Dauphine | PSL
UNIVERSITÉ PARIS

Méthodes numériques

Introduction à l'analyse numérique et au calcul scientifique

Guillaume Legendre

(version provisoire du 7 janvier 2025)

Au regretté logo de l'université (2009-2019)

Avant-propos

Ce document est une version augmentée et regroupée des notes de deux cours enseignés à l'université Paris-Dauphine, respectivement en deuxième année de licence de Mathématiques et Informatique appliquées à l'Économie et à l'Entreprise (MI2E) et en première année de master de Mathématiques Appliquées. Ces enseignements se composent à la fois de cours magistraux et de séances de travaux dirigés et de travaux pratiques.

Leur but est de présenter plusieurs méthodes numériques de base utilisées pour la résolution des systèmes linéaires, des équations non linéaires, des équations différentielles et aux dérivées partielles, pour le calcul numérique d'intégrales ou encore pour l'approximation de fonctions par interpolation polynomiale, ainsi que d'introduire aux étudiants les techniques d'analyse (théorique) de ces dernières. Certains aspects pratiques de mise en œuvre sont également évoqués et l'emploi des méthodes est motivé par des problèmes « concrets ». La présentation et l'analyse des méthodes se trouvent complétées par un travail d'implémentation et d'application réalisé par les étudiants avec les logiciels MATLAB® [1] et GNU OCTAVE [2].

Il est à noter que ce support de cours comporte plusieurs passages qui ne sont pas traités dans le cours devant les étudiants (ce dernier fixant le programme de l'examen), ou tout au moins pas de manière aussi détaillée. Il contient également deux annexes de taille relativement conséquente, l'une consacrée à des rappels d'algèbre, l'autre à des rappels d'analyse, qui constituent les pré-requis à une bonne compréhension des deux premières parties du cours. Les courtes notes biographiques, qui apparaissent en bas de page à chaque première fois que le nom d'une personne est cité, sont pour partie tirées de WIKIPEDIA [3].

Je tiens enfin à adresser mes remerciements à tous les étudiants ayant décelé des erreurs, à Matthieu Hillairet pour sa relecture d'une partie du manuscrit et ses remarques, à André Casadevall, Djalil Chafaï, Maxime Chupin, Olga Mula, Pierre Lissy, Nicolas Salles, Julien Salomon, Anders Thorin et Gabriel Turinici pour leurs suggestions et enfin à Donald Knuth pour l'invention de TEX et à Leslie Lamport pour celle de LATEX.

Guillaume Legendre
Paris, janvier 2018.

Quelques références bibliographiques

Pour approfondir les thèmes abordés dans ces pages, voici une sélection de plusieurs ouvrages de référence, que l'on pourra consulter avec intérêt en complément du cours. Par ailleurs, afin de faciliter l'accès à la littérature de langue anglaise, des traductions des termes spécifiques à ce cours sont proposées tout au long du manuscrit, généralement lors de l'introduction de l'objet ou de la notion en question.

Ouvrages rédigés en français

[AD08] L. AMODEI et J.-P. DEDIEU. *Analyse numérique matricielle*, *Mathématiques pour le master/SMAI*. Dunod, 2008.

[AK02] G. ALLAIRE et S. M. KABER. *Algèbre linéaire numérique*, *Mathématiques pour le deuxième cycle*. Ellipses, 2002.

[Cia98] P. G. CIARLET. *Introduction à l'analyse numérique matricielle et à l'optimisation – cours et exercices corrigés*, *Mathématiques appliquées pour la maîtrise*. Dunod, 1998.

[Dem06] J.-P. DEMAILLY. *Analyse numérique et équations différentielles*, *Grenoble Sciences*. EDP Sciences, 2006.

1. MATLAB est une marque déposée de The MathWorks, Inc., `http://www.mathworks.com/`.
2. GNU OCTAVE est distribué sous licence GNU GPL, `http://www.gnu.org/software/octave/`.
3. WIKIPEDIA, *the free encyclopedia*, `http://www.wikipedia.org/`.

[Fil09] F. FILBET. *Analyse numérique – Algorithme et étude mathématique*. Dunod, 2009.

[LT00a] P. LASCAUX et R. THÉODOR. *Analyse numérique matricielle appliquée à l'art de l'ingénieur. 1. Méthodes directes*. Dunod, 2000.

[LT00b] P. LASCAUX et R. THÉODOR. *Analyse numérique matricielle appliquée à l'art de l'ingénieur. 2. Méthodes itératives*. Dunod, 2000.

[QSS07] A. QUARTERONI, R. SACCO et F. SALERI. *Méthodes numériques. Algorithmes, analyse et applications*. Springer, 2007. DOI : 10.1007/978-88-470-0496-2.

Ouvrages rédigés en anglais

[Act90] F. S. ACTON. *Numerical methods that (usually) work*. The Mathematical Association of America, 1990.

[Atk89] K. ATKINSON. *An introduction to numerical analysis*. John Wiley & Sons, second edition, 1989.

[Axe94] O. AXELSSON. *Iterative solution methods*. Cambridge University Press, 1994. DOI: 10.1017/CBO9780511624100.

[CLR+09] T. H. CORMEN, C. E. LEISERSON, R. L. RIVEST, and C. STEIN. *Introduction to algorithms*. MIT Press, third edition, 2009.

[DB08] G. DAHLQUIST and Å. BJÖRCK. *Numerical methods in scientific computing. Volume I*. SIAM, 2008. DOI: 10.1137/1.9780898717785.

[Gau12] W. GAUTSCHI. *Numerical analysis*. Birkhäuser Basel, second edition, 2012. DOI: 10.1007/978-0-8176-8259-0.

[GGK14] W. GANDER, M. J. GANDER, and F. KWOK. *Scientific computing. An introduction using Maple and MATLAB*, volume 11 of *Texts in computational science and engineering*. Springer, 2014. DOI: 10.1007/978-3-319-04325-8.

[GV13] G. H. GOLUB and C. F. VAN LOAN. *Matrix computations*. Johns Hopkins University Press, fourth edition, 2013.

[Hig02] N. J. HIGHAM. *Accuracy and stability of numerical algorithms*. SIAM, second edition, 2002. DOI: 10.1137/1.9780898718027.

[IK94] E. ISAACSON and H. B. KELLER. *Analysis of numerical methods*. Dover, 1994.

[LeV07] R. J. LEVEQUE. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007. DOI: 10.1137/1.9780898717839.

[MM05] K. W. MORTON and D. F. MYERS. *Numerical solution of partial differential equations*. Cambridge University Press, second edition, 2005. DOI: 10.1017/CBO9780511812248.

[Par98] B. N. PARLETT. *The symmetric eigenvalue problem*, *Classics in applied mathematics*. SIAM, 1998. DOI: 10.1137/1.9781611971163.

[PTV+07] W. H. PRESS, S. A. TEUKOLSKY, W. T. VETERLING, and B. P. FLANNERY. *Numerical recipes: the art of scientific computing*. Cambridge University Press, third edition, 2007.

[QV97] A. QUARTERONI and A. VALLI. *Numerical approximation of partial differential equations*, volume 23 of *Springer series in computational mathematics*. Springer, corrected second printing edition, 1997. DOI: 10.1007/978-3-540-85268-1.

[SB02] J. STOER and R. BULIRSCH. *Introduction to numerical analysis*, volume 12 of *Texts in applied mathematics*. Springer, third edition, 2002. DOI: 10.1007/978-0-387-21738-3.

[SM03] E. SÜLI and D. F. MAYERS. *An introduction to numerical analysis*. Cambridge University Press, 2003. DOI: 10.1017/CBO9780511801181.

[Ste01] G. W. STEWART. *Matrix algorithms. Volume II: eigensystems*. SIAM, 2001. DOI: 10.1137/1.9780898718058.

[Ste98] G. W. STEWART. *Matrix algorithms. Volume I: basic decompositions*. SIAM, 1998. DOI: 10.1137/1.9781611971408.

[TB97] L. N. TREFETHEN and D. BAU, III. *Numerical linear algebra*. SIAM, 1997.

[Tho95] J. W. THOMAS. *Numerical partial differential equations: finite difference methods*, volume 22 of *Texts in applied mathematics*. Springer, 1995. DOI: 10.1007/978-1-4899-7278-1.

[Var00] R. S. VARGA. *Matrix iterative analysis*, volume 27 of *Springer series in computational mathematics*. Springer, second edition, 2000. DOI: 10.1007/978-3-642-05156-2.

[Wil65] J. H. WILKINSON. *The algebraic eigenvalue problem*, *Numerical mathematics and scientific computation*. Oxford University Press, 1965.

[You71] D. M. YOUNG. *Iterative solution of large linear systems*. Academic Press, 1971.

Table des matières

Chapitre 1

Généralités sur l'analyse numérique et le calcul scientifique

L'*analyse numérique* (*numerical analysis* en anglais) est une branche des mathématiques appliquées s'intéressant au développement d'outils et de méthodes numériques pour le calcul d'approximations de solutions de problèmes de mathématiques [1] qu'il serait difficile, voire impossible, d'obtenir par des moyens analytiques [2]. Son objectif est notamment d'introduire des procédures calculatoires détaillées susceptibles d'être mises en œuvre par des calculateurs (électroniques, mécaniques ou humains) et d'analyser leurs caractéristiques et leurs performances. Elle possède des liens étroits avec deux disciplines à la croisée des mathématiques et de l'informatique. La première est l'*analyse des algorithmes* (*analysis of algorithms* en anglais), elle-même une branche de la *théorie de la complexité* [3] (*computational complexity theory* en anglais), qui fournit une mesure de l'efficacité d'une méthode en quantifiant le nombre d'*opérations élémentaires* [4], ou parfois la quantité de ressources informatiques (comme le temps de calcul, le besoin en mémoire...), qu'elle requiert pour la résolution d'un problème donné. La seconde est le *calcul scientifique* (*scientific computing* en anglais), qui consiste en l'étude de l'implémentation de méthodes numériques dans des architectures d'ordinateurs et leur application à la résolution effective de problèmes issus de la physique, de la biologie, des sciences de l'ingénieur ou encore de l'économie et de la finance.

Si l'introduction et l'utilisation de méthodes numériques précèdent de plusieurs siècles l'avènement des ordinateurs [5], c'est néanmoins avec l'apparition de ces outils modernes, vers la fin des années 1940 et le début des années 1950, que le calcul scientifique connut un essor sans précédent et que l'analyse numérique devint une domaine à part entière des mathématiques. La possibilité d'effectuer un grand nombre d'opérations arithmétiques très rapidement et simplement ouvrit en effet la voie au développement à de nouvelles classes de méthodes nécessitant d'être rigoureusement analysées pour s'assurer de l'exactitude et de la pertinence des résultats qu'elles fournissent. À ce titre, les travaux pionniers de Turing [6], avec notamment l'article [Tur48] sur l'analyse des effets des erreurs d'arrondi sur la factorisation LU, et de Wilkinson [7], dont on peut citer l'ouvrage [Wil94] initialement

1. Les problèmes considérés peuvent virtuellement provenir de tous les domaines d'étude des mathématiques pures ou appliquées. La théorie des nombres, la combinatoire, les algèbres abstraite et linéaire, la géométrie, les analyses réelle et complexe, la théorie de l'approximation et l'optimisation, pour ne citer qu'elles, possèdent toutes des aspects calculatoires. Sont ainsi couramment traitées numériquement l'évaluation d'une fonction en un point, le calcul d'intégrales, ou encore la résolution d'équations, ou de systèmes d'équations, algébriques, transcendantes, différentielles ordinaires ou aux dérivées partielles (déterministes ou stochastiques), de problèmes aux valeurs et vecteurs propres, d'interpolation ou d'optimisation (avec ou sans contraintes).

2. Pour compléter cette première définition, on ne peut que recommander la lecture de l'essai de L. N. Trefethen intitulé *The definition of numerical analysis*, publié dans la revue SIAM News en novembre 1992 et reproduit par la suite dans une annexe de l'ouvrage [Tre00].

3. Cette branche des mathématiques et de l'informatique théorique s'attache à connaître la difficulté intrinsèque d'une réponse algorithmique à un problème posé de façon mathématique et à définir en conséquence une *classe de complexité* pour ce problème.

4. La notion d'« opération élémentaire » est ici laissée nécessairement floue et entendue un sens plus large que celui qu'on lui attribue habituellement en arithmétique.

5. Le lecteur intéressé est renvoyé à l'ouvrage de Goldstine [Gol77], qui retrace une grande partie des développements de l'analyse numérique en Europe entre le seizième et le dix-neuvième siècle.

6. Alan Mathison Turing (23 juin 1912 - 7 juin 1954) était un mathématicien et informaticien anglais, spécialiste de la logique et de la cryptanalyse. Il est l'auteur d'un article fondateur de la science informatique, dans lequel il formalisa les notions d'algorithme et de calculabilité et introduisit le concept d'un calculateur universel programmable, la fameuse *« machine de Turing »*, qui joua un rôle majeur dans la création des ordinateurs.

7. James Hardy Wilkinson (27 septembre 1919 - 5 octobre 1986) était un mathématicien anglais. Il fut l'un des pionniers, et demeure une

publié en 1963, constituent deux des premiers éléments d'une longue succession de contributions sur le sujet.

Dans ce premier chapitre, nous revenons sur plusieurs principes qui, bien que n'ayant *a priori* pas toujours de rapport direct avec les méthodes numériques, interviennent de manière fondamentale dans leur mise en œuvre et leur application à la résolution de problèmes.

1.1 Différentes sources d'erreur dans une méthode numérique

Les solutions de problèmes calculées par une méthode numérique sont affectées par des erreurs que l'on peut principalement classer en trois catégories :

- les *erreurs d'arrondi* dans les opérations arithmétiques, qui proviennent des erreurs de représentation dues au fait que tout calculateur travaille *en précision finie*, c'est-à-dire dans un sous-ensemble discret du corps des réels $\mathbb{R}$, l'arithmétique naturelle étant alors approchée par une arithmétique de *nombres à virgule flottante* (voir la section 1.3) ;
- les *erreurs sur les données*, imputables à une connaissance imparfaite des données du problème que l'on cherche à résoudre, comme lorsqu'elles sont issues de mesures physiques soumises à des contraintes expérimentales ;
- les *erreurs de troncature, d'approximation ou de discrétisation*, introduites par les *schémas de résolution numérique* utilisés, comme le fait de tronquer le développement en série infini d'une solution analytique pour permettre son évaluation, d'arrêter d'un processus itératif dès qu'un itéré satisfait un critère donné avec une tolérance prescrite, ou encore d'approcher la solution d'une équation aux dérivées partielles en un nombre fini de points.

On peut également envisager d'ajouter à cette liste les erreurs qualifiées d'« humaines », telles les erreurs de programmation, ou causées par des dysfonctionnements des machines réalisant les calculs [8].

Le présent chapitre est en grande partie consacré aux erreurs d'arrondi, aux mécanismes qui en sont à l'origine, à leur propagation, ainsi qu'à l'analyse de leurs effets sur le résultat d'une suite de calculs. L'étude des erreurs de troncature, d'approximation ou de discrétisation constitue pour sa part un autre sujet majeur traité par l'analyse numérique. Elle sera abordée à plusieurs reprises dans ce cours, lors de l'étude de diverses méthodes itératives (chapitres 3, 4 et 5), de techniques d'interpolation polynomiale (chapitre 6) ou de formules de quadrature (chapitre 7).

Pour mesurer l'erreur entre la solution fournie par une méthode numérique et la solution du problème que l'on cherche à résoudre (on parle encore d'estimer la *précision* de la méthode), on introduit les notions d'*erreur absolue* et *relative*.

Définition 1.1 *Soit $\hat{x}$ une approximation d'un nombre réel x. On définit l'**erreur absolue** entre ces deux scalaires par*

$$|x - \hat{x}|,$$

*et, lorsque x est non nul, l'**erreur relative** par*

$$\frac{|x - \hat{x}|}{|x|}.$$

De ces deux quantités, c'est souvent la seconde que l'on privilégie pour évaluer la précision d'un résultat, en raison de son *invariance par changement d'échelle* : la mise à l'échelle $x \to \alpha x$ et $\hat{x} \to \alpha \hat{x}$, $\alpha \neq 0$, laisse en effet l'erreur relative inchangée.

Notons que ces définitions se généralisent de manière immédiate à des variables vectorielles ou matricielles en substituant des normes aux valeurs absolues (on parle de *normwise errors* en anglais). Par exemple, pour des vecteurs $\boldsymbol{x}$ et $\hat{\boldsymbol{x}}$ de $\mathbb{R}^n$, on a ainsi l'expression $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|$ pour l'erreur absolue et $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|/\|\boldsymbol{x}\|$ pour l'erreur relative, où $\|\cdot\|$ désigne une norme vectorielle donnée. Dans ces derniers cas, les erreurs sont également couramment évaluées *par composante* ou *par élément* (*componentwise errors* en anglais) dans le cadre de l'*analyse de sensibilité*

grande figure, de l'analyse numérique.

