + (\gamma_2 + j\delta_2)z^{n-3} + \cdots + (\gamma_{n-1} + j\delta_{n-1}) + (\Phi + j\Psi). \end{aligned}$$

L'identification des termes est immédiate, on obtient :

$$\begin{aligned} &\gamma_0 = \alpha_0 && \delta_0 = \beta_0 \\ &\gamma_1 = \alpha_1 + \rho\gamma_0 - \sigma\delta_0 && \delta_1 = \beta_1 + \rho\delta_0 + \sigma\gamma_0 \\ &\cdots \\ &\gamma_q = \alpha_q + \rho\gamma_{q-1} - \sigma\delta_{q-1} && \delta_q = \beta_q + \rho\delta_{q-1} + \sigma\gamma_{q-1} \\ &\cdots \\ &\Phi = \alpha_n + \rho\gamma_{n-1} - \sigma\delta_{n-1} && \Psi = \beta_n + \rho\delta_{n-1} + \sigma\gamma_{n-1}. \end{aligned}$$

Bien entendu, comme dans la méthode de Bairstow, $\Phi$ et $\Psi$ sont des fonctions de $\rho$ et $\sigma$, et nous voulons obtenir $\rho$ et $\sigma$ tels que :

$$\Phi(\rho,\sigma)=0 \quad \text{et} \quad \Psi(\rho,\sigma)=0.$$

Connaissant une première approximation $\rho_0$ et $\sigma_0$ on va chercher à calculer $h$ et $l$ en développant $\Phi$ et $\Psi$ au premier ordre :

$$0=\Phi+h\frac{\partial\Phi}{\partial\rho}+l\frac{\partial\Phi}{\partial\sigma} \qquad \text{pour } \rho=\rho_0 \text{ et } \sigma=\sigma_0,$$
$$0=\Psi+h\frac{\partial\Psi}{\partial\rho}+l\frac{\partial\Psi}{\partial\sigma} \qquad \text{pour } \rho=\rho_0 \text{ et } \sigma=\sigma_0.$$

La résolution du système linéaire donne $h$ et $l$ :

$$h=\left[\Phi\frac{\partial\Psi}{\partial\sigma}-\Psi\frac{\partial\Phi}{\partial\sigma}\right]/\det,$$
$$l=\left[\Psi\frac{\partial\Phi}{\partial\rho}-\Phi\frac{\partial\Psi}{\partial\sigma}\right]/\det,$$
$$\text{avec} \quad \det=\frac{\partial\Phi}{\partial\rho}\cdot\frac{\partial\Psi}{\partial\sigma}-\frac{\partial\Phi}{\partial\sigma}\cdot\frac{\partial\Psi}{\partial\rho},$$

ces valeurs étant toutes calculées pour $\rho=\rho_0$ et $\sigma=\sigma_0$.

Le problème consiste donc à obtenir : $\Phi$, $\Psi$, $\frac{\partial\Phi}{\partial\rho}$, $\frac{\partial\Phi}{\partial\sigma}$, $\frac{\partial\Psi}{\partial\rho}$ et $\frac{\partial\Psi}{\partial\sigma}$. Il reste à calculer les dérivées ; on calcule ensemble les dérivées par rapport à $\rho$ puis les dérivées par rapport à $\sigma$. On a donc :

$$\frac{\partial\gamma_k}{\partial\rho}=\gamma_{k-1}+\rho\frac{\partial\gamma_{k-1}}{\partial\rho}-\sigma\frac{\partial\delta_{k-1}}{\partial\rho} \qquad \frac{\partial\delta_k}{\partial\rho}=\delta_{k-1}+\rho\frac{\partial\delta_{k-1}}{\partial\rho}+\sigma\frac{\partial\gamma_{k-1}}{\partial\rho}\cdot$$

On pose :

$$\frac{\partial\gamma_k}{\partial\rho}=t_{k-1} \quad \text{et} \quad \frac{\partial\delta_k}{\partial\rho}=u_{k-1}$$

ce qui permet d'écrire :

$$t_{k-1}=\gamma_{k-1}+\rho t_{k-2}-\sigma u_{k-2} \quad \text{et} \quad u_{k-1}=\delta_{k-1}+\rho u_{k-2}-\sigma t_{k-2},$$

avec $t_0=\gamma_0$ et $u_0=\delta_0$. Ces relations aboutissent à :

$$\frac{\partial\Phi}{\partial\rho}=\gamma_{n-1}+\rho t_{n-2}-\sigma u_{n-2} \quad \text{et} \quad \frac{\partial\Psi}{\partial\rho}=\delta_{n-1}+\rho u_{n-2}-\sigma t_{n-2}.$$

De la même manière, on calcule les dérivées partielles par rapport à $\sigma$ :

$$\frac{\partial\gamma_k}{\partial\sigma}=-\delta_{k-1}-\sigma\frac{\partial\delta_{k-1}}{\partial\sigma}+\rho\frac{\partial\gamma_{k-1}}{\partial\sigma} \qquad \frac{\partial\delta_k}{\partial\sigma}=\gamma_{k-1}+\sigma\frac{\partial\gamma_{k-1}}{\partial\sigma}+\rho\frac{\partial\delta_{k-1}}{\partial\sigma}$$

On pose :

$$\frac{\partial\gamma_k}{\partial\sigma}=v_{k-1} \quad \text{et} \quad \frac{\partial\delta_k}{\partial\sigma}=w_{k-1},$$

ce qui permet d'écrire :

$$v_{k-1} = -\delta_{k-1} - \sigma w_{k-2} + \rho v_{k-2} \quad \text{et} \quad w_{k-1} = \gamma_{k-1} + \sigma v_{k-2} + \rho w_{k-2}$$

avec $v_0 = -\delta_0$ et $w_0 = \gamma_0$. Ces relations aboutissent à :

$$\frac{\partial \Phi}{\partial \sigma} = -\delta_{n-1} - \sigma w_{n-2} + \rho v_{n-2} \quad \text{et} \quad \frac{\partial \Psi}{\partial \sigma} = \gamma_{n-1} + \sigma v_{n-2} + \rho w_{n-2}.$$

En partant de $\rho = \rho_0$ et $\sigma = \sigma_0$, on va calculer la racine $z_0$. Une fois cette racine connue on est en mesure de calculer les coefficients du polynôme $Q_{n-1}(z)$ puis d'en calculer une racine. On aboutira à la fin du calcul au produit de deux monômes dont on calculera les racines. Attention au calcul de la dernière racine car le dernier monôme se présente de la façon suivante : $(\gamma_0 + j\delta_0)z + (\gamma_1 + j\delta_1)$.

### 4.6. Résolution d'un système non linéaire

***a*** – On peut toujours écrire ce système sous la forme :

$$x_p = \Phi_p(x_1, x_2, x_3, \ldots, x_n) \quad \text{avec} \quad p = 1, 2, 3, \ldots, n.$$

pour cela il suffit d'ajouter $x_p$ dans chaque membre de l'équation :

$$F_p(x_1, x_2, x_3, \ldots, x_n) + x_p = x_p = \Phi_p(x_1, x_2, x_3, \ldots, x_n) \quad \text{avec} \quad q = 1, 2, 3, \ldots, n.$$

***b*** – On a les équations :

$$x_p^{(k)} = \Phi_p\left(x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}\right),$$

dont la solution exacte notée avec une étoile s'écrit :

$$x_p^* = \Phi_p\left(x_1^*, x_2^*, x_3^*, \ldots, x_n^*\right) = \Phi_p\left(x_1^{(k)} + \xi_1, x_2^{(k)} + \xi_2, x_3^{(k)} + \xi_3, \ldots, x_n^{(k)} + \xi_n\right).$$

Le développement au premier ordre des fonctions $\Phi_p$ conduit au résultat suivant :

$$x_p^* = \Phi_p\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}\right) + \sum_{j=1}^{n} \xi_j \frac{\partial \Phi_p}{\partial x_j} = x_p^{(k+1)} + \sum_{j=1}^{n} \xi_j \frac{\partial \Phi_p}{\partial x_j},$$

d'où :

$$x_p^* - x_p^{(k+1)} = \sum_{j=1}^{n} \xi_j \frac{\partial \Phi_p}{\partial x_j}.$$

$\xi_p$ est l'erreur sur $x_p$ au tour $k$ tandis que $x_p^* - x_p^{(k+1)}$ est l'erreur au tour $k+1$. Désignons cette erreur au tour $k$ par $e_p^{(k)}$. Nous avons la relation entre les erreurs :

$$e_p^{(k+1)} = \sum_{j=1}^{n} e_j^{(k)} \frac{\partial \Phi_p}{\partial x_j},$$

que nous pouvons écrire en fonction de la matrice jacobienne :

$$E^{(k+1)} = M^{(k)} E^{(k)}.$$

Il faut faire attention au fait que les matrices $M^{(k)}$ voient leurs valeurs changer au cours des calculs. Cependant le processus sera convergent si toutes les matrices $M^{(k)}$ ont tous les modules de leurs valeurs propres inférieurs à 1. Il n'est pas aisé de réaliser cette condition.

Pour améliorer la vitesse de convergence, on peut utiliser l'epsilon-algorithme vectoriel quand bien même la suite des vecteurs $E^{(k)}$ divergerait.

## 4.7. Résolution d'un système non linéaire

***a*** – Si $(X, Y)$ est solution du système :

$$f(x, y) = 0 \quad \text{et} \quad g(x, y) = 0,$$

il l'est également de la forme quadratique :

$$\Phi(x, y) = f^2(x, y) + g^2(x, y)$$

car on a $f(X, Y) = 0$ et par conséquent $f^2(X, Y) = 0$ et de même pour la fonction $g$. Une remarque s'impose si $(X, Y)$ annule $\Phi(x, y)$, $(X, Y)$ n'est pas obligatoirement solution du système primitif, car $\Phi'(x, y) = 2ff' + 2gg'$ et ceci montre que l'on peut avoir d'autres racines que $f = 0$ et $g = 0$ et quatre combinaisons sont possibles avec les dérivées. Il faudra donc bien vérifier ce point en reportant le couple $(X, Y)$ dans les équations de départ.

***b*** – Partant d'une approximation $(x_0, y_0)$ on va chercher d'abord à améliorer la valeur de $x$ et l'on écrit :

$$\Phi(x_0 + \xi, y_0) = \Phi(x_0, y_0) + \xi \frac{\partial \Phi}{\partial x} + \frac{\xi^2}{2!} \cdot \frac{\partial^2 \Phi}{\partial x^2} \quad \text{pour } x = x_0 \quad \text{et } y = y_0.$$

On cherche le minimum en écrivant que $\frac{\partial \Phi}{\partial \xi} = 0$ ; cette relation impose que :

$$\frac{\partial \Phi}{\partial x} = -\xi \frac{\partial^2 \Phi}{\partial x^2},$$

d'où la valeur de $\xi$ :

$$\xi = -\frac{\dfrac{\partial \Phi}{\partial x}}{\dfrac{\partial^2 \Phi}{\partial x^2}},$$

par ailleurs, on peut toujours effectuer un calcul direct des quantités figurant dans l'expression de $\xi$ :

$$\frac{\partial \Phi}{\partial x} = 2f \frac{\partial f}{\partial x} + 2g \frac{\partial g}{\partial x}$$

$$\text{et} \quad \frac{\partial^2 \Phi}{\partial x^2} = 2f \frac{\partial^2 f}{\partial x^2} + 2g \frac{\partial^2 g}{\partial x^2} + 2\left[\frac{\partial f}{\partial x}\right]^2 + 2\left[\frac{\partial g}{\partial x}\right]^2.$$

On sait donc calculer $x_1 = x_0 + \xi$.

***c*** – Rien de fondamental n'est changé et il suffit de remplacer $x$ par $y$ et $\xi$ par $\eta$ dans les équations :

$$\eta = -\frac{\dfrac{\partial \Phi}{\partial y}}{\dfrac{\partial^2 \Phi}{\partial y^2}}$$

avec :

$$\frac{\partial \Phi}{\partial y} = 2f \frac{\partial f}{\partial y} + 2g \frac{\partial g}{\partial y}$$

$$\text{et :} \quad \frac{\partial^2 \Phi}{\partial y^2} = 2f \frac{\partial^2 f}{\partial y^2} + 2g \frac{\partial^2 g}{\partial y^2} + 2\left[\frac{\partial f}{\partial y}\right]^2 + 2\left[\frac{\partial g}{\partial y}\right]^2.$$

On sait donc calculer $y_1 = y_0 + \eta$.

***d*** – On recommence l'ensemble des opérations avec les nouvelles valeurs $x_1$ et $y_1$ qui se substituent aux valeurs $x_0$ et $y_0$. On calcule d'abord $x_2$ puis $y_2$, mais rien n'empêche de faire l'inverse. On poursuit les calculs jusqu'à obtention de la précision permise par la machine sans omettre de réintroduire les dernières valeurs dans les équations primitives.

## 4.8. Ordre d'un processus itératif

La méthode dite par itérations fonctionne directement sur le schéma proposé : $x_{k+1} = \Phi(x_k)$. La méthode de Newton est un peu moins directe : $x_{k+1} = x_k - \Phi(x_k)/\Phi'(x_k) = \Psi(x_k)$ mais le schéma fonctionnel s'applique sur la fonction $\Psi(x_k)$ associée.

On écrit $e_{n+1} = X - x_{n+1} = X - f(x_n)$. Il convient de faire disparaître $x_n$ de cette relation pour répondre à la question posée en notant que $x_n = X - e_n$ ; ainsi on écrit :

$$e_{n+1} = X - f(X - e_n).$$

En développant $f(x)$ au voisinage de la racine $X$, on obtient alors :

$$e_{n+1} = X - \left[ f(X) - \frac{e_n}{1!} \cdot \frac{\mathrm{d}f}{\mathrm{d}x} + \frac{e_n^2}{2!} \cdot \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} - \frac{e_n^3}{3!} \cdot \frac{\mathrm{d}^3 f}{\mathrm{d}x^3} + \cdots \right],$$

l'expression se simplifie en remarquant que $f(X) = X$, soit encore :

$$e_{n+1} = +\frac{e_n}{1!} \cdot \frac{\mathrm{d}f}{\mathrm{d}x} - \frac{e_n^2}{2!} \cdot \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} + \frac{e_n^3}{3!} \cdot \frac{\mathrm{d}^3 f}{\mathrm{d}x^3} + \cdots + (-1)^{k+1} \frac{e_n^k}{k!} \cdot \frac{\mathrm{d}^k f}{\mathrm{d}x^k} + \cdots \quad (\text{pour } x = X),$$

qui est bien un polynôme en $e_n$ dont les coefficients sont :

$$\alpha = \frac{\mathrm{d}f}{\mathrm{d}x} \qquad \beta = -\frac{1}{2!}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} \qquad \gamma = \frac{1}{3!}\frac{\mathrm{d}^3 f}{\mathrm{d}x^3} \quad \text{pour } x = X.$$

### *1. Application à la méthode de Newton*

***a*** – On a $\Phi(x) = 0$ et $f(x) = x - [\Phi(x)/\Phi'(x)]$ et l'on calcule les coefficients $\alpha$, $\beta$ et $\gamma$.

$$\alpha = -\frac{\Phi'^2 - \Phi\Phi''}{\Phi'^2} + 1 = \frac{\Phi\Phi''}{\Phi'^2} \quad \text{pour } x = X, \quad \text{donc } \alpha = 0 \quad \text{car } \Phi(X) = 0.$$

$$\beta = -\frac{1}{2!} \cdot \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = \beta = -\frac{1}{2!} \cdot \frac{\mathrm{d}}{\mathrm{d}x} \cdot \left( \frac{\Phi\Phi''}{\Phi'^2} \right) = -\frac{1}{2!}\frac{\Phi''}{\Phi'} \quad \text{pour } x = X.$$

D'où l'expression de l'erreur :

$$e_{n+1} = -\frac{e_n^2}{2!} \cdot \frac{\Phi''(X)}{\Phi'(X)} \cdot$$

***b*** – L'application sera contractante si

$$|e_{n+1}| < \frac{e_n^2}{2!} \left| \frac{\Phi''(X)}{\Phi'(X)} \right| \cdot$$

***2. Généralisation de la méthode de Newton*** – On a :

$$f(x) = x - \frac{g(x)\Phi(x)}{g(x)\Phi'(x) + h(x)\Phi(x)}.$$

Après quelques calculs qu'il convient de mener avec soin, on obtient les résultats suivants :

$$\alpha = \frac{\mathrm{d}f}{\mathrm{d}x} = 0$$
$$\beta = -\frac{1}{2!} \cdot \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = -\frac{1}{2!}\left(\frac{\Phi''}{\Phi'} + \frac{2h}{g}\right) \quad \text{pour } x = X.$$

Le processus sera du troisième ordre en choisissant les fonctions $h$ et $g$ de telle sorte que le coefficient $b$ soit annulé, c'est-à-dire lorsque $\Phi''/\Phi' = -2h/g$.

## 4.9. Résolution d'un système de deux équations à deux inconnues. Méthode de Kacmarz (1937)

***1. a*** – Les coordonnées du point $A_1(x_1, y_1)$ sont données par les équations :

$$a_1(y_1 - y_0) - b_1(x_1 - x_0) = 0 \quad \text{et} \quad a_1 x_1 + b_1 y_1 + c_1 = 0$$

dont la résolution donne :

$$x_1 = x_0 - \frac{a_1 R_1}{a_1^2 + a_1^2} \quad \text{et} \quad y_1 = y_0 - \frac{b_1 R_1}{a_1^2 + a_1^2} \quad \text{avec} \quad R_1 = a_1 x_0 + b_1 y_0 + c_1.$$

***b*** – Par le même procédé, on obtient les coordonnées de $A_2(x_2, y_2)$ :

$$x_2 = x_1 - \frac{a_2 R_2}{a_2^2 + a_2^2} \quad \text{et} \quad y_2 = y_1 - \frac{b_2 R_2}{a_2^2 + a_2^2} \quad \text{avec} \quad R_2 = a_2 x_1 + b_2 y_1 + c_2.$$

***2.*** – On linéarise les équations comme il a été montré dans le cours, on écrit :

$$f(x_0 + h, y_0 + l) = f(x_0, y_0) + h\frac{\partial f}{\partial x} + l\frac{\partial f}{\partial y} = 0$$
$$g(x_0 + h, y_0 + l) = g(x_0, y_0) + h\frac{\partial g}{\partial x} + l\frac{\partial g}{\partial y} = 0$$

et par identification on obtient :

$$a_1 = \frac{\partial f}{\partial x} \qquad b_1 = \frac{\partial f}{\partial y} \qquad a_2 = \frac{\partial g}{\partial x} \qquad b_2 = \frac{\partial g}{\partial y}.$$

De même que $a_1 x_0 + b_1 y_0 + c_1 \neq 0$ si le point $(x_0, y_0)$ n'est pas la racine, de même $f(x_0, y_0) \neq 0$ si le point $(x_0, y_0)$ n'est pas la racine. On continue de désigner par $R_1$ la valeur de $f(x_0, y_0)$ : $R_1 = f(x_0, y_0)$. Rien n'est changé pour l'indice 2. On a les relations :

$$x_{k+1} = x_k - \frac{a_1 R_1}{a_1^2 + a_1^2} \quad \text{et} \quad y_{k+1} = y_k - \frac{b_1 R_1}{a_1^2 + a_1^2} \quad \text{avec} \quad R_1 = f(x_k, y_k)$$
$$x_{k+2} = x_{k+1} - \frac{a_2 R_2}{a_2^2 + a_2^2} \quad \text{et} \quad y_{k+2} = y_{k+1} - \frac{b_2 R_2}{a_2^2 + a_2^2} \quad \text{avec} \quad R_2 = g(x_{k+1}, y_{k+1}).$$

***Exemple***

$$f(x,y) = x^2 + 4y^2 - 36 = 0$$
$$g(x,y) = x^2 + y^2 - 12x + 27 = 0.$$

On calcule aisément :

$$a_1 = \frac{\partial f}{\partial x} = 2x \qquad b_1 = \frac{\partial f}{\partial y} = 8y$$
$$a_2 = \frac{\partial g}{\partial x} = 2x - 12 \qquad b_2 = \frac{\partial g}{\partial y} = 2y.$$

# 5. Éléments de calcul matriciel

## 5.1. Voir le cours p. 70

## 5.2. Voir le cours p. 378

## 5.3. Résolution d'un système linéaire volumineux

La matrice $A_1$ possède $l_1$ lignes et $k_1$ colonnes, la matrice $A_2$ possède $l_1$ lignes et $N-k_1$ colonnes, la matrice $A_3$ possède $N - l_1$ lignes et $k_1$ colonnes et enfin la matrice $A_4$ possède $N - l_1$ lignes et $N - k_1$ colonnes. Les vecteurs $X_1$ et $B_1$ sont à $l_1$ lignes, et les vecteurs $X_2$ et $B_2$ à $N - l_1$ lignes.

Pour que l'opération proposée soit possible, il faut que $l_1 = k_1$, donc que les matrices $A_1$ et $A_4$ soient carrées. La première équation donne :

$$X_1 = A_1^{-1}[B_1 - A_2X_2],$$

valeur que l'on substitue dans la seconde équation :

$$A_3A_1^{-1}[B_1 - A_2X_2] + A_4X_2 = B_2.$$

Cette équation peut encore être transformée de la manière suivante :

$$\left[A_4 - A_3A_1^{-1}A_2\right] X_2 = B_2 - A_3A_1^{-1}B_1,$$

ou encore :

$$\left[A_3A_1^{-1}A_2 - A_4\right] X_2 = A_3A_1^{-1}B_1 - B_2.$$

Pour calculer $X_2$, on réalise la suite des opérations :

1. Inversion de $A_1$ en $A_1^{-1}$,
2. Multiplication de $A_1^{-1}$ et $A_2$ soit $A_1^{-1}A_2$,
3. Multiplication de $A_3$ par le résultat précédent, soit $A_3A_1^{-1}A_2$,
4. Soustraction de la matrice $A_4$ au résultat précédent, soit $A_3A_1^{-1}A_2 - A_4$,
5. Inversion de la matrice précédente, soit $[A_3A_1^{-1}A_2 - A_4]^{-1}$,
6. Calcul de $A_1^{-1}B_1$,
7. Multiplication de $A_3$ par le résultat précédent, soit $A_3A_1^{-1}B_1$,
8. Soustraction de $B_2$ au résultat précédent, soit $A_3A_1^{-1}B_1 - B_2$.
9. Multiplication du résultat précédent par le résultat du 5.

On obtient $X_2$, et l'on peut donc calculer $A_2X_2$, expression que l'on retranche de $B_1$, il ne reste plus qu'à multiplier $A_1^{-1}$ par ce dernier résultat pour avoir la valeur de $X_1$.

Le programme `grosys.c` effectue les calculs proposés par cette procédure.

## 5.4. Résolution d'un système linéaire par la méthode itérative de Jacobi

***a*** – Transformation de la matrice $A'$. On divise chaque ligne par l'élément qui est sur sa diagonale, si un des éléments de la diagonale est nul, il suffit de permuter deux lignes pour réaliser l'opération : elle est toujours possible puisque la matrice $A'$ est inversible, sinon cela voudrait dire que le déterminant est nul. Après cette transformation, désignons les éléments de $A$ par $a_{lk}$ et l'on a : $a_{lk} = a'_{lk}/a'_{ll}$ si $l \neq k$ et $a_{ll} = 1$. On obtient la matrice $M$ aisément, ses éléments sont : $m_{lk} = a_{lk}$ si $l \neq k$ et $m_{ll} = 0$. Les composantes de $B$ s'écrivent simplement : $b_1 = b'_1/a'_{ll}$.

***b*** – On a la relation : $[I - M]X = B$, soit encore $X = B + MX$ expression qui se transforme en processus itératif, à savoir :

$$X_{k+1} = B + MX_k$$

que l'on peut initialiser avec $X_0 = 0$. On obtient la suite :

$$\begin{aligned}
X_1 &= B,\\
X_2 &= B + MX_1 = B + MB = [I + M]B,\\
X_3 &= B + MX_2 = B + MB + M^2B,\\
&\cdots\cdots\\
X_{n+1} &= B + MX_n = B + MB + \cdots + M^nB = [I + M + \cdots + M^n]B.
\end{aligned}$$

Si $n$ est infini ou simplement très grand, on peut alors écrire que :

$$I + M + \cdots + M^n + \cdots = \frac{I}{I - M}$$

et si l'on désigne par $Z$ la limite de $X_n$, on retrouve naturellement que :

$$[I - M]Z = IB = B,$$

mais $Z$ ne sera la solution effective que si l'algorithme est convergent, c'est-à-dire si les valeurs propres de $M$ sont de module inférieur à 1. Ceci n'est pas tout à fait exact, car même si la suite des $X_n$ diverge, on peut néanmoins faire appel à l'epsilon-algorithme vectoriel pour ramener les choses en l'ordre. C'est ce que réalise le programme `jacobi0.c`.

## 5.5. Résolution d'un système linéaire dépendant d'une matrice symétrique (méthode de Choleski)

***1. a*** – On écrit : $A = LS$, la décomposition étant explicitée dans le chapitre 5.
On obtient : $AX = LSX = B$ d'où $SX = L^{-1}B = Y$.

***b*** – L'explicitation de la dernière équation donne :

$$\begin{vmatrix} 1 & s_{12} & s_{13} & s_{14} & \cdots \\ 0 & 1 & s_{23} & s_{24} & \cdots \\ 0 & 0 & 1 & s_{34} & \cdots \\ 0 & 0 & & 1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{vmatrix} \cdot \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \cdots \end{vmatrix} = \begin{vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \cdots \end{vmatrix}.$$

On obtient les valeurs successives de $x_n$, $x_{n-1}$, $x_{n-2}$ et ainsi de suite jusqu'à $x_1$ en écrivant :

$$\begin{aligned} x_n &= y_n \\ x_{n-1} &= y_{n-1} - s_{n-1,n}x_n \\ &\cdots\cdots \\ x_{n-k} &= y_{n-k} - \sum_{p=0}^{k-1} s_{n-k,n-p}x_{n-p} \quad \text{avec } k = 0, 1, 2, \ldots, n-1. \end{aligned}$$

***c*** – On écrit sans difficulté :

$$\begin{vmatrix} l_{11} & 0 & 0 & 0 & \ldots & 0 \\ l_{21} & l_{22} & 0 & 0 & \ldots & 0 \\ l_{31} & l_{32} & l_{33} & 0 & \ldots & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix} \cdot \begin{vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \ldots \end{vmatrix} = \begin{vmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ \ldots \end{vmatrix}$$

d'où l'on déduit :

$$y_1 = \frac{b_1}{l_{11}}$$

$$\text{et} \quad y_k = \frac{1}{l_{kk}} \left[ b_k - \sum_{p=1}^{k-1} l_{kp}y_p \right] \quad \text{avec } k = 1, 2, 3, \ldots, n.$$

***d*** – On écrit simplement :

$$\begin{vmatrix} l_{11} & 0 & 0 & 0 & \ldots & 0 \\ l_{21} & l_{22} & 0 & 0 & \ldots & 0 \\ l_{31} & l_{32} & l_{33} & 0 & \ldots & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix} \cdot \begin{vmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ \lambda_{21} & 1 & 0 & 0 & \ldots & 0 \\ \lambda_{31} & \lambda_{32} & 1 & 0 & \ldots & 0 \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix} \cdot \begin{vmatrix} d_{11} & 0 & 0 & 0 & \ldots & 0 \\ 0 & d_{22} & 0 & 0 & \ldots & 0 \\ 0 & 0 & d_{33} & 0 & \ldots & 0 \\ 0 & 0 & 0 & d_{44} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix}$$

ce qui permet d'écrire :

$$l_{11} = d_{11}, \quad l_{21} = \lambda_{21}d_{11}, \quad \ldots \quad l_{j1} = \lambda_{j1}d_{11}.$$

En définitive, la colonne 1 est divisée par $d_{11} = l_{11}$, la colonne 2 par $d_{22} = l_{22}$ et ainsi de suite. On réécrit la matrice $A$ de la façon suivante :

$A = LS = (\Lambda D)S$ et comme $D$ est diagonale, $A = \Lambda(DS)$.

***2. a*** – Puisque la matrice $A$ est symétrique, nous écrivons que la matrice $A$ est égale à sa transposée, soit : $A = A^T$. À partir de $A = LS = \Lambda(DS)$, on écrit la matrice transposée de $A$ : $A^T = S^T D\Lambda^T$ ; car $D^T = D$. On en conclut que : $\Lambda = S^T$ et que $\Lambda^T = S$.

***b*** – Puisque $\Lambda = S^T$, on en déduit que : $A = A^T = S^T DS = \Lambda D\Lambda^T$. On peut encore écrire : $A = \Lambda\sqrt{D}\sqrt{D}\Lambda^T = MM^T$ en posant $M = \Lambda\sqrt{D}$ et on a également : $\delta_{ll} = \sqrt{d_{ll}}$, pour $l = 1, 2, 3, \ldots, n$.

**c** – L'équation $A = MM^T$ donne les relations suivantes : $m_{11}^2 = a_{11}$ puis $m_{21}m_{11} = a_{12}$ et plus généralement : $m_{l1}m_{11} = a_{1l}$ pour $l = 1, 2, \ldots, n$. De là nous déduisons la première ligne et la première colonne :

$$m_{11} = \pm\sqrt{a_{11}} \quad \text{puis} \quad m_{l1} = \frac{a_{1l}}{m_{11}} \quad \text{pour} \quad l = 2, 3, 4, \ldots, n.$$

On passe à la deuxième ligne et la deuxième colonne au moyen des équations suivantes : $m_{22}^2 = a_{22} - m_{21}^2$ et plus généralement : $m_{k2}m_{22} + m_{k1}m_{21} = a_{2k}$ pour $k = 1, 2, \ldots, n$, dont nous déduisons :

$$m_{22} = \pm\sqrt{a_{22} - m_{21}^2} \quad \text{et} \quad m_{k2} = \frac{a_{2k} - m_{k1}m_{21}}{m_{22}}.$$

Ce calcul se généralise aisément :

$$m_{kk} = \pm\sqrt{a_{kk} - \sum_{p=1}^{k-1} m_{pk}^2} \quad \text{et} \quad m_{lk} = \frac{a_{lk} - \sum_{q=1}^{l-1} m_{qk}m_{qk}}{m_{ll}}.$$

**d** – Après avoir calculé les $m_{lk}$ on calcule les $y_k$ du paragraphe c en écrivant : $Y = M^{-1}B$. On résout ensuite $X = M^{-1}Y$ en faisant usage des relations établies au paragraphe b. Bien entendu, dans les relations établies aux paragraphes b et c, on remplace les $s_{kl}$ par les $m_{kl}$ et les $l_{kl}$ par les $m_{kl}$.

On trouve sur le Web(*) le programme `choleski.c` qui réalise l'algorithme étudié.

## 5.6. Résolution d'un système linéaire par la méthode du gradient conjugué

**Nota** - Pour éviter de commettre quelques confusions, tout au long de ce problème, les vecteurs sont notés en caractères gras.

**1.** – Soit $\mathbf{e_i}$ une base orthonormale à $N$ dimensions,

$$\text{le vecteur } \mathbf{p} \text{ de composantes } p_k \text{ s'écrit : } \quad \mathbf{p} = \sum_{i=0}^{N-1} p_i \mathbf{e_i}$$

$$\text{de même le vecteur } \mathbf{q} \text{ s'écrit : } \quad \mathbf{q} = \sum_{i=0}^{N-1} q_i \mathbf{e_i}$$

$$\text{le produit } A\mathbf{q} \text{ s'écrit : } \quad A\mathbf{q} = \sum_{l=0}^{N-1} \sum_{k=0}^{N-1} a_{lk} q_k \mathbf{e_l}$$

et le produit scalaire s'écrit :

$$(\mathbf{p}, A\mathbf{q}) = \sum_{l=0}^{N-1} \sum_{k=0}^{N-1} a_{lk} q_k p_l = (\mathbf{q}, A\mathbf{p}) = (A\mathbf{p}, \mathbf{q}) = (A\mathbf{q}, \mathbf{p}).$$

**2. a** – Comme $A$ est une matrice définie positive, $(A\mathbf{p_j}, \mathbf{p_j}) > 0$ si $\mathbf{p_j} \neq 0$ et $(A\mathbf{p_j}, \mathbf{p_j}) = 0$ si $\mathbf{p_j} = 0$.

* `http://www.edpsciences.com/guilpin/`

***b*** – Les $\{\mathbf{p_i}\}$ sont linéairement indépendants et forment une base de l'espace à $N$ dimensions. En effet, $(A\mathbf{p_i}, \mathbf{p_j}) = 0$ si $i \neq j$ et $(A\mathbf{p_i}, \mathbf{p_i}) > 0$. S'ils étaient linéairement dépendants, il existerait au moins un $k$ qui donnerait $(A\mathbf{p_i}, \mathbf{p_k}) > 0$ ce qui est contraire à l'hypothèse.

***c*** – $\mathbf{h}$ étant un vecteur de l'espace à $N$ dimensions qui est solution du système $A\mathbf{X} = \mathbf{B}$, il peut s'exprimer comme une combinaison linéaire des vecteurs de base $\mathbf{p_j}$ :

$$\mathbf{h} = \sum_{i=0}^{N-1} c_i \mathbf{p_i}.$$

Calculons les $c_i$ :

$$(A\mathbf{h}, \mathbf{p_k}) = (\mathbf{B}, \mathbf{p_k}) = \left( A \sum_{i=0}^{N-1} c_i \mathbf{p_i}, \mathbf{p_k} \right) = c_k (A\mathbf{p_k}, \mathbf{p_k}),$$

et donc on a :

$$c_k = \frac{(\mathbf{B}, \mathbf{p_k})}{(A\mathbf{p_k}, \mathbf{p_k})}.$$

Comme

$$\mathbf{h} = \sum_{k=0}^{N-1} \frac{(\mathbf{B}, \mathbf{p_k})}{(A\mathbf{p_k}, \mathbf{p_k})} \mathbf{p_k},$$

on forme la suite de vecteurs $\mathbf{x_{j+1}} = \mathbf{x_j} + c_j \mathbf{p_j}$ en commençant par $\mathbf{x_0} = \mathbf{c_0 p_0}$

***3. a*** – On a $(A\mathbf{p_0}, \mathbf{p_1}) = 0 = (\mathbf{p_0}, A\mathbf{p_1})$ avec $\mathbf{p_1} = \mathbf{v_1} + \alpha_{10}\mathbf{p_0}$, soit encore : $(\mathbf{p_0}, A\mathbf{v_1} + \alpha_{10}\mathbf{p_0}) = 0 = (\mathbf{p_0}, A\mathbf{v_1}) + \alpha_{10}(\mathbf{p_0}, A\mathbf{p_0})$ car $\mathbf{p_0} = \mathbf{v_0}$. On en déduit :

$$\alpha_{10} = -\frac{(\mathbf{p_0}, A\mathbf{v_1})}{(\mathbf{p_0}, A\mathbf{p_0})}.$$

Plus généralement, le procédé d'orthogonalisation consiste à écrire :

$$\mathbf{p_{k+1}} = \mathbf{v_{k+1}} + \sum_{j=0}^{k} \alpha_{k+1,j} \mathbf{p_j},$$

ici encore les produits scalaires : $(A\mathbf{p_{k+1}}, \mathbf{p_j}) = (\mathbf{p_{k+1}}, A\mathbf{p_j}) = 0$, avec $j = 0, 1, \ldots, k$, permettent de calculer les $(k+1)$ coefficients $\alpha_{k+1,j}$ et l'on trouve :

$$\alpha_{k+1,j} = -\frac{(A\mathbf{v_{k+1}}, \mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})}.$$

Ensuite, on peut écrire tout simplement que $\mathbf{v_k}$ est une combinaison linéaire des $\mathbf{p_j}$ pour $j = 1, 2, \ldots, k$ : $\mathbf{v_k} = \sum_{j=0}^{k} \mathrm{d}_{k,j} \mathbf{p_j}$, il s'ensuit que :

$$(\mathbf{v_k}, A\mathbf{p_l}) = (A\mathbf{v_k}, \mathbf{p_l}) = \sum_{j=0}^{k} \mathrm{d}_{k,j} (A\mathbf{p_j}, \mathbf{p_l}) = 0 \quad \text{si } j > l,$$

car les $\{\mathbf{p_i}\}$ sont A-orthogonaux.

***b*** – Si $A = I$, alors les vecteurs A-orthogonaux deviennent orthogonaux, car :

$$(A\mathbf{p_k}, \mathbf{p_l}) = (\mathbf{p_k}, \mathbf{p_l}) = 0 \quad \text{si} \quad k \neq l.$$

Ensuite, on pose :

$$\mathbf{p'_{k+1}} = \mathbf{p'_k} + \sum_{j=0}^{k-1} \gamma_{k+1,j}\mathbf{p'_j} + \gamma_{k+1,k}\mathbf{v_{k+1}}. \tag{I.1}$$

Dans le cas où $j < k+1$, le produit scalaire $(\mathbf{p'_{k+1}}, \mathbf{p'_j})$ s'exprime ainsi :

$$(\mathbf{p'_{k+1}}, \mathbf{p'_j}) = 0 = \gamma_{k+1,j}(\mathbf{p'_j}, \mathbf{p'_j}) + \gamma_{k+1,k}(\mathbf{v_{k+1}}, \mathbf{p'_j}),$$

d'où :

$$\gamma_{k+1,j} = \gamma_{k+1,k}\frac{(\mathbf{v_{k+1}}, \mathbf{p'_j})}{(\mathbf{p'_j}, \mathbf{p'_j})}.$$

On obtient $\gamma_{k+1,k}$ en faisant $(\mathbf{p'_{k+1}}, \mathbf{p'_k}) = 0 = (\mathbf{p'_k}, \mathbf{p'_k}) + \gamma_{k+1,k}(\mathbf{v_{k+1}}, \mathbf{p'_k})$, soit :

$$\gamma_{k+1,k} = -\frac{(\mathbf{p'_k}, \mathbf{p'_k})}{(\mathbf{v_{k+1}}, \mathbf{p'_k})},$$

de là on tire l'expression de $\gamma_{k+1,j}$ :

$$\gamma_{k+1,j} = \frac{(\mathbf{p'_k}, \mathbf{p'_k})}{(\mathbf{v_{k+1}}, \mathbf{p'_k})} \cdot \frac{(\mathbf{v_{k+1}}, \mathbf{p'_j})}{(\mathbf{p'_j}, \mathbf{p'_j})}.$$

***4. a*** – La suite $\{\mathbf{r_i}\}$ est orthogonale. Posons : $\mathbf{v_{k+1}} = A\mathbf{p_k}$ et $\mathbf{p'_k} = \mathbf{r_k}$, puis remplaçons dans l'équation (I.1) :

$$\mathbf{r_{k+1}} = \mathbf{r_k} + \frac{r_k^2}{(A\mathbf{p_k}, \mathbf{r_k})}\left\{\sum_{j=0}^{k-1}\frac{(A\mathbf{p_k}, \mathbf{r_j})}{(\mathbf{r_j}, \mathbf{r_j})}\mathbf{r_j} - A\mathbf{p_k}\right\},$$

mais la somme est nulle puisque $k > j$. Par conséquent :

$$\mathbf{r_{k+1}} = \mathbf{r_k} - A\mathbf{p_k}\frac{r_k^2}{(A\mathbf{p_k}, \mathbf{r_k})},$$

qui constitue bien une suite orthogonale puisqu'elle obéit au processus de formation de l'équation (I.1). Par un procédé analogue, on démontre que la suite des $\mathbf{p_k}$ est A-orthogonale. Pour cela, on pose dans l'équation (H.7 p. 457) $\mathbf{v_k} = \mathbf{r_k}$ puis partant de :

$$\mathbf{p_{i+1}} = \mathbf{v_{i+1}} + \sum_{j=0}^{i}\alpha_{i+1,j}\mathbf{p_j}, \quad \text{avec} \quad i+2 < N, \quad \text{et} \quad \alpha_{i+1,j} = -\frac{(A\mathbf{v_{i+1}}, \mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})},$$

on obtient l'expression :

$$\mathbf{p_{i+1}} = \mathbf{r_{i+1}} - \sum_{j=0}^{i}\frac{(A\mathbf{r_{i+1}}, \mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})}\mathbf{p_j} = \mathbf{r_{i+1}} - \sum_{j=0}^{i}\frac{(\mathbf{r_{i+1}}, A\mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})}\mathbf{p_j}.$$

Montrons que $(A\mathbf{r_i}, \mathbf{p_k}) = (A\mathbf{p_k}, \mathbf{r_i}) = 0$ si $i > k+1$ :

En effet, la suite des $\{\mathbf{p_{i+1}}\}$ est construite par A-orthogonalisation de $\mathbf{r_0}, \mathbf{r_1}, \mathbf{r_2}, \ldots$, laquelle a été construite par orthogonalisation de $\mathbf{r_0}, A\mathbf{p_0}, A\mathbf{p_1}, \ldots, A\mathbf{p_i}$. Comme cette dernière suite a été formée à partir de l'expression (H.8 p. 457), alors l'expression $(\mathbf{v_k}, A\mathbf{p_i}) = 0$ avec $i > k$ est vraie en faisant $A = I$ et en remplaçant $\mathbf{v_{k+1}}$ par $A\mathbf{p_k}$ et $\mathbf{p_i}$ par $\mathbf{r_i}$ ; on peut écrire alors :

$$(\mathbf{v_{k+1}}, A\mathbf{p_i}) = 0 \quad \text{si } i > k+1 \quad \text{soit encore } (\mathbf{v_{k+1}}, \mathbf{p_i}) = 0 \quad \text{et } (A\mathbf{p_k}, \mathbf{r_i}) = 0,$$

soit en définitive :

$$\mathbf{p_{i+1}} = \mathbf{r_{i+1}} - \frac{(\mathbf{r_{i+1}}, A\mathbf{p_i})}{(A\mathbf{p_i}, \mathbf{p_i})}\mathbf{p_i}.$$

***b*** – Le maximum de tours d'itération est $N$, ordre du système linéaire.

***5. a*** – $E^2(\mathbf{x}) = [A(\mathbf{h}-\mathbf{x}), \mathbf{h}-\mathbf{x}]$ qui n'est rien d'autre que le carré A-scalaire du résidu donc comme $A$ est définie positive, cela entraîne que $E^2(\mathbf{x}) > 0$.

***b*** – Développons le calcul de l'erreur :

$$\begin{aligned} E^2(\mathbf{x_j} + l_j\mathbf{p_j}) &= [A(\mathbf{h} - \mathbf{x_j} - l_j\mathbf{p_j}), \mathbf{h} - \mathbf{x_j} - l_j\mathbf{p_j}] \\ &= [A(\mathbf{h}-\mathbf{x_j}), \mathbf{h}-\mathbf{x_j}] - [A(\mathbf{h}-\mathbf{x_j}), l_j\mathbf{p_j}] - [A(l_j\mathbf{p_j}), \mathbf{h}-\mathbf{x_j}-l_j\mathbf{p_j}] \\ &= E^2(\mathbf{x_j}) - l_j[A(\mathbf{h}-\mathbf{x_j}), \mathbf{p_j}] - l_j(A\mathbf{p_j}, \mathbf{h}-\mathbf{x_j}-l_j\mathbf{p_j}). \end{aligned}$$

Puisque $A\mathbf{h} = \mathbf{B}$, on pose $\mathbf{r_j} = \mathbf{B} - A\mathbf{x_j} = A\mathbf{h} - A\mathbf{x_j}$, il s'ensuit que :

$$E^2(\mathbf{x_j} + l_j\mathbf{p_j}) = E^2(\mathbf{x_j}) - 2l_j(\mathbf{r_j}, \mathbf{p_j}) + l_j^2(A\mathbf{p_j}, \mathbf{p_j}).$$

Dérivons cette dernière expression par rapport à $l_j$ pour obtenir le minimum de $E^2$ :

$$\frac{\mathrm{d}E^2(\mathbf{x_j} + l_j\mathbf{p_j})}{\mathrm{d}l_j} = -2(\mathbf{r_j}, \mathbf{p_j}) + 2l_j(A\mathbf{p_j}, \mathbf{p_j}) = 0,$$

on obtient alors la valeur de $l_j$ :

$$l_j = \frac{(\mathbf{r_j}, \mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})}.$$

Explicitons le dénominateur :

$$(\mathbf{r_j}, \mathbf{p_j}) = (\mathbf{B}, \mathbf{p_j}) - (\mathbf{x_j}, A\mathbf{p_j}).$$

Montrons que le dernier produit scalaire est nul ; en effet, $\mathbf{x_j}$ est une combinaison linéaire des $\{\mathbf{p_i}\}$, donc $(\mathbf{r_j}, \mathbf{p_j}) = (\mathbf{B}, \mathbf{p_j})$, et l'expression des $l_j$ devient :

$$l_j = \frac{(\mathbf{B}, \mathbf{p_j})}{(A\mathbf{p_j}, \mathbf{p_j})} = c_j.$$

On voit que chaque itération $j$ rend $E^2(\mathbf{x})$ minimum selon l'axe des $\mathbf{p_j}$ passant par le point $x_j$. Le minimum minimorum est atteint pour $\mathbf{x} = \mathbf{h}$ et alors $E^2(\mathbf{x}) = 0$.

***6.*** – Dans le cas où la matrice n'est pas définie positive, alors on peut multiplier à gauche les deux membres de l'équation $A\mathbf{x} = \mathbf{B}$ par la matrice transposée $A^T$. Ainsi :

$A^T A\mathbf{x} = A^T\mathbf{B}$, dans la mesure où la matrice $A$ n'est pas singulière, on est ramené à l'étude précédente.