8. Il a été fait grand cas du bogue de division de l'unité de calcul en virgule flottante du fameux processeur Pentium® d'Intel®, découvert peu après le lancement de ce dernier sur le marché en 1994. En réalisant des tests informatiques pour ses recherches sur les nombres premiers, Thomas Nicely, de l'université de Lynchburg (Virginie, USA), constata que la division de 1 par 824633702441 renvoyait un résultat erroné. Il apparut plus tard que cette erreur était due à l'algorithme de division implanté sur le microprocesseur. Pour plus de détails, on pourra consulter [Ede97].

et de l'*analyse d'erreur* (voir respectivement les sections 1.4 et 1.5). Pour des vecteurs $\boldsymbol{x}$ et $\hat{\boldsymbol{x}}$ de $\mathbb{R}^n$, une mesure de l'erreur relative par composante est

$$\max_{1\leq i\leq n} \frac{|x_i - \hat{x}_i|}{|x_i|}.$$

1.2 Quelques notions d'algorithmique

Une méthode numérique repose sur l'emploi d'un (ou de plusieurs) *algorithme(s)* (*algorithm(s)* en anglais), notion ancienne, apparue bien avant les premiers ordinateurs, avec laquelle le lecteur est peut-être déjà familier et que nous abordons plus en détail ci-après.

1.2.1 Algorithme

De manière informelle, on peut définir un algorithme comme un énoncé décrivant, à l'aide d'un enchaînement déterminé d'opérations élémentaires (par exemple arithmétiques ou logiques), une démarche systématique permettant la résolution d'un problème donné en un nombre *fini ou infini* [9] d'étapes.

Un exemple d'algorithme : l'algorithme d'Euclide. Décrit dans le septième livre des *Éléments* d'Euclide [10], cet algorithme permet de déterminer le plus grand commun diviseur de deux entiers naturels. Il est basé sur la propriété suivante : *on suppose que $a \geq b$ et on note r le reste de la division euclidienne de a par b ; alors le plus grand commun diviseur de a et b est le plus grand commun diviseur de b et r*. En pratique, on divise le plus grand des deux nombres entiers par le plus petit, puis le plus petit des deux par le reste de la première division euclidienne. On répète ensuite le procédé jusqu'à ce que le reste de la division, qui diminue sans cesse, devienne nul. Le plus grand commun diviseur cherché est alors le dernier reste non nul (ou le premier diviseur, si le premier reste est nul).

La description d'un algorithme fait intervenir différentes *structures algorithmiques*, qui peuvent être des *structures de contrôle* (*control structures* en anglais) relatives à l'enchaînement des opérations élémentaires (comme des instructions d'affectation ou conditionnelles, des boucles, des appels de fonctions, des commandes de saut, de sortie ou d'arrêt, etc.), ou bien des *structures de données* (*data structures* en anglais), qui vont contenir et organiser les données (sous forme de listes, de tableaux, d'arbres, de piles ou de files selon le problème considéré) afin d'en permettre un traitement efficace.

Un algorithme est dit *séquentiel* (*sequential* en anglais) lorsque les instructions qui le forment sont exécutées les unes après les autres. Il est dit *parallèle* (*parallel* en anglais) lorsqu'elles s'exécutent concurremment. D'autre part, un algorithme exploitant des tâches s'exécutant plusieurs unités de calcul reliées par un réseau de communication est dit *réparti* ou *distribué* (*distributed* en anglais).

On dénombre trois principaux paradigmes présidant à la stratégie mise en œuvre dans un algorithme pour la résolution d'un problème. Tout d'abord, le principe *diviser pour régner* (*divide and conquer* en anglais) consiste à scinder le problème en sous-problèmes de même nature mais dont les données sont de taille plus petite, puis à résoudre ces sous-problèmes, pour combiner les résultats obtenus et construire une solution au problème posé. Il implique une approche récursive de la résolution du problème considéré.

La *programmation dynamique* (*dynamic programming* en anglais), introduite dans les années 1940 et 1950 par Bellman [11], possède aussi le caractère récursif de la précédente stratégie, tout en palliant un défaut majeur de cette dernière, conduisant dans certains cas à résoudre plusieurs fois des sous-problèmes identiques. Pour ce faire, une fois obtenue la solution d'un sous-problème, celle-ci est « mémorisée » (on parle de *mémoïsation*, du mot anglais *memoization*) de manière à être simplement rappelée chaque fois que le sous-problème sera de nouveau rencontré, évitant ainsi un recalcul.

9. D'un point de vue pratique, c'est-à-dire pour être utilisé par le biais d'un programme informatique, un algorithme doit nécessairement s'achever après avoir effectué un nombre fini d'opérations. Dans un contexte abstrait cependant, le nombre d'opérations réalisées peut être infini, tout en restant néanmoins dénombrable.

10. Euclide (Ευκλείδης en grec, v. 325 avant J.-C. - v. 265 avant J.-C.) était un mathématicien de la Grèce antique ayant probablement vécu en Afrique. Il est l'auteur des *Éléments*, un traité de mathématiques et de géométrie qui est considéré comme l'un des textes fondateurs des mathématiques modernes.

11. Richard Ernest Bellman (26 août 1920 - 19 mars 1984) était un mathématicien américain. Il est l'inventeur de la programmation dynamique et fit de nombreuses contributions aux théories de la décision et du contrôle optimal.

Enfin, l'approche *gloutonne* (*greedy* en anglais), destinée à la résolution de problèmes d'optimisation, vise à construire une solution en faisant une suite de choix qui sont chacuns localement optimaux.

1.2.2 Codage

Le *codage* (on utilise aussi souvent l'anglicisme *implémentation*) d'un algorithme consiste en l'écriture de la suite d'opérations élémentaires le composant dans un langage de programmation. Une première étape en vue de cette tâche est d'écrire l'algorithme en *pseudo-code*, c'est-à-dire d'en donner une description compacte et informelle qui utilise les conventions structurelles des langages de programmation [12] tout en s'affranchissant de certains détails techniques non essentiels à la bonne compréhension de l'algorithme, tels que la syntaxe, les déclarations de variables, le passage d'arguments lors des appels à des fonctions ou des routines externes, etc. À ce titre, les algorithmes 1 et 2 ci-après proposent deux manières de calculer un produit de matrices rectangulaires compatibles.

Algorithme 1 : Algorithme pour le calcul du produit $C = AB$ des matrices A de $M_{m,p}(\mathbb{R})$ et B de $M_{p,n}(\mathbb{R})$ (version « ijk »).

Entrée(s) : les tableaux contenant les matrices A et B.
Sortie(s) : le tableau contenant la matrice C.
pour $i = 1$ *à* m **faire**
  pour $j = 1$ *à* n **faire**
    $C(i,j) = 0$
    pour $k = 1$ *à* p **faire**
      $C(i,j) = C(i,j) + A(i,k) * B(k,j)$
    fin
  fin
fin

Pour chaque problème posé de façon mathématique, il peut exister plusieurs algorithmes le résolvant. Certains se distinguent par la nature et/ou le nombre des opérations élémentaires qui les constituent, alors que d'autres vont au contraire effectuer strictement les mêmes opérations élémentaires en ne différant que par la façon d'enchaîner ces dernières. Afin d'illustrer ce dernier point, comparons les algorithmes 1 et 2. On remarque tout d'abord qu'ils ne se différencient que par l'ordre de leurs boucles, ce qui ne change évidemment rien au résultat obtenu. D'un point de vue informatique cependant, on voit qu'on accède aux éléments des matrices A et B selon leurs lignes ou leurs colonnes, ce qui ne se fait pas à la même vitesse selon la manière dont les matrices sont stockées dans la mémoire. En particulier, la boucle interne de l'algorithme 1 correspond à un produit scalaire entre une ligne de la matrice A et une colonne de la matrice B. Dans chacun de ces deux algorithmes présentés, on peut encore modifier l'ordre des boucle en i et j pour obtenir d'autres implémentations parmi les six qu'il est possible de réaliser.

Algorithme 2 : Algorithme pour le calcul du produit $C = AB$ des matrices A de $M_{m,p}(\mathbb{R})$ et B de $M_{p,n}(\mathbb{R})$ (version « kij »).

Entrée(s) : les tableaux contenant les matrices A et B.
Sortie(s) : le tableau contenant la matrice C.
pour $i = 1$ *à* m **faire**
  pour $j = 1$ *à* n **faire**
    $C(i,j) = 0$
  fin
fin
pour $k = 1$ *à* p **faire**
  pour $i = 1$ *à* m **faire**
    pour $j = 1$ *à* n **faire**
      $C(i,j) = C(i,j) + A(i,k) * B(k,j)$
    fin
  fin
fin

12. On notera en particulier que le signe = en pseudo-code ne représente pas l'égalité mathématique, mais l'affectation de la valeur d'une variable à une autre.

1.2.3 Efficacité et complexité

Tout calculateur est soumis à des limitations physiques touchant à sa capacité de calcul, c'est-à-dire le nombre maximal d'opérations élémentaires, arithmétiques ou logiques, pouvant être effectuées à chaque seconde [13], ainsi qu'à la mémoire disponible, c'est-à-dire la quantité d'information à disposition ou à laquelle il est possible d'accéder à tout moment en un temps raisonnable. Typiquement, le temps d'exécution d'un algorithme mis en œuvre sur une machine informatique pour la résolution d'un problème dépendra :

- de la qualité du codage de l'algorithme dans un langage de programmation et de l'optimisation du code en langage machine du programme exécutable généré par le compilateur à partir de l'analyse du code source ;
- de l'organisation et de l'architecture des unités de calcul, ainsi que de la répartition de la mémoire de la machine ;
- des données du problème ;
- de la *complexité* (*computational complexity* en anglais) de l'algorithme.

Pour mesurer l'efficacité d'un algorithme en s'affranchissant des facteurs matériels et de l'instance du problème considéré, on a coutume d'utiliser sa seule complexité. Celle-ci est le plus souvent donnée par le nombre d'opérations de base que l'on doit effectuer (on parle alors de complexité *temporelle*) et la quantité de mémoire requise (on parle dans ce cas de complexité *spatiale*) pour résoudre un problème dont les données sont de taille fixée. Évidemment, le type des opérations de base peut varier d'un problème à l'autre ; par exemple, ce seront essentiellement des opérations arithmétiques (addition, soustraction, multiplication, division, etc.) pour des applications en calcul scientifique, mais, comme pour un tri de données, il pourra aussi s'agir d'une comparaison ou d'un déplacement d'éléments. De la même manière, la mesure de la taille des données doit refléter la quantité, la nature et la structure de l'information manipulée. Par exemple, pour des opérations sur les matrices, on emploie le (ou les) entier(s) donnant la taille des matrices en jeu ; dans le cas d'un graphe, ce sont les nombres de sommets ou de nœuds et d'arêtes ou d'arcs que l'on utilise.

Dans la suite, nous ne nous intéressons qu'à la complexité temporelle des algorithmes, que nous désignons par la fonction T. Nous supposons pour simplifier que la taille des données est représentée par un unique entier naturel n et que la complexité de l'algorithme étudié est une fonction de n, ayant pour valeur le nombre d'opérations élémentaires effectuées, sans que l'on différencie ces dernières [14].

Analyse de complexité du calcul du produit de deux matrices carrées. On considère l'évaluation du produit de deux matrices d'ordre n à coefficients dans un anneau, $\mathbb{R}$ par exemple. Pour le réaliser, on a *a priori*, c'est-à-dire en utilisant la définition (A.1), besoin de n^3 multiplications et $n^2(n-1)$ additions, soit $2n^3-n^2$ opérations arithmétiques. Par exemple, dans le cas de matrices d'ordre 2,

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}, \tag{1.1}$$

il faut ainsi faire huit multiplications et quatre additions. Plus explicitement, on a

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21}, \quad & c_{12} &= a_{11}b_{12} + a_{12}b_{22}, \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21}, \quad & c_{22} &= a_{21}b_{12} + a_{22}b_{22}. \end{aligned}$$

Il est cependant possible d'effectuer le produit (1.1) avec moins de multiplications. En effet, en faisant appel aux formules découvertes par Strassen [15] en 1969, qui consistent en l'introduction des quantités

$$\begin{aligned} q_1 &= (a_{11} + a_{22})(b_{11} + b_{22}), \\ q_2 &= (a_{21} + a_{22})b_{11}, \\ q_3 &= a_{11}(b_{12} - b_{22}), \\ q_4 &= a_{22}(-b_{11} + b_{21}), \\ q_5 &= (a_{11} + a_{12})b_{22}, \\ q_6 &= (-a_{11} + a_{21})(b_{11} + b_{12}), \\ q_7 &= (a_{12} - a_{22})(b_{21} + b_{22}), \end{aligned}$$

13. On admet implicitement dans la suite qu'une opération élémentaire s'effectue en un temps constant.

14. En pratique cependant, et bien que cela varie avec l'unité de calcul employée, on s'accorde à considérer que le coût d'une addition équivaut à celui d'une soustraction, et qu'une addition est moins coûteuse qu'une multiplication, qui est elle-même moins coûteuse qu'une division.

15. Volker Strassen (né le 29 avril 1936) est un mathématicien allemand. Il est célèbre pour ses travaux sur la complexité algorithmique, avec l'algorithme de Strassen pour la multiplication rapide de matrices carrées et l'algorithme de Schönhage–Strassen pour la multiplication rapide de grands entiers, et en théorie algorithmique des nombres, avec le test de primalité de Solovay–Strassen.

telles que

$$c_{11} = q_1 + q_4 - q_5 + q_7, \quad c_{12} = q_3 + q_5,$$
$$c_{21} = q_2 + q_4, \quad c_{22} = q_1 - q_2 + q_3 + q_6,$$

on utilise sept multiplications et dix-huit additions et soustractions.

Cette construction ne dépendant pas du fait que les éléments multipliés commutent entre eux ou non, on peut l'appliquer à des matrices décomposées par blocs. Ainsi, si A, B et C sont des matrices d'ordre n, avec n un entier pair, partitionnées en blocs d'ordre $\frac{n}{2}$,

$$\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

les blocs C_{ij}, $1 \leq i, j \leq 2$, du produit C peuvent être calculés comme précédemment, en substituant aux coefficients les blocs correspondant. L'*algorithme de Strassen* [Str69] consiste à appliquer récursivement ce procédé jusqu'à ce que les blocs soient des scalaires, selon une stratégie de type diviser pour régner. Pour cela, il faut que l'entier n soit une puissance de 2, cas auquel on peut toujours se ramener en ajoutant des colonnes et des lignes de zéros aux matrices A, B et C. Pour des matrices d'ordre $n = 2^m$, $m \in \mathbb{N}^*$, le nombre $T(n)$ de multiplications et d'additions requises par l'algorithme de Strassen vérifie

$$T(n) = T(2^m) = 7\,T(2^{m-1}) + 18\,(2^{m-1})^2,$$

la somme de deux matrices d'ordre n nécessitant n^2 additions. Un raisonnement par récurrence montre alors que

$$T(2^m) = 7^m\,T(1) + 18\sum_{k=0}^{m-1} 7^k 4^{m-k-1} \leq 7^m(T(1)+6),$$

dont on déduit que

$$T(n) \leq C\,n^{\log_2(7)},$$

avec C une constante strictement positive [16] et $\log_2(7) \approx 2,807$.

En pratique, la constante C fait que cette technique de multiplication n'est avantageuse que pour une valeur de n suffisamment grande, qui dépend par ailleurs du codage et de l'architecture de l'unité de calcul utilisée, principalement en raison du caractère récursif de l'algorithme qui implique le stockage de sous-matrices. D'autre part, le prix à payer pour la diminution asymptotique du nombre d'opérations est une stabilité numérique bien moindre que celle de la méthode « standard » de multiplication. Sur ce point particulier, on pourra consulter l'article [Hig90].

L'exemple précédent montre que la question de la complexité d'un algorithme revêt toute son importance lorsque la taille des données est très grande. C'est donc le comportement asymptotique de cette quantité qu'il convient de déterminer et l'on peut par conséquent simplifier l'analyse en ne s'intéressant qu'au *terme dominant* (relativement à n) dans l'expression de la fonction $T(n)$. Ce terme fixe ce qu'on nomme l'*ordre de croissance* (*order of growth* en anglais) du coût de l'algorithme en fonction de la taille des données du problème. On fait alors appel aux *notations de Bachmann* [17]*–Landau* [18] [Bac94, Lan09], provenant de la comparaison asymptotique, pour caractériser la complexité d'un algorithme. Plus précisément, étant donné une fonction f à valeurs positives définie sur $\mathbb{N}$, on écrit que (voir [Knu76]) :

- $T(n) \in O(f(n))$ s'il existe une constante C strictement positive et un entier n_0 tels que, pour tout $n > n_0$, $T(n) \geq C\,f(n)$;
- $T(n) \in \Omega(f(n))$ s'il existe une constante C strictement positive et un entier n_0 tels que, pour tout $n > n_0$, $C\,f(n) \leq T(n)$ (on notera que cette définition diffère de celle [19] de l'article [HL14], dans lequel le symbole fut introduit, qui a cours en théorie des nombres) ;
- $T(n) \in \Theta(f(n))$ s'il existe deux constantes C_1 et C_2 strictement positives et un entier n_0 tels que, pour tout $n > n_0$, $C_1\,f(n) \leq T(n) \leq C_2\,f(n)$.