La procédure présentée est concrétisée par le programme `gradconj.c`.

## 5.7. Calcul direct des coefficients du polynôme caractéristique

***a*** – Nous avons :

$$P_n(\lambda) = P_n(0) + \lambda P'_n(0) + \frac{\lambda^2}{2!}P''_n(0) + \cdots + \frac{\lambda^n}{n!}P^n_n(0) = \sum_{j=0}^{n} \alpha_j \lambda^{n-j},$$

il s'ensuit que :

$$\alpha_k = \frac{P^k_n(0)}{k!}\,.$$

***b*** – À présent, dérivons l'expression $P_n(\lambda) = [\lambda I_n - A]$. Nous avons :

$$P'_n(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}[\lambda I_n - A] = \sum_{j=1}^{n}[\lambda I_{n-1} - A_{jj}],$$

expression dans laquelle on fait $\lambda = 0$, soit :

$$P'_n(0) = \sum_{j=1}^{n}[-A_{jj}] = (-1)^{n-1}\sum_{j=1}^{n}[A_{jj}],$$

ensuite, on a :

$$P''_n(0) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\sum_{j=1}^{n}[\lambda I_{n-1} - A_{jj}] = 2(-1)^{n-2}\sum_{j=1}^{n}\sum_{k>j}^{n}[A_{jj,kk}].$$

***c*** – Cette expression se généralise de la façon suivante :

$$\frac{1}{m!}P^{(m)}_n(0) = (-1)^{n-m}\sum_{j=1}^{n}\sum_{k>j}^{n}\cdots\sum_{p>q}^{n}\sum_{m>p}^{n}[A_{jj,kk,\ldots,pp,mm}],$$

pour aboutir aux expressions terminales :

$$\frac{1}{(n-1)!}P^{(n-1)}_n(0) = -\sum_{j=1}^{n}\sum_{k>j}^{n}\cdots\sum_{p>q}^{n}\cdots\sum_{n-1}^{n}[A_{jj,kk,\ldots,pp,n-1n-1)}] = -\sum_{j=1}^{n}a_{ii},$$

où l'on a noté par $a_{lk}$ les coefficients de la matrice $A$.

***d*** – Par un examen direct de (H.10 p. 459), on peut écrire :

$$\frac{1}{n!}P^{(n)}_n(0) = 1.$$

***e*** – On obtiendra les valeurs propres de $A$ après avoir obtenu les coefficients du polynôme caractéristique. Pour obtenir ceux-ci, il peut être habile de confectionner un programme récursif (mais cela n'est pas indispensable) puisque le calcul d'un déterminant d'ordre $n$ est une combinaison linéaire de $n$ déterminants d'ordre $n-1$.

### 5.8. Calcul des valeurs propres d'une matrice réelle et symétrique par la méthode de Jacobi

***a*** – On a $AX = \lambda X$ puis on fait $X = TZ$ ce qui donne $ATZ = \lambda TZ$. Multiplions à gauche cette dernière équation par $T^{-1}$ : $T^{-1}ATZ = T^{-1}\lambda TZ = \lambda Z$. On voit que la matrice $T^{-1}AT$ a les mêmes valeurs propres que $A$ mais pas les mêmes vecteurs propres.

***b*** – Puisque $A$ est symétrique, ses valeurs propres sont réelles, donc $A$ est diagonalisable.

***c*** – $T$ est symétrique par construction, $T^{-1}$ est symétrique aussi car $(T^{-1})_{ij} = K_{ij}/\Delta$ où $K_{ij}$ est le cofacteur et le $\Delta$ déterminant qui se réduit à un déterminant d'ordre deux :

$$\begin{vmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{vmatrix}$$

donc qui vaut $-1$. On vérifie bien que $TT = I$ (matrice unité d'ordre $n$).

***d*** – Écriture des éléments de $(AT)$, seules les colonnes $p$ et $q$ sont changées dans la matrice $A$ :

| colonne $p$ | | colonne $q$ |
|---|---|---|
| $a_{1p}\cos(\varphi) + a_{1q}\sin(\varphi)$ | ... | $a_{1p}\sin(\varphi) - a_{1q}\cos(\varphi)$ |
| $a_{2p}\cos(\varphi) + a_{2q}\sin(\varphi)$ | ... | $a_{2p}\sin(\varphi) - a_{2q}\cos(\varphi)$ |
| ..... | | |
| $a_{lp}\cos(\varphi) + a_{lq}\sin(\varphi)$ | ... | $a_{lp}\sin(\varphi) - a_{lq}\cos(\varphi)$ |
| ..... | | |
| $a_{np}\cos(\varphi) + a_{nq}\sin(\varphi)$ | ... | $a_{np}\sin(\varphi) - a_{nq}\cos(\varphi)$ |

***e*** – Les éléments de $(TAT)$ sont les mêmes que ceux de $(AT)$ à l'exception d'une part des lignes $p$ et $q$ qui sont remplacées respectivement par la colonne $p$ et la colonne $q$ et des éléments à l'intersection des lignes $p$ et $q$ et des colonnes $p$ et $q$ d'autre part ; ces derniers deviennent :

$$\begin{aligned}
b_{pp} &= a_{pp}\cos^2(\varphi) + a_{pq}\sin(\varphi)\cos(\varphi) + a_{qp}\sin(\varphi)\cos(\varphi) + a_{qq}\sin^2(\varphi) \\
b_{qq} &= a_{pp}\sin^2(\varphi) - a_{pq}\sin(\varphi)\cos(\varphi) - a_{qp}\sin(\varphi)\cos(\varphi) + a_{qq}\cos^2(\varphi) \\
b_{pq} &= a_{pp}\sin(\varphi)\cos(\varphi) - a_{pq}\cos^2(\varphi) + a_{pq}\sin^2(\varphi) - a_{qq}\sin(\varphi)\cos(\varphi) \\
b_{qp} &= b_{pq},
\end{aligned}$$

puisque $A$ est symétrique, $B = (TAT)$ est aussi symétrique.

***f*** – On veut que $b_{qp} = b_{pq} = 0$. Soit : $(a_{pp} - a_{qq})\sin(\varphi)\cos(\varphi) = a_{pq}[\cos^2(\varphi) - \sin^2(\varphi)]$, ce qui donne :

$$\tan(2\varphi) = \frac{2a_{pq}}{a_{pp} - a_{qq}} = \tau, \quad \text{l'angle } \varphi \text{ appartenant à } \quad \left(-\frac{\pi}{4}, \frac{\pi}{4}\right).$$

Il convient à présent de calculer $\sin(\varphi)$ et $\cos(\varphi)$ en fonction de $\tau$. Nous avons :

$$\sin(2\varphi) = \tau\cos(2\varphi)$$

que nous élevons au carré et nous obtenons :

$$\sin(2\varphi) = \frac{\varepsilon\tau}{\sqrt{1+\tau^2}} \quad \text{et} \quad \cos(2\varphi) = \frac{1}{\sqrt{1+\tau^2}} \quad \text{avec } \varepsilon = \text{ signe de } \tau,$$

de là on tire :

$$\sin(\varphi) = \varepsilon\sqrt{\frac{1}{2}\left\{1 - \frac{1}{\sqrt{1+\tau^2}}\right\}} \quad \text{et} \quad \cos(\varphi) = \sqrt{\frac{1}{2}\left\{1 + \frac{1}{\sqrt{1+\tau^2}}\right\}}.$$

***g*** – Comme la matrice $B$ reste symétrique, on peut opérer sur elle au moyen des transformations $TBT$ dont le résultat est toujours une matrice symétrique.

Adoptons une notation pour simplifier l'écriture : $B_{pq}^{k+1} = T_{pq}B_{pq}^{k}T_{pq}$ pour $p < q$, $q$ variant de 2 à $n$. $k$ est l'ordre d'itération, en effet, ayant réalisé une fois l'opération « la nouvelle matrice $B$ » n'est pas diagonale et l'on doit réitérer l'opération jusqu'à ce que les éléments non diagonaux soient suffisamment nuls à une précision définie à l'avance. La fin des itérations dépend aussi de la matrice $A$ et de son ordre.

Dans le cas particulier où $a_{pp} = a_{qq}$, la méthode ne tombe cependant pas en défaut, car on calcule directement la valeur des fonctions trigonométriques :

$$\sin(\varphi) = \varepsilon\frac{\sqrt{2}}{2} \qquad \cos(\varphi) = \frac{\sqrt{2}}{2} \quad \text{avec } \varepsilon = \text{ signe de } \tau.$$

***h – Convergence de la méthode*** – La somme des carrés des valeurs propres de $A$ est un invariant dont on sait calculer la valeur :

$$S^2 = \sum_{k=1}^{n} \lambda_k^2 = \text{Trace } (AA) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{lk}^2,$$

on cesse les itérations lorsque la somme des carrés des éléments diagonaux de la matrice $B^k$ est suffisamment proche de $S^2$.

***i*** – Nous avons les relations :

$$b_{qq} + b_{pp} = a_{pp} + a_{qq} \tag{I.2}$$

$$b_{pp} - b_{qq} = (a_{pp} - a_{qq})\cos(2\varphi) + 2a_{pq}\sin(2\varphi) \tag{I.3}$$

$$2b_{pq} = (a_{pp} - a_{qq})\sin(2\varphi) - 2a_{pq}\cos(2\varphi) \tag{I.4}$$

puis on fait (I.3) au carré plus (I.4) au carré et l'on obtient :

$$(b_{pp} - b_{qq})^2 + 4b_{pq}^2 = (a_{pp} - a_{qq})^2 + 4a_{pq}^2,$$

et l'on combine cette dernière équation à (I.2) élevé au carré, puis en tenant compte du fait que $b_{pq} = 0$, on peut écrire :

$$b_{qq}^2 + b_{pp}^2 = a_{qq}^2 + a_{pp}^2 + 2a_{pq}^2.$$

La somme des termes diagonaux augmente de $2a_{pq}^2$. Mais, comme la somme de tous les termes au carré est une constante, la méthode provoque l'accroissement de la somme des carrés des éléments de la diagonale et la diminue hors de cette diagonale de la même quantité. La méthode est donc convergente sans condition.

Le programme `jacobi1.c` met en œuvre cet algorithme.

### 5.9. Calcul des valeurs propres d'une matrice par la méthode de Souriau (1948)

***a*** – Nous avons :

$$B_q = A^q + a_1 A^{q-1} + a_2 A^{q-2} + \cdots + a_q I$$
$$B_{q+1} = A^{q+1} + a_1 A^q + a_2 A^{q-1} + \cdots + a_{q+1} I = A(A^q + a_1 A^{q-1} + a_2 A^{q-2} + \cdots + a_q) + a_{q+1} I$$

d'où la relation de récurrence : $B_{q+1} = AB_q + a_{q+1} I$, et si $q = 0, B_1 = AB_0 + a_1 I$, avec la relation de définition qui impose : $B_0 = I$.

***b*** – La démonstration est dans le cours. On a :

$$a_k = -\frac{1}{k}\left[S^k + a_1 S^{k-1} + a_2 S^{k-2} + \cdots + a_{k-1} S^1\right],$$
$$\text{avec} \quad S^m = \sum_{j=1}^{n} \lambda_j^m = \text{Trace}\,(A^m),$$

où les $\{\lambda_j\}$ sont les valeurs propres de $A$. Donc :

$$\begin{aligned} a_k &= -\frac{1}{k}\left\{\text{Trace}\,(A^k) + a_1 \text{Trace}\,(A^{k-1}) + a_2 \text{Trace}\,(A^{k-2}) + \cdots + a_{k-1}\text{Trace}\,(A)\right\} \\ &= -\frac{1}{k}\text{Trace}\,\left\{A^k + a_1 A^{k-1} + a_2 A^{k-2} + \cdots + a_{k-1} A\right\} \\ &= -\frac{1}{k}\text{Trace}\,\left\{A(A^{k-1} + a_1 A^{k-2} + a_2 A^{k-3} + \cdots + a_{k-1} I)\right\} = -\frac{1}{k}\text{Trace}\,\left\{AB^{k-1}\right\}. \end{aligned}$$

***Remarque*** : $B_n = A^n + a_1 A^{n-1} + a_2 A^{n-2} + \cdots + a_n I = 0$ d'après le théorème de Cayley-Hamilton.

Le programme `souriau.c` exploite cette procédure.

## 6. Interpolation

Les exercices proposés sont traités dans le cours p. 89 et suivantes.

## 7. Intégration des équations différentielles dans le champ réel

### 7.1. Étude d'un pendule de longueur variable

***a*** – Appliquons le théorème du moment cinétique, nous obtenons :

$$\frac{\mathrm{d}}{\mathrm{d}t}(l^2\theta) = -gl\sin(\theta),$$

d'où :

$$l\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + 2v\frac{\mathrm{d}\theta}{\mathrm{d}t} + g\sin(\theta) = 0.$$

***b*** – Lorsque $\theta$ est petit nous pouvons écrire :

$$l\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + 2v\frac{\mathrm{d}\theta}{\mathrm{d}t} + g\theta = 0,$$

et comme $l = l_0 + vt$, nous pouvons exprimer l'équation différentielle en fonction de la variable $l$ ($\mathrm{d}l = v\,\mathrm{d}t$) :

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}l^2} + \frac{2}{l}\frac{\mathrm{d}\theta}{\mathrm{d}l} + \frac{g}{lv^2}\theta = 0,$$

sur le plan numérique, on est amené à résoudre le système :

$$\begin{aligned}\frac{\mathrm{d}\theta}{\mathrm{d}l} &= \varphi\\ \frac{\mathrm{d}\varphi}{\mathrm{d}l} &= -\frac{2}{l}\varphi - \frac{g}{lv^2}\theta\end{aligned}$$

avec les conditions initiales au temps $t = 0$ : $\theta = \dfrac{\pi}{200}$ rd et $\dfrac{\mathrm{d}\theta}{\mathrm{d}t} = 0$ rd s$^{-1}$. Le programme `pendule0.c` réalise cette procédure.

## 7.2. Système électromécanique dépendant d'une équation de Mathieu (1835–1890)

La charge $q(t)$ sur le condensateur obéit à l'équation différentielle classique :

$$L\frac{\mathrm{d}^2q}{\mathrm{d}t^2} + R\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{q}{C} = 0,$$

soit en effectuant le remplacement de la valeur de $C$ :

$$LC_0\frac{\mathrm{d}^2q}{\mathrm{d}t^2} + RC_0\frac{\mathrm{d}q}{\mathrm{d}t} + q[1 + \varepsilon\sin(\omega_1 t)] = 0.$$

Il suffit de poser :

$$\begin{aligned}\frac{\mathrm{d}q}{\mathrm{d}t} &= i\\ \frac{\mathrm{d}i}{\mathrm{d}t} &= -\frac{R}{L}i - \frac{1}{LC_0}[1 + \varepsilon\sin(\omega_1 t)].\end{aligned}$$

Il convient de choisir le pas temporel $h$ petit devant les périodes $T_0 = 2\pi\sqrt{LC_0}$ et $T_1 = 2\pi/\omega_1$. Compte tenu des grandeurs numériques proposées, on doit trouver un courant oscillant qui croît indéfiniment. Il ne faut pas croire que les principes de la thermodynamique sont violés, le moteur qui fait varier le condensateur fournit l'énergie.

On trouvera le programme `mathieu.c` qui concrétise cet algorithme.

## 7.3. L'équation de Van der Pol (1889–1959)

Il s'agit d'une importante équation de la physique concernant la synchronisation des systèmes oscillants. Elle présente aujourd'hui toujours un intérêt dans la mesure où c'est une équation non linéaire. Son étude s'effectue dans l'espace des phases (l'abscisse est la coordonnée d'espace et l'ordonnée la dérivée par rapport au temps de la coordonnée d'espace) et l'équation différentielle s'écrit :

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} - 2\varepsilon(1 - y^2)\frac{\mathrm{d}y}{\mathrm{d}t} + y = 0.$$

On décompose cette équation du deuxième ordre en un système de deux équations du premier ordre :

$$\frac{\mathrm{d}y}{\mathrm{d}t} = x \quad \text{et} \quad \frac{\mathrm{d}x}{\mathrm{d}t} = 2\varepsilon(1-y^2)x - y.$$

Au temps $t_0 = 0$, on se donne $x_0$ et $y_0$ quelconques, valeurs que l'on fournit au programme de manière conversationnelle car il convient d'observer les courbes de l'espace des phases pour différents couples de $x_0$ et $y_0$ ainsi que pour différentes valeurs de $\varepsilon$. Il suffit d'utiliser les programmes proposés dans le cours en choisissant convenablement le pas $\mathrm{d}t$ ; il est nécessaire de l'adapter pour chaque valeur de $\varepsilon$.

On doit observer que, pour une valeur fixée de $\varepsilon$, quelles que soient les conditions initiales ($x_0$ et $y_0$), toutes les trajectoires admettent le même cycle limite, que le point figuratif initial ($x_0$ et $y_0$) se situe à l'intérieur ou à l'extérieur de ce cycle limite. Cela signifie que, si un système obéit à cette équation différentielle, au bout d'un certain temps (transitoire), ses oscillations seront synchrones.

Le programme `vanderp.c` réalise les calculs proposés.

### 7.4. Intégration d'une équation différentielle du premier ordre par une méthode itérative (prédiction-correction)

***a*** – L'équation différentielle suivante :

$$\mathrm{d}y/\,\mathrm{d}x = f(x,y)$$

pour $x$ appartenant à $(a,b)$ admet une solution unique qui prend la valeur $y_0$ pour la valeur $x_0 = a$ si les conditions des théorèmes d'Arzelà et de Cauchy-Liptchitz sont vérifiées (*cf.* le cours).

***b*** – Le résultat est immédiat : $y_{m+1} = y_{m-1} + 2hf(x_m, y_m)$.

***c*** – On pose :

$$\begin{aligned} u &= f(x_0, y_0) \\ v &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}u\right) \\ w &= f(x_0 + h, y_0 + hv) \\ \text{et l'on obtient :} \quad y_1 &= y_0 + \frac{h}{6}(u + 4v + w)\,. \end{aligned}$$

***d*** – On écrit la relation de définition :

$$y_{m+1}^{(2)} = y_m^{(1)} + \frac{h}{2}\left[f\left(x_m, y_m^{(1)}\right) + f\left(x_{m+1}, y_{m+1}^{(1)}\right)\right]$$

pour $m > 1$. On obtient la formule d'itération :

$$y_{m+1}^{(k+1)} = y_m^{(k)} + \frac{h}{2}\left[f\left(x_m, y_m^{(k)}\right) + f\left(x_{m+1}, y_{m+1}^{(k)}\right)\right].$$

***e*** – On cesse les itérations lorsque :

$$\left|\frac{y_{m+1}^{(k+1)} - y_{m+1}^{(k)}}{y_{m+1}^{(k+1)}}\right| < 10^{-\alpha},$$

$10^{-\alpha}$ étant la précision relative retenue.

***f*** – Éventuellement, dans le voisinage du point $x_m$, on adapte le pas $h$ de telle sorte que la relation $h < 2/M$ avec $M = \sup|\partial f/\partial x|$ soit vérifiée. On peut estimer $M$ de deux façons, la première consiste à dériver formellement la fonction $f$ et la seconde à calculer numériquement la valeur de la dérivée.

***g*** – Pour la prédiction on écrit :

$$\int_{y_{m-1}^{(k)}}^{y_{m+1}^{(k)}} \mathrm{d}y = y_{m+1}^{(k)} - y_{m-1}^{(k)} = \int_{x_{m-1}}^{x_{m+1}} f[x, y(x)]\,\mathrm{d}x.$$

Il reste à calculer l'intégrale de droite, que l'on désigne par $I_m$, et l'erreur que l'on commet sur l'estimation. Pour simplifier l'écriture, on pose $\varphi(x_m) = f[x_m, y(x_m)]$, on a alors $I_m \approx 2h\varphi(x_m)$ qui est l'approximation donnée par la formule de prédiction. Développons-la autour du point $x_{m-1}$.

$$I_m \approx 2h\varphi(x_{m-1}+h) = 2h\varphi(x_{m-1}) + 2h^2\varphi'(x_{m-1}) + h^3\varphi''(x_{m-1}) + \cdots$$

Désignons par $\Phi(x)$ la primitive de $\varphi(x)$, son introduction va nous permettre de calculer l'erreur sur $I$. Pour cela écrivons l'expression rigoureuse de $I$ puis effectuons le développement de l'expression trouvée autour du point $x_{m-1}$ :

$$I_m = \Phi(x_{m-1}+2h) - \Phi(x_{m-1}) \approx 2h\varphi(x_{m-1}) + 2h^2\varphi'(x_{m-1}) + \frac{4}{3}h^3\varphi''(x_{m-1}).$$

Maintenant on obtient l'erreur $e_m$ sur $I_m$ :

$$e_m = \frac{h^3}{3}f''(\xi)$$

avec $\xi$ appartenant à $(x_{m-1}, x_{m+1})$. Comme d'habitude, on cherche une majoration de $|f''(x)|$ dans l'intervalle considéré. Pour ce qui concerne l'erreur de correction, on a :

$$y_{m+1} = y_m + \frac{h}{2}[f(x_m, y_m) + f(x_{m+1}, y_{m+1})],$$

et l'on reconnaît la formule des trapèzes sur laquelle nous avons déjà effectué le calcul d'erreur :

$$e'_m = \frac{h^3}{12}f''(\xi),$$

avec $\xi$ appartenant à $(x_{m-1}, x_{m+1})$. On note que les deux erreurs sont localement en $h^3$.

Le programme `predcor.c` réalise cet algorithme.

## 7.5. Étude d'un phénomène transitoire obéissant à une équation différentielle du premier ordre

***1. a*** – On trouve tout de suite la solution particulière $y = 1/10$, puis la solution de l'équation sans second membre $y(t) = a\exp(-50t)$, et enfin la solution générale :

$$y(t) = a\exp(-50t) + \frac{1}{10},$$

et la condition initiale permet de déterminer $a$, soit : $a = 1/10$, ce qui donne

$$y(t) = \frac{1}{10}[\exp(-50t) + 1],$$

quand $t$ varie de 0 à l'infini, $y(t)$ varie de 0,2 à 0,1.

***b*** – La méthode d'Euler s'écrit simplement : $y_{k+1} = y_k + hf(y_k, t_k)$ soit encore :

$$y_{k+1} = -1{,}5y_k + 0{,}25 \quad \text{puisque} \quad h = 0{,}05.$$

Pour $t = 1$, on obtient $y = 332{,}63$ et pour $t = 2$ on trouve $y = 1\,105\,733{,}3$ alors que le calcul direct de la solution analytique donne la valeur 0,1 pour $t = 1$ et $t = 2$ ; il s'agit de la valeur asymptotique de la solution. Pour obtenir une valeur numérique raisonnable, il faut prendre un pas $h$ petit devant la constante de temps $\tau = 1/50$, disons par exemple $h = \tau/10 = 0{,}002$. Alors pour $t = 1$, on trouve $y = 1{,}000\,000\,005 \times 10^{-1}$.