On pourra remarquer que chacune de ces notations ignore la constante en facteur du terme dominant (parfois appelée la *constante d'ordre*), celle-ci jouant un rôle bien moins significatif que l'ordre de croissance si la taille des données est suffisamment grande. En pratique, la notation $O(\cdot)$ est la plus communément utilisée, la notation $\Theta(\cdot)$ la remplaçant lorsque l'on a besoin d'une estimation plus précise.

16. Dans [Str69], il est établi que $T(n) \leq 4,7\,n^{\log_2(7)}$.

17. Paul Gustav Heinrich Bachmann (22 juin 1837 - 31 mars 1920) était un mathématicien allemand qui s'intéressa principalement à la théorie des nombres.

18. Edmund Georg Hermann Landau (14 février 1877 - 19 février 1938) était un mathématicien allemand qui travailla dans les domaines de la théorie des nombres et de l'analyse complexe.

19. La définition originelle est en effet $T(n) \in \Omega(f(n))$ s'il existe une constante C strictement positive telle que, pour tout entier naturel n_0, il existe un entier $n > n_0$ tel que $C\,f(n) \leq T(n)$.

Par ailleurs, pour certains problèmes et certains algorithmes, la complexité peut dépendre non seulement de la taille des données, mais également des données elles-mêmes. Lorsque c'est le cas, l'analyse prend différentes formes, selon que l'on considère la complexité

- dans le *meilleur des cas* (*best case* en anglais), c'est-à-dire pour une configuration faisant qu'elle est la plus petite possible ;
- dans le *pire des cas* (*worst case* en anglais), c'est-à-dire pour une configuration faisant qu'elle est la plus grande possible ;
- *en moyenne* (*average case* en anglais), ce qui nécessite de la définir au travers d'un modèle probabiliste de distribution des données, mais permet en contrepartie de caractériser un comportement en quelques sorte « générique » de l'algorithme par rapport aux données.

Analyse de complexité du tri rapide. Le *tri rapide* (*quicksort* en anglais) est un algorithme de tri en place de liste inventé par Tony Hoare [20] [Hoa61a, Hoa61b, Hoa62] et fondé sur le principe diviser pour régner. Il consiste, à partir du choix arbitraire d'un élément appelé *pivot*, en un réarrangement de toute liste donnée en deux sous-listes contenant respectivement les éléments inférieurs (placés à gauche) et supérieurs [21] (placés à droite) au pivot. Cette opération, dite de *partitionnement*, est alors appliquée récursivement jusqu'à ce que la liste soit triée, c'est-à-dire que les sous-listes courantes contiennent un ou aucun élément. À chaque étape, un choix de pivot possible est celui du premier (ou dernier) élément de la liste à trier.

Conduisons une analyse de la complexité du tri rapide pour une liste possédant n éléments. Pour cela, remarquons que l'étape de partitionnement appliquée à une (sous-)liste contenant k éléments à trier, $k = 2, \ldots, n$, entraîne $k-1$ comparaisons et un certain nombre de permutations, induisant un coût linéaire en l'entier k, et supposons que $T(1) = T(0) = 1$.

Considérons tout d'abord le pire des cas, c'est-à dire celui pour lequel chaque choix de pivot scinde une liste de k éléments en une sous-liste contenant $k-1$ éléments et une autre n'en contenant aucun (en choisissant comme pivot le premier ou dernier élément, ceci correspond au cas d'une liste déjà triée, le pivot étant alors le plus petit ou le plus grand éléments de chaque sous-liste contenant au moins deux éléments). Il en découle que

$$\begin{aligned} T(n) = T(0) + T(n-1) + C\,n = 2\,T(0) + T(n-2) + C\,(n-1+n) = 3\,T(0) + T(n-3) + C\,(n-2+n-1+n) \\ = \cdots = n\,T(0) + C\sum_{k=0}^{n-2}(n-k), \end{aligned}$$

d'où

$$T(n) = n\,T(0) + C\left(n(n-2) - \frac{1}{2}(n-1)(n-2)\right) = \frac{C}{2}\,n^2 + \left(1 - \frac{C}{2}\right)n - C.$$

La complexité de l'algorithme dans ce cas est donc en $O(n^2)$.

Dans le meilleur des cas, chaque choix de pivot partage chaque liste en deux sous-listes contenant approximativement le même nombre d'éléments, c'est-à-dire

$$T(n) = 2\,T\left(\frac{n}{2}\right) + C\,n = \cdots = 2^k\,T(1) + C\,mn,$$

avec $n = 2^m$, $m \in \mathbb{N}^*$. On trouve par conséquent

$$T(n) = n\,T(1) + \frac{C}{\ln(2)}\,n\ln(n),$$

et la complexité est en $O(n\,ln(n))$.

Enfin, pour l'analyse de complexité en moyenne, faisons l'hypothèse que le choix de pivot renvoie avec la même probabilité n'importe quelle valeur présente dans la liste à trier. Il vient alors

$$T(n) = C\,n + \frac{1}{n}\sum_{k=1}^{n}(T(k-1) + T(n-k)) = C\,n + \frac{2}{n}\sum_{k=1}^{n}T(k-1),$$

soit encore

$$n\,T(n) = C\,n^2 + 2\sum_{k=1}^{n}T(k-1).$$

20. Charles Antony Richard Hoare (généralement appelé Tony Hoare ou C. A. R. Hoare, né le 11 janvier 1934) est un informaticien anglais. Il inventa en 1960 un algorithme de tri rapide encore très utilisé et fut le premier à avoir écrit un compilateur complet pour le langage Algol 60. Il est aussi à l'origine de la *logique de Hoare*, qui sert à la vérification des programmes, et d'un langage formel permettant de spécifier l'interaction de processus concurrents.

21. En cas d'égalité avec le pivot, on peut placer l'élément dans l'une ou l'autre des sous-listes.

Pour $n > 1$, on a aussi

$$(n-1)\,T(n-1) = C\,(n-1)^2 + 2\sum_{k=1}^{n-1} T(k-1),$$

et, par soustraction,

$$n\,T(n) = (n+1)\,T(n-1) + C\,(2n-1),$$

d'où

$$\frac{T(n)}{n+1} = \frac{T(n-1)}{n} + \frac{C\,(2n-1)}{n(n+1)}.$$

On arrive à

$$\frac{T(n)}{n+1} = \frac{T(n-2)}{n-1} + C\left(\frac{2n-3}{(n-1)n} + \frac{2n-1}{n(n+1)}\right) = \cdots = \frac{T(1)}{2} + C\sum_{k=2}^{n}\frac{2k-1}{k(k+1)},$$

d'où

$$T(n) \underset{n\to+\infty}{\sim} 2C\,n\ln(n).$$

La complexité est là encore en $O(n\ln(n))$.

Bien que l'algorithme de tri rapide possède une complexité en $O(n^2)$ dans le pire des cas, ce comportement est rarement observé en pratique, le pivot pouvant être choisi aléatoirement (selon une loi uniforme) de façon à éviter les problèmes liés aux listes déjà presque triées. Indiquons qu'un autre algorithme de tri efficace est celui du *tri par tas* (*heapsort* en anglais) [Wil64], dont la complexité dans le pire des cas est en $O(n\ln(n))$.

On dit que la complexité d'un algorithme est *constante* si elle est bornée par une constante, qu'elle est *linéaire* si elle est en $O(n)$, *quadratique* (et plus généralement *polynomiale*) si elle est en $O(n^2)$ (en $O(n^p)$ pour un certain entier strictement positif p) et *exponentielle* si elle est en $O(2^{n^p})$ pour un certain entier strictement positif p. On estime qu'un algorithme est *praticable* si sa complexité est polynomiale.

On considère généralement qu'un algorithme est plus efficace qu'un autre pour la résolution d'un problème si sa complexité dans le pire des cas est d'ordre de croissance moindre. Bien entendu, en raison de la constante d'ordre et de la présence de termes non dominants dans l'expression de la fonction de complexité, ce jugement peut être erroné pour des données de petite taille, mais il sera toujours vrai pour des données de taille suffisamment grande.

On terminera en indiquant que l'analyse de complexité du (ou des) algorithme(s) qui la compose(nt) est une partie importante de l'étude d'une méthode numérique, la recherche dans ce domaine consistant en la conception de moyens de résolution efficaces.

1.3 Arithmétique à virgule flottante

La mise en œuvre d'une méthode numérique sur une machine amène un certain nombre de difficultés d'ordre pratique, qui sont principalement liées à la nécessaire représentation approchée des nombres réels en mémoire. Avant de décrire plusieurs des particularités de l'*arithmétique à virgule flottante* (*floating-point arithmetic* en anglais) en usage sur la majorité des ordinateurs et calculateurs actuels, les principes de représentation des nombres réels et de leur stockage en machine sont rappelés. Une brève présentation du modèle d'arithmétique à virgule flottante le plus en usage actuellement, la *norme IEEE 754*, clôt la section.

1.3.1 Système de numération

Les nombres réels sont les éléments d'un corps archimédien complet totalement ordonné [22] noté $\mathbb{R}$, constitué de nombres dits *rationnels*, comme 76 ou $-\frac{4}{3}$, et de nombres dits *irrationnels*, comme $\sqrt{2}$ ou π. On peut les représenter grâce à un *système de numération positionnel* relatif au choix d'une *base* (*base* ou encore *radix* en anglais) β, $\beta \in \mathbb{N}$, $\beta \geq 2$, en utilisant que

$$x \in \mathbb{R} \iff x = s\sum_{i=-p}^{q} b_i\,\beta^i, \tag{1.2}$$

22. Le lecteur est renvoyé à la section B.1 de l'annexe B pour plus détails sur ces propriétés.

où s est le *signe* de x ($s = \pm 1$), $p \in \mathbb{N} \cup \{+\infty\}$, $q \in \mathbb{N}$ et les coefficients b_i, $-p \leq i \leq q$, prennent leurs valeurs dans l'ensemble $\{0, \ldots, \beta - 1\}$. On écrit alors [23] conventionnellement

$$x = s\, b_q b_{q-1} \ldots b_0, b_{-1} \ldots b_{-p\,\beta}, \tag{1.3}$$

où la virgule [24] est le *séparateur* entre la partie entière et la partie fractionnaire du réel x, l'indice β final précisant simplement que la représentation du nombre est faite relativement à la base β. Le système de numération est dit positionnel au sens où la position du chiffre b_i, $-p \leq i \leq q$, par rapport au séparateur indique par quelle puissance de l'entier β il est multiplié dans le développement (1.2). Lorsque $\beta = 10$, on a affaire au système de numération *décimal* communément employé, puisque l'on manipule généralement les nombres réels en utilisant implicitement leur représentation décimale (d'où l'omission de l'indice final). En pratique cependant, le choix $\beta = 2$, donnant lieu au système *binaire*, est le plus courant [25]. Dans ce dernier cas, les coefficients b_i, $-p \leq i \leq q$, du développement (1.2) peuvent prendre les valeurs 0 et 1 et sont appelés *chiffres binaires*, de l'anglais *binary digits* dont l'abréviation est le mot *bits*.

Le développement (1.2) peut posséder une infinité de termes non triviaux, c'est notamment le cas pour les nombres irrationnels, et la représentation (1.3) lui correspondant est alors qualifiée d'*infinie*. Une telle représentation ne pouvant être écrite, on a coutume d'indiquer les chiffres omis par des points de suspension, par exemple

$$\pi = 3,14159265358979323846\ldots_{10}.$$

Par ailleurs, la représentation d'un nombre rationnel dans une base donnée est dite *périodique* lorsque l'écriture contient un bloc de chiffres se répétant à l'infini. On a, par exemple,

$$\frac{1}{3} = 0,333333333333\ldots_{10},\ \frac{1}{7} = 0,142857142857\ldots_{10},\ \frac{7}{12} = 0,583333333333\ldots_{10}.$$

Il est possible de noter cette répétition de chiffres à l'infini en plaçant des points de suspension après plusieurs occurrences du bloc en question, comme on l'a fait ci-dessus. Cette écriture peut paraître claire lorsqu'une seule décimale est répétée une dizaine de fois, mais d'autres notations, plus explicites, font le choix de placer, de manière classique, la partie entière du nombre rationnel à gauche du séparateur et la partie fractionnaire non périodique suivie du bloc récurrent de la partie fractionnaire périodique, marqué d'un trait tiré au-dessus ou au-dessous ou bien placé entre crochets, à droite du séparateur. On a ainsi, pour les exemples précédents,

$$\frac{1}{3} = 0,\overline{3}_{10},\ \frac{1}{7} = 0,\overline{142857}_{10},\ \frac{7}{12} = 0,58\overline{3}_{10}.$$

Ajoutons que la représentation d'un nombre dans un système de numération n'est pas forcément unique, tout nombre pouvant s'écrire avec un nombre fini de chiffres ayant plusieurs représentations, dont une représentation infinie non triviale. On peut en effet accoler une répétition finie ou infinie du chiffre 0 à la représentation finie d'un nombre réel pour obtenir d'autres représentations, mais on peut également diminuer le dernier chiffre non nul de la représentation d'une unité et faire suivre ce chiffre d'une répétition infinie du chiffre $\beta - 1$ pour obtenir une représentation infinie non triviale. Le nombre 1 possède à ce titre les représentations finies 1_{10}, $1,0_{10}$, $1,00000_{10}$, parmi d'autres, et les deux représentations infinies $1,\overline{0}_{10}$ et $0,\overline{9}_{10}$. Pour parer à ce problème d'unicité, il est courant de ne pas retenir la représentation infinie d'un nombre lorsqu'une représentation finie existe et d'interdire que les chiffres de cette représentation finie soient tous nuls à partir d'un certain rang.

Exemples d'écriture de nombres réels dans les systèmes binaire et décimal.

- $10011,011_2 = 2^4 + 2^1 + 2^0 + 2^{-2} = 16 + 2 + 1 + \frac{1}{4} + \frac{1}{8} = 19,375_{10}$;
- $0,\overline{01}_2 = \sum_{i=1}^{+\infty} 2^{-2i} = \frac{1}{4}\sum_{i=0}^{+\infty}\left(\frac{1}{4}\right)^i = \frac{1}{4}\frac{1}{1-\frac{1}{4}} = \frac{1}{3} = 0,\overline{3}_{10}$;

23. Pour bien illustrer le propos, on a considéré l'exemple d'un réel x pour lequel p et q sont tous deux strictement plus grand que 1.

24. Il est important de noter que le symbole utilisé comme séparateur par les anglo-saxons, et notamment les langages de programmation ou les logiciels MATLAB et GNU OCTAVE, est le point et non la virgule.

25. Sur certains ordinateurs plus anciens, le système *hexadécimal*, c'est-à-dire tel que $\beta = 16$, est parfois utilisé. Un nombre s'exprime dans ce cas à l'aide des seize symboles 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E et F.

- $0,\overline{0011}_2 = 3\sum_{i=1}^{+\infty} 2^{-4i} = \frac{3}{16}\sum_{i=0}^{+\infty}\left(\frac{1}{16}\right)^i = \frac{1}{5} = 0,2_{10}$.

Le dernier des exemples ci-dessus montre qu'à la représentation finie d'un nombre réel dans le système décimal peut parfaitement correspondre une représentation infinie non triviale dans le système binaire. Nous verrons dans la prochaine section que ceci a des conséquences au niveau des calculs effectués sur une machine. En revanche, il est facile de voir qu'un nombre ayant une représentation finie dans le système binaire aura également une représentation finie dans le système décimal.

1.3.2 Représentation des nombres réels en machine

La mémoire d'une machine étant constituée d'un support physique, sa capacité est, par construction, limitée. Pour cette raison, le nombre de valeurs (entières, réelles, etc.) représentables, stockées en machine sous la forme d'ensembles de chiffres affectés à des *cellules-mémoire* [26] portant le nom de *mots-mémoire*, est fini. Pour les nombres réels, il existe essentiellement deux systèmes de représentation : celui des *nombres à virgule fixe* et celui des nombres à virgule flottante.

Tout d'abord, supposons que l'on dispose de N cellules-mémoire pour stocker un nombre réel non nul. Une manière naturelle de faire est de réserver une cellule-mémoire pour son signe, $N-r-1$ cellules-mémoire pour les chiffres situés à droite du séparateur (la partie entière du nombre) et r cellules-mémoire restantes pour les chiffres situés à droite du séparateur, l'entier r étant fixé, c'est-à-dire

$$s\, b_{n-r-1}\dots b_0, b_{-1}\dots b_{-r\beta}, \tag{1.4}$$

ce qui revient à convenir d'une position immuable et tacite du séparateur. Les nombres réels ainsi représentables sont dits à virgule fixe (*fixed-point numbers* en anglais). Ils sont principalement utilisés lorsque le processeur de la machine (un microcontrôleur par exemple) ne possède pas d'unité de calcul pour les nombres à virgule flottante ou bien quand ils permettent de diminuer le temps de traitement et/ou d'améliorer l'exactitude des calculs. Cependant, l'absence de « dynamique » dans le choix de placement du séparateur limite considérablement la plage de valeurs représentables par un nombre à virgule fixe, sauf à disposer d'un grand nombre de cellules-mémoire.

Ce défaut peut néanmoins être aisément corrigé en s'inspirant de la notation scientifique des nombres réels. L'idée est d'écrire tout nombre réel représentable sous la forme symbolique

$$s\, m\, \beta^e, \tag{1.5}$$

où m est un réel positif, composé d'au plus t chiffres en base β, appelé *significande* (*significand* en anglais), ou plus communément *mantisse* [27], et e est un entier signé, appelé *exposant*, compris entre deux bornes $e_{\min}$ et $e_{\max}$ (on a généralement $e_{\min} < 0$ et $e_{\max} > 0$). En autorisant l'exposant e à changer de valeur, on voit qu'on laisse le séparateur (la virgule ou le point selon la convention) « flotter » et une même valeur du significande peut alors servir à représenter des nombres réels dont la valeur absolue est arbitrairement grande ou petite. Les nombres ainsi définis sont dits à virgule flottante (*floating-point numbers* en anglais), et l'entier t est la *précision* ou encore *nombre de chiffres significatifs* du nombre à virgule flottante.

On remarquera qu'un même nombre peut posséder plusieurs écritures dans ce système de représentation. Par exemple, en base 10 et avec une précision égale à 3, on peut représenter $\frac{1}{10}$ par $100\,10^{-3}$, $10,0\,10^{-2}$, $1,00\,10^{-1}$, $0,10\,10^{0}$, $0,010\,10^{1}$ ou bien $0,001\,10^{2}$. La notion d'exposant n'est pas intrinsèque et dépend de conventions adoptées sur le significande, comme la place du séparateur dans ce dernier. On parle de représentation *normalisée* lorsque le premier chiffre, encore appelé le *chiffre de poids fort*, du significande est non nul, ce qui assure, une fois la position du séparateur fixée, que tout réel non nul [28] représentable ne possède qu'une seule représentation. En base binaire, une conséquence intéressante est que le premier bit du significande d'un nombre à virgule flottante normalisé est toujours égal à 1. On peut alors décider de ne pas le stocker physiquement et on parle de *bit de poids fort implicite* ou *caché* (*implicit or hidden leading bit* en anglais) du significande.

26. Quand $\beta = 2$, on notera que la taille d'une cellule-mémoire est de un bit.

27. En toute rigueur, ce terme désigne la différence entre un nombre et sa partie entière, et c'est en ce sens que l'on parle de la mantisse d'un logarithme décimal. C'est probablement le rapport étroit entre le logarithme décimal et la notation scientifique d'un nombre qui est à l'origine du glissement de sens de ce mot.

28. La restriction imposée fait que le nombre 0 n'est pas représentable par un nombre à virgule flottante normalisé.

Dans toute la suite, on suppose que le séparateur est placé entre le premier et le deuxième chiffre du significande. Le significande d'un nombre à virgule flottante vérifie par conséquent

$$0 \leq m \leq (1-\beta^{-t})\beta,$$

et $m \geq 1$ si le nombre est normalisé. Il est alors facile de vérifier que le plus petit (resp. grand) nombre réel positif atteint par un nombre à virgule flottante normalisé est

$$\beta^{e_{\min}} \text{ (resp. } (1-\beta^{-t})\,\beta^{e_{\max}+1}). \tag{1.6}$$

L'ensemble des nombres à virgule flottante construit à partir d'une représentation normalisée est un ensemble fini de points de la droite réelle [29], qui ne sont par ailleurs pas équirépartis sur cette dernière [30]. On note parfois $\mathbb{F}(\beta, t, e_{\min}, e_{\max})$ l'union de ces nombres avec le singleton $\{0\}$. L'écart entre un nombre à virgule flottante normalisé x non nul et son plus proche voisin se mesure à l'aide de l'*epsilon machine*, $\varepsilon_{\text{mach}} = \beta^{1-t}$, qui est la distance entre le nombre 1 et le nombre à virgule flottante le plus proche qui lui est supérieur. On a l'estimation suivante.

Lemme 1.2 *La distance entre un nombre à virgule flottante normalisé x non nul et nombre à virgule flottante normalisé adjacent est au moins $\beta^{-1}\varepsilon_{\text{mach}}|x|$ et au plus $\varepsilon_{\text{mach}}|x|$.*

DÉMONSTRATION. On peut, sans perte de généralité, supposer que le réel x est strictement positif et l'on pose $x = m\beta^e$ avec $1 \leq m < \beta$. Le nombre à virgule flottante supérieur à x lui étant le plus proche est $x + \beta^{e-t+1}$, d'où

$$\beta^{-1}\varepsilon_{\text{mach}}x = \beta^{-t}x < \frac{x}{m\beta^{t-1}} = x + \beta^{e-t+1} - x = \beta^{e-t+1} \leq (m\,\beta^{1-t})\beta^{e-t+1} = \varepsilon_{\text{mach}}x.$$

Si l'on considère le nombre à virgule flottante adjacent et inférieur à x, celui-ci vaut $x - \beta^{e-t+1}$ si $x > \beta^e$, ce qui fournit à le même majorant que précédemment, et $x - \beta^{e-t}$ si $x = \beta^e$, auquel cas on a

$$x - x + \beta^{e-t} = \beta^{-1}\beta^{-t+1}\beta^e = \beta^{-1}\varepsilon_{\text{mach}}x.$$

□

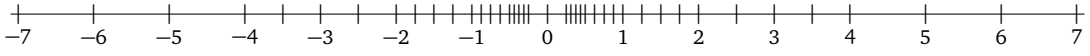
Tout ensemble de nombres à virgule flottante normalisés peut être complété [31] par des nombres *dénormalisés* ou *sous-normaux* (*denormalized or subnormal numbers* en anglais). Ces derniers permettent de représenter des nombres réels dont la valeur absolue est aussi petite que $\beta^{e_{\min}-t+1}$, en abandonnant l'hypothèse sur le chiffre de poids fort du significande (et donc au détriment de la précision de la représentation). En représentation binaire, le bit de poids fort n'étant plus implicite, on réserve une valeur spéciale de l'exposant (celle qui correspondrait à $e_{\min} - 1$) pour la représentation de ces nombres.

Arrondi

Le caractère fini des ensembles de nombres à virgule flottante pose notamment le problème de la représentation en machine d'un nombre réel *quelconque* donné, que l'on a coutume de résoudre en remplaçant, le cas échéant, ce nombre par un autre admettant une représentation à virgule flottante dans le système considéré.

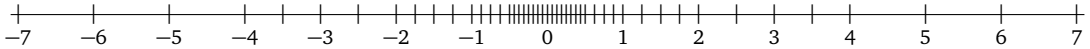
29. Par construction, cet ensemble est constitué des nombres rationnels exactement représentables dans le système de numération utilisé.

30. Ils sont en effet plus denses près du plus petit (resp. grand) nombre positif (resp. négatif) non nul représentable. Par exemple, pour $\beta = 2$, $t = 3$, $e_{\min} = -2$ et $e_{\max} = 2$, les nombres à virgule flottante positifs représentables sont 0 et les nombres normalisés $\frac{1}{4}$, $\frac{5}{16}$, $\frac{3}{8}$, $\frac{7}{16}$, $\frac{1}{2}$, $\frac{5}{8}$, $\frac{3}{4}$, $\frac{7}{8}$, 1, $\frac{5}{4}$, $\frac{3}{2}$, $\frac{7}{4}$, 2, $\frac{5}{2}$, 3, $\frac{7}{2}$, 4, 5, 6, 7, d'où la répartition suivante des éléments de $\mathbb{F}(2,3,-2,2)$



sur la droite réelle. On remarque que la distance entre deux nombres à virgule flottante consécutifs est multipliée par β (doublée dans cet exemple) à chaque fois que l'on passe une puissance de β (2 dans cet exemple).

31. Si l'on reprend l'exemple précédent, les nombres à virgule flottante dénormalisés positifs sont $\frac{1}{16}$, $\frac{1}{8}$, $\frac{3}{16}$ et on a alors la répartition suivante pour l'ensemble des nombres à virgule flottante



sur la droite réelle.

Pour un nombre réel dont la valeur absolue est comprise entre les bornes (1.6), une première manière de faire consiste à tout d'abord écrire le nombre sous la forme (1.5) pour ne conserver ensuite que les t premiers chiffres de sa mantisse. On parle alors de *troncature* ou d'*arrondi vers zéro* (*chopping* ou *rounding towards zero* en anglais), qui est un premier exemple d'*arrondi dirigé*. On peut aussi substituer au nombre réel le nombre à virgule flottante qui lui est le plus proche ; c'est l'*arrondi au plus proche* (*rounding to nearest* en anglais). Lorsque le nombre se situe à égale distance des deux nombres à virgule flottante qui l'entourent, on choisit la valeur de l'arrondi en faisant appel à des arrondis dirigés. On peut alors prendre

- le nombre à virgule flottante le plus petit (resp. grand), c'est l'*arrondi par défaut* (resp. *excès*) (*rounding half down* (resp. *up*) en anglais) ;
- le nombre à virgule flottante le plus petit (resp. grand) en valeur absolue, c'est l'*arrondi vers zéro* (resp. *vers l'infini*) (*rounding half towards zero* (resp. *away from zero*) en anglais) ;
- le nombre à virgule flottante dont le dernier chiffre de la mantisse est pair (resp. impair), c'est, par abus de langage, l'*arrondi au chiffre pair* (resp. *impair*) (*rounding half to even* (resp. *odd*) en anglais). Cette dernière méthode est employée afin d'éliminer le biais pouvant survenir en arrondissant selon les autres règles.

Dans le cas d'un nombre réel non nul dont la valeur absolue n'appartient pas à l'intervalle défini par (1.6), il n'est pas possible d'effectuer un remplacement correspondant à un arrondi. Ce dépassement de la capacité de stockage est appelé *débordement vers l'infini* (*overflow* en anglais) si la valeur absolue du nombre est trop grande ou *débordement vers zéro* (*underflow* en anglais) si elle est trop petite. L'occurrence d'un débordement vers l'infini est un problème sérieux [32], notamment lorsque le nombre qui le provoque est le résultat d'une opération, et devrait, en toute rigueur, conduire à l'interruption du calcul en cours. Un débordement vers zéro est moins grave et un remplacement par 0 (la valeur la plus proche) est en général effectué, mais cette solution n'est cependant pas toujours satisfaisante [33]. Quand l'ensemble des nombres à virgule flottante contient des nombres dénormalisés, l'arrondi peut prendre une valeur non nulle comprise entre 0 et la dernière valeur représentable par un nombre normalisé [34] et le débordement vers zéro est dit *progressif* (*gradual or graceful underflow* en anglais).

Dans toute la suite, on note $\mathrm{fl}(x)$ l'arrondi au plus proche d'un nombre réel x, définissant ainsi une application de $[(\beta^{-t}-1)\,\beta^{e_{\max}+1}, -\beta^{e_{\min}}] \cup \{0\} \cup [\beta^{e_{\min}}, (1-\beta^{-t})\,\beta^{e_{\max}+1}]$ dans $\mathbb{F}(\beta, t, e_{\min}, e_{\max})$. Si x est un nombre à virgule flottante, on a clairement $\mathrm{fl}(x) = x$. On vérifie également la propriété de monotonie suivante

$$x \leq y \implies \mathrm{fl}(x) \leq \mathrm{fl}(y),$$

pour tous nombres réels x et y pour lesquels l'arrondi est défini. L'erreur d'arrondi (*round-off error* ou *rounding error* en anglais) sur un nombre x est la différence $x - \mathrm{fl}(x)$.

Le résultat suivant montre qu'un nombre réel x, pour lequel l'arrondi est défini, est approché avec une erreur relative en valeur absolue ne dépassant pas la valeur $u = \frac{1}{2}\,\beta^{1-t} = \frac{1}{2}\,\varepsilon_{\mathrm{mach}}$, appelée *précision machine* (*machine precision* ou *unit round-off* [35] en anglais).

Théorème 1.3 *Soit x un nombre réel tel que $\beta^{e_{\min}} \leq |x| \leq (1-\beta^{-t})\beta^{e_{\max}+1}$. Alors, on a*

$$\mathrm{fl}(x) = x(1+\delta),\ |\delta| < u. \tag{1.7}$$

DÉMONSTRATION. On peut, sans perte de généralité, supposer que le réel x est strictement positif. En écrivant x sous la forme

$$x = \mu\,\beta^{e-t+1},\ \beta^{t-1} \leq \mu < \beta^{t},$$

32. Une illustration des conséquences désastreuses auxquelles peut conduire une mauvaise gestion d'un dépassement de capacité est celle du vol inaugural d'Ariane 5 le 4 juin 1996, durant lequel la fusée explosa à peine quarante secondes après son décollage de Kourou en Guyane française, détruisant ainsi sa charge utile (quatre sondes spatiales) d'une valeur totale de 370 millions de dollars. Une enquête (voir J.-L. Lions *et al.*, *Ariane 5: flight 501 failure*, Ariane 501 inquiry board report, 1996) mit à jour un dysfonctionnement du système de guidage inertiel, causé par la conversion d'un nombre à virgule flottante stocké sur 64 bits donnant la vitesse horizontale de la fusée en un entier signé stocké sur 16 bits. L'entier obtenu étant plus grand que 32767, la plus grande valeur entière signée représentable avec 16 bits, l'échec de conversion déclencha une exception non traitée (suite à une erreur de programmation) qui fût interprétée comme une déviation de la trajectoire. La violente correction demandée par le système de guidage provoqua alors un dérapage de la fusée de sa trajectoire, entraînant son auto-destruction préventive. Il s'avére que, pour des raisons d'économies sur le coût des préparatifs, aucune simulation n'avait été effectuée avant le vol, le système de navigation étant le même que celui d'Ariane 4, fusée moins puissante et donc moins rapide qu'Ariane 5, et réputé fiable...

33. On peut en effet imaginer que le nombre incriminé puisse ensuite être multiplié par un très grand nombre ; si un remplacement par zéro a lieu, le résultat final sera nul...

34. En d'autres mots, en présence de nombres dénormalisés, on a l'intéressante propriété suivante : si $x \neq y$, la valeur calculée de $x - y$ ne peut être nulle. Le dépassement de capacité progressif assure ainsi l'existence et l'unicité dans $\mathbb{F}$ de l'opposé d'un nombre à virgule flottante.

35. Cette dernière appellation provient du fait que le nombre u représente la plus grande erreur relative commise sur les nombres réels arrondis à 1.

on observe que x se trouve entre les deux nombres à virgule flottante adjacents $y_1 = \lfloor \mu \rfloor \beta^{e-t+1}$ et $y_2 = \lceil \mu \rceil \beta^{e-t+1}$ (ou $y_2 = \frac{\lceil \mu \rceil}{\beta} \beta^{e-t}$ si $\lceil \mu \rceil = \beta^t$), où $\lfloor \mu \rfloor$ (resp. $\lceil \mu \rceil$) désigne la partie entière par défaut (resp. par excès) du réel μ. Par conséquent, $\mathrm{fl}(x) = y_1$ ou y_2 et l'on a

$$|x - \mathrm{fl}(x)| \leq \frac{|y_2 - y_1|}{2} \leq \frac{\beta^{e-t+1}}{2},$$

d'où

$$\left| \frac{x - \mathrm{fl}(x)}{x} \right| \leq \frac{\frac{\beta^{e-t+1}}{2}}{\mu \beta^{e-t+1}} \leq \frac{\beta^{1-t}}{2} = u.$$

La dernière inégalité est stricte sauf si $\mu = \beta^{t-1}$, auquel cas $\mathrm{fl}(x) = x$. L'inégalité dans (1.7) est donc stricte. □

On peut établir que les arrondis dirigés satisfont une inégalité identique à (1.7) avec $|\delta| < 2u$. La version modifiée suivante du précédent résultat est parfois utile pour l'analyse d'erreur.

Théorème 1.4 *Sous les hypothèses du théorème 1.3, on a*

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \ |\delta| \leq u.$$

DÉMONSTRATION. En reprenant la preuve du théorème 1.3, on constate que $|y_i| \geq \beta^e$, $i = 1, 2$. On a par conséquent

$$\frac{|x - \mathrm{fl}(x)|}{|\mathrm{fl}(x)|} \leq \frac{\frac{\beta^{e-t+1}}{2}}{\beta^e} = \frac{\beta^{1-t}}{2} = u,$$

dont on déduit le résultat. □

Les erreurs d'arrondi sont inévitables et parfois présentes avant même qu'une seule opération ait eu lieu, puisque la représentation en machine des données d'un problème peut nécessiter de les arrondir. Prises isolément, ces erreurs sont généralement bénignes, mais leur propagation et leur accumulation [36] au cours d'une série de calculs, notamment lorsque l'on cherche à résoudre un problème mal conditionné (voir la sous-section 1.4.2) et/ou que l'on utilise un algorithme numériquement instable (voir la sous-section 1.5.2), peuvent faire perdre toute signification au résultat numérique obtenu.