***2. a*** – On obtient :

$$\int_{t_1}^{t_2} y'(t)\exp(Dt)\,\mathrm{d}t = -D\int_{t_1}^{t_2} y(t)\exp(Dt)\,\mathrm{d}t + \int_{t_1}^{t_2} s(t)\exp(Dt)\,\mathrm{d}t.$$

Intégrons par parties le premier membre, il devient :

$$\int_{t_1}^{t_2} y'(t)\exp(Dt)\,\mathrm{d}t = [y(t)\exp(Dt)]_{t_1}^{t_2} - D\int_{t_1}^{t_2} y(t)\exp(Dt)\,\mathrm{d}t,$$

il s'ensuit que :

$$y(t_2)\exp(Dt_2) - y(t_1)\exp(Dt_1) = \int_{t_1}^{t_2} s(t)\exp(Dt)\,\mathrm{d}t,$$

ou encore :

$$y(t_2) = y(t_1)\exp[-D(t_2 - t_1)] + \int_{t_1}^{t_2} s(t)\exp[D(t - t_2)]\,\mathrm{d}t,$$

qui est l'expression recherchée.

***b*** – À présent,

$$s(t) = \sum_{k=0}^{N} \frac{(t - t_1)^k}{k!\mu^k} a_k \quad \text{avec} \quad \mu = t_2 - t_1.$$

Alors $I$ s'écrit :

$$I = \int_{t_1}^{t_2} \sum_{k=0}^{N} \frac{(t - t_1)^k}{k!\mu^k} a_k \exp[D(t - t_2)]\,\mathrm{d}t = \sum_{k=0}^{N} a_k \int_{t_1}^{t_2} \frac{(t - t_1)^k}{k!\mu^k} \exp[D(t - t_2)]\,\mathrm{d}t,$$

et en posant

$$C_k = \int_{t_1}^{t_2} \frac{(t - t_1)^k}{k!\mu^k} \exp[D(t - t_2)]\,\mathrm{d}t,$$

on a :

$$I = \sum_{k=0}^{N} a_k C_k.$$

On calcule immédiatement $C_0$ :

$$C_0 = \int_{t_1}^{t_2} \exp[D(t - t_2)]\, \mathrm{d}t = \frac{1}{D}[1 - \exp(-D\mu)],$$

puis on établit la relation de récurrence entre les $C_k$, en intégrant l'expression de $C_k$ par parties :

$$k!\mu^k C_k = \int_{t_1}^{t_2} (t - t_1)^k \exp[D(t - t_2)]\, \mathrm{d}t$$

en posant $u = \exp[D(t - t_2)]$ et $\mathrm{d}v = (t - t_1)^k\, \mathrm{d}t$. Tous les calculs faits on trouve :

$$C_k = \frac{\mu}{(k+1)!} - \mu D C_{k+1} \quad \text{ou} \quad C_{k+1} = -\frac{C_k}{\mu D} + \frac{1}{D(k+1)!} = \frac{1}{D}\left(\frac{1}{(k+1)!} - \frac{C_k}{\mu}\right).$$

***Étude du cas* N = 1** – On a $T_1 = t_2$ et $T_0 = t_2 - (t_2 - t_1) = t_1$, il s'ensuit que nous avons alors :

$$s(t) = a_0 + \frac{t - t_1}{t_2 - t_1} a_1.$$

Par ailleurs nous pouvons écrire : $s_0 = s(T_0) = a_0$ et $s_1 = s(T_1) = a_0 + a_1$, relations qui donnent $a_0$ et $a_1$ en fonction de $s_0$ et $s_1$ : $a_0 = s_0$ et $a_1 = s_1 - s_0$, on obtient en définitive l'équation (I.5) :

$$y(T_1) = y(T_0) \exp(-D\mu) + (C_0 - C_1) s_0 + C_1 s_1. \tag{I.5}$$

Application numérique : $C_0 = 0{,}018\,36$ et $C_1 = 0{,}012\,66$, puis $y(1) = 0{,}099\,955$.

***Étude du cas* N = 2** – Comme on a $T_2 = t_2$, $T_1 = t_1$ et $T_0 = 2t_1 - t_2$, il s'ensuit que nous pouvons écrire :

$$s(t) = a_0 + \frac{t - t_1}{t_2 - t_1} a_1 \cdot$$

Les calculs étant menés comme précédemment on obtient :

$$a_0 = s_1 \quad \text{et} \quad a_1 = (s_2 - s_0)/2 \quad \text{et} \quad a_2 = s_2 + s_0 - 2s_1,$$

d'où l'équation (I.6) :

$$y(T_2) = y(T_1) \exp(-D\mu) + \left(C_2 - \frac{C_1}{2}\right) s_0 + (C_0 - 2C_2)\, s_1 + \left(C_2 + \frac{C_1}{2}\right) s_2. \tag{I.6}$$

***3. Généralisation de l'équation (I.6)*** – On reprend la relation (I.5) dans laquelle on effectue le changement $s_1 = s(T_1) = s[y(T_1), T_1]$. On fait donc d'abord $s_1 = s_0$ pour obtenir $y(T_1)$ que l'on reporte dans l'expression de $s_1$. Alors, on peut itérer jusqu'à ce qu'on obtienne une précision souhaitée à l'avance $10^{-r}$ par exemple en erreur relative. On connaît donc $T_1$ et $y(T_1)$ à l'issue de ce calcul. On va poursuivre de la même façon avec l'expression (I.6) pour obtenir $t_2$ et $y(T_2)$.

Au départ, on choisira $s_2 = s_1$ et l'on calculera alors $y(T_2)$, valeur qui sera reportée dans l'expression de $s_2$. On poursuivra les itérations de la même manière que celle indiquée au début de ce paragraphe.

***Description de la procédure d'itération*** – On établit une plate-forme de calcul avec la relation (I.5) que l'on itère suffisamment, puis on continue les calculs en itérant la relation (I.6).

## 7.6. Problème de la poursuite

Désignons par $x(t)$, $y(t)$ et $v_0$ les coordonnées cartésiennes et la vitesse du jardinier, puis par $X(t)$, $Y(t)$ et $V_0$ les coordonnées cartésiennes et la vitesse du chien. Nous pouvons écrire :

$$\left[\frac{\mathrm{d}X}{\mathrm{d}t}\right]^2 + \left[\frac{\mathrm{d}Y}{\mathrm{d}t}\right]^2 = V_0^2$$

d'où :

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \pm\sqrt{V_0^2 - \left[\frac{\mathrm{d}Y}{\mathrm{d}t}\right]^2} \quad \text{et} \quad \frac{\mathrm{d}Y}{\mathrm{d}t} = \pm\sqrt{V_0^2 - \left[\frac{\mathrm{d}X}{\mathrm{d}t}\right]^2},$$

ainsi que :

$$\left[\frac{\mathrm{d}x}{\mathrm{d}t}\right]^2 + \left[\frac{\mathrm{d}y}{\mathrm{d}t}\right]^2 = v_0^2.$$

Par ailleurs, nous savons que le chien se dirige en permanence vers le jardinier, nous pouvons alors écrire :

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = \frac{Y-y}{X-x} \quad \text{d'où} \quad \frac{\mathrm{d}Y}{\mathrm{d}t} = \frac{Y-y}{X-x}\frac{\mathrm{d}X}{\mathrm{d}t}.$$

On écrit l'équation horaire du jardinier :

$$\begin{aligned} x &= a\cos[f(t)] && \text{et} \quad y = b\sin[f(t)] \\ \frac{\mathrm{d}x}{\mathrm{d}t} &= -af'(t)\sin[f(t)] && \text{et} \quad \frac{\mathrm{d}y}{\mathrm{d}t} = bf'(t)\cos[f(t)] \end{aligned}$$

d'où

$$\left[\frac{\mathrm{d}x}{\mathrm{d}t}\right]^2 + \left[\frac{\mathrm{d}y}{\mathrm{d}t}\right]^2 = v_0^2 = f'^2(t)[a^2+b^2].$$

De là on tire que $f'(t)$ est une constante et par conséquent que $f(t) = \omega t + \varphi$, ce qui donne :

$$\omega = \sqrt{\frac{v_0^2}{a^2+b^2}}$$

et $\varphi = 0$ (condition initiale). Donc, à $t = 0$ : $y = -b\cos(\omega t)$ et $x = a\sin(\omega t)$.

Il faudra prendre garde au signe de $\mathrm{d}Y/\,\mathrm{d}t$ ainsi que celui de $X$ dans l'exécution des calculs et l'on devra prévoir des tests pour affecter les bons signes.

# 8. Intégration des équations aux dérivées partielles

## 8.1. L'équation de Laplace, les isothermes d'un réfrigérateur

Le travail ne présente pas de réelles difficultés, et l'on trouvera un programme qui traite ce problème et qui s'appelle `refrig.c`.

## 8.2. Refroidissement d'une sphère homogène

Le problème admet la symétrie sphérique et nous employons donc l'équation de Laplace en coordonnées sphériques dans laquelle nous ne conservons que les termes qui dépendent uniquement de la coordonnée spatiale $r$. Une remarque importante s'impose : nous ne pouvons pas calculer directement la valeur de la température au centre de la sphère car une discontinuité artificielle est introduite par le formalisme. En effet, le champ de température est parfaitement continu, mais le choix du système de coordonnées introduit une discontinuité au centre puisque l'équation fait apparaître un terme en $1/r$. On pallie cet ennui en choisissant pour $T(0,t)$ la valeur obtenue à l'instant $t - \delta t$ au point adjacent (premier voisin), soit $T(0,t) = T(0, t - \delta t)$, $\delta t$ étant le pas dans le temps.

# 9. Les transformées de Fourier

## 9.1. Étude d'un vélocimètre

Pour des particules se déplaçant selon l'axe du réseau à vitesse constante, le spectre en fréquence de la lumière transmise, appelée $I(x)$, est le produit des transformées de Fourier de la fonction d'éclairement d'une part et de la fonction d'absorption d'autre part.

Il s'agit donc dans un premier temps de calculer la transformée de Fourier d'un réseau de $N$ trous alignés et équidistants. En choisissant l'origine au centre du réseau et en appelant $F(k)$ la transformée de Fourier d'un trou, on aura :

$$I(x) \xrightarrow{TF} \Phi(\chi) = F(\chi) \sum_N \exp(2\pi j k \chi b) \quad \text{avec :} \quad F(\chi) = \frac{\sin(\pi\chi a)}{\pi\chi},$$

expression dans laquelle $a$ est le diamètre du trou. Tout le calcul fait :

$$\Phi(\chi) = \frac{\sin(\pi\chi a)}{\pi k} \cdot \frac{\sin(\pi\chi bN)}{\sin(\pi\chi b)},$$

on retrouve l'expression classique des réseaux, et grosso modo, dans l'espace de Fourier, le terme $[\sin(\pi\chi bN)]/[\sin(\pi\chi b)]$ élevé au carré donne un « spectre de raies » et le terme $[\sin(\pi\chi a)]/[\pi\chi]$ élevé au carré donne l'image de l'enveloppe de ces raies. Rappelons que $\chi$ est un nombre d'ondes (inverse d'une longueur).

En réalité, ce qui nous intéresse est le signal dépendant du temps $t$ et de travailler dans l'espace de Fourier correspondant qui est celui des fréquences $\nu$.

Dans notre problème, comme les longueurs sont parcourues à la vitesse constante $v$, il est commode d'effectuer la transformation sur le plan dimensionnel : $[l] = v[t]$ et comme $[X] = 1/[l] = 1/(v[t]) = [\nu]/v$ dans l'équation donnant $\Phi(X)$ il suffit donc de remplacer $X$ par $\nu/v$, en désignant par $\Psi(\nu)$ la nouvelle fonction on obtient alors l'intensité de l'éclairement :

$$\Psi(\nu) = I_0 \left[ \frac{\sin(\pi\nu a/v)}{\pi\nu/v} \cdot \frac{\sin(\pi\nu b/vN)}{\sin(\pi\nu b/v)} \right]^2.$$

L'examen de cette expression permet de répondre aux questions :

1. Le premier minimum nul de la modulation existe quand $\sin(\pi\nu a/v) = 0$, c'est-à-dire lorsque $\nu = v/a$.
2. « Deux raies » consécutives sont séparées de la distance $\alpha$ déterminée par deux maximums relatifs et consécutifs de la fonction « raie », elle est déterminée par la différence de deux valeurs consécutives qui annulent le dénominateur, on a donc :

$$\alpha = v/b.$$

3. La finesse des raies, c'est-à-dire la largeur à mi-hauteur, est inversement proportionnelle à $N$, elle est donc de l'ordre de $c\alpha/N$, et quand $N$ est grand devant $1, c$ vaut $\sqrt{6}$.

## 9.2. Calcul numérique des transformées de Fourier pour un nombre de données $N = b^s$

On envisage le cas où le nombre de données est une puissance de trois : $N = 3^s$. On écrit la définition de la $TF$ :

$$\Phi_k = f_0 W_N^0 + f_1 W_N^k + f_2 W_N^{2k} + f_3 W_N^{3k} + \cdots f_{N-1} W_N^{k(N-1)}$$

puis on regroupe les termes en trois sommes partielles :

$$\begin{aligned}\Phi_k = {} & f_0 W_N^0 + f_3 W_N^{3k} + f_6 W_N^{6k} + \cdots \quad (= A_k) \\ & + f_1 W_N^k + f_4 W_N^{4k} + f_7 W_N^{7k} + \cdots \quad (= B_k W_N^k) \\ & + f_2 W_N^{2k} + f_5 W_N^{5k} + f_8 W_N^{8k} + \cdots \quad (= C_k W_N^{2k})\end{aligned}$$

d'où on tire : $\Phi_k = A_k + B_k W_N^k + C_k W_N^{2k}$. Ces expressions s'écrivent encore :

$$\begin{aligned}\Phi_k = {} & f_0 W_{N/3}^0 + f_3 W_{N/3}^k + f_6 W_{N/3}^{2k} + \cdots \\ & + \left\{ f_1 W_{N/3}^0 + f_4 W_{N/3}^k + f_7 W_{N/3}^{2k} + \cdots \right\} W_N^k \\ & + \left\{ f_2 W_N^0 + f_5 W_N^k + f_8 W_N^{2k} + \cdots \right\} W_N^{2k}.\end{aligned}$$

On voit que $A_k$, $B_k$ et $C_k$ sont les transformées de Fourier de trois fois moins de points que n'en comporte l'échantillon initial. Cependant, les dernières relations ne peuvent être utilisées que pour $k = 0, 1, 2, \ldots, N/3 - 1$, reste donc à régler les problèmes de périodicité :

$$\Phi_{k+N/3} = A_k + B_k W_N^{k+N/3} + C_k W_N^{2(k+N/3)},$$

équation qui se transforme de la manière suivante :

$$\Phi_{k+N/3} = A_k + B_k W_N^k \left( -\frac{1}{2} + j\frac{\sqrt{3}}{2} \right) + C_k W_N^{2k} \left( -\frac{1}{2} - j\frac{\sqrt{3}}{2} \right);$$

un calcul analogue donne :

$$\Phi_{k+2N/3} = A_k + B_k W_N^k \left( -\frac{1}{2} - j\frac{\sqrt{3}}{2} \right) + C_k W_N^{2k} \left( -\frac{1}{2} + j\frac{\sqrt{3}}{2} \right).$$

La formule d'adressage des données s'écrit :

$$j = R_N(k) = R_N(k - 3^q) + \frac{N}{3^{q+1}} \quad \text{avec} \quad 3^q \leq k < 3^{q+1} \text{et} \quad R_N(0) = 0.$$

Dans le cas où $N = b^s$, on combine linéairement $b$ transformées de Fourier de $b$ fois moins de points que dans l'échantillon initial. La première relation permet les calculs pour $k = 0, 1, 2, \ldots, N/b$. Il reste à établir les $b - 1$ relations pour obtenir une transformée complète ce qui ne présente pas de difficulté. La formule d'adressage devient :

$$j = R_N(k) = R_N(k - b^q) + \frac{N}{b^{q+1}} \quad \text{avec} \quad b^q \leq k < b^{q+1} \quad \text{et} \quad R_N(0) = 0.$$

# 10. Introduction aux méthodes de Monte-Carlo

Les exercices sont traités dans le cours p. 287 et suivantes.

# 11. Éléments de calcul des probabilités

## 11.1. L'ordre statistique

***a*** –

$$\begin{aligned} S_n &= \frac{k}{n} \quad \text{si} \quad x_k < x \leq x_{k+1} \quad \text{avec} \quad k = 1, 2, 3, \ldots, n-1, \\ S_n &= 0 \quad \text{si} \quad x \leq x_1, \\ S_n &= 1 \quad \text{si} \quad x > x_n. \end{aligned}$$

$S_n$ prend des valeurs sur $(0, 1)$ et c'est une fonction non décroissante de $x$. C'est bien la fonction de répartition de la variable aléatoire $x$.

***b*** – $P(\xi_k < x) = F(x)$ expression qui est vraie pour tout $\xi_i$. Il s'ensuit que $P(\xi_k < x) = p$ pour toutes les composantes.

***c*** –

$$P\left(S_n(x) = \frac{m}{n}\right) = C_n^m [F(x)]^m [1 - F(x)]^{n-m} \quad \text{avec} \quad C_n^m = \frac{n!}{m!(n-m)!}.$$

Cette expression est donnée par la loi binomiale car toutes les composantes obéissent à la même distribution (répartition), elles sont toutes équiréparties (idem les faces d'un dé), et $S_n(x)$ est la fréquence de succès ; toutes ces conditions sont celles du schéma de Bernoulli.

***d*** – Il suffit de placer $x$ dans la série des $x_j^{(n)}$. Ainsi, $P(x_j^{(n)} < x)$ est la probabilité pour qu'il y ait au moins $j$ composantes de $V$ inférieures à $x$. Cela revient au même de dire que $S_n(x)$ peut prendre les valeurs $\left\{\frac{j}{n}, \frac{j+1}{n}, \ldots, 1\right\}$ ou encore que $S_n(x)$ ne peut pas être inférieur à $\frac{j}{n}$.

Ici encore la probabilité d'obtenir la valeur $\frac{j}{n}$ est donnée par la relation :

$$P\left(S_n(x) = \frac{j}{n}\right) = C_n^j [F(x)]^j [1 - F(x)]^{n-j},$$

mais de plus, $j$ prend les valeurs de $j$ à $n$. Les événements étant indépendants :

$$\Phi_{nj}(x) = \sum_{m=j}^{n} P\left(S_n(x) = \frac{m}{n}\right),$$

soit encore :

$$\Phi_{nj}(x) = \sum_{m=j}^{n} C_n^m [F(x)]^m [1 - F(x)]^{n-m}.$$

**e** – On part de l'intégrale :

$$\Phi_j^{(n)}(x) = \frac{n!}{(j-1)!(n-j)!} \int_0^{F(x)} t^{j-1}(1-t)^{n-j}\,\mathrm{d}t,$$

que l'on va calculer par une succession d'intégrations par parties. Pour cela posons :

$$\mathrm{d}u = t^{j-1}\,\mathrm{d}t$$

d'où :

$$u = \frac{t^j}{j}, \quad \text{puis } v = (1-t)^{n-j} \quad \text{d'où} \quad \mathrm{d}v = -(n-j)(1-t)^{n-j-1}\,\mathrm{d}t,$$

on obtient alors :

$$\Phi_j^{(n)}(x) = C_n^j[F(x)]^j[1-F(x)]^{n-j} + \frac{n!(n-j)}{j!(n-j)!} \int_0^{F(x)} t^j(1-t)^{n-j-1}\,\mathrm{d}t,$$

on poursuit l'intégration par parties et l'on vérifie que l'on retrouve bien l'expression de $\Phi_{nj}(x)$. En dérivant cette dernière fonction, on obtient la fonction de distribution recherchée :

$$f_{nj}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Phi_{nj}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^j[1-F(x)]^{n-j}.$$

À partir de cette relation, il est facile de calculer la moyenne des $\{x\}$ qui tombent dans « l'urne $j$ », soit :

$$\langle x_j^{(n)} \rangle = \frac{n!}{(j-1)!(n-j)!} \int_{-\infty}^{+\infty} f(x)[F(x)]^{j-1}[1-F(x)]^{n-j} x\,\mathrm{d}x.$$

***f* – *Application*** – $F(x)$ est la loi uniforme, soit :

$F(x) = x$ pour $x$ appartenant à $(0,1)$ et vaut zéro ailleurs.
$f(x) = 1$ pour $x$ appartenant à $(0,1)$ et vaut zéro ailleurs.

On utilise les dernières relations établies :

$$\langle x_j^{(n)} \rangle = \frac{n!}{(j-1)!(n-j)!} \int_0^1 x^{j-1}[1-x]^{n-j} x\,\mathrm{d}x.$$

$$\langle x_1^{(5)} \rangle = \frac{1}{6},$$

$$\langle x_2^{(5)} \rangle = \frac{2}{6},$$

$$\langle x_3^{(5)} \rangle = \frac{3}{6},$$

$$\langle x_4^{(5)} \rangle = \frac{4}{6},$$

$$\langle x_5^{(5)} \rangle = \frac{5}{6}.$$

On remarque que les valeurs sont « symétriques » par rapport à la valeur centrale.

# 12. Lois

## 12.1. L'exercice est sans difficulté

## 12.2. L'exercice est traité dans le cours p. 320

## 12.3. L'exercice est traité p. 328 et p. 331

## 12.4. L'exercice est traité p. 336

## 12.5. L'exercice n'offre pas de difficulté

## 12.6. Loi de Poisson. Distribution de Cauchy

***a*** – On doit avoir $\int_{-\infty}^{+\infty} p(x)\,\mathrm{d}x = k\int_{-\infty}^{+\infty} \exp(-a|x|)\,\mathrm{d}x = 1$ ;
soit encore : $2k\int_{0}^{\infty} \exp -a|x|\,\mathrm{d}x = 1$. D'où $k = a/2$.

***b*** – $\varphi(t) = \int_{-\infty}^{+\infty} \exp(j2\pi xt)p(x)\,\mathrm{d}x$ expression dans laquelle $t$ appartient à $(-\infty, +\infty)$. Soit ici :

$$\varphi(t) = \frac{a}{2}\int_{-\infty}^{+\infty} \exp(j2\pi xt)\exp(-a|x|)\,\mathrm{d}x = \frac{a}{2}\int_{0}^{\infty}[\exp(j2\pi xt) + \exp(-j2\pi xt)]\exp(-ax)\,\mathrm{d}x$$
$$= \frac{a^2}{a^2 + 4\pi^2t^2} = \frac{1}{1 + \frac{4\pi^2t^2}{a^2}}\cdot$$

***c*** – Calcul de la moyenne :

$$m = \frac{a}{2}\int_{-\infty}^{+\infty} x\exp(-a|x|)\,\mathrm{d}x = 0$$

car la fonction sous le signe somme est impaire. Calcul de la variance :

$$D = \frac{a}{2}\int_{-\infty}^{+\infty} x^2\exp(-a|x|)\,\mathrm{d}x = a\int_{0}^{\infty} x^2\exp(-ax)\,\mathrm{d}x.$$

On intègre deux fois par parties cette dernière expression, et l'on trouve : $D = 2/a^2$.