1.3.3 Arithmétique en précision finie

L'ensemble des nombres représentables sur une machine étant introduit, il faut définir sur celui-ci une arithmétique reproduisant de manière aussi fidèle que possible celle existant sur $\mathbb{R}$. On parle alors d'arithmétique en précision finie, les principales différences avec l'arithmétique « exacte » (*i.e.*, en précision infinie) provenant du caractère discret et fini de l'ensemble des nombres manipulés.

Une conséquence directe de cette arithmétique sur les algorithmes des méthodes numériques est que leurs résultats sont entachés d'erreurs [37], que l'on devrait être capable de mesurer afin de garantir la pertinence des calculs. C'est l'objet de l'analyse d'erreur (voir la section 1.5) que de les estimer et d'identifier leur(s) origine(s). C'est dans cette perspective que nous allons maintenant mettre en avant certaines des particularités de l'arithmétique en précision finie.

36. Un exemple célèbre de désastre dû à une erreur d'arrondi est celui de l'échec d'interception par un missile Patriot américain d'un missile Scud irakien visant des baraquements militaires situés à Dhahran en Arabie Saoudite durant la guerre du Golfe, le 25 février 1991, qui eut pour conséquence la mort de 28 soldats américains et près d'une centaine de blessés. Un rapport (voir United States General Accounting Office, *Patriot missile defense: software problem led to system failure at Dhahran, Saudi Arabia*, GAO/IMTEC-92-26 report, 1992) imputa cette défaillance à une imprécision dans le calcul de la date depuis le démarrage du système de la batterie de missiles Patriot. Plus précisément, la date était mesurée en dixièmes de seconde par l'horloge interne du système, stockée dans un registre sous la forme d'un entier et obtenue en multipliant cet entier par une approximation de $\frac{1}{10}$ stockée sur 24 bits (l'écriture de $\frac{1}{10}$ dans le système binaire est en effet $0,0\overline{0011}_2$; la valeur tronquée effectivement stockée était donc $0,00011001100110011001100_2$, ce qui introduit une erreur égale à $0,0000000000000000000000\overline{0011}_2$, soit encore $0,00000095367431640625_{10}$). La batterie de missiles ayant été en service depuis une centaine d'heures au moment de l'attaque, l'erreur accumulée causée par l'arrondi fut d'environ $0,34$ seconde, temps pendant lequel un missile Scud parcourt plus de 500 mètres, rendant de fait son interception impossible.

37. Le résultat d'un calcul est aussi affecté par la présence de perturbations sur les données, qui peuvent aussi bien être le fruit d'arrondis ayant eu lieu lors de leur stockage en machine qu'être causées par une connaissance seulement approximative de celles-ci (c'est le cas lorsque les données sont le résultat de calculs antérieurs ou qu'elles proviennent d'estimations statistiques, de mesures expérimentales imparfaites, etc.).

Un modèle d'arithmétique à virgule flottante

Le modèle d'arithmétique, dû à Wilkinson [Wil60], classiquement utilisé pour l'analyse d'erreur d'un algorithme possède la propriété suivante : en désignant par le symbole « op » n'importe quelle opération arithmétique de base (addition, soustraction, multiplication, division), soit x et y deux nombres à virgule flottante tels que le résultat $x \,\mathrm{op}\, y$ ne provoque pas de dépassement de capacité, alors on a

$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1+\delta),\ |\delta| \leq u,\ \mathrm{op} = +,\ -,\ *,\ /. \tag{1.8}$$

Ci-dessus, on a utilisé la notation $\mathrm{fl}(\cdot)$ appliquée à une expression arithmétique pour représenter la valeur *effectivement* fournie par la machine pour l'évaluation de cette expression. La propriété assure alors que cette valeur est « aussi bonne » que l'arrondi du résultat exact, au sens où l'erreur relative possède la même borne dans les deux cas. Il n'est cependant pas demandé d'avoir $\delta = 0$ si le résultat exact appartient à $\mathbb{F}$, alors que cette dernière condition est satisfaite par l'arrondi. Malgré cet inconvénient, ce modèle décrit la plupart des arithmétiques à virgule flottante utilisées en pratique et est considéré comme standard. Ajoutons que des contraintes matérielles [38] sont nécessaires pour que la condition (1.8) soit vérifiée dans le cas de l'addition et de la soustraction et qu'il est également courant de supposer que l'erreur d'arrondi sur le calcul d'une racine carrée vérifie une inégalité semblable, c'est-à-dire

$$\mathrm{fl}(\sqrt{x}) = \sqrt{x}(1+\delta),\ |\delta| \leq u.$$

Multiplication et addition fusionnées

Certaines machines ont la possibilité d'effectuer une multiplication et une addition *fusionnées* (*fused multiply-add* en anglais), c'est-à-dire une multiplication suivie d'une addition ou d'une soustraction, comme si elle correspondait à *une seule* opération en arithmétique à virgule flottante, et donc en ne procédant qu'à un arrondi, d'où l'erreur suivante sur le résultat :

$$\mathrm{fl}(x \pm y * z) = (x \pm y * z)(1+\delta),\ |\delta| \leq u.$$

Cette opération peut avantageusement être mise à profit pour améliorer la rapidité et la précision du calcul du produit scalaire de deux vecteurs de taille n (on ne commet alors que n erreurs d'arrondi consécutives au lieu de $2n-1$) ou de l'application de la méthode de Horner (voir la sous-section 5.7.2) pour l'évaluation d'un polynôme en un point, par exemple.

Perte d'associativité et de distributivité

Aux problèmes déjà posés par les arrondis s'ajoute le fait que ces derniers rendent plusieurs des propriétés fondamentales de l'arithmétique « exacte » caduques en arithmétique à virgule flottante, ce qui tend à faire de l'analyse d'erreur de calculs effectués sur une machine un travail passablement compliqué. Ainsi, si l'addition et la multiplication de nombres à virgule flottante sont bien commutatives, elles ne sont en général plus associatives, comme le montre l'exemple qui suit. La distributivité de la multiplication par rapport à l'addition est également perdue.

Exemple de non-associativité de l'addition en arithmétique à virgule flottante. Considérons la somme des trois nombres à virgule flottante suivants

$$x = 0,1234567\,10^0,\ y = 0,4711325\,10^4 \text{ et } z = -y,$$

dans un système avec une mantisse à sept chiffres en base 10. Si l'on réalise le calcul $x + (y+z)$, il vient

$$\mathrm{fl}(y+z) = 0,\ \mathrm{fl}(x + \mathrm{fl}(y+z)) = x = 0,1234567\,10^0.$$

En revanche, si l'on effectue d'abord l'addition entre x et y, on trouve

$$\mathrm{fl}(x+y) = 0,4711448\,10^4,\ \mathrm{fl}(\mathrm{fl}(x+y)+z) = 0,0000123\,10^4 = 0,123\ 10^0.$$

38. Plus précisément, on doit avoir recours à l'utilisation de *chiffres de garde* durant le calcul (voir [Gol91]). Lorsque ce n'est pas le cas, le modèle satisfait seulement

$$\mathrm{fl}(x \pm y) = x(1+\delta_1) \pm y(1+\delta_2),\ |\delta_i| \leq u,\ i = 1,2,$$
$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1+\delta),\ |\delta| \leq u,\ \mathrm{op} = *,\ /.$$

Si x, y et z sont trois nombres à virgule flottante, $x \neq 0$, indiquons encore que l'égalité $xy = xz$ n'implique pas $y = z$ ou que le produit $x\left(\frac{y}{x}\right)$ ne vaut pas forcément y en arithmétique en précision finie. De la même manière, les implications de stricte comparaison suivantes

$$x < y \implies x + z < y + z, \quad y < z,\ x > 0 \implies xy < xz,$$

ne seront vérifiées qu'à condition d'être affaiblies en remplaçant les inégalités strictes dans les membres de droite par des inégalités larges.

Soustraction exacte

Il est intéressant de noter que la soustraction de deux nombres à virgule flottante suffisamment proches est toujours exacte. On dispose en effet du résultat suivant, dû à Sterbenz [Ste74], valable dans toute base de représentation et pour un système d'arithmétique à virgule flottante utilisant au moins un chiffre de garde.

Théorème 1.5 *Soit x et y deux nombres à virgule flottante tels que $\frac{1}{2}\, y \leq x \leq 2\, y$. Alors, si le résultat $x - y$ n'entraîne par de débordement vers zéro, on a* $\mathrm{fl}(x - y) = x - y$.

Lorsque le système permet un débordement vers zéro progressif, l'hypothese du théorème sur la différence $x - y$ peut être levée. Ce dernier résultat s'avère essentiel pour prouver la stabilité (voir la section 1.5) de certains algorithmes, comme celui de l'exemple suivant.

Exemple du calcul de l'aire d'un triangle connaissant les longueurs de ses côtés. Soit un triangle dont les longueurs des côtés sont données par les réels a, b et c. Son aire A est donnée par la *formule de Héron* [39]

$$A = \sqrt{s(s-a)(s-b)(s-c)},$$

dans laquelle s est le demi-périmètre du triangle, $s = \frac{1}{2}(a + b + c)$. L'implémentation directe de cette formule tend à fournir un très mauvais résultat numérique quand le triangle est dit « en épingle », ce qui est, par exemple, le cas lorsque le nombre c est très petit devant a et b. L'erreur d'arrondi sur la valeur de s peut alors être du même ordre que c, ce qui conduit à des calculs particulièrement inexacts des quantités $s - a$ et $s - b$. Pour remédier à ce problème, Kahan [40] proposa de renommer a, b et c de manière à ce que $a \leq b \leq c$ et d'utiliser la formule

$$A = \frac{1}{4}\sqrt{(a + (b + c))(c - (a - b))(c + (a - b))(a + (b - c))},$$

dans laquelle le placement des parenthèses est fondamental (voir [Kah83]). Lorsque le théorème 1.5 est valide, on montre que l'erreur relative sur le résultat fourni par cette dernière égalité est bornée par un petit multiple de la précision machine u, ce qui assure la stabilité (directe) de la méthode (voir la sous-section 1.5.2).

Arithmétique complexe

On peut déduire de la propriété (1.8) des résultats similaires pour les opérations élémentaires réalisées sur des nombres complexes en considérant qu'un nombre à virgule flottante complexe est composé de deux nombres à virgule flottante réels représentant respectivement ses parties réelle et imaginaire. Ainsi, en posant $x = a + \mathrm{i}b$ et $y = c + \mathrm{i}b$, où $\mathrm{i}^2 = -1$ et a, b, c et d sont des nombres réels, et en observant que

$$x \pm y = a \pm c + \mathrm{i}(b \pm d),\ xy = ac - bd + \mathrm{i}(ad + bc),\ \text{et}\ \frac{x}{y} = \frac{ac + bd}{c^2 + d^2} + \mathrm{i}\,\frac{bc - ad}{c^2 + d^2}, \tag{1.9}$$

on établit le résultat suivant (voir la preuve du Lemme 3.5 de [Hig02] pour une démonstration).

39. Héron d'Alexandrie (Ἥρων ο Αλεξανδρεύς en grec, v. 10 - v. 70) était un ingénieur, mécanicien et mathématicien grec du premier siècle après J.-C.. On lui attribue plusieurs découvertes mathématiques, dont une formule de calcul de l'aire d'un triangle à partir des longueurs de ses côtés, ainsi qu'une méthode récursive d'approximation de la racine carrée de n'importe quel nombre positif. Il a cependant été suggéré que la première était connue d'Archimède, tandis que la seconde était apparement déjà utilisée par les babyloniens.

40. William Morton Kahan (né le 5 juin 1933) est un mathématicien et informaticien canadien. Surnommé *« le père de la virgule flottante »*, il est notamment à l'origine de la norme IEEE 754 et l'auteur d'un algorithme de somme basé sur un principe de compensation des erreurs d'arrondi.

Lemme 1.6 *Supposons que le modèle d'arithmétique à virgule flottante réelle satisfasse la propriété* (1.8) *et soit x et y deux nombres à virgule flottante complexes tels que les résultats des opérations élémentaires* (1.9) *ne provoquent pas de dépassement de capacité. Alors, on a les relations suivantes* [41]

$$\begin{aligned} \mathrm{fl}(x \pm y) &= (x+y)(1+\delta),\ |\delta| \leq u, \\ \mathrm{fl}(xy) &= xy\,(1+\delta),\ |\delta| \leq \sqrt{2}\,\gamma_2, \\ \mathrm{fl}\left(\frac{x}{y}\right) &= \frac{x}{y}\,(1+\delta),\ |\delta| \leq \sqrt{2}\,\gamma_4, \end{aligned}$$

où δ est un nombre complexe et [42]

$$\gamma_n = \frac{nu}{1-nu} \text{ pour } n \in \mathbb{N}^* \text{ tel que } nu < 1. \tag{1.10}$$

On notera que, en vertu de (1.9), une addition ou une soustraction entre nombres complexes est deux fois (et une multiplication au moins quatre fois) plus coûteuse que son analogue pour des nombres réels.

1.3.4 La norme IEEE 754

Historiquement, la représentation interne et le comportement des nombres à virgule flottante variaient d'un ordinateur à l'autre et le portage de programmes nécessitait parfois une profonde reprise de ces derniers, jusqu'à ce qu'un standard soit proposé par l'Institute of Electrical and Electronics Engineers. La norme IEEE 754-1985 (*IEEE Standard for Binary Floating-Point Arithmetic*), introduite en 1985 et aussi connue sous le nom de IEC 60559:1989 (*Binary Floating-Point Arithmetic for Microprocessor Systems*), spécifie plusieurs formats de représentation binaire des nombres réels à virgule flottante (normalisés et dénormalisés), ainsi qu'un ensemble d'opérations sur ces nombres, de valeurs spéciales et de modes d'arrondi. Elle est aujourd'hui [43] la plus employée pour les calculs avec des nombres à virgule flottante.

Elle définit un format *simple précision* sur un mot-mémoire de 32 bits (1 bit de signe, 8 bits d'exposant, 23 bits de significande, avec bit de poids fort implicite pour ce dernier), un format *double précision* sur un mot-mémoire de 64 bits (1 bit de signe, 11 bits d'exposant, 52 bits de significande, avec bit de poids fort implicite pour ce dernier), un format *simple précision étendue* (rarement utilisé) sur un mot-mémoire d'au moins 43 bits et un format *double précision étendue* sur un mot-mémoire d'au moins 79 bits. Les nombres à virgule flottante en simple (resp. double) précision standard correspondent aux éléments de l'ensemble $\mathbb{F}(2, 24, -126, 127)$ (resp. $\mathbb{F}(2, 53, -1022, 1023)$) et peuvent décrire des nombres réels dont la valeur absolue est comprise entre $2^{-126} \approx 1,175494351\,10^{-38}$ et $(1-2^{-24})\,2^{128} \approx 3,4028235\,10^{38}$ (resp. $2^{-1022} \approx 2,2250738585072020\,10^{-308}$ et $(1-2^{-53})\,2^{1024} \approx 1,7976931348623157\,10^{308}$). Les formats étendus permettent quant à eux d'intégrer à la norme des représentations dont la précision est supérieure à la précision courante, servant habituellement dans les calculs intermédiaires. Le codage du signe se fait par la valeur 0 du bit correspondant si le nombre est positif et 1 s'il est négatif (le nombre 0 étant signé) et l'exposant est *biaisé*, c'est-à-dire qu'on le décale afin le stocker sous la forme d'un entier non signé.

Les valeurs spéciales sont deux « infinis », `+Inf` et `-Inf`, représentant respectivement $+\infty$ et $-\infty$, renvoyés comme résultat d'un dépassement de capacité comme la division par zéro d'une quantité non nulle, le zéro signé, qui correspond aux inverses des infinis et représente [44] le nombre 0, et la valeur `NaN` (de l'anglais *"not a number"*, « pas un nombre » en français), produite par le résultat d'une opération arithmétique invalide [45]. Elles permettent

41. Notons que l'on peut obtenir une meilleure estimation pour la multiplication, à savoir $|\delta| \leq \sqrt{5}\,u$ (voir [BPZ07]).

42. Dans toute la suite, nous supposons la condition $nu < 1$ implicitement vérifiée pour toute valeur de l'entier n envisagée, ceci étant toujours le cas en arithmétique IEEE en simple ou en double précision (voir la sous-section 1.3.4).

43. La version actuelle, publiée en août 2008, de cette norme est IEEE 754-2008. Elle inclue la quasi-totalité de la norme originale IEEE 754-1985, ainsi que la norme IEEE 854-1987 (*IEEE Standard for Radix-Independent Floating-Point Arithmetic*).

44. Par convention, le test $+0 = -0$ est vrai.

45. On l'obtient pour toutes les opérations qui sont des formes indéterminées mathématiques, comme les divisions $0/0$, $+\infty/(+\infty)$, $+\infty/(-\infty)$, $-\infty/(+\infty)$ et $-\infty/(-\infty)$, les multiplications $0(+\infty)$ et $0(-\infty)$, les additions $+\infty+(-\infty)$ et $-\infty+(+\infty)$ (ainsi que les soustractions équivalentes) ou encore les opérations sur les réels dont le résultat est complexe (racine carrée ou logarithme d'un nombre négatif par exemple). La valeur `NaN` est en quelque sorte un *élément absorbant* : toute opération arithmétique la faisant intervenir ou toute fonction mathématique lui étant appliquée la renvoie comme résultat. On dit encore que cette valeur *se propage*.

de définir une arithmétique sur un système *fermé*, au sens où chaque opération renvoit un résultat qui peut être représenté, même si ce dernier n'est pas mathématiquement défini [46].

Enfin, les modes d'arrondi proposés sont au nombre de quatre (l'arrondi au plus proche, qui est celui utilisé par défaut, avec une stratégie d'arrondi au chiffre pair quand un choix est nécessaire, l'arrondi vers zéro et les arrondis vers $\pm\infty$). La norme garantit que les opérations arithmétiques sur les nombres à virgule flottante sont effectuées de manière exacte et que le résultat est arrondi selon le mode choisi. L'arithmétique à virgule flottante qu'elle définit satisfait donc la propriété (1.8), en assurant de plus que $\delta = 0$ lorsque le résultat de l'opération considérée est exactement représentable.

1.4 Propagation des erreurs et conditionnement

Nous avons vu comment des erreurs d'arrondi étaient produites par l'exécution inexacte d'opérations arithmétiques par une machine, mais elles ne sont pas les seules à affecter une méthode numérique. En effet, les données du problème que l'on cherche à résoudre peuvent elles-mêmes contenir des erreurs dont les sources peuvent être diverses (voir la section 1.1). Pour être en mesure d'apprécier la pertinence de la valeur numérique du résultat d'un calcul, il est donc fondamental de pouvoir estimer comment des perturbations, si petites soient-elles, influent sur ce résultat. Cette analyse de sensibilité de la solution d'un problème par rapport à des changements dans les données est l'un des aspects fondamentaux de l'étude de la *propagation des erreurs* et constitue, tout comme l'analyse d'erreur, une étape préliminaire essentielle à l'utilisation d'une méthode numérique. L'objet de cette section est d'introduire des outils généraux permettant de la conduire.

1.4.1 Propagation des erreurs dans les opérations arithmétiques

Nous considérons tout d'abord le cas le plus simple de propagation d'erreurs : celle ayant lieu dans des opérations arithmétiques élémentaires comme la multiplication, la division, l'addition ou la soustraction. Pour l'étudier, nous allons supposer que les opérations sont effectuées de manière *exacte*, mais que leurs opérandes contiennent des erreurs.

Cas de la multiplication

On considère les valeurs $x + \delta x$ et $y + \delta y$, représentant deux réels non nuls x et y respectivement entachés des erreurs absolues $|\delta x|$ et $|\delta y|$. En supposant que les erreurs relatives $\left|\frac{\delta x}{x}\right|$ et $\left|\frac{\delta y}{y}\right|$ sont suffisamment petites en valeur absolue pour que les termes d'ordre deux en ces quantités soient négligeables, on obtient

$$(x+\delta x)(y+\delta y) = xy\left(1+\frac{\delta x}{x}+\frac{\delta y}{y}+\frac{\delta x\,\delta y}{xy}\right) \approx xy\left(1+\frac{\delta x}{x}+\frac{\delta y}{y}\right).$$

L'erreur relative sur le résultat est donc approximativement égale à la somme des erreurs relatives sur les opérandes, ce qui est parfaitement acceptable. Pour cette raison, la multiplication est considérée comme une opération *bénigne* du point de vue de la propagation des erreurs.

Cas de la division

Pour la division, on trouve, sous les mêmes hypothèses que précédemment,

$$\frac{x+\delta x}{y+\delta y} = \frac{x}{y}\left(1+\frac{\delta x}{x}\right)\left(1-\frac{\delta y}{y}+\left(\frac{\delta y}{y}\right)^2-\dots\right) \approx \frac{x}{y}\left(1+\frac{\delta x}{x}-\frac{\delta y}{y}\right).$$

Dans ce cas, l'erreur relative sur le résultat est de l'ordre de la différence entre les erreurs relatives sur les opérandes, ce qui est encore une fois tout à fait acceptable.

46. Dans ce cas précis, les valeurs sont d'ailleurs accompagnées de *signaux d'exception* (*exception flags* en anglais) qu'on peut choisir d'activer pour interrompre le calcul.

Cas de l'addition et de la soustraction

Les nombres réels x et y pouvant être positifs ou négatifs, on ne va s'intéresser qu'à l'addition. On a, en supposant que la somme $x+y$ est non nulle,

$$(x+\delta x)+(y+\delta y)=(x+y)\left(1+\frac{x}{x+y}\frac{\delta x}{x}+\frac{y}{x+y}\frac{\delta y}{y}\right).$$

Lorsque les opérandes sont de même signe, l'erreur relative sur le résultat est majorée par

$$\max\left\{\left|\frac{\delta x}{x}\right|,\left|\frac{\delta y}{y}\right|\right\},$$

et reste donc du même ordre les erreurs relatives sur les opérandes. En revanche, si ces derniers sont de signes opposés, au moins l'un des facteurs $\left|\frac{x}{x+y}\right|$ et $\left|\frac{y}{x+y}\right|$ est plus grand que 1 et au moins l'une des erreurs relatives sur les opérandes est amplifiée, de manière d'autant plus importante que ces opérandes sont presque égaux en valeur absolue, donnant alors lieu au phénomène d'*annulation catastrophique* (*catastrophic cancellation* en anglais).

Un exemple d'annulation catastrophique. Les racines réelles de l'équation algébrique du second degré $a\,x^2+b\,x+c=0$, avec $a\neq 0$, sont respectivement données par les formules $\frac{-b+\sqrt{b^2-4ac}}{2a}$ et $\frac{-b-\sqrt{b^2-4ac}}{2a}$. Dans le cas où $b^2 \gg |4ac|$, on a $\sqrt{b^2-4ac}\approx|b|$ et le calcul de l'une des deux racines (celle dont la valeur absolue est la plus petite) sera affecté par une annulation, dont l'effet est de mettre en avant l'erreur d'arrondi résultant de l'évaluation inexacte de la quantité $\sqrt{b^2-4ac}$. Il est cependant simple d'éviter cette annulation : il suffit de déterminer la racine dont la valeur absolue est la plus grande et d'utiliser ensuite la *relation de Viète*[47] selon laquelle le produit des racines est égal à $\frac{c}{a}$ pour obtenir l'autre racine (voir l'algorithme 3).

Algorithme 3 : Algorithme pour le calcul des racines réelles de l'équation algébrique du second degré $a\,x^2+b\,x+c=0$.

Entrée(s) : les coefficients a, b et c.
Sortie(s) : les racines réelles ξ_1 et ξ_2.
$d = b*b-4*a*c$
si $d \geq 0$ **alors**
  $\xi_1 = -\text{sign}(b)*(\text{abs}(b)+\text{sqrt}(d))/(2*a)$
  $\xi_2 = c/(a*\xi_1)$
fin

Le calcul des racines peut également subir une annulation lorsque celles-ci sont presque égales, c'est-à-dire quand $b^2\approx 4ac$. Dans ce cas, il n'existe pas de façon de garantir le résultat autre que le recours à une arithmétique en précision étendue pour l'évaluation de b^2-4ac.

L'annulation amplifiant, de manière parfois très conséquente, des imprécisions sur les valeurs des opérandes d'une addition ou d'une soustraction (causées par exemple par des erreurs d'arrondi accumulées au fil d'opérations effectuées antérieurement), son effet est potentiellement dévastateur, mais également très difficile à anticiper. Il est néanmoins important de comprendre qu'elle n'est pas forcément une fatalité. Tout d'abord, les données sont parfois connues exactement. D'autre part, l'impact d'une annulation sur un calcul dépend de la contribution du résultat intermédiaire qu'elle affecte au résultat final. Par exemple, si l'on cherche à évaluer la quantité $x+(y-z)$, où x, y et z sont trois nombres réels tels que $x \gg y \approx z > 0$, alors l'erreur due à une annulation ayant lieu lors de la soustraction $y-z$ n'est généralement pas notable dans le résultat obtenu.

1.4.2 Analyse de sensibilité et conditionnement d'un problème

L'étude de la propagation des erreurs sur de simples opérations arithmétiques a montré que la sensibilité d'un problème à une perturbation des données pouvait présenter deux tendances opposées, de petites variations

47. François Viète (ou François Viette, Franciscus Vieta en latin, 1540 - 23 février 1603) était un juriste, conseiller du roi de France et mathématicien français, considéré comme le fondateur de l'algèbre moderne. En parallèle de ses fonctions au service de l'État, il développa une œuvre mathématique importante en algèbre, en trigonométrie, en géométrie, en cryptanalyse et en astronomie.

des données pouvant entraîner, selon les cas, de petits ou de grands changements sur la solution. Ceci nous amène à introduire un cadre d'étude général, dans lequel on identifie la résolution d'un problème donné à une application définie sur l'ensemble des données possibles pour le problème et à valeurs dans l'ensemble des solutions correspondantes.

Problème bien posé

Dans tout la suite, nous allons nous intéresser à la résolution d'un problème de la forme suivante : *connaissant* $\boldsymbol{d}$*, trouver* $\boldsymbol{s}$ *tel que*

$$F(\boldsymbol{s},\boldsymbol{d}) = 0, \tag{1.11}$$

où F désigne une relation fonctionnelle liant la solution $\boldsymbol{s}$ à la donnée $\boldsymbol{d}$ du problème, ces dernières variables étant supposées appartenir à des espaces vectoriels normés sur $\mathbb{R}$ ou $\mathbb{C}$.

Un tel problème est dit *bien posé au sens de Hadamard*[48] *[Had02]* (*well-posed in the sense of Hadamard* en anglais) si, pour une donnée $\boldsymbol{d}$ fixée, une solution $\boldsymbol{s}$ existe, qu'elle est unique et qu'elle dépend continûment de la donnée $\boldsymbol{d}$. La première de ces conditions semble être la moindre des choses à exiger du problème que l'on cherche à résoudre : il faut qu'il admette au moins une solution. La seconde condition exclut de la définition les problèmes possédant plusieurs, voire une infinité de, solutions, car une telle multiplicité cache une indétermination du modèle sur lequel est basé le problème. La dernière condition, qui est la moins évidente *a priori*, est absolument fondamentale dans la perspective de l'utilisation de méthodes numériques de résolution. En effet, si de petites incertitudes sur les données peuvent conduire à de grandes variations sur la solution, il sera quasiment impossible d'obtenir une approximation valable de cette dernière par un calcul dans une arithmétique en précision finie.

Deux exemples de problèmes bien posés pour la résolution desquels des méthodes numériques sont présentées dans le présent document sont ceux de la résolution d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec A une matrice, réelle ou complexe, d'ordre n inversible et $\boldsymbol{b}$ un vecteur de $\mathbb{R}^n$ ou de $\mathbb{C}^n$ donnés, traitée dans les chapitres 2 et 3, et de la détermination des racines d'une équation algébrique, abordée dans la section 5.7.

Conditionnement

De manière abstraite, tout problème bien posé de la forme (1.11) peut être assimilé à une *boîte noire*[49], ayant pour entrée une donnée $\boldsymbol{d}$, qui sera typiquement constituée de nombres[50], et pour sortie une solution $\boldsymbol{s}$ déterminée de manière unique par les données, qui sera elle aussi généralement représentée par un ensemble de nombres. En supposant que la donnée et la solution du problème ne font intervenir que des quantités réelles (la discussion qui suit s'étendant sans difficulté au cas complexe), on est alors en mesure de décrire la résolution du problème par une application φ, définie de l'ensemble des données convenables pour le problème, supposé être un sous-ensemble $\mathscr{D}$ de $\mathbb{R}^n$, et à valeurs dans (un sous-ensemble de) $\mathbb{R}^m$, avec m et n des entiers naturels non nuls, telle que

$$\varphi(\boldsymbol{d}) = \boldsymbol{s},\ \boldsymbol{d} \in \mathscr{D},$$

soit encore, en reprenant la notation précédemment introduite pour le problème,

$$F(\varphi(\boldsymbol{d}),\boldsymbol{d}) = 0,\ \boldsymbol{d} \in \mathscr{D}.$$

L'application φ est généralement non linéaire, mais, en raison de l'hypothèse sur le caractère bien posé du problème, elle est au moins continue. On observe par ailleurs qu'elle n'est pas forcément définie de manière unique : il peut en effet exister plusieurs façons de résoudre un même problème.

Nous allons nous intéresser à la sensibilité de l'application φ par rapport à une (petite) perturbation *admissible* de la donnée ou, plus précisément, à l'estimation de la variation de la solution $\boldsymbol{s}$ due à une modification $\delta\boldsymbol{d}$ de la donnée $\boldsymbol{d}$ telle que $\boldsymbol{d} + \delta\boldsymbol{d}$ appartienne à $\mathscr{D}$. Pour ce faire, on définit le *conditionnement absolu* (*absolute condition number* en anglais) de l'application φ (ou du problème) au point (ou en la donnée) $\boldsymbol{d}$ par

$$\kappa_{\text{abs}}(\varphi,\boldsymbol{d}) = \lim_{\varepsilon\to 0} \sup_{\|\delta\boldsymbol{d}\|\leq\varepsilon} \left\{ \frac{\|\varphi(\boldsymbol{d}+\delta\boldsymbol{d}) - \varphi(\boldsymbol{d})\|}{\|\delta\boldsymbol{d}\|} \right\},$$

48. Jacques Salomon Hadamard (8 décembre 1865 - 17 octobre 1963) était un mathématicien français, connu pour ses travaux en théorie des nombres et en cryptologie.

49. Ce terme désigne un système (que ce soit un objet, un organisme, un mode d'organisation sociale, etc.) connu uniquement en termes de ses entrées, de ses sorties et de sa fonction de transfert, son fonctionnement interne restant totalement inaccessible.

50. On pourrait aussi bien considérer des problèmes impliquant des espaces plus généraux, en particulier des espaces de fonctions, mais on remarquera que, dans la pratique, ceux-ci se trouvent toujours réduits à des espaces de dimension *finie*.

où $\|\cdot\|$ désigne indifféremment une norme sur $\mathbb{R}^m$ ou $\mathbb{R}^n$ et où la limite de la borne supérieure de la quantité est comprise comme sa borne supérieure sur l'ensemble des perturbations *infinitésimales* $\delta\boldsymbol{d}$. Lorsque l'application φ est différentiable au point $\boldsymbol{d}$, le conditionnement absolu s'exprime en fonction de la dérivée de φ au point $\boldsymbol{d}$. Par définition de la différentielle de φ au point $\boldsymbol{d}$ et en introduisant la *matrice jacobienne* $J_\varphi(\boldsymbol{d})$ de φ en $\boldsymbol{d}$,

$$J_\varphi(\boldsymbol{d}) = \begin{pmatrix} \dfrac{\partial \varphi_1}{\partial d_1}(\boldsymbol{d}) & \dots & \dfrac{\partial \varphi_1}{\partial d_n}(\boldsymbol{d}) \\ \vdots & & \vdots \\ \dfrac{\partial \varphi_m}{\partial d_1}(\boldsymbol{d}) & \dots & \dfrac{\partial \varphi_m}{\partial d_n}(\boldsymbol{d}) \end{pmatrix},$$

on a (voir [Ric66])

$$\kappa_{\text{abs}}(\varphi, \boldsymbol{d}) = \|J_\varphi(\boldsymbol{d})\|,$$

où $\|\cdot\|$ désigne la norme matricielle sur $M_{m,n}(\mathbb{R})$ induite par les normes vectorielles choisies sur $\mathbb{R}^m$ et $\mathbb{R}^n$ (voir la proposition A.155).