***d*** – La distribution de Student à $m$ degrés de liberté s'écrit :

$$l_m(x) = \frac{\Gamma[(m+1)/2]}{\sqrt{\pi m}\Gamma(m/2)}\left[1 + \frac{x^2}{m}\right]^{-(m+1)/2} \qquad \text{pour } x \in (-\infty; +\infty),$$

où $\Gamma(x)$ est la fonction factorielle. Dans le cas où $m = 1$, on obtient :

$$q(x) = \frac{1}{\pi}\cdot\frac{1}{1+x^2} \quad \text{(cette fonction s'appelle fonction de Lorentz).}$$

***e*** – On a

$$\exp(-2\pi|t|) \quad \overset{TF}{\longleftrightarrow} \quad \frac{1}{\pi(1+x^2)}$$

expression donnée dans le cours à propos des transformées de Fourier. De là on écrit : $\Phi(t) = \exp(-2\pi|t|)$.

***f*** – Les variables aléatoires $\xi_j$ sont indépendantes et obéissent à la même fonction de distribution $q(x)$ ; la variable aléatoire $\frac{\xi_j}{n}$ obéit à une fonction de distribution $q(x/n)$, ainsi, la variable aléatoire $z$ a sa fonction caractéristique $\Psi(t)$ égale au produit des fonctions caractéristiques $\psi_j(t)$ de toutes les variables indépendantes $\xi_j/n$. Nous avons donc :

$$\psi_j(t) = \Phi(t/n) = \exp\left(-2\pi|t/n|\right),$$
$$\text{ainsi :} \quad \Psi(t) = \prod_{i=1}^{n} \exp\left(-2\pi|t/n|\right) = \exp\left(-2\pi|t|\right).$$

La fonction de distribution de $z$ est donc

$$q(x) = \frac{1}{\pi(1+x^2)}\cdot$$

***g*** – Calcul de la moyenne de $\xi_j$ :

$$M = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{x}{(1+x^2)}\,\mathrm{d}x.$$

C'est une fonction impaire, donc on pourrait dire que $M = 0$ ; cependant, $M = [\log_e(1+x^2)]_{-\infty}^{+\infty}$, la moyenne n'a pas de limite et n'existe donc pas au sens ordinaire. On dit que $M = 0$ en valeur principale au sens de Cauchy.

## 12.7. Loi binomiale négative. Loi de Poisson dont le paramètre suit une loi du $\chi^2$. Urne de Pólya

***1. a*** – La probabilité d'avoir tiré exactement $k$ échecs qui précèdent le tirage du $r^{\mathrm{e}}$ succès est $f(k,r,p)$.

***b*** – Nous avons $A = B \cdot C$, et comme les événements $B$ et $C$ sont indépendants, nous pouvons écrire : $P(A) = P(B)P(C)$ avec :

$$P(B) = C_{r+k-1}^{r-1} p^{r-1} p^k = C_{r+k-1}^{k} p^{r-1} p^k = C_{n-1}^{k} p^{r-1} p^k,$$
$$\text{et} \quad P(C) = p,$$
$$\text{d'où} \quad P(A) = C_{r+k-1}^{k} p^r p^k = f(k,r,p) = (-1)^k \binom{-r}{k} p^r p^k.$$

***c*** – Transformons l'égalité $p^r p^{-r} = p^r(1-q)^{-r} = 1$, on trouve : $p^r \sum_{k=0}^{\infty} C_{k+r-1}^{k} q^k = 1$.

***d*** – Par analogie avec la transformation réalisée sur la loi binomiale, on remplace $q$ par $qt$ pour obtenir la loi de génération des moments, soit : $f(t) = p^r(1-qt)^{-r}$, que l'on développe encore :

$$f(t) = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k (qt)^k$$

expression à partir de laquelle on va calculer les deux premiers moments non centrés :

$$m_1 = f'(1) = p^r r q p^{-r-1} = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k k q^k,$$

$$\text{d'où :} \quad m_1 = r\frac{q}{p}\cdot$$

On poursuit les calculs :

$$f''(1) = p^r r(r-1)q^2 p^{-r-2} = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k k(k-1)q^k = m_2 - m_1,$$

ce qui donne :

$$m_2 = m_1 + r(r+1)\frac{q^2}{p^2},$$

de là on tire l'écart type :

$$\sigma^2 = m_2 - m_1^2 = \frac{rq}{p^2}\cdot$$

Pour mémoire, on rappelle que, pour la loi binomiale, on obtient le résultat :

$$\sigma^2 = npq.$$

***e*** – On écrit : $\varphi(t) = M[\exp(jkt)]$, $M(\ )$ représentant l'opérateur moyenne, ce qui donne :

$$\varphi(t) = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k q^k \exp(jkt) = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k [q\exp(jt)]^k = p^r[1 - q\exp(jt)]^{-r}.$$

***2. a*** – On rappelle quelques résultats concernant la loi de Poisson :

$$P(\xi = k) = \frac{\exp(-\lambda)\lambda^k}{k!} \quad \text{avec} \ \langle\xi\rangle = \lambda \quad \text{et} \ \sigma^2 = \lambda.$$

***b*** – Il convient de vérifier que $\int_0^\infty g(x)\,\mathrm{d}x = 1$, pour cela calculons :

$$I = \int_0^\infty \frac{\alpha^\nu}{\Gamma(\nu)} x^{\nu-1}\exp(-\alpha x)\,\mathrm{d}x = \frac{\alpha^\nu}{\Gamma(\nu)}\int_0^\infty x^{\nu-1}\exp(-\alpha x)\,\mathrm{d}x,$$

effectuons le changement de variable $y = \alpha x$, on obtient :

$$I = \frac{\alpha}{\Gamma(\nu)}\int_0^\infty y^{\nu-1}\exp(-x)\frac{\mathrm{d}x}{\alpha} = \frac{\alpha}{\Gamma(\nu)}\cdot\frac{\Gamma(\nu)}{\alpha} = 1,$$

par ailleurs, l'expression sous le signe somme est toujours non négative, $g(x)$ est donc bien une densité de probabilité.

Calcul de

$$\Pi(\xi = k) = \int_0^\infty \exp(-\lambda)\frac{\lambda^k}{k!} g(\lambda)\, \mathrm{d}\lambda,$$

soit :

$$= \frac{\alpha^\nu}{k!\Gamma(\nu)} \int_0^\infty \lambda^{k+\nu-1} \exp[-(\alpha+1)\lambda]\, \mathrm{d}\lambda = \frac{\Gamma(\nu+k)}{k!\Gamma(\nu)} \cdot \frac{\alpha^\nu}{(\alpha+1)^{\nu+k}},$$

soit encore :

$$= \frac{\alpha^\nu}{(\alpha+1)^{\nu+k}} C^\nu_{k+\nu-1} = (-1)^\nu \frac{\alpha^\nu}{(\alpha+1)^{\nu+k}} C^\nu_{-k},$$

en définitive :

$$\Pi(\xi = k) = \left[\frac{1}{(\alpha+1)}\right]^k C^\nu_{k+\nu-1} \left[\frac{\alpha}{(\alpha+1)}\right]^\nu,$$

expression dans laquelle on pose : $p = 1/(\alpha+1)$ et $q = \alpha/(\alpha+1)$, ainsi on aboutit à :

$$\Pi(\xi = k) = p^k C^\nu_{k+\nu-1} q^\nu$$

qui est bien l'expression de la loi binomiale négative.

***3. a*** – Supposons que l'on ait tiré d'abord toutes les boules noires ensuite toutes les boules rouges, la probabilité de ce tirage s'écrirait dans ce cas particulier :

$$\Pi_{n_1 n_2} = \frac{b}{b+r} \cdot \frac{b+c}{b+r+c} \cdots \frac{b+(n_1-1)c}{b+r+(n_1-1)c} \cdot \frac{r}{b+r+n_1 c} \cdots \frac{r+(n_2-1)c}{b+r+(n_1+n_2-1)c}.$$

Maintenant, en examinant cette expression, on s'aperçoit que l'on peut réaliser n'importe quelle combinaison d'un des produits du numérateur et d'un des produits du numérateur sans que la probabilité soit changée ; et, si l'on considère un autre ordre pour les $n_1$ boules noires et $n_2$ boules rouges, le calcul reste le même, mais les termes sont arrangés dans un autre ordre. Donc pourvu que $n_1$ et $n_2$ ne soit pas changés, toutes les séquences où figurent $n_1$ boules noires et $n_2$ boules rouges ont la même probabilité quel que soit l'ordre de tirage des boules.

***b*** – Par suite, la probabilité de tirer $n_1$ boules noires parmi $n = n_1 + n_2$ boules est donc :

$$P(n_1, n) = C^{n_1}_n \Pi_{n_1 n_2} = \frac{n!}{n_1! n_2!} \Pi_{n_1 n_2} = \frac{n!}{n_1! n_2!} \cdot \frac{b}{b+r} \cdot \frac{b+c}{b+r+c} \cdots$$
$$\times \frac{b+(n_1-1)c}{b+r+(n_1-1)c} \cdot \frac{r}{b+r+n_1 c} \cdots \frac{r+(n_2-1)c}{b+r+(n_1+n_2-1)c}.$$

On divise chaque élément binôme du numérateur et chaque élément binôme du dénominateur de $\Pi_{n_1 n_2}$ par le nombre $c$, et l'on obtient :

$$P(n_1, n) = \frac{C^{n_1}_{n_1+b/c-1} C^{n_2}_{n_2+r/c-1}}{C^n_{n+b/c+r/c-1}} = \frac{C^{n_1}_{-b/c} C^{n_2}_{-r/c}}{C^n_{-b/c-r/c}}.$$

**c** – Le changement de notation donne :

$$P(n_1, n) = \frac{C_{-p/\gamma}^{n_1} C_{-q/\gamma}^{n_2}}{C_{-1/\gamma}^{n}} \cdot$$

**d** – Lorsque $n \longrightarrow \infty$ et que $p$ et $\gamma$ tendent vers zéro de telle sorte que :

$$np \longrightarrow \lambda \quad \text{et} \quad n\gamma \longrightarrow \frac{1}{\rho},$$

ce qui entraîne aussi que $n_2 \gg n_1$, c'est-à-dire que $n_2$ est de l'ordre de $n$ ; on peut étudier le comportement des différents facteurs :

$$C_{-p/\gamma}^{n_1} \longrightarrow C_{-\lambda\rho}^{n_1},$$

par ailleurs :

$$C_{-1/\gamma}^{n_1} \approx C_{-1/\gamma}^{n_2} \left[\frac{1+\gamma n_2}{\gamma n_2}\right]^{n_1} = C_{-1/\gamma}^{n_2} [1+\rho]^{n_1},$$

ensuite on calcule l'expression suivante :

$$W = \frac{C_{-q/\gamma}^{n_2}}{C_{-1/\gamma}^{n_2}} = \frac{C_{-q/\gamma}^{n_2}}{C_{-1/\gamma}^{n_2}} = \frac{q(q+\gamma)\ldots[q+\gamma(n_2-1)]}{(1+\gamma)\ldots[1+\gamma(n_2-1)]} \cdot$$

Cette dernière expression est plus délicate à étudier et désignons par $a_k$ le terme général dans le dernier membre :

$$a_k = \frac{q+\gamma k}{1+\gamma k} = 1 - \frac{p}{1+\gamma k}$$

ce terme tend vers 1 quel que soit $k$ par valeurs inférieures à 1. Prenons le logarithme de $W$, nous avons :

$$\log_e(W) = \sum_{j=1}^{n_2} \log_e\left(1 - \frac{p}{1+\gamma k}\right) \approx -\sum_{j=1}^{n_2} \frac{p}{1+\gamma k} = -p \sum_{j=1}^{n_2} \frac{1}{1+\gamma k}$$

$$\log_e(W) \approx -p\frac{1}{\gamma} \log_e(1+\gamma_{n_2}) = -\frac{p}{\gamma} \log_e\left(1+\frac{1}{\rho}\right)$$

or $\frac{p}{\gamma} = \lambda\rho$, de là on tire : $W = [\frac{\rho}{1+\rho}]^{\rho\lambda}$, en définitive, on aboutit au résultat :

$$P(n_1, n) \longrightarrow C_{-\rho\lambda}^{n_1} \left[\frac{\rho}{1+\rho}\right]^{\rho\lambda} \left[\frac{1}{1+\rho}\right]^{n_1} \cdot$$

## 12.8. La loi binomiale négative (reprise et applications)

Les questions a, b, c et d ont fait l'objet d'une réponse lors de l'étude du problème précédent.

**e** – Voici les résultats du calcul :

$$m_1 = \frac{172}{150} = 1{,}15, \quad \text{puis} \quad m_2 = \frac{536}{149} = 3{,}597, \quad \text{et} \quad s^2 = 2{,}282.$$

**Tableau I.4.**

| $k$ | expériences | $p_{k^*}$ | $p_k$ | $(p_{k^*} - p_k)^2/p_{k^*}$ |
|---|---|---|---|---|
| 0 | 70 | 0,467 | 0,318 | $0{,}475 \times 10^{-1}$ |
| 1 | 38 | 0,253 | 0,364 | $0{,}487 \times 10^{-1}$ |
| 2 | 17 | 0,133 | 0,209 | $0{,}434 \times 10^{-1}$ |
| 3 | 10 | 0,067 | 0,080 | $0{,}252 \times 10^{-2}$ |
| 4 | 9 | 0,060 | 0,023 | $0{,}228 \times 10^{-1}$ |
| 5 | 3 | 0,020 | 0,005 | $0{,}45 \times 10^{-2}$ |
| 6 | 2 | 0,013 | 0,001 | $0{,}111 \times 10^{-1}$ |
| 7 | 1 | 0,006 | 0,000 1 | $0{,}6 \times 10^{-2}$ |
| 8 et plus | 0 | 0,0 | 0,0 | 0,0 |

***f*** – Bien que $m_1$ et $s^2$ soient notablement différents, envisageons la distribution de Poisson en prenant $\lambda = m_1$, on obtient le tableau I.4.

Pour la valeur expérimentale du $\chi^2$ ; on obtient 31, la consultation des tables du $\chi^2$ à $(9-1)$ degrés de liberté (attention la case $k = 0$ compte...) montre que la probabilité de pouvoir dépasser ce nombre est inférieure à 1 chance sur 1 000. Force nous est de rejeter l'hypothèse d'une distribution de Poisson **avec moins d'une chance sur 1 000 de la rejeter à tort.**

***g*** – On a : $p = m_1/\sigma^2 = 0{,}502\,4$, par suite $q = 0{,}497\,6$ et $r = 1{,}158$. Réécrivons la loi binomiale négative :

$$p^r(1-q)^{-r} = p^r \sum_{k=0}^{\infty} C_{k+r-1}^k q^k = 1.$$

On calcule $p^r = 0{,}450\,7$, il reste à calculer les différents coefficients du binôme que l'on multiplie par $p^r$ (*cf.* Tab. I.5).

**Tableau I.5.**

| $k$ | expériences | $p_{k^*}$ | $p_k$ | $10^3(p_{k^*} - p_k)^2/p_{k^*}$ |
|---|---|---|---|---|
| 0 | 70 | 0,467 | 0,450 7 | 0,581 |
| 1 | 38 | 0,253 | 0,259 6 | 0,172 |
| 2 | 17 | 0,133 | 0,139 4 | 0,308 |
| 3 | 10 | 0,067 | 0,073 | 0,537 |
| 4 | 9 | 0,060 | 0,037 8 | 8,066 |
| 5 | 3 | 0,020 | 0,023 | 0,45 |
| 6 | 2 | 0,013 | 0,014 | 0,008 |
| 7 | 1 | 0,006 | 0,008 | 0,666 |
| 8 et plus | 0 | 0,0 | 0,0 | 0,0 |

En définitive, on trouve $\chi^2 = 1{,}618$ à $(9-3)$ degrés de liberté (on a calculé $p$, $q$ et $r$). Si l'on regarde les tables du $\chi^2$ on voit qu'il y a 95 chances sur 100 de dépasser la valeur calculée. Il n'y a pas de raison de ne pas conserver cette loi.

## 12.9. Loi de Poisson. Durée de vie d'un système simple. Fiabilité

***a*** –

1. $\lambda = 0{,}5 \quad P_1 = 0{,}5\exp(-0{,}5) = 0{,}303$
2. $\lambda = 2 \quad P_1 = 2\exp(-2) = 0{,}271$
3. $\lambda = 2 \quad P_0 = \exp(-2) = 0{,}135$
4. $\lambda = 2 \quad P_2 = 2\exp(-2) = 0{,}271 \quad$ puis $P = 1 - P_0 - P_1 - P_2 = 0{,}323$
5. $\lambda = 4 \quad P_4 = \dfrac{4^4}{4!}\exp(-4) = 0{,}195.$

***b*** – $P_k(t) = [(\lambda t)^k/k!]\exp(-\lambda t)$. On écrit : $F(t) = P(\xi \leq t)$, et la probabilité pour qu'il n'y ait aucun événement dans l'intervalle de temps $t$ débutant à l'instant $t_i$ est $P_0(t) = 1 - F(t) = \exp(-\lambda t)$, ce qui donne $F(t) = 1 - \exp(-\lambda t)$ et la densité de probabilité $f(t) = \lambda\exp(-\lambda t)$.

***Calcul de la moyenne***

$$a = \int_0^\infty t\lambda\exp(-\lambda t)\,\mathrm{d}t,$$

après une intégration par parties on obtient : $a = 1/\lambda$.

***Calcul de $\sigma^2$***

$$D = \int_0^\infty t^2\lambda\exp(-\lambda t)\,\mathrm{d}t = \frac{2}{\lambda^2},$$

d'où

$$\sigma^2 = D - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}\cdot$$

***Une propriété remarquable*** – On écrit le théorème concernant le produit d'événements : $P(A\cdot B) = P(A)\cdot P(B|A)$, d'où la probabilité conditionnelle $P(B|A) = P(A\cdot B)/P(A) = \Phi(t)$. Comme $P(A\cdot B) = P(\tau < \xi < t+\tau) = F(t+\tau) - F(\tau)$ et que par ailleurs :

$$P(A) = P(\xi > \tau) = 1 - P(\xi \leq \tau) = 1 - F(\tau),$$

on obtient en définitive :

$$\Phi(t) = \frac{F(t+\tau) - F(\tau)}{1 - F(\tau)}\cdot$$

Enfin, remplaçons $F(t)$ par sa valeur : $\Phi(t) = 1 - \exp(-\lambda t) = F(t)$.

## 12.10. Durée de vie d'un système. Loi de Weibull. Fiabilité

***a*** – Nous pouvons écrire :

$$F(t) = P(\xi < t) = \frac{n_0 - n(t)}{n_0} = 1 - \frac{n(t)}{n_0}, \quad \text{d'où :} \quad \frac{n(t)}{n_0} = 1 - F(t).$$

À présent ajoutons et retranchons $n_0$ au numérateur de $\lambda(t)$ :

$$\lambda(t) = \frac{-n_0 n(t) + n_0 - n(t + \Delta t)}{\Delta t n(t)} = \frac{F(t) - F(t + \Delta t)}{\Delta t n(t)} = -\frac{\dfrac{\mathrm{d}F(t)}{\mathrm{d}t}}{\dfrac{n(t)}{n_0}}.$$

***b*** – De ce résultat on déduit l'équation différentielle du premier ordre à laquelle obéit la répartition $F(t)$ en fonction de $\lambda(t)$ :

$$\frac{\mathrm{d}F(t)}{\mathrm{d}t} + \lambda(t)[1 - F(t)] = 0.$$

***c*** – L'intégration donne : $\log_e[1 - F(t)] = \int_0^t \lambda(x)\,\mathrm{d}x$, d'où :

$$F(t) = 1 - \exp\left(-\int_0^t \lambda(x)\,\mathrm{d}x\right).$$

***d*** – À présent, $\lambda(t) = \lambda_0 \alpha t^{\alpha-1}$, il s'ensuit que : $F(t) = 1 - \exp(-\lambda_0 t^\alpha)$, d'où la loi de Weibull :

$$f(t) = \lambda_0 \alpha t^{\alpha-1} \exp(-\lambda_0 t^\alpha).$$

***e*** – Calcul de $\bar{a}$, nous avons :

$$\bar{a} = \int_0^\infty t \lambda_0 \alpha t^{\alpha-1} \exp(-\lambda_0 t^\alpha)\,\mathrm{d}t,$$

après une intégration par parties on trouve :

$$\bar{a} = \lambda_0^{-1/\alpha} \Gamma\left(1 + \frac{1}{\alpha}\right).$$

Calcul de $\bar{D}$ moment non centré du deuxième ordre, il s'effectue selon le même procédé :

$$\bar{D} = \int_0^\infty t^2 \lambda_0 \alpha t^{\alpha-1} \exp(-\lambda_0 t^\alpha)\,\mathrm{d}t,$$

$$\bar{D} = \lambda_0^{-2/\alpha} \Gamma\left(1 + \frac{2}{\alpha}\right).$$

on en déduit :

$$\sigma^2 = \lambda_0^{-2/\alpha} \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right)\right].$$

La densité de probabilité établie au § b du problème précédent est un cas particulier de la loi de Weibull dans laquelle il suffit de faire $\alpha = 1$.

## 12.11. L'inégalité de Kolmogorov

***a*** – L'inégalité de Bienaymé-Tchebycheff est démontrée dans le cours.

***b*** – Partons de l'égalité :

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\,\mathrm{d}x = \int_{|x|\leq a} x^2 f(x)\,\mathrm{d}x + \int_{|x|>a} x^2 f(x)\,\mathrm{d}X,$$

la première intégrale du second membre est majorée par $a^2 \int_{|x|\leq a} f(x)\ \mathrm{d}x$, d'où l'inégalité demandée :

$$E(X^2) \leq a^2 \int_{|x|\leq a} f(x)\,\mathrm{d}x + \int_{|x|>a} x^2 f(x)\,\mathrm{d}x,$$

expression dans laquelle $a$ est compris entre 0 et $B$.