En pratique [51], c'est souvent la notion de perturbation ou d'erreur *relative* qui est pertinente et l'on utilise alors, en supposant que $\boldsymbol{d}$ et $\boldsymbol{s} = \varphi(\boldsymbol{d})$ sont tous deux non nuls, la quantité

$$\kappa(\varphi, \boldsymbol{d}) = \lim_{\varepsilon \to 0} \sup_{\|\delta \boldsymbol{d}\| \leq \varepsilon} \left\{ \frac{\|\varphi(\boldsymbol{d} + \delta\boldsymbol{d}) - \varphi(\boldsymbol{d})\|}{\|\varphi(\boldsymbol{d})\|} \left(\frac{\|\delta \boldsymbol{d}\|}{\|\boldsymbol{d}\|} \right)^{-1} \right\}, \tag{1.12}$$

appelée *conditionnement relatif* (*relative condition number* en anglais) de l'application φ (ou du problème) au point (ou en la donnée) $\boldsymbol{d}$. Lorsque l'application φ est différentiable, ce conditionnement s'exprime en termes de la matrice jacobienne de φ en $\boldsymbol{d}$ et l'on a

$$\kappa(\varphi, \boldsymbol{d}) = \|J_\varphi(\boldsymbol{d})\| \, \frac{\|\boldsymbol{d}\|}{\|\varphi(\boldsymbol{d})\|}. \tag{1.13}$$

Le conditionnement vise à donner une mesure de l'influence d'une perturbation de la donnée d'un problème bien posé sur sa solution. Ainsi, on dit que le problème (1.11) est *bien conditionné* (*well-conditioned* en anglais) *pour la donnée* $\boldsymbol{d}$ si le nombre $\kappa(\varphi, \boldsymbol{d})$ (ou $\kappa_{\text{abs}}(\varphi, \boldsymbol{d})$ le cas échéant) est « petit » (typiquement de l'ordre de l'unité à quelques centaines), ce qui signifie encore que la variation observée de la solution est grossièrement du même ordre de grandeur que la perturbation de la donnée qui en est à l'origine. Au contraire [52], si le conditionnement est « grand » (typiquement de l'ordre du million et plus), le problème est *mal conditionné* (*ill-conditioned* en anglais).

On notera de plus que, dans toutes ces définitions, on a considéré le problème comme *exactement résolu*, c'est-à-dire que l'application φ est évaluée avec une précision *infinie*. Le conditionnement est par conséquent une propriété *intrinsèque* du problème et ne dépend d'aucune considération algorithmique sur sa résolution. Ceci étant, nous verrons dans la section 1.5 qu'il intervient de manière fondamentale dans l'analyse de stabilité et de précision d'une méthode numérique. Ajoutons que si la valeur du conditionnement dépend de la norme retenue dans sa définition, son ordre de grandeur restera plus ou moins le même quel que soit ce choix, les normes sur un espace vectoriel de dimension finie étant équivalentes.

La notion de conditionnement trouve son origine dans la *théorie des perturbations*. Pour comprendre ceci, considérons un problème pour lequel l'application φ est une fonction réelle d'une variable réelle (on a dans ce cas $m = n = 1$), par ailleurs supposée régulière et faisons l'hypothèse d'une donnée d et d'une solution $s = \varphi(d)$ toutes deux non nulles. En notant δd la perturbation de d et en la supposant suffisamment petite pour que les termes d'ordre supérieur à un dans le développement de Taylor de $\varphi(d + \delta d)$ au point d soient négligeables, on trouve que

$$\varphi(d + \delta d) - \varphi(d) \approx \varphi'(d)\, \delta d, \tag{1.14}$$

51. Ceci est particulièrement vrai dans le domaine d'étude de l'analyse numérique, l'usage d'une arithmétique à virgule flottante en précision finie introduisant, comme on l'a vu, des erreurs relatives plutôt qu'absolues.

52. La séparation entre problèmes bien et mal conditionnés n'est pas systématique. Les valeurs indicatives du conditionnement données ici s'entendent dans le contexte particulier de la résolution numérique d'un problème dans une arithmétique à virgule flottante définie par la norme IEEE (voir encore la sous-section 1.5.2). D'autre part, le conditionnement d'un problème dépendant de sa donnée, un même type de problème peut, selon les cas, être bien ou mal conditionné. Par exemple, on a vu dans la sous-section 1.4.1 que la soustraction de deux nombres réels de même signe est d'autant plus mal conditionnée que ces nombres sont proches l'un de l'autre. De la même façon, nous verrons que le problème de l'évaluation d'un polynôme en un point est d'autant plus mal conditionné que ce point est proche d'une des racines du polynôme.

d'où

$$|\varphi(d+\delta d)-\varphi(d)| \approx \left|\varphi'(d)\right| \, |\delta d| \, .$$

Si l'on s'intéresse à la relation entre l'erreur relative sur la solution et l'erreur relative sur la donnée, on a encore que

$$\left|\frac{\varphi(d+\delta d)-\varphi(d)}{\varphi(d)}\right| \approx \left|\varphi'(d)\,\frac{d}{\varphi(d)}\right| \left|\frac{\delta d}{d}\right|,$$

ces égalités approchées devenant exactes lorsqu'on fait tendre δd vers 0 et que l'on passe à la limite. Il apparaît alors clairement que le facteur d'amplification de l'erreur absolue (resp. relative) sur la donnée correspond à la quantité que l'on a défini comme étant le conditionnement absolu (resp. relatif) de φ au point d, c'est-à-dire [53]

$$\kappa(\varphi,d)_{\text{abs}} = \left|\varphi'(d)\right| \;\; (\text{resp. } \kappa(\varphi,d) = \left|\varphi'(d)\,\frac{d}{\varphi(d)}\right|).$$

Traitons maintenant le cas où les entiers m et n sont arbitraires ; on a alors

$$\boldsymbol{s} = \varphi(\boldsymbol{d}), \text{ avec } \boldsymbol{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix} \in \mathbb{R}^n, \; \boldsymbol{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_m \end{pmatrix} \in \mathbb{R}^m \text{ et } s_i = \varphi_i(\boldsymbol{d}), \; i = 1,\ldots,m.$$

Supposons que les composantes φ_i, $i = 1,\ldots,m$, soient des fonctions régulières des variables d_j, $j = 1,\ldots,n$. L'analyse de sensibilité consiste alors à considérer tour à tour chacune de ces applications comme une fonction d'une seule variable, à perturber la variable en question et à estimer la variation produite. Sous une hypothèse de petites perturbations, on a, au premier ordre, la relation

$$\varphi(\boldsymbol{d}+\delta\boldsymbol{d})-\varphi(\boldsymbol{d}) \approx J_\varphi(\boldsymbol{d})\,\delta\boldsymbol{d}, \tag{1.15}$$

analogue à (1.14), d'où, après avoir fait le choix de normes,

$$\|\varphi(\boldsymbol{d}+\delta\boldsymbol{d})-\varphi(\boldsymbol{d})\| \leq \|J_\varphi(\boldsymbol{d})\|\,\|\delta\boldsymbol{d}\|,$$

au moins en un sens approché, et

$$\frac{\|\varphi(\boldsymbol{d}+\delta\boldsymbol{d})-\varphi(\boldsymbol{d})\|}{\|\varphi(\boldsymbol{d})\|} \leq \|J_\varphi(\boldsymbol{d})\|\,\frac{\|\boldsymbol{d}\|}{\|\varphi(\boldsymbol{d})\|}\,\frac{\|\delta\boldsymbol{d}\|}{\|\boldsymbol{d}\|}.$$

Ces deux inégalités sont *optimales*, au sens où elles deviennent des égalités pour pour une perturbation $\delta\boldsymbol{d}$ appropriée, ce qui justifie les définitions du conditionnement données plus haut. Il faut néanmoins noter que cette analyse, calquée sur le précédent cas scalaire, ne donne qu'une vision *globale* de l'approche perturbative et conduit, pour certains problèmes, à des majorations d'erreurs trop grossières pour correctement rendre compte ce qui est observé en pratique. Le passage aux normes estompe en effet quelque peu le fait que les composantes de la donnée $\boldsymbol{d}$ puissent être d'ordres de grandeur très différents et que les perturbations sont, en général, *relatives par composante* (notamment si des erreurs d'arrondis en sont à l'origine), c'est-à-dire qu'elles satisfont une relation du type

$$\left|\delta d_j\right| \leq \delta \, \left|d_j\right|, \; j = 1,\ldots,n, \; \delta > 0,$$

qui est une condition *a priori* plus contraignante que

$$\|\delta\boldsymbol{d}\| \leq \delta\,\|\boldsymbol{d}\|.$$

Il est néanmoins possible d'adapter de différentes manières la définition du conditionnement afin de prendre en compte la nature des perturbations et de caractériser ainsi plus finement la sensibilité d'un problème. Parmi celles-ci, mentionnons le *conditionnement relatif par composante* (*componentwise relative condition number* en anglais) de l'application φ (ou du problème) au point $\boldsymbol{d}$, par opposition à la définition classique (1.12) du conditionnement

53. Lorsque φ est une fonction positive, on remarque que la seconde quantité coïncide, à la valeur absolue près, avec l'*élasticité* de φ au point d, utilisée en économie pour mesurer un rapport de cause et d'effet.

relatif que l'on qualifie parfois de *normwise relative condition number* dans la littérature anglo-saxonne. On le définit par

$$\kappa_c(\varphi,\boldsymbol{d}) = \lim_{\varepsilon\to 0}\sup_{\|\delta\boldsymbol{d}\|\leq\varepsilon}\left\{\max_{1\leq i\leq m}\left|\frac{\varphi_i(\boldsymbol{d}+\delta\boldsymbol{d})-\varphi_i(\boldsymbol{d})}{\varphi_i(\boldsymbol{d})}\right|\left(\max_{1\leq j\leq n}\left|\frac{\delta d_j}{d_j}\right|\right)^{-1}\right\},$$

et il vaut encore, lorsque l'application φ est différentiable,

$$\kappa_c(\varphi,\boldsymbol{d}) = \left\|\operatorname{diag}(\varphi_1(\boldsymbol{d}),\ldots,\varphi_m(\boldsymbol{d}))^{-1}J_\varphi(\boldsymbol{d})\operatorname{diag}(d_1,\ldots,d_n)\right\|_\infty = \max_{1\leq i\leq m}\left\{\sum_{j=1}^{n}\left|\frac{\partial\varphi_i}{\partial d_j}(\boldsymbol{d})\frac{d_j}{\varphi_i(\boldsymbol{d})}\right|\right\}, \tag{1.16}$$

où $\operatorname{diag}(\varphi_1(\boldsymbol{d}),\ldots,\varphi_m(\boldsymbol{d}))$ et $\operatorname{diag}(d_1,\ldots,d_n)$ désignent des matrices diagonales d'ordre m et n ayant pour éléments diagonaux respectifs les quantités $\varphi_i(\boldsymbol{d})$, $i=1,\ldots,m$, et d_j, $j=1,\ldots,n$. On notera que l'on a supposé qu'aucune des composantes des vecteurs $\boldsymbol{d}$ et $\varphi(\boldsymbol{d})$ n'était nulle, mais il est possible de généraliser la définition pour inclure de telles éventualités sans difficulté.

Nous terminons sur la notion de *conditionnement relatif mixte* (*mixed relative condition number* en anglais) d'un problème, faisant le lien entre une erreur relative mesurée en norme infinie sur la solution et une perturbation relative par composante de la donnée $\boldsymbol{d}$,

$$\kappa_m(\varphi,\boldsymbol{d}) = \lim_{\varepsilon\to 0}\sup_{\|\delta\boldsymbol{d}\|\leq\varepsilon}\left\{\frac{\|\varphi(\boldsymbol{d}+\delta\boldsymbol{d})-\varphi(\boldsymbol{d})\|_\infty}{\|\varphi(\boldsymbol{d})\|_\infty}\left(\max_{1\leq j\leq n}\left|\frac{\delta d_j}{d_j}\right|\right)^{-1}\right\}.$$

Lorsque l'application φ est différentiable, on a l'identité

$$\kappa_m(\varphi,\boldsymbol{d}) = \frac{\left\|J_\varphi(\boldsymbol{d})\operatorname{diag}(d_1,\ldots,d_n)\right\|_\infty}{\|\varphi(\boldsymbol{d})\|_\infty}.$$

Cette dernière définition du conditionnement intervient notamment pour la résolution de systèmes linéaires (voir le quatrième exemple ci-après).

Pour plus détails sur les différentes définitions et propriétés du conditionnement, on pourra consulter l'article [GK93].

Quelques exemples

Les différentes définitions du conditionnement permettent d'étudier, de manière *ad hoc*, la sensibilité du problème que l'on cherche à résoudre numériquement par rapport à de petites perturbations de ses données. Pour illustrer ce propos, nous allons, sur des exemples classiques, définir un conditionnement adapté au problème traité et caractériser des cas pour lesquels le problème est mal conditionné.

Opérations arithmétiques. Commençons par reprendre, à l'aune de la notion de conditionnement, l'analyse de propagation d'erreurs dans les opérations arithmétiques de base menée dans la sous-section 1.4.1. Chaque opération arithmétique considérée ayant pour données deux opérandes (que l'on va ici supposer réels), nous allons modéliser son évaluation par une application définie de $\mathbb{R}^2$ dans $\mathbb{R}$.

Pour la multiplication, on a alors

$$\varphi_*(x,y)=xy,\ \frac{\partial\varphi_*}{\partial x}(x,y)=y \text{ et } \frac{\partial\varphi_*}{\partial y}(x,y)=x,$$

d'où $\kappa(\varphi_*,(x,y)) = \left\|\begin{pmatrix}1 & 1\end{pmatrix}^\top\right\|$, avec $\|\cdot\|$ une norme sur $\mathbb{R}^2$. Ainsi, on a par exemple $\kappa_1(\varphi_*,(x,y))=2$, $\kappa_2(\varphi_*,(x,y))=\sqrt{2}$ ou $\kappa_\infty(\varphi_*,(x,y))=1$, et l'on retrouve que la multiplication est une opération bien conditionnée quelles que soient les valeurs des opérandes.

Pour la division, on a de la même manière

$$\varphi_/(x,y)=\frac{x}{y},\ \frac{\partial\varphi_/}{\partial x}(x,y)=\frac{1}{y} \text{ et } \frac{\partial\varphi_/}{\partial y}(x,y)=-\frac{1}{y^2},$$

d'où $\kappa(\varphi_/,(x,y)) = \left\| \begin{pmatrix} 1 & -1 \end{pmatrix}^\top \right\|$, dont on déduit que la division est une opération toujours bien conditionnée.

En revanche, dans le cas de l'addition et de la soustraction, il vient

$$\varphi_\pm(x,y) = x \pm y,\ \frac{\partial \varphi_\pm}{\partial x}(x,y) = 1 \text{ et } \frac{\partial \varphi_\pm}{\partial y}(x,y) = \pm 1,$$

et alors $\kappa(\varphi_\pm,(x,y)) = \dfrac{1}{|x \pm y|} \left\| \begin{pmatrix} x & y \end{pmatrix}^\top \right\|$. On a par exemple $\kappa_1(\varphi_\pm,(x,y)) = \dfrac{|x|+|y|}{|x \pm y|}$, cette quantité pouvant prendre des valeurs arbitrairement grandes dès que $|x \pm y| \ll |x| + |y|$. L'addition de deux nombres de signes opposés (ou la soustraction de deux nombres de même signe) est par conséquent une opération potentiellement mal conditionnée.

Évaluation d'une fonction polynomiale en un point. Soit une fonction polynomiale p_n, associée à un polynôme réel non identiquement nul de degré n, $n \geq 1$,

$$p_n(x) = a_n\, x^n + a_{n-1}\, x^{n-1} + \cdots + a_1\, x + a_0, \tag{1.17}$$

que l'on cherche à évaluer en un point z de $\mathbb{R}$. La donnée et la solution de ce problème sont alors le réel z et un vecteur $\boldsymbol{a}$ de $\mathbb{R}^{n+1}$ dont les composantes sont les coefficients a_i, $i = 0,\ldots,n$, du polynôme d'une part et la valeur réelle $p_n(z) = \sum_{i=0}^n a_i\, z^i$ d'autre part. Faisons dans un premier temps l'hypothèse que seule la donnée de z est sujette à des perturbations. Dans ce cas, on représente l'évaluation du polynôme par une application $\varphi_{|_a}$ de $\mathbb{R}$ dans $\mathbb{R}$ telle que $\varphi_{|_a}(z) = p_n(z)$, dont le conditionnement relatif au point z, en supposant que z n'est ni nul, ni racine du polynôme, est

$$\kappa(\varphi_{|_a}, z) = \left| p_n'(z) \frac{z}{p_n(z)} \right|.$$

On remarque que le problème d'évaluation est mal conditionné au voisinage d'une racine du polynôme, la valeur de ce conditionnement tendant vers l'infini lorsque z tend vers un zéro de la fonction p_n. Envisageons à présent une perturbation des coefficients du polynôme et introduisons une application $\varphi_{|_z}$ définie de $\mathbb{R}^{n+1}$ dans $\mathbb{R}$ telle que $\varphi_{|_z}(\boldsymbol{a}) = p_n(z)$. Dans ce cas, on a coutume de mesurer la sensibilité du problème d'évaluation au moyen du conditionnement relatif par composante. En supposant que les réels a_i, $i = 0,\ldots,n$, sont non nuls, en observant que l'on a

$$\frac{\partial \varphi_{|_z}}{\partial a_i}(\boldsymbol{a}) = z^i,\ i = 0,\ldots,n,$$

et puisque z n'est pas racine du polynôme, on obtient, en utilisant la définition (1.16), que

$$\kappa_c(\varphi_{|_z}, \boldsymbol{a}) = \frac{\sum_{i=0}^n \left| a_i\, z^i \right|}{|p_n(z)|}.$$

Là encore, ce nombre peut être arbitrairement grand, notamment au voisinage des racines de p_n. L'évaluation d'un polynôme est donc d'autant plus sensible à des incertitudes sur les valeurs de ses coefficients qu'il est évalué près d'une de ses racines.