Si $|x| > a$, il reste néanmoins que $|x| < B$. En désignant par $F(x)$ la fonction de répartition de la variable aléatoire $x$, on obtient les relations suivantes :

$$F(x) = 1 \quad \text{si } x > B$$
$$F(x) = 0 \quad \text{si } x < -B,$$

on peut écrire la suite d'inégalités :

$$\int_{a<|x|<B} x^2 f(x)\,\mathrm{d}x < B^2 \int_{a<|x|<B} f(x)\,\mathrm{d}x < B^2 \int_{|x|>a} f(x)\,\mathrm{d}x$$

d'où il ressort que la dernière intégrale n'est rien d'autre que $B^2 P(|X| > a)$. En remarquant que : $\int_{|x|\leq a} f(x)\,\mathrm{d}x < 1$, on en conclut que :

$$E(X^2) \leq a^2 + B^2 P(|X| > a),$$

soit encore :

$$P(|X| > a) \geq \frac{E(X^2) - a^2}{B^2}\cdot$$

## 12.12. Loi quasi normale

***a*** – On a $E(Y) = 0$ l'écart quadratique moyen $E(Y^2) = \sigma^2$, $E(Y^3) = 0$ et $E(Y^4) = 3\sigma^4$.

***b*** – On a :

$$f(x) = \frac{1}{\sqrt{2\pi}}(ax^2 + bx - 3a)^2 \exp\left(-\frac{x^2}{2}\right),$$

où $f(x)$ est une fonction manifestement non négative, la condition $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$ impose la relation : $6a^2 + b^2 = 1$. Ensuite on calcule les différents moments :

$$\begin{aligned} E(X) &= 6ab - 6ab = 0 \\ E(X^2) &= 6a^2 + 3b^2 \\ E(X^3) &= 12ab \\ E(X^4) &= 24a^2 + 15b^2. \end{aligned}$$

La condition $E(X^4) = 3E(X^2)^2$ impose une autre relation :

$$24a^2 + 15b^2 = 3\left(6a^2 + 3b^2\right)^2,$$

à laquelle nous combinons la condition de normalité $6a^2 + b^2 = 1$ pour obtenir :

$$36a^4 - 14a^2 + 1 = 0,$$

d'où nous tirons :

$$a^2 = 0{,}295 \qquad \text{ou encore} \qquad a^2 = 0{,}094.$$

On peut choisir entre les deux valeurs de $a^2$, car $b^2$ doit être positif, ce qui entraîne que $a^2 < 1/6 = 0{,}166$ donc $a^2 = 0{,}094$.

Dans ces conditions, $b^2 = 0{,}346$ et $\sigma = 0{,}642$.

## 13. La fonction caractéristique

### 13.1. $\left(\frac{\sin x}{x}\right)^n$ tend vers une gaussienne quand n croît positivement

***a*** – La fonction caractéristique d'une variable gaussienne centrée réduite s'écrit :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(j2\pi xt) \exp\left(-\frac{x^2}{2}\right)\,\mathrm{d}x = \exp\left(-2\pi^2 t^2\right)$$

c'est, bien sûr, une gaussienne.

La variable $\eta$ est la somme de deux variables gaussiennes, centrées, réduites et indépendantes, alors sa fonction caractéristique est $\psi(t) = \varphi^2(t) = \exp(-4\pi^2 t^2)$. $\psi(t)$ est la transformée de Fourier de $1/(2\sqrt{\pi})\exp\left(-t^2/4\right)$ et l'on retrouve bien le théorème d'addition des variances : $\sigma^2 = 2$ (*cf.* chapitre 21).

***b*** – À présent il suffit de faire une transformation linéaire $x = (y - m)/\sigma$ dans la fonction de distribution, on écrit :

$$\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(j2\pi yt) \exp\left(-\frac{[y-m]^2}{2\sigma^2}\right)\,\mathrm{d}y,$$

et l'on pose $z = (y - m)/\sigma$ ce qui donne $y = \sigma z + m$ et $\mathrm{d}y = \sigma\,\mathrm{d}z$. On obtient donc :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp[j2\pi(\sigma z + m)t] \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$$

$$= \frac{\exp(j2\pi tm)}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(j2\pi\sigma zt) \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z,$$

$$\varphi(t) = \exp(j2\pi tm)\varphi(\sigma t) = \exp(j2\pi tm) \exp(-2\pi^2\sigma^2 t^2), \tag{I.7}$$

théorème déjà établi dans le cours.

La fonction caractéristique de la variable $\eta$ qui est la somme de deux variables gaussiennes indépendantes est donc le produit des fonctions caractéristiques de chaque variable :

$$\Psi(t) = \exp[j2\pi t(m_1 + m_2)]\varphi(\sigma_1 t)\varphi(\sigma_2 t) = \exp[j2\pi t(m_1 + m_2)] \exp[-2\pi^2 t^2(\sigma_1^2 + \sigma_2^2)]$$

on voit qu'il suffit de poser $m = m_1 + m_2$ et $\sigma^2 = \sigma_1^2 + \sigma_2^2$ pour retrouver l'expression du théorème (I.7). Donc la distribution de $\eta$ est :

$$f(y) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left(-\frac{[y - m_1 - m_2]^2}{2(\sigma_1^2 + \sigma_2^2)}\right).$$

***c*** – La généralisation n'est pas très difficile, on obtient :

$$f(z) = \frac{1}{\sqrt{2\pi \sum_{k=1}^{n} \sigma_k^2}} \exp\left(-\frac{\left[z - \sum_{k=1}^{n} m_k\right]^2}{2\sum_{k=1}^{n} \sigma_k^2}\right).$$

***d*** – Il s'agit de la fonction fente et la fonction caractéristique s'écrit : $\varphi(t) = [\sin(\pi t)]/(\pi t)$. La fonction caractéristique de la variable $\xi/n$ donne : $\psi(t) = [\sin(\pi t/n)]/(\pi t/n)$ (attention, il ne s'agit pas d'une dilatation d'abscisse...) et la fonction caractéristique de la variable aléatoire $\eta = \dfrac{1}{n}\sum_{k=1}^{n} x_k$ s'écrit :

$$\Psi(t) = \left[\frac{\sin(\pi t/n)}{\pi t/n}\right]^n.$$

Développons

$$\log_e[\Psi(t)] = n \log_e\left(\frac{\sin(\pi t/n)}{\pi t/n}\right) = n \log_e\left(1 - \frac{\pi^2 t^2}{6n^2}\right) \approx -\frac{\pi^2 t^2}{6n},$$

autrement dit, $\Psi(t) = A \exp\left[-t^2/(2\sigma^2)\right]$, ce qui veut dire que la densité de probabilité est gaussienne. Cela mérite un commentaire supplémentaire, en effet, on voit très bien que la valeur $\sigma$ croît comme $\sqrt{n}$, mais cela est conforme au théorème d'addition des $\sigma_k$ des variables indépendantes. Plus la somme des variables est élevée, plus $\sigma$ croît. Il peut être alors utile d'effectuer un changement de variable pour conserver une échelle raisonnable.

Il résulte de ces calculs que l'on connaît le comportement asymptotique de la fonction $[\sin(x)/x]^n$ qui tend donc vers une gaussienne. Lorsque $n$ est un ordre de grandeur plus grand que 1, disons 10 environ, l'écart entre les deux fonctions en valeur absolue est de quelques millièmes. Ce théorème n'apparaît pas explicitement dans les ouvrages d'analyse. Le programme `card0.c` réalise les opérations proposées.

# 14. La loi du $\chi^2$ et la loi de Student

## 14.1. Le test de Pearson

***a*** – On développe la fonction caractéristique de plusieurs variables :

$$\varphi(t_1, t_2, t_3, \ldots, t_m) = M[\exp(j\,\{t_1X_1 + t_2X_2 + t_3X_3, \ldots, +t_mX_m\})]$$

en série de MacLaurin-Taylor, au préalable on s'intéresse au développement de :

$$\exp(jt_qX_q) = \sum_{k=0}^{\infty} j^k \frac{X_q^k}{k!} t_q^k = 1 + j\frac{X_q}{1!}t^q - \frac{X_q^2}{2!}t_q^2 + \cdots$$

puis au produit :

$$\prod_{q=1}^{m} \exp(jt_qX_q) = \prod_{q=1}^{m} \left(1 + j\frac{X_q}{1!}t_q - \frac{X_q^2}{2!}t_q^2 + \cdots\right) = 1 + j\sum_{q=1}^{m} X_qt_q - \frac{1}{2}\sum_{q=1}^{m} X_q^2t_q^2 + \cdots,$$

et enfin à la fonction caractéristique :

$$\varphi(t_1, t_2, t_3, \ldots, t_m) = M\left[\prod_{q=1}^{m} \exp(jt_qX_q)\right] = 1 + jM[X_1]t_1 + jM[X_2]t_2 + \cdots$$
$$+ jM[X_m]t_m - \frac{1}{2}M[X_1^2]t_1^2 - \frac{1}{2}M[X_2^2]t_2^2 - \cdots - \frac{1}{2}M[X_m^2]t_m^2 - \cdots$$

Nous avons le développement direct de $\varphi(t_1, t_2, t_3, \ldots, t_m)$ :

$$\varphi(t_1, t_2, t_3, \ldots, t_m) = \varphi + \sum_{q=1}^{m} \frac{\partial\varphi}{\partial t_q}t_q + \frac{1}{2}\sum_{q=1}^{m} \frac{\partial^2\varphi}{\partial t_q^2}t_q^2 + \sum_{q=1}^{m}\sum_{p\neq q}^{m} \frac{\partial\varphi}{\partial t_q} \cdot \frac{\partial\varphi}{\partial t_p}t_qt_p + \cdots$$
$$\text{pour} \quad \{t_q\} = 0.$$

Par identification, on obtient :

$$\frac{\partial\varphi}{\partial t_q} = jM[X_q] \quad \text{et} \quad \frac{\partial^2\varphi}{\partial t^2} = -M[X_q^2] \quad \text{pour} \quad \{t_q\} = 0.$$

***b*** – Si l'on spécifie l'ordre dans lequel les cas doivent se succéder :

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_r = n_r) = p_1^{n_1}p_2^{n_2}, \ldots, p_r^{n_r},$$

En revanche, si l'ordre n'a pas d'importance, il y a $n!$ combinaisons de $n$ éléments ; seulement, il doit y avoir $n_1$ éléments dans $I_1$, $n_2$ éléments dans $I_2$, $n_r$ éléments dans $I_r$. Il y a donc $n_1!$ combinaisons dans $I_1$ et $n_r!$ combinaisons dans $I_r$ lesquelles ont déjà été factorisées dans $n!$ ; donc le coefficient qu'il convient d'appliquer au cas où l'ordre est spécifié est :

$$\frac{n!}{n_1!n_2!\ldots n_r!}$$

de là on tire la loi de distribution multinomiale :

$$P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_r = n_r) = \frac{n!}{n_1!n_2!\ldots n_r!}p_1^{n_1}p_2^{n_2}, \ldots, p_r^{n_r}.$$

**c** – Nous avons $Y_q = \sum_{k=1}^{n} Z_k^q$ = nombre de fois que $X_k$ appartient à $I_q$. La fonction caractéristique associée à cette variable s'écrit :

$$\varphi_q(t_1, t_2, t_3, \ldots, t_r) = M\left[\exp\left(j\sum_{k=1}^{r} Z_k^q t_k\right)\right],$$

et comme $P(Z_k^q = 1) = p_k$ et comme $P(Z_k^q = 0) = 1 - p_k$, on peut écrire :

$$\varphi_q(t_1, t_2, t_3, \ldots, t_r) = \sum_{k=1}^{r} p_k \exp(jt_k).$$

La fonction caractéristique d'une somme de variables aléatoires indépendantes est le produit des fonctions caractéristiques de chacune des variables aléatoires, donc, ici, comme elles ont toutes la même fonction de distribution, on peut écrire :

$$\varphi(t_1, t_2, t_3, \ldots, t_r) = \left[\sum_{k=1}^{r} p_k \exp(jt_k)\right]^n.$$

**d** – Il suffit de dériver la fonction caractéristique pour obtenir les moments désirés :

$$m_{1q} = \frac{1}{j}\frac{\partial\varphi}{\partial t_q} = \frac{1}{j} np_q j \exp(jt_q)\left[\sum_{k=1}^{r} p_k \exp(jt_k)\right]^{n-1} \quad \text{pour} \quad \{t_q\} = 0.$$

Comme $\sum_{k=1}^{r} p_k = 1$, on trouve : $m_{1q} = np_q$ (ce résultat aurait pu être obtenu directement). Ensuite, par le même style de calculs, on montre que :

$$m_{2q} = -\frac{\partial^2\varphi}{\partial t_q^2} = n^2 p_q^2 + np_q - np_q^2,$$

d'où on déduit l'écart type :

$$\sigma^2 = np_q(1 - p_q).$$

$\sigma$ est proportionnel à $\sqrt{np_q}$ donc $U_q$ est une variable centrée réduite indépendante de $n$.

**e** – Écrivons la fonction caractéristique de $U_q$ :

$$\begin{aligned}
\varphi(t_1, t_2, t_3, \ldots, t_r) &= M\left[\exp\left(j\sum_{q=1}^{r} U_q t_q\right)\right] = M\left[\exp\left(j\sum_{q=1}^{r} \frac{Y_q - np_q}{\sqrt{np_q}} t_q\right)\right] \\
&= \exp\left(-j\sum_{q=1}^{r} t_q\sqrt{np_q}\right) M\left[\exp\left(j\sum_{q=1}^{r} \frac{Y_q}{\sqrt{np_q}} t_q\right)\right] \\
&= \exp\left(-j\sum_{q=1}^{r} t_q\sqrt{np_q}\right)\left[\sum_{q=1}^{r} p_q \exp\left(j\frac{t_q}{\sqrt{np_q}}\right)\right]^n \\
&= \left[\exp\left(-j\sum_{q=1}^{r} t_q\sqrt{\frac{p_q}{n}}\right)\sum_{q=1}^{r} p_q \exp\left(j\frac{t_q}{\sqrt{np_q}}\right)\right]^n.
\end{aligned}$$

Posons

$$A = \exp\left(-j\sum_{q=1}^{r} t_q\sqrt{\frac{p_q}{n}}\right) \quad \text{et} \quad B = \sum_{q=1}^{r} B_q = \sum_{q=1}^{r} p_q \exp\left(j\frac{t_q}{\sqrt{np_q}}\right),$$

puis développons chacune de ces quantités pour $n$ tendant vers l'infini :

$$A = 1 - j\sum_{q=1}^{r} t_q\sqrt{\frac{p_q}{n}} - \frac{1}{2}\left[\sum_{q=1}^{r} t_q\sqrt{\frac{p_q}{n}}\right]^2 + \cdots$$

$$B_q = p_q\left(1 + j\frac{t_q}{\sqrt{np_q}} - \frac{1}{2}\cdot\frac{t_q^2}{np_q} + \cdots\right)$$

ensuite :

$$B = \sum_{q=1}^{r} p_q + j\sum_{q=1}^{r}\frac{t_q}{\sqrt{np_q}}p_q - \frac{1}{2}\sum_{q=1}^{r}\frac{t_q^2}{np_q}p_q + \cdots = 1 + j\sum_{q=1}^{r} t_q\sqrt{\frac{p_q}{n}} - \frac{1}{2n}\sum_{q=1}^{r} t_q^2 + \cdots$$

Revenons au produit $AB$ :

$$AB = 1 - \frac{1}{2n}\left[\sum_{q=1}^{r} t_q^2 - \left\{\sum_{q=1}^{r} t_q\sqrt{p_q}\right\}^2\right] + \cdots,$$

reste à élever ce dernier résultat à la puissance $n$ :

$$[AB]^n = \left[1 - \frac{1}{2n}\left(\sum_{q=1}^{r} t_q^2 - \left\{\sum_{q=1}^{r} t_q\sqrt{p_q}\right\}^2\right) + \cdots\right]^n = \exp\left(-\frac{1}{2}\sum_{q=1}^{r} t_q^2 - \left\{\sum_{q=1}^{r} t_q\sqrt{p_q}\right\}^2\right),$$

on en conclut que la variable $U$ (vecteur) obéit à une loi gaussienne.

$$\sum_{q=1}^{r} U_q\sqrt{p_q} = \sum_{q=1}^{r}\frac{Y_q - np_q}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{q=1}^{r} Y_q - \sqrt{n}\sum_{q=1}^{r} p_q = \frac{1}{\sqrt{n}}\left[\sum_{q=1}^{r} Y_q - n\right] = 0.$$

***Remarque importante*** – Les variables $U_q$ obéissent donc à une liaison linéaire. Cela ne signifie pas pour autant qu'elles ne sont pas indépendantes ; en effet, les probabilités sont théoriques car les variables obéissent à une loi de distribution théorique, et l'apparition d'une variable ne dépend aucunement de l'apparition de celles qui précèdent leur tirage.

***f*** – Chacune des composantes **indépendantes** $U_q$ tend donc vers la loi normale, alors la variable $\chi^2 = \sum_{q=1}^{r} U_q^2$ est bien la somme des carrés de variables gaussiennes, centrées, réduites et indépendantes (à une constante multiplicative près). Comme la loi statistique théorique impose une contrainte c'est-à-dire que $\sum_{q=1}^{r} U_q\sqrt{p_q} = 0$ ou que $\sum_{q=1}^{r} Y_q = n$, on perd un degré de liberté, il reste donc $(r-1)$ degrés de liberté.

## 14.2. Le test de Student

***1re approche*** – L'inégalité de Bienaymé-Tchebycheff donne :

$$P(|\langle x\rangle - m| > t\sigma) \leq \frac{1}{t^2},$$

et l'on désire que cette probabilité soit égale à 0,1, ce qui donne $t = \sqrt{10}$.

Par suite : $|\langle x\rangle - m| = s\sqrt{10} = 11{,}54$ cm

***2^e approche*** – Nous avons :

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{r}(x_i - \langle x\rangle)^2 \quad \text{avec} \quad \langle x\rangle = \frac{1}{n}\sum_{i=1}^{r} x_i,$$

et par ailleurs :

$$(x_i - \langle x\rangle)^2 = (x_i - m)^2 - m^2 + 2x_i + \langle x\rangle m^2 - 2x_i\langle x\rangle,$$

enfin :

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - m)^2 - n(m - \langle x\rangle)^2\right].$$

***Remarque*** : On a aussi $\sigma^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - n\langle x\rangle^2)$ et comme $\sigma^2$ ne dépend pas de l'origine des $x_i$, on peut aussi écrire :

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - m)^2 - n(\langle x\rangle - m)^2\right].$$

***a*** – L'opération orthogonale s'effectue sur les vecteurs orthogonaux $(x_i - m)$. Il y a conservation du carré scalaire, soit :

$$\sum_{i=1}^{n} x_i' = \sum_{i=1}^{n}(x_i - m)^2.$$

***b*** – La dernière équation de la transformation donne :

$$x_n' = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - m) = \frac{1}{\sqrt{n}}(n\langle x\rangle - nm),$$

d'où :

$$\langle x\rangle - m = \frac{x_n'}{\sqrt{n}}\cdot$$

***c*** – Comme

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - m)^2 - n(\langle x\rangle - m)^2\right],$$

on obtient :

$$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i'^2 - x_n^2\right) = \frac{1}{n-1}\sum_{i=1}^{n} x_i'^2.$$

***d*** – Les variables $(x_i - m)$ sont normales, centrées, indépendantes et de même variance, il en est de même des variables $x_i'$ obtenues par la transformation orthogonale qui conserve les longueurs et les angles.

***e*** – On a :

$$t = \frac{\sqrt{n}(\langle x\rangle - m)}{\sigma} = \frac{x'_n}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x'^2_i}},$$

cette dernière expression est bien la définition d'une variable de Student à $(n-1)$ degrés de liberté (*cf.* le chapitre 22). Autrement dit, $P[t < t_\alpha(n-1)] = \alpha$ ou encore :

$$P[t > t_\alpha(n-1)] = 1 - \alpha.$$

***f*** – Puisque $P(|t| > t_0) = 1 - \alpha$, cela entraîne :

$$P\left(\frac{\sqrt{n}|\langle x\rangle - m|}{\sigma} > t_0\right) = 1 - \alpha,$$

soit encore :

$$P\left(\frac{\sqrt{n}|\langle x\rangle - m|}{\sigma} > t_0\right) = P\left(|\langle x\rangle - m| > t_0\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

La condition

$$|\langle x\rangle - m| > t_0\frac{\sigma}{\sqrt{n}} \quad \left(= t_{\alpha/2}(n-1)\frac{\sigma}{\sqrt{n}}\right)$$

a la probabilité $(1-\alpha)$ de se réaliser. **Attention aux valeurs absolues qui doublent l'intervalle !**

***g*** – On veut que $|\langle x\rangle - m|$ appartienne à $I$ avec 90 chances sur 100, donc :

$$P(|\langle x\rangle - m| < I) = P\left(|\langle x\rangle - m| < t_{0,05}(n-1)\frac{\sigma}{\sqrt{n}}\right) = 0{,}10.$$

Comme $t_{0,05}(19) = 1{,}729$, on en déduit que $I = 1{,}729\dfrac{3,65}{\sqrt{20}} = 1{,}4$ (il s'agit de cm).

***3e approche*** – Si l'on admet que $t$ obéit à une distribution gaussienne, nous avons : $P(|t| < t_0) = \alpha$ avec $t_0 = 1{,}645$, il s'ensuit que $I' = 1{,}645 \times 3{,}65/\sqrt{20} = 1{,}34$ (cm).

La morale de cette histoire tient en peu de mots : l'inégalité de Bienaymé-Tchebycheff convient pour toutes les lois de distribution c'est donc la plus pénalisante avec une erreur de 11,5 cm qui, il faut bien le dire, est peu significative ; le choix *a priori* d'une distribution gaussienne est très raisonnable puisque les données sont au nombre de 20, cependant, on va trouver une erreur un peu optimiste (1,34 cm) car le nombre de degrés de liberté n'est quand même pas très élevé ; le calcul rigoureux effectué avec la distribution de Student donne le bon résultat : une erreur de 1,4 cm ce qui est peu au-dessus de l'erreur gaussienne. Sans avoir recours aux tables, on peut d'ores et déjà affirmer que la loi de Student tendant vers la loi de Gauss quand le nombre de degrés de liberté tend vers l'infini donnera le résultat : $t_{0,05}(\infty) = 1{,}645$. Au-delà d'un nombre de degrés de liberté de l'ordre de 50, la distribution gaussienne remplit bien son office.

# 15. Systèmes à plusieurs variables aléatoires

## 15.1. Système linéaire surdéterminé. Matrice de corrélation

Voici les résultats :

$$\begin{aligned} \langle x\rangle &= 4 & \sigma_x^2 &= 2{,}2 \\ \langle y\rangle &= 3{,}89 & \sigma_y^2 &= 4{,}26 \\ \langle z\rangle &= 2{,}24 & \sigma_z^2 &= 0{,}75 \end{aligned}$$

***a*** – Calcul de

$$r_{xz} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i z_i - \langle x\rangle\langle z\rangle}{\sigma_x \sigma_z},$$

on trouve $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i z_i = 9{,}06$, d'où :

$$r_p = 0{,}078.$$

***b*** – Calcul de

$$r_{xy} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i y_i - \langle x\rangle\langle y\rangle}{\sigma_x \sigma_y},$$

on trouve $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i y_i = 15{,}4$, d'où :

$$r_q = -0{,}052.$$

***c*** – Calcul de

$$r_{yz} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_i y_i - \langle z\rangle\langle y\rangle}{\sigma_z \sigma_y},$$

on trouve $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_i y_i = 10{,}17$, d'où :

$$r_q = 0{,}818.$$

Étude des liaisons de corrélation : on teste l'hypothèse $r = 0$, et l'on calcule :

$$\alpha_{xz} = \frac{|r_{xz}|\sqrt{n-2}}{\sqrt{1-r_{xz}^2}} = 0{,}22,$$

$$\alpha_{xy} = \frac{|r_{xy}|\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = 0{,}15,$$

$$\alpha_{yz} = \frac{|r_{yz}|\sqrt{n-2}}{\sqrt{1-r_{yz}^2}} = 4{,}02.$$

Si $\alpha < t_\alpha(n-2)$, on rejette alors **l'hypothèse d'une liaison de corrélation avec la probabilité $\alpha$ de la rejeter à tort.** On a $t_{0,05}(n-2) = 1{,}86$ donc on accepte l'hypothèse d'une absence d'une liaison de corrélation entre les variables $x$ et $z$ d'une part et $x$ et $z$ d'autre part. Seule la liaison entre les variables $y$ et $z$ est significative. En clair voici ce que signifie cette étude :

1. il n'y a pas d'impôt perçu sur les personnes ;
2. la taille d'une famille n'est pas liée à sa richesse ;
3. les impôts perçus ne dépendent que de la quantité détenue.

# 16. Critères de conformité

## 16.1. Étalonnage d'un appareil de mesure

***a*** – Calcul de la moyenne :

$$m = \int_a^b \frac{1}{b-a} x \, \mathrm{d}x = \frac{1}{2}(b+a),$$

puis calcul de :

$$D = \int_a^b \frac{1}{b-a} x^2 \, \mathrm{d}x = \frac{1}{3}\left(b^2 + ab + a^2\right),$$

d'où l'on tire :

$$\sigma^2 = D - m^2 = \frac{(b-a)^2}{12} \quad \text{puis} \quad \sigma = \frac{b-a}{2\sqrt{3}},$$

on remarque bien que $a$ et $b$ jouent des rôles identiques.

***b*** – On obtient le système :

$$\begin{aligned} b + a &= 2m \\ b - a &= \sigma 2\sqrt{3} \end{aligned}$$

sa résolution donne : $a = m - \sigma\sqrt{3}$ et $b = m + \sigma\sqrt{3}$. Numériquement, nous trouvons :

$$\begin{aligned} m &= 60{,}25 \\ D &= 4\,079 \\ \sigma^2 &= 448{,}93 \quad \text{et} \quad \sigma = 21{,}19 \\ a &= 23{,}55 \\ b &= 96{,}95. \end{aligned}$$

De là on tire la probabilité de tomber entre $a$ et $b$ : $f(x) = 1/(b-a) = 0{,}013\,624$ d'où le tableau I.6, page suivante, des probabilités théoriques.

Nombre de degrés de liberté $= 8 - 2 - 1 = 5$, les trois contraintes étant le calcul de $a$ et $b$ ainsi que la normalisation de la distribution.

$$\chi^2_{\text{obs}} = n \sum_{i=1}^{K} \frac{(p_i - p_i^*)^2}{p_i^*} = 21{,}57$$

où $K$ est le nombre de cases de regroupement.

**Tableau I.6.**

| | 23,55 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 96,95 |
|---|---|---|---|---|---|---|---|---|---|
| $m_i$ | 21 | 72 | 66 | 38 | 51 | 56 | 64 | 32 | |
| $m_i p_i^*$ | 35,15 | 54,5 | 54,5 | 54,5 | 54,5 | 54, 5 | 54,5 | 37,87 | |

En consultant les tables, on trouve une probabilité inférieure à une chance sur mille de dépasser cette valeur, on ne prend pas beaucoup de risques en rejetant l'hypothèse d'une distribution uniforme.

## 16.2. Tests d'hypothèse

***a*** – Les résultats des calculs sont présentés dans tableau I.7.

**Tableau I.7.**

| $m_i$ | intervalles | $p_i^*$ | $p_i^*$ cumulés | var. c. r. | $\Phi(x)$ | $\Delta\Phi(x)$ |
|---|---|---|---|---|---|---|
| | 18 | | | −2,815 | 0,002 4 | |
| 2 | | 0,025 | 0,025 | | | 0,014 5 |
| | 20 | | | −2,185 | 0,014 5 | |
| 3 | | 0,037 5 | 0,062 5 | | | 0,045 6 |
| | 22 | | | −1,154 | 0,060 1 | |
| 9 | | 0,112 5 | 0,175 | | | 0,118 |
| | 24 | | | −0,923 | 0,178 1 | |
| 12 | | 0,15 | 0,325 | | | 0,208 5 |
| | 26 | | | −0,292 | 0,386 6 | |
| 27 | | 0,337 5 | 0,662 5 | | | 0,250 1 |
| | 28 | | | 0,339 | 0,636 7 | |
| 16 | | 0,200 | 0,862 5 | | | 0,197 2 |
| | 30 | | | 0,970 | 0,833 9 | |
| 7 | | 0,087 5 | 0,950 | | | 0,099 4 |
| | 32 | | | 1,601 | 0,933 3 | |
| 2 | | 0,025 | 0,975 | | | 0,053 9 |
| | 34 | | | 2,232 | 0,987 2 | |
| 2 | | 0,025 | 1,0 | | | 0,012 8 |
| | 36 | | | 2,863 | 0,997 9 | |

**Notation -** le terme « variable centrée réduite » a été abrégé en var. c. r. Par ailleurs, $\Phi(x)$ désigne la fonction de répartition. Pour évaluer les moments, on a pris le milieu des intervalles de regroupement.

On trouve alors : $m = 26,92$ m, puis $D = 735$ m$^2$ et enfin $\sigma = 3,17$ m.

***b*** – Les deux premières cases de regroupement ainsi que les deux dernières ne possèdent pas assez de représentants, donc on va regrouper les deux premières cases en une seule d'une part, et les deux dernières en une seule case d'autre part.

***c*** – Le tableau I.8 donne les résultats des calculs.

**Tableau I.8.**

| $m_i$ | intervalles | $p_i^*$ | $p_i$ | $\lvert\Delta p_i\rvert\, 10^3$ | $p_i^*$ cumulés | $p_i$ cumulés | $\lvert\Delta\Phi\rvert 10^3$ |
|---|---|---|---|---|---|---|---|
| | 18 | | | | | | |
| 5 | | 0,062 5 | 0,060 1 | 2,4 | 0,062 5 | 0,060 1 | 2 |
| | 22 | | | | | | |
| 9 | | 0,112 5 | 0,118 | 5,5 | 0,175 | 0,178 1 | 3 |
| | 24 | | | | | | |
| 12 | | 0,15 | 0,208 5 | 58,5 | 0,325 | 0,386 6 | 62 |
| | 26 | | | | | | |
| 27 | | 0,337 5 | 0,250 1 | 87,4 | 0,662 5 | 0,636 7 | 26 |
| | 28 | | | | | | |
| 6 | | 0,200 | 0,197 2 | 2,8 | 0,862 5 | 0,833 9 | 28 |
| | 30 | | | | | | |
| 7 | | 0,087 5 | 0,099 4 | 11,9 | 0,95 | 0,933 3 | 17 |
| | 32 | | | | | | |
| 4 | | 0,050 | 0,066 7 | 16,7 | 1,0 | 1,0 | 0,0 |
| | 36 | | | | | | |

***d*** – $\chi^2_{\text{obs}} = n\sum_{i=1}^{K}(p_i - p_i^*)^2/p_i^* = 4{,}24$ où $K$ est le nombre de cases de regroupement, $K = 7$.

***e*** – Comme la loi normale impose deux contraintes auxquelles il faut ajouter la contrainte de normalisation, le nombre de degrés de liberté est donc $K - 3 = 4$. La consultation des tables du $\chi^2(4)$ donne, par interpolation linéaire, environ 37 chances sur 100 de pouvoir dépasser ce seuil. Il est clair que l'on conserve l'hypothèse d'une loi normale car les données ne contredisent pas cette hypothèse.

***f*** – Le test de Kolmogorov donne : $v = \sup|\Delta\Phi| = 0{,}062$. Donc $l = v\sqrt{n} = 0{,}555$. La probabilité que l'écart maximal entre les deux répartitions soit non inférieur à 0,555 est 92 chances sur 100. Ce test ne comportant pas de degré de liberté est notablement plus optimiste que le critère du $\chi^2$.

# 17. Étude des dépendances dans le cas linéaire

## 17.1. Test d'indépendance stochastique du tirage d'un échantillon

***Partie 1. a*** – La probabilité de tirer une sous-chaîne de longueur $\mu$ est : $P = (1/2)^\mu$, le théorème de référence est celui des probabilités composées.

***b*** – Calcul de la moyenne

$$m = \sum_{\mu=0}^{\infty} \mu \left(\frac{1}{2}\right)^{\mu} = 2.$$

Calcul de

$$D = \sum_{\mu=0}^{\infty} \mu^2 \left(\frac{1}{2}\right)^{\mu} = 6$$

d'où $\sigma^2 = 2$ et $\sigma = \sqrt{2}$.

***c*** – Longueur moyenne des chaînes : $\lambda_K = Km = 2K$.

***d*** – Soit $P$ la probabilité que $\nu_j \geq M_0$. Soit $Q = 1 - P$ la probabilité qu'il n'y ait aucune chaîne telle que $\nu_j \geq M_0$. La probabilité $p$ pour que $\nu_j \geq M_0$ est donnée par la relation :

$$p = \sum_{k=M_0}^{\infty} \left(\frac{1}{2}\right)^{k} = \left(\frac{1}{2}\right)^{M_0} \left(1 + \frac{1}{2} + \frac{1}{4} + \cdots\right) = \left(\frac{1}{2}\right)^{M_0 - 1},$$

en faisant usage du théorème des probabilités totales. La probabilité pour que $\nu_j$ soit inférieure à $M_0$ est donc : $q = 1 - p = 1 - (1/2)^{M_0 - 1}$. Comme les $K$ sous-chaînes sont indépendantes, la probabilité de ne pas réaliser au moins un événement est donc : $Q = (1 - p)^K$, d'où il ressort que :

$$P = 1 - Q = 1 - (1 - p)^K = 1 - \left[1 - \left(\frac{1}{2}\right)^{M_0 - 1}\right]^K.$$

Si $K$ est assez grand ou encore si $(1/2)^{M_0 - 1}$ est assez petit devant l'unité, on peut écrire que :

$$P = K \left(\frac{1}{2}\right)^{M_0 - 1}.$$

***Application numérique*** – $K = 10$ et $M_0 = 8$, on calcule alors : $P = 0{,}078$ par la formule approchée et $P = 0{,}075\,4$ par la formule rigoureuse.

***e*** – Avec la formule approchée : $P = 0{,}05$ et $Q = 0{,}95$, et :

$$0{,}95 \leq \left[1 - \left(\frac{1}{2}\right)^{M_0 - 1}\right]^K$$

ce qui entraîne que $K(1/2)^{M_0 - 1} \leq 0{,}05$, et comme $K = n/2$, on obtient :

$$M_0 < \frac{\log_e(20n)}{\log_e(2)}.$$

Avec la formule rigoureuse :

$$\log_e(0{,}95) \geq K \log_e \left[1 - \left(\frac{1}{2}\right)^{M_0 - 1}\right].$$

À partir de maintenant, pour obtenir $M_0$ il faut réaliser des approximations qui vont conduire au même résultat que celui précédemment obtenu.

***Application numérique*** – $n = 100$. On trouve alors $M_0 < 11$. Avec l'expression donnée dans le cours on a $M_0 = 10$.

***Partie 2. a*** – Comme $K$ est grand devant l'unité et que $\lambda_K = Km = 2K$, on en déduit que $\lambda_K$ est grand devant l'unité.

Nous avons : $\xi = \sum_{j=1}^{K} l_j$ les $l_j$ étant indépendants, $\Sigma^2 = K\sigma^2 = 2K$ d'où : $\Sigma = \sqrt{2K}$. Le théorème central limite nous dit que la distribution de $\xi$ est gaussienne ; à partir de là on peut écrire la distribution de $\xi$ :

$$f(x) = \frac{1}{\Sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\lambda_K)^2}{2\Sigma^2}\right],$$

d'où l'on déduit que :

$$P(\xi < \nu_0) = \frac{1}{\Sigma\sqrt{2\pi}} \int_{-\infty}^{\nu_0} \exp\left[-\frac{(x-\lambda_K)^2}{2\Sigma^2}\right] \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\nu_0-\lambda_K}{\Sigma}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}x,$$

puis encore :

$$P(\xi > \nu_0) = 0{,}05 = \frac{1}{\sqrt{2\pi}} \int_{\frac{\nu_0 - 2K_{\min}}{\sqrt{2K_{\min}}}}^{\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}x,$$

d'où il ressort :

$$\frac{\nu_0 - 2K_{\min}}{\sqrt{2K_{\min}}} \geq 1{,}65.$$

La résolution de l'équation du second degré donne une seule racine positive :

$$K_{\min} = \frac{\left(-1{,}65 + \sqrt{2{,}723 + 4\nu_0}\right)^2}{8},$$

par conséquent on doit avoir $K \geq K_{\min}$.

***Application numérique*** – $n = 100$. On trouve alors $K_{\min} = 42$. L'expression donnée dans le cours fournit le nombre 41. Soit $d$ la différence numérique entre les deux tests :

$$d = \left|\frac{1}{2}n + \frac{1}{2} - \frac{\alpha}{2}\sqrt{n-1} - \frac{1}{8}\left\{-\alpha + \sqrt{\alpha^2 + 4n}\right\}^2\right|,$$

avec $\alpha = 1{,}65$ en effectuant les développements usuels :

$$d = \left|\frac{1}{2}n + \frac{1}{2} - \frac{\alpha}{2}\sqrt{n} - \frac{1}{8}\left\{-\alpha + 2\sqrt{n}\right\}^2\right| = \left|\frac{1}{2}n + \frac{1}{2} - \frac{\alpha}{2}\sqrt{n} - \frac{n}{2}\left\{1 - \frac{\alpha}{\sqrt{n}}\right\}^2\right| = \frac{1}{2}.$$

***b*** – Chacun des deux tests a fait l'objet d'un calcul concernant l'exploitation d'une certaine distribution. Si les deux tests sont utilisés ensemble, il faut alors calculer la probabilité conditionnelle de la réalisation du second test sachant que le premier s'est réalisé. Les choses peuvent être corrigées dans le sens suivant : on sait, à présent, que toutes les chaînes ont une longueur inférieure à $M_0$. Approximativement, la moyenne est donnée par l'expression :

$$m' = \sum_{k=1}^{M_0-1} k\left(\frac{1}{2}\right)^k \approx 2 - \frac{M_0}{2^{M_0-1}}\cdot$$

Pour $M_0 \geq 16$, l'erreur est de l'ordre de $5\,10^{-4}$ par rapport à la solution asymptotique. Par ailleurs, on calcule le moment du second ordre :

$$D' = 2 + \frac{4M_0 - 3M_0^2}{2^{M_0-1}} - \frac{M_0^2}{2^{2M_0-2}}\cdot$$

À partir de là, on retrouve le théorème central limite, et l'on peut effectuer les développements traditionnels.

## 17.2. Loi F(m, l). Étude du rapport de deux variances : loi de Fisher-Snedecor

***1.*** – Distribution de $\xi^2 = 1/(m-1)\sum_{i=1}^{m}\xi_i^2$, chaque $\xi_i^2$ indépendant obéit à la distribution :

$$p_0(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$

chaque $\xi_i^2$ obéit à la distribution :

$$p(y) = \frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{y}{2}\right)$$

si $y \geq 0$ et $p(y) = 0$ si $y < 0$. Calculons la fonction caractéristique de $\xi_i^2$ :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}}\int_0^{+\infty} p_k(x)\exp(jtx)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_0^{+\infty}\frac{\mathrm{d}x}{\sqrt{x}\exp\left[(1-2jt)\frac{x}{2}\right]}$$

$$= \frac{1}{\sqrt{\pi(1-2jt)}}\int_0^{+\infty} x^{-1/2}\exp(-x)\,\mathrm{d}x = \Gamma(1/2)\frac{1}{\sqrt{\pi(1-2jt)}} = (1-2jt)^{-1/2}.$$

Il s'ensuit que la fonction caractéristique de la variable aléatoire $(m-1)\xi^2$ s'écrit :

$$\Phi(t) = (1-2jt)^{-m/2} \quad \text{pour } t \in (-\infty, +\infty).$$

Pour obtenir la fonction de distribution, il n'est pas utile de calculer la transformée de Fourier inverse, pour cela il suffit de remarquer que :

$$A\int_0^{+\infty} x^{(m/2)-1}\exp\left(-\frac{x}{2}\right)\exp(jtx)\,\mathrm{d}x = (1-2jt)^{-m/2},$$

$A$ étant un coefficient de normalisation. Donc la densité de probabilité de la variable aléatoire $(m-1)\xi^2$ s'écrit :

$$p(x) = Ax^{(m/2)-1} \exp\left(-\frac{x}{2}\right) \quad \text{pour } x \in (0, +\infty),$$

avec

$$A = \frac{1}{2^{m/2}\Gamma(m/2)} \cdot$$

d'où l'expression :

$$p(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{(m/2)-1} \exp\left(-\frac{x}{2}\right) \cdot$$

Nous avons donc :

$$\frac{1}{2^{m/2}\Gamma(m/2)} x^{(m/2)-1} \exp\left(-\frac{x}{2}\right) \quad \xrightarrow{TF} \quad (1-2jt)^{-m/2}.$$

La fonction caractéristique $\psi(t)$ de la variable aléatoire $\xi^2$ s'exprime en fonction de celle de $(m-1)\xi^2$, c'est-à-dire :

$$\psi(t) = \Phi\left(\frac{t}{m-1}\right) = \left(1 - 2j\frac{t}{m-1}\right)^{-m/2} \quad \text{pour } t \in (-\infty, +\infty).$$

Il suffit d'appliquer le théorème de dilatation d'abscisse sur les transformées de Fourier pour obtenir la densité de probabilité $p(z)$ de la variable $\xi^2$ :

$$\begin{aligned} f(x) &\xrightarrow{TF} \varphi(t) \\ af(ax) &\xrightarrow{TF} \varphi\left(\frac{t}{|a|}\right) \end{aligned}$$

avec $a = (m-1)$, ce qui donne :

$$p_\xi(z, m) = \frac{(m-1)^{m/2}}{2^{m/2}\Gamma(m/2)} z^{(m/2)-1} \exp\left(-\frac{(m-1)z}{2}\right) \cdot$$

**2. *a*** – La loi de distribution de $\eta$ est immédiate :

$$p_\eta(z, n) = \frac{(n-1)^{n/2}}{2^{n/2}\Gamma(n/2)} z^{(n/2)-1} \exp\left(-\frac{(n-1)z}{2}\right) \cdot$$

***b*** – La distribution de $w = u/v$ qui est le rapport de deux variables positives et indépendantes $u$ et $v$ de distribution $p_u(u)$ et $p_v(v)$ est donnée par l'expression établie en cours :

$$p_z(z) = \int_0^\infty p_u(xz) p_x(x) x \, \mathrm{d}x.$$

***c*** –

$$p_z(z) = B \int_0^\infty (zx)^{(m/2)-1} \exp\left(-\frac{(m-1)zx}{2}\right) x^{(n/2)-1} \exp\left(-\frac{(n-1)x}{2}\right) x \, \mathrm{d}x$$

avec

$$B = \frac{(n-1)^{n/2}}{2^{n/2}\Gamma(n/2)} \cdot \frac{(m-1)^{m/2}}{2^{m/2}\Gamma(m/2)} z^{(m/2)-1}.$$

En poursuivant les calculs :

$$p_z(z) = Bz^{(m/2)-1} \int_0^\infty x^{(m/2)-1} \exp\left(-\frac{(m-1)zx}{2}\right) x^{(n/2)-1} \exp\left(-\frac{(n-1)x}{2}\right) x \, \mathrm{d}x$$

$$= Bz^{(m/2)-1} \int_0^\infty x^{m/2+(n/2)-1} \exp\left\{-\frac{x}{2}[n-1+(m-1)z]\right\} \mathrm{d}x,$$

alors, en posant $r = (x/2)[n-1+(m-1)z]$, on obtient

$$x = \frac{2r}{n-1+(m-1)z} \qquad \text{et} \qquad \mathrm{d}x = \frac{2\,\mathrm{d}r}{n-1+(m-1)z}$$

puis :

$$p_z(z) = Bz^{(m/2)-1} \int_0^\infty \left[\frac{2r}{n-1+(m-1)z}\right]^{m/2+(n/2)-1} \exp(-r) \frac{2\,\mathrm{d}r}{n-1+(m-1)z}$$

$$= Bz^{(m/2)-1} [(n-1)+(m-1)z]^{-\frac{(m+n)}{2}} (2)^{\frac{(m+n)}{2}} \int_0^\infty r^{\frac{(m+n)}{2}-1} \exp(-r)\, \mathrm{d}r$$

soit encore :

$$p_z(z) = Bz^{(m/2)-1}[n-1+(m-1)z]^{\frac{(m+n)}{2}} (2)^{(m+n)/2} \Gamma\left(\frac{m+n}{2}\right).$$

Réécrivons tous les termes :

$$p_z(z) = \frac{(n-1)^{n/2}}{\Gamma(n/2)} \cdot \frac{(m-1)^{m/2}}{\Gamma(m/2)} z^{(m/2)-1} \left[(n-1)\left(1+\frac{m-1}{n-1}z\right)\right]^{-\frac{(m+n)}{2}} \Gamma\left(\frac{m+n}{2}\right).$$

Nous allons encore nous livrer à quelques petites manœuvres en posant $y = (m-1)/(n-1)z$, mais attention aux changements de variables dans les lois de distribution car $\mathrm{d}y/\,\mathrm{d}z = (m-1)/(n-1)$, en définitive, on trouve :

$$p_z(y) = \frac{(n-1)^{n/2}(m-1)^{m/2}}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n-1}{m-1}\right)^{(m/2)-1}$$

$$\times y^{(m/2)-1}(n-1)^{-\frac{(m+n)}{2}}(1+y)^{-\frac{(m+n)}{2}} \Gamma\left(\frac{m+n}{2}\right)\left(\frac{n-1}{m-1}\right),$$

soit encore :

$$p_z(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(n/2)\Gamma(m/2)} y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}}.$$

***d*** – Calculons la moyenne de $y$ que l'on note $\langle y \rangle$ :

$$\langle y \rangle = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(n/2)\Gamma(m/2)} \int_0^\infty y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}} y \, \mathrm{d}y = B\left(\frac{m}{2}+1, \frac{n}{2}-1\right) = \frac{m}{n-2} \cdot$$

***e*** – Calcul de

$$D(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(n/2)\Gamma(m/2)} \int_0^\infty y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}} y^2 \, \mathrm{d}y = B\left(\frac{m}{2}+2, \frac{n}{2}-2\right) = \frac{m(m+2)}{(n-2)(n-4)},$$

d'où l'on tire

$$\sigma^2 = \frac{2m(m+n-2)}{(n-2)^2(n-4)} \cdot$$

***f*** – Le nombre de degrés de liberté : chaque variable a $m$ et $n$ degrés de liberté et il faut dénombrer les contraintes ; il y en a une par variable pour le calcul de $\sigma^2$. Donc la loi est à $(m-1, n-1)$ degrés de liberté.

***g*** – Montrons que $F_{1-\alpha}(m,n) = \dfrac{1}{F_\alpha(n,m)}$ (attention $m$ et $n$ ne jouent pas des rôles symétriques) :

$$\alpha = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(n/2)\Gamma(m/2)} \int_0^{B_{\min}} y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}} \, \mathrm{d}y = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(n/2)\Gamma(m/2)} \int_{B_{\max}}^\infty y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}} \, \mathrm{d}y,$$

après avoir abandonné le coefficient devant le signe somme qui n'est d'aucun intérêt ; effectuons un changement de variable dans la seconde intégrale à savoir $y = \dfrac{1}{x}$, on trouve :

$$\alpha' = \int_{B_{\max}}^\infty y^{\frac{m}{2}-1}(1+y)^{-\frac{(m+n)}{2}} \, \mathrm{d}y = \int_0^{\frac{1}{B_{\max}}} x^{\frac{m}{2}-1}(1+x)^{-\frac{(m+n)}{2}} \, \mathrm{d}x,$$

$B_{\min}B_{\max} = 1$, ou encore : $1/B_{\max} = F_\alpha(n,m)$ mais par ailleurs :

$$B_{\min} = F_\alpha(m,n) \quad \text{et} \quad B_{\max} = F_{1-\alpha}(m,n)$$

et par conséquent :

$$\frac{1}{B_{\max}} = F_\alpha(n,m) = \frac{1}{F_{1-\alpha}(m,n)} \cdot$$

# 18. Analyse de régression-corrélation

## 18.1. La distribution gaussienne à deux dimensions. Coefficient de corrélation

***a*** – On a évidemment : $x = \dfrac{\xi - m_x}{\sigma_x}$ et $y = \dfrac{\xi - m_y}{\sigma_y}$.

***b*** – On doit avoir

$$\int_{-\infty}^{+\infty}\!\!\int \Phi(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1 = A\int_{-\infty}^{+\infty}\!\!\int \exp\left(-\frac{x^2 - 2\alpha xy + y^2}{\beta^2}\right)\mathrm{d}x\,\mathrm{d}y,$$

pour calculer cette intégrale, on effectue un changement de variables :

$$\beta s = x - \alpha y \qquad \text{et} \qquad \beta r = y\sqrt{1-\alpha^2},$$

qui conduit au jacobien :

$$J = \frac{\partial(x,y)}{\partial(s,r)} = \frac{\beta^2}{\sqrt{1-\alpha^2}}\cdot$$

Ainsi, on obtient :

$$1 = A\int_{-\infty}^{+\infty}\!\!\int \exp(-s^2)\exp(-r^2)\frac{\beta^2}{\sqrt{1-\alpha^2}}\,\mathrm{d}r\,\mathrm{d}s = A\frac{\beta^2}{\sqrt{1-\alpha^2}}\pi\cdot$$

***c*** – La fonction de répartition s'écrit simplement :

$$\Psi(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y}\varphi(u,v)\,\mathrm{d}u\,\mathrm{d}v$$

où

$$\varphi(u,v) = \frac{\sqrt{1-\alpha^2}}{\beta^2\pi}\exp\left[-\frac{1}{\beta^2}(x-2\alpha y)^2\right]\exp\left[-\frac{1}{\beta^2}y^2(1-\alpha^2)\right],$$

expression à partir de laquelle on tire :

$$\begin{aligned}\Psi_x(x) &= \int_{-\infty}^{+\infty}\varphi(x,v)\,\mathrm{d}v = A\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{\beta^2}(x-2\alpha y)^2\right]\exp\left[-\frac{1}{\beta^2}y^2(1-\alpha^2)\right]\mathrm{d}y\\ &= \frac{\sqrt{1-\alpha^2}}{\beta\sqrt{\pi}}\exp\left[-\frac{1}{\beta^2}x^2(1-\alpha^2)\right].\end{aligned}$$

***d*** – Pour que $\Psi_x(x)$ soit la loi normale réduite, il faut que $\beta^2 = 2(1-\alpha^2)$. Comme $x$ et $y$ jouent des rôles symétriques, forcément, $y$ obéit à la loi normale réduite si $x$ obéit à la loi normale réduite.

***e*** – Calcul de la covariance :

$$K_{xy} = \frac{1}{2\pi}\frac{1}{\sqrt{1-\beta^2}}\iint_{-\infty}^{+\infty} xy \exp\left(-\frac{x^2 - 2\alpha xy + y^2}{2(1-\alpha^2)}\right) \mathrm{d}x\, \mathrm{d}y,$$

et on effectue un changement de variable identique à celui du § b seulement $\beta$ est remplacé par sa valeur du § d. On établit que : $K_{xy} = \alpha$, c'est la covariance de $x$ et $y$ mais aussi le coefficient de corrélation car les variables sont centrées et réduites ce qui impose que les écarts quadratiques sont égaux à 1.

Dans le cas où $\alpha = 0$, $\Psi(x, y) = \Psi_x(x)\Psi_y(y)$ et dans le cas de la loi normale, l'absence de corrélation entraîne l'indépendance des variables ce qui était affirmé en cours, mais non démontré.

***f*** – L'expression de la droite de régression de $y$ en $x$ s'écrit : $y = \alpha x$ car il s'agit de variables centrées.

***g*** – L'équation caractéristique associée aux variables $x$ et $y$ est la transformée de Fourier à deux dimensions de la fonction de distribution $\varphi(u, v)$ c'est-à-dire :

$$\chi(u, v) = \iint_{-\infty}^{+\infty} \exp\left(-\frac{x^2 - 2\alpha xy + y^2}{\beta^2}\right) \exp[2\pi j(xu + yv)]\, \mathrm{d}x\, \mathrm{d}y,$$

et pour effectuer l'intégration, on effectue les changements de variables suivants :

$$r\sqrt{\pi} = \frac{x - \alpha y}{\beta} \quad \text{et} \quad s\sqrt{\pi} = \frac{y}{\sqrt{2}},$$

$$\text{soit :} \quad \begin{aligned} x &= \beta r\sqrt{\pi} + \alpha s\sqrt{2\pi} \\ y &= s\sqrt{2\pi}, \end{aligned}$$

transformations qui conduisent au jacobien (compte tenu du fait que $\beta^2 = 2(1 - \alpha^2)$) :

$$J = \frac{\partial(x, y)}{\partial(r, s)} = 2\pi\sqrt{1-\alpha^2}.$$

Ainsi, on obtient :

$$\chi(u, v) = A^* \int_{-\infty}^{+\infty} \exp(-\pi r^2) \exp[2\pi j(u\beta r\sqrt{\pi})]\, \mathrm{d}r \int_{-\infty}^{+\infty} \exp(-\pi s^2) \exp[2\pi j(\alpha u + v)s\sqrt{2\pi})]\, \mathrm{d}s.$$

enfin :

$$\chi(u, v) = \exp[-2\pi^2(u^2 + 2\alpha uv + v^2)].$$

## 18.2. Étude d'un changement de phase. Analyse de régression

***a*** – L'équation de la droite de régression $S_1$ : $Y_1(x) = \bar{a}_1 + \bar{b}_1(x - \bar{x}_1)$ où :

$$\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_i \qquad \bar{a}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} y_i \qquad \bar{b}_1 = \frac{\sum_{i=1}^{N_1} x_i y_i - N_1 \bar{x}_1 \bar{y}_1}{N_1 \sigma_{x_1}^2},$$

avec

$$\sigma_{x_1}^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2 \cdot$$

Par ailleurs on connaît $x_0$, donc on calcule $Y_1(x_0) = \bar{a}_1 + \bar{b}_1(x_0 - \bar{x}_1)$ ; ensuite, on en déduit l'intervalle de confiance sur cette valeur estimée :

$$|\Delta Y_1(x_0)| \leq t_{\alpha/2}(N_1 - 2)\frac{S}{\sqrt{N_1}}\sqrt{1 + \frac{(x_0 - \bar{x}_1)^2}{\sigma_{x_1}^2}},$$

avec

$$S^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}[y_i - Y_1(x_i)]^2,$$

ici, comme d'habitude, $t_{\alpha/2}(N_1 - 2)$ est la distribution de Student à $(N_1 - 2)$ degrés de liberté avec le niveau de confiance $(1 - \alpha)$. Attention aux valeurs absolues.

***b*** – Nous avons le même résultat pour la droite $S_2$ : $Y_2(x) = \bar{a}_2 + \bar{b}_2(x - \bar{x}_2)$ pour cela, il suffit de remplacer l'indice 1 par l'indice 2. En définitive, on obtient :

$$|\Delta H| = |\Delta Y_2(x_0) + \Delta Y_1(x_0)|.$$

## 18.3. Mesure d'une température par spectroscopie

***a*** – Posons $Z_J = \log_e [I_J/(2J + 1)]$ ainsi : $Z_J = A - B[J(J + 1)/T]$. La pente de la droite de régression est donc $\beta = -B/T$ et par ailleurs elle est donnée par le calcul statistique :

$$\beta = \frac{\sum_{i=1}^{n} X_i Z_i - n\bar{X}\bar{Z}}{n\sigma_X^2}, \quad \text{avec } X_i = i(i + 1)$$

et

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

***b*** – On a $T = -B/\beta$ dont on déduit : $|\Delta T| = \frac{B}{\beta^2}|\Delta\beta| = \frac{T}{|\beta|}|\Delta\beta|$ avec :

$$|\Delta\beta| = t_{\alpha/2}(n - 2)\frac{S}{\sigma_X\sqrt{n}},$$

où

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}[Z_i - Y(x_i)]^2.$$

### 18.4. Solubilité du nitrate de sodium dans l'eau

***a*** – On écrit l'équation de la droite de régression : $Y(x) = \bar{a} + \bar{b}(x - \bar{x})$ où :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{a} = \frac{1}{N_1}\sum_{i=1}^{n} y_i \qquad \bar{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n\sigma_x^2} = \bar{r}\frac{\sigma_y}{\sigma_x},$$

avec $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ et $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ et aussi :

$$\bar{r} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n\sigma_x\sigma_y},$$

$\bar{r}$ étant le coefficient de corrélation. On trouve les résultats numériques suivants :

$$\bar{x} = 26, \quad \bar{a} = \bar{y} = 90{,}14, \quad \sigma_x = 21{,}24, \quad \sigma_y = 18{,}51 \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 2\,736{,}5$$

ensuite on calcule : $\bar{b} = 0{,}871$, $\bar{r} = 0{,}999$. Il n'y a aucune raison de rejeter une liaison de corrélation puisqu'il y a quasiment une dépendance linéaire. La droite de régression s'écrit donc : $Y(x) = 90{,}14 + 0{,}871(x - 26) = 67{,}5 + 0{,}871x$.

Calculons à présent $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - Y(x_i)]^2 = 0{,}805$ d'où $S = 0{,}805$.

***b*** – Pour $x_0 = 40$, voici les résultats trouvés : $y = 102{,}34$ avec une incertitude :

$$|\Delta y| = \pm t_{\alpha/2}(7)\frac{S}{\sqrt{n}}\sqrt{1 + \frac{(x_0 - \bar{x})^2}{\sigma_x^2}} = \pm 0{,}68 \quad \text{avec} \quad \alpha = 0{,}10.$$

Autrement dit, on a 90 chances sur 100 de tomber dans l'intervalle $|\Delta y|$.

## 19. Les fractions continues

### 19.1. Calcul d'une fonction donnée sous forme d'une fraction continue

***a*** – D'abord on calcule : $P_1/Q_1 = A_0$, puis $P_2/Q_2 = [A_0A_1 + (x - a_0)]/A_1$, et enfin :

$$\frac{P_3}{Q_3} = A_0 + \frac{(x - a_0)A_2}{A_1A_2 + (x - a_1)} = \frac{A_0A_1A_2 + A_0(x - a_1) + A_2(x - a_0)}{A_1A_2 + (x - a_1)}.$$

À partir de cette dernière expression, la démonstration des relations cherchées s'effectue par récurrence ; les relations sont donc vraies pour $m = 3$, montrons qu'elles sont vraies pour $m + 1$. Il suffit de remplacer $A_{m-1}$ par $A_{m-1} + (x - a_{m-1})/A_m$, dans les relations :

$$\begin{aligned} P_m &= A_{m-1}P_{m-1} + (x - a_{m-2})P_{m-2} \\ Q_m &= A_{m-1}Q_{m-1} + (x - a_{m-2})Q_{m-2} \end{aligned}$$

mais il est commode de former le rapport $P_{m+1}/Q_{m+1}$ :

$$\begin{aligned} \frac{P_{m+1}}{Q_{m+1}} &= \frac{A_{m-1}A_m + (x - a_{m-1})P_{m-1} + (x - a_{m-2})P_{m-2}A_m}{A_{m-1}A_m + (x - a_{m-1})Q_{m-1} + (x - a_{m-2})Q_{m-2}A_m} \\ &= \frac{A_m[A_{m-1}P_{m-1} + (x - a_{m-2})P_{m-2}] + (x - a_{m-1})P_{m-1}}{A_m[A_{m-1}Q_{m-1} + (x - a_{m-2})Q_{m-2}] + (x - a_{m-1})Q_{m-1}} \\ &= \frac{A_mP_m + (x - a_{m-1})P_{m-1}}{A_mQ_m + (x - a_{m-1})Q_{m-1}}. \end{aligned}$$

***b*** – Nous avons : $f(x) = f(x_0) + [(x - x_0)/1!]f'(x_0)$ au premier ordre donc :

$$A_0 = f(x_0) \quad \text{et} \quad A_1 = \frac{1}{f'(x_0)}\cdot$$

***c*** – Utilisons la relation (que l'on ne demande pas de démontrer) :

$$A_{k+1} = \frac{(k+1)A_{k-1}}{(k-1) + A_{k-1}\frac{\mathrm{d}A_k}{\mathrm{d}x_0}},$$

on calcule :

$$A_2 = \frac{2}{\frac{\mathrm{d}A_1}{\mathrm{d}x_0}} = -2\frac{f''}{f'^2},$$

puis :

$$A_3 = \frac{3f''^2}{-3f'f''^2 + 2f'^2f'''} = \frac{c_2^2}{c_1[c_2^2 - c_1c_3]}\cdot$$

***d*** – Étude de la fonction $y = \log_e(1+x)$. On écrit : $\mathrm{d}f/\,\mathrm{d}x = 1/(1+x_0)$, puis :

$$\frac{\mathrm{d}^2f}{\mathrm{d}x^2} = -\frac{1}{(1+x_0)^2} \quad \text{et} \quad \frac{\mathrm{d}^3f}{\mathrm{d}x^3} = \frac{2}{(1+x_0)^3},$$

et d'une façon générale :

$$\frac{\mathrm{d}^nf}{\mathrm{d}x^n} = \frac{(n-1)!}{(1+x_0)^n}\cdot$$

Ensuite, on calcule : $A_0 = \log_e(1+x_0), A_1 = (1+x_0), A_2 = 2, A_3 = 3(1+x_0)$.

Pour calculer $\log_e(x)$, il suffit de faire $x_0 = 0$. Les relations de récurrence donnent : $P_0 = 1, Q_0 = 0, P_1 = A_0$ et $Q_1 = 1$. En définitive, on obtient :

$$\log_e(1+x) \approx \frac{x(9+x)}{9+4x},$$

et pour $x = 1$, on trouve $\log_e(2) \approx 0{,}769$.

***e*** – Le développement de $\log_e(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4}$ (au même ordre que pour la question précédente) donne $\log_e(2) \approx 0{,}583$.

L'erreur absolue sur le résultat donné par la formule de MacLaurin est 0,2 tandis qu'elle est de 0,08 sur le résultat donné par la fraction continue.

# 20. Éléments de traitement du signal

## 20.1. Détermination d'une constante de temps

***a*** – On a $V(t) = A\exp(-\alpha t)$ avec $\alpha = \dfrac{1}{RC}\cdot$

***Approche n° 1*** – On note les données $(V_i, t_i)$ pour $i = 0, 1, 2, \ldots, N$. On prend le logarithme de l'expression précédente, ce qui permet de linéariser le problème :

$$\log_e(V_i) = \xi_i = \log_e(A) - \alpha t_i = a - \alpha t_i,$$

$a$ et $\alpha$ sont les inconnues du système linéarisé. $\alpha$ est donc déterminé par la pente de la droite de régression de $\xi$ en $t$, à savoir :

$$\alpha = \frac{(N+1)\bar{\xi}\bar{t} - \sum_{i=0}^{N} \xi_i t_i}{(N+1)s_t^2},$$

où $\bar{\xi}$ et $\bar{t}$ sont respectivement la moyenne des $\xi$ et des $t$, et $s_t^2$ est l'écart quadratique moyen des $t$. On trouve :

$$\bar{\xi} = 3{,}050, \quad \bar{t} = 5 \quad \text{et} \quad \sum_{i=0}^{N} \xi_i t_i = 133{,}35 \quad \text{et enfin} \quad s_t^2 = 10{,}0,$$

on trouve que $\alpha = -0{,}312\,6$ et que $z = s/(s_t\sqrt{N+1})$ avec $s^2 = (1/N)[\xi_i - \hat{Y}(t_i)]^2$, puis on obtient : $s = 0{,}010\,42$ et $z = 0{,}014\,05$. Enfin, pour terminer, on trouve dans les tables de Student : $t_{0,15}(9) = 1{,}156$ après avoir effectué une interpolation linéaire. De là on déduit la valeur numérique de $\delta\alpha = zt_{0,15}(9) = \pm 0{,}016\,2$. Autrement dit, nous avons 70 chances sur 100 de trouver $\alpha$ dans l'intervalle :

$$-0{,}328\,84 < \alpha < -0{,}296\,44.$$

***Approche n° 2*** – Nous avons $V = A + B\exp(\beta t)$ et il nous faut calculer $A$, $B$ et $\beta$. On sait que les abscisses $t$ sont en progression arithmétique de raison $\Delta t$ et de premier terme $t_0$, ainsi, les équations discrétisées s'écrivent :

$$V_i = A + B\exp(\beta t_i) = A + B\exp[\beta(t_0 + i\Delta t)]$$

en posant $u = \exp(\beta\Delta t)$ et $p_i = A + B\exp[\beta(t_0 + i\Delta t)]$ on obtient les équations :

$$\begin{aligned} V_i &= A + Bp_i \\ V_{i+1} &= A + Bp_i u \\ V_{i+2} &= A + Bp_i u^2 \\ V_{i+3} &= A + Bp_i u^3 \quad \text{etc.} \end{aligned}$$

À présent, on note $\Delta_i$ les différences premières de $V_i$ à savoir :

$$\Delta_i = \Delta b_i = b_{i+1} - b_i \quad \text{pour} \quad i = 0, 1, 2, \ldots, n-1,$$

ce qui nous permet d'écrire :

$$\begin{aligned} \Delta_i &= Bp_i(u-1) \\ \Delta_{i+1} &= Bp_i(u-1)u \end{aligned}$$

équations entre lesquelles on élimine $Bp_i$ ce qui donne : $u\Delta_i = \Delta_{i+1}$, on a $(n-2)$ équations de ce type où $i$ prend les valeurs $0, 1, 2, \ldots, n-2$. À présent, on détermine $u$ par la méthode des moindres carrés en posant $\varepsilon_i = u\Delta_i - \Delta_{i+1}$, et l'on minimise la somme des carrés des résidus :

$$E^2 = \sum_{i=0}^{n-2} \varepsilon_i^2 = \sum_{i=0}^{n-2} [u\Delta_i - \Delta_{i+1}]^2,$$

en la dérivant par rapport à $u$, soit :

$$\frac{\mathrm{d}}{\mathrm{d}u}\sum_{i=0}^{n-2}[u\Delta_i - \Delta_{i+1}]^2 = 0,$$

ce qui donne :

$$u = \frac{\displaystyle\sum_{i=0}^{n-2}\Delta_i\Delta_{i+1}}{\displaystyle\sum_{i=0}^{n-2}\Delta_i^2}.$$

Comme $u = \exp(\beta\Delta t)$, on tire immédiatement : $\beta = \frac{1}{\Delta t}\log_e(u)$. À présent, on obtient $A$ et $B$ en écrivant $n+1$ équations (incompatibles) : $V_i = A + Bp_i$ auxquelles on applique la méthode des moindres carrés. On pose :

$$\varepsilon'_i = A + Bp_i - V_i$$

$$\text{et} \quad E'^2 = \sum_{i=0}^{n}\varepsilon'^2_i = \sum_{i=0}^{n}[A + Bp_i - V_i]^2.$$

On minimise cette dernière expression par rapport à $A$ et $B$ ce qui nous donne un système linéaire de deux équations à deux inconnues (*cf.* la droite de régression), on obtient en définitive :

$$A = \frac{\displaystyle\sum_{i=0}^{n}V_i\sum_{i=0}^{n}p_i^2 - \sum_{i=0}^{n}V_ip_i\sum_{i=0}^{n}p_i}{\displaystyle(n+1)\sum_{i=0}^{n}p_i^2 - \sum_{i=0}^{n}p_i\sum_{i=0}^{n}p_i},$$

$$B = \frac{\displaystyle(n+1)\sum_{i=0}^{n}V_ip_i - \sum_{i=0}^{n}V_i\sum_{i=0}^{n}p_i}{\displaystyle(n+1)\sum_{i=0}^{n}p_i^2 - \sum_{i=0}^{n}p_i\sum_{i=0}^{n}p_i}.$$

***Application numérique*** – On trouve : $b = -0{,}304\,2$ puis $A = -0{,}756\,4$ et $B = 101{,}5$.

Le module de la différence entre $\alpha$ et $\beta$, noté $\delta\alpha$, est 0,008 4. La valeur correspondante de $t_{\alpha/2}(9)$ s'obtient à partir des calculs de la première partie :

$$t_{\gamma/2}(9) = \frac{\delta\alpha}{z} = 0{,}6.$$

La consultation de la table de Student donne environ $\gamma/2 = 0{,}25$, soit de l'ordre de 50 chances sur 100 de tomber dans l'intervalle $\delta\alpha$. On notera que cet intervalle est environ deux fois plus petit que le précédent calculé dans la première partie pour lequel correspondait une probabilité de 70 chances sur 100, c'est donc tout à fait normal de trouver une probabilité plus faible (50 chances sur 100) de tomber dans $\delta\alpha$.

# Index