Détermination des racines d'une équation algébrique. On considère de nouveau la fonction polynomiale p_n de l'exemple précédent, que l'on suppose telle que $a_n \neq 0$ et $a_0 \neq 0$ et que l'on normalise de manière à ce que $a_n = 1$ (ceci ne modifiera pas substantiellement le conditionnement du problème). Soit ξ une *racine simple* du polynôme [54], c'est-à-dire une solution de l'équation $p_n(x) = 0$ telle que

$$p_n'(\xi) \neq 0.$$

Le problème de détermination de la racine ξ connaissant p_n ayant pour donnée un vecteur $\boldsymbol{a}$, formé de l'ensemble des coefficients a_i, $i = 0,\ldots,n-1$, et pour solution la racine ξ, on introduit une application φ, définie

54. On note que, par hypothèse sur les coefficients, la racine ξ est non nulle.

sur $\mathbb{R}^n$ à valeurs dans $\mathbb{C}$, telle que $\varphi(\boldsymbol{a}) = \xi$. Là encore, c'est la notion de conditionnement relatif par composante qui est utilisée en pratique. On obtient alors le *conditionnement de la racine* ξ suivant

$$\mathrm{cond}(\xi) = \kappa_c(\varphi, \boldsymbol{a}) = \frac{1}{|\xi|} \sum_{i=0}^{n-1} \left| \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a})\, a_i \right|.$$

Pour calculer les dérivées partielles de l'application φ, on utilise l'identité

$$(\varphi(\boldsymbol{a}))^n + a_{n-1}\,(\varphi(\boldsymbol{a}))^{n-1} + \cdots + a_1\,\varphi(\boldsymbol{a}) + a_0 = 0.$$

Par différentiation, on a, pour tout entier i compris entre 0 et $n-1$,

$$n\,(\varphi(\boldsymbol{a}))^{n-1} \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a}) + (n-1) a_{n-1}\,(\varphi(\boldsymbol{a}))^{n-2} \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a}) + \cdots + i a_i\,(\varphi(\boldsymbol{a}))^{i-1} \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a}) + (\varphi(\boldsymbol{a}))^i + \cdots + a_1 \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a}) = 0,$$

ce qui se réécrit encore

$$p_n'(\xi) \frac{\partial \varphi}{\partial a_i}(\boldsymbol{a}) + \xi^i = 0.$$

La racine ξ étant supposée simple, on trouve finalement

$$\mathrm{cond}(\xi) = \frac{1}{\left|\xi\, p_n'(\xi)\right|} \sum_{i=0}^{n-1} \left| a_i\, \xi^i \right|.$$

Examinons à présent l'exemple célèbre du *polynôme de Wilkinson*, qui est un polynôme de degré n ayant pour racine les entiers $1, 2, \ldots, n$, c'est-à-dire $p_n(x) = \prod_{\mu=1}^{n} (x - \xi_\mu)$, avec $\xi_\mu = \mu$, $\mu = 1, \ldots, n$. Il a été établi dans [Gau73] que

$$\min_{1 \le \mu \le n} \mathrm{cond}(\xi_\mu) = \mathrm{cond}(\xi_1) \underset{n \to +\infty}{\sim} n^2 \text{ et } \max_{1 \le \mu \le n} \mathrm{cond}(\xi_\mu) \underset{n \to +\infty}{\sim} \frac{1}{(2-\sqrt{2})n\pi} \left(\frac{\sqrt{2}+1}{\sqrt{2}-1} \right)^n.$$

Ceci montre que, lorsque n est grand, la racine conduisant au problème de détermination le plus mal conditionné (qui est l'entier le plus proche de $\frac{n}{\sqrt{2}}$) possède un conditionnement dont la valeur croît exponentiellement avec n (à titre indicatif, celui-ci vaut approximativement $5,3952\,10^{13}$ pour $n = 20$ et $5,5698\,10^{28}$ pour $n = 40$).

Cet exemple frappant illustre combien les racines d'un polynôme écrit sous la forme (1.17) peuvent être sensibles à petites perturbations des coefficients du polynôme. Il est ainsi mal avisé d'essayer de calculer l'ensemble des valeurs propres d'une matrice par recherche des racines du polynôme caractéristique associé, efficacement obtenu par la *méthode de Le Verrier*[55] [Le 40] par exemple, cette approche pouvant se révéler particulièrement imprécise dès que l'ordre de la matrice dépasse quelques dizaines en raison des erreurs sur le calcul des coefficients du polynôme. Il est alors préférable d'employer des méthodes transformant, par une suite de similitudes, la matrice à diagonaliser de manière à y faire « apparaître » les valeurs propres (nous renvoyons le lecteur à la section 4.5.1 pour la présentation d'une telle méthode).

Résolution d'un système linéaire. Considérons à présent à la résolution du système linéaire

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{1.18}$$

où A est une matrice réelle d'ordre n inversible et $\boldsymbol{b}$ est un vecteur de $\mathbb{R}^n$, que l'on peut voir comme une application φ de $\mathbb{R}^{n^2+n}$ dans $\mathbb{R}^n$ qui associe aux données A et $\boldsymbol{b}$ le vecteur $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ de $\mathbb{R}^n$ solution de (1.18). Soit le système linéaire perturbé

$$(A + \delta A)(\boldsymbol{x} + \delta \boldsymbol{x}) = \boldsymbol{b} + \delta \boldsymbol{b},$$

55. Urbain Jean Joseph Le Verrier (11 mars 1811 - 23 septembre 1877) était un astronome, mathématicien et homme politique français. Il devint célèbre lorsque la planète Neptune, dont il avait calculé les caractéristiques comme cause hypothétique des anomalies du mouvement orbital d'Uranus, fut effectivement observée le 23 septembre 1846.

où δA est une matrice d'ordre n, telle que $\|A^{-1}\|\|\delta A\| < 1$ (ce qui implique que la matrice $A+\delta A$ est inversible), et $\delta \boldsymbol{b}$ est un vecteur de $\mathbb{R}^n$. On montre alors (voir la proposition A.169) que

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{1}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|A^{-1}\|\|\delta \boldsymbol{b}\|}{\|A^{-1}\boldsymbol{b}\|}+\|A^{-1}\|\|\delta A\|\right),$$

où $\|\cdot\|$ désigne à la fois une norme vectorielle sur $\mathbb{R}^n$ et la norme matricielle qui lui est subordonnée (voir la proposition A.155). Pour simplifier l'analyse, nous allons supposer que les perturbations des données sont telles que

$$\|\delta A\| \leq \delta \|A\|,\ \|\delta \boldsymbol{b}\| \leq \delta \|\boldsymbol{b}\|, \text{ avec } \delta > 0.$$

On a alors

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{\delta}{1-\delta \|A^{-1}\|\|A\|}\left(\frac{\|A^{-1}\|\|\boldsymbol{b}\|}{\|A^{-1}\boldsymbol{b}\|}+\|A^{-1}\|\|A\|\right),$$

l'égalité au premier ordre en δ étant obtenue pour $\delta A = \delta \|A\|\|A^{-1}\boldsymbol{b}\|\boldsymbol{v}\boldsymbol{w}^\top$ et $\delta \boldsymbol{b} = -\delta \|\boldsymbol{b}\|\boldsymbol{w}$, où $\|\boldsymbol{w}\| = 1$, $\|A^{-1}\boldsymbol{w}\| = \|A^{-1}\|$ et $\boldsymbol{v}$ est un élément dual du dual de $\boldsymbol{x}$ par rapport à la norme $\|\cdot\|$ (voir la définition A.147). On en déduit que le conditionnement du problème est

$$\kappa(\varphi,(A,\boldsymbol{b})) = \frac{\|A^{-1}\|\|\boldsymbol{b}\|}{\|A^{-1}\boldsymbol{b}\|}+\|A^{-1}\|\|A\|,$$

et qu'il vérifie l'encadrement

$$\operatorname{cond}(A) \leq \kappa(\varphi,(A,\boldsymbol{b})) \leq 2\operatorname{cond}(A),$$

où la quantité $\operatorname{cond}(A) = \|A\|\|A^{-1}\|$ est appelée le *conditionnement de la matrice A* (voir la sous-section A.5.4 de l'annexe A).

Pour obtenir une estimation parfois plus satisfaisante de la sensibilité du problème aux perturbations des données, on peut supposer que ces dernières sont la forme

$$\left|(\delta A)_{ij}\right| \leq \delta \left|a_{ij}\right|,\ |\delta b_i| \leq \delta \left|b_j\right|,\ i,j = 1,\ldots,n, \text{ avec } \delta > 0,$$

et envisager une analyse mixte. En introduisant, pour toute matrice M de $M_n(\mathbb{R})$ et tout vecteur $\boldsymbol{v}$ de $\mathbb{R}^n$, les notations $|M|$ et $|\boldsymbol{v}|$ pour désigner la matrice et le vecteur de composantes respectives $|M|_{ij} = \left|m_{ij}\right|$, $1 \leq i,j \leq n$, et $|\boldsymbol{v}|_i = |v_i|$, $1 \leq i \leq n$, et en supposant que $\left\||A^{-1}|\,|\delta A|\right\|_\infty < 1$, on obtient que

$$\frac{\|\delta \boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq \frac{\delta}{1-\delta\ \||A^{-1}|\,|A|\|_\infty}\frac{\left\||A^{-1}|\,|\boldsymbol{b}|+|A^{-1}|\,|A|\,|\boldsymbol{x}|\right\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

L'égalité au premier ordre en δ est atteinte pour les perturbations

$$\delta A = \delta \operatorname{diag}\left(\operatorname{sign}((A^{-1})_{k1}),\ldots,\operatorname{sign}((A^{-1})_{kn})\right)|A|\operatorname{diag}\left(\operatorname{sign}(x_1),\ldots,\operatorname{sign}(x_n)\right)$$

et

$$\delta \boldsymbol{b} = -\delta \operatorname{diag}\left(\operatorname{sign}((A^{-1})_{k1}),\ldots,\operatorname{sign}((A^{-1})_{kn})\right)|\boldsymbol{b}|,$$

où l'indice k est tel que $\left\||A^{-1}|(|\boldsymbol{b}|+|A|\,|\boldsymbol{x}|)\right\|_\infty = \left(|A^{-1}|(|\boldsymbol{b}|+|A|\,|\boldsymbol{x}|)\right)_k$, d'où la valeur suivante pour le conditionnement mixte du problème

$$\kappa_m(\varphi,(A,\boldsymbol{b})) = \frac{\left\||A^{-1}|\,|\boldsymbol{b}|+|A^{-1}|\,|A|\,|\boldsymbol{x}|\right\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

La plus grande valeur atteinte par ce conditionnement pour une matrice A donnée est $2\left\||A^{-1}|\,|A|\right\|_\infty$, la quantité

$$\left\||A^{-1}|\,|A|\right\|_\infty, \tag{1.19}$$

appelée *conditionnement de Bauer–Skeel de la matrice A* [Bau66, Ske79], pouvant être arbitrairement petite par rapport à $\operatorname{cond}_\infty(A)$ en raison d'une propriété d'invariance par *équilibrage des lignes de la matrice A*. On montre en effet facilement que

$$\left\||A^{-1}|\,|A|\right\|_\infty = \min\left\{\operatorname{cond}_\infty(DA)\,|\,D \text{ matrice diagonale inversible d'ordre } n\right\}.$$

Un exemple canonique de matrices ayant un mauvais conditionnement est celui des matrices dites de Hilbert [56]. Une *matrice de Hilbert* H_n d'ordre n, $n \in \mathbb{N}^*$, est une matrice carrée de terme général $(H_n)_{ij} = (i+j-1)^{-1}$, $1 \leq i, j \leq n$, intervenant dans des problèmes d'approximation polynomiale au sens des moindres carrées de fonctions arbitraires [57]. Elle correspond [58] à la *matrice de Gram* [59] associée à la famille $\left(1, x, x^2, \ldots, x^n\right)$ de fonctions puissances d'une variable réelle x relativement au produit scalaire de $L^2([0,1])$, c'est par conséquent une matrice symétrique définie positive (donc inversible). On observe dans le tableau 1.1 que ces matrices sont extrêmement mal conditionnées. En particulier, on peut montrer, grâce à un résultat dû à Szegő [60] [Sze36], qu'on a, asymptotiquement,

$$\mathrm{cond}_2(H_n) \underset{n\to+\infty}{\sim} \frac{(\sqrt{2}+1)^{4(n+1)}}{2^{15/4}\sqrt{n\pi}}.$$

n	$\mathrm{cond}_2(H_n)$
1	1
2	$1{,}92815\ 10^{1}$
5	$4{,}76607\ 10^{5}$
10	$1{,}60263\ 10^{13}$
20	$2{,}45216\ 10^{28}$
50	$1{,}42294\ 10^{74}$
100	$3{,}77649\ 10^{150}$
200	$3{,}57675\ 10^{303}$

TABLE 1.1 – Valeurs numériques arrondies du conditionnement de la matrice de Hilbert H_n relativement à la norme $\|\cdot\|_2$ pour quelques valeurs de l'entier n. On rappelle (voir notamment le théorème A.168) que la valeur de ce conditionnement est égale au produit de la plus grande valeur propre de H_n par la plus grande valeur propre de son inverse. Pour obtenir la matrice ${H_n}^{-1}$, on a utilisé une forme explicite bien connue (voir par exemple [Tod61]) et non cherché à la calculer numériquement (ce qui serait désastreux !).

Cette croissance exponentielle de la valeur du conditionnement en fonction de l'ordre de la matrice rend la résolution d'un système linéaire associé numériquement impossible : pour des calculs effectués en arithmétique en double précision, la solution obtenue n'a plus aucune pertinence dès que $n \geq 20$ (voir la sous-section 1.5.2).

Les *matrices de Vandermonde* [61] d'ordre n

$$V_n = \begin{pmatrix} 1 & x_0 & \ldots & {x_0}^{n-1} \\ 1 & x_1 & \ldots & {x_1}^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & \ldots & {x_{n-1}}^{n-1} \end{pmatrix}, (x_0, \ldots, x_{n-1})^\top \in \mathbb{C}^n,$$

56. David Hilbert (23 janvier 1862 - 14 février 1943) était un mathématicien allemand, souvent considéré comme l'un des plus grands mathématiciens du vingtième siècle. Il a créé ou développé un large éventail d'idées fondamentales, comme la théorie des invariants, l'axiomatisation de la géométrie ou les fondements de l'analyse fonctionnelle.

57. Dans l'article [Hil94], Hilbert pose la question suivante : *« Étant donné un intervalle réel* $[a, b]$*, est-il possible de trouver un polynôme à coefficients entiers p non trivial, tel que la valeur de l'intégrale*

$$\int_a^b p(x)^2 \, \mathrm{d}x$$

soit inférieure à un nombre strictement positif choisi arbitrairement ? ». Pour y répondre, il établit une formule exacte pour le déterminant d'une matrice de Hilbert H_n et étudie son comportement asymptotique lorsque n tend vers l'infini. Il conclut par l'affirmative à la question posée si la longueur de l'intervalle est strictement inférieure à 4.

58. On vérifie en effet que $(H_n)_{ij} = \displaystyle\int_0^1 x^{i+j-2} \, \mathrm{d}x$, $1 \leq i, j \leq n$.

59. Jørgen Pedersen Gram (27 juin 1850 - 29 avril 1916) était un actuaire et mathématicien danois. Il fit d'importantes contributions dans les domaines des probabilités, de l'analyse numérique et de la théorie des nombres.

60. Gábor Szegő (20 janvier 1895 - 7 août 1985) était un mathématicien hongrois. S'intéressant principalement à l'analyse, il est l'auteur de résultats fondamentaux sur les matrices de Toeplitz et les polynômes orthogonaux.

61. Alexandre-Théophile Vandermonde (28 février 1735 - 1er janvier 1796) était un musicien, mathématicien et chimiste français. Son nom est aujourd'hui surtout associé à un déterminant.

dont les coefficients de chaque ligne présentent une progression géométrique, possèdent également la réputation d'être mal conditionnées. Elles apparaissent naturellement dans divers domaines des mathématiques, comme des problèmes d'interpolation polynomiale (voir la section 6.2) ou des problèmes de moments en statistiques, et le comportement de leur conditionnement en fonction du nombre et de la répartition des points (supposés distincts [62]) x_i, $i = 0, \ldots, n-1$, a pour cette raison été particulièrement étudié. Les estimations ci-apprès sont par exemple établies dans [Gau75] :

- $\mathrm{cond}_\infty(V_n) \underset{n\to+\infty}{\sim} \frac{1}{\pi} e^{-\frac{\pi}{4}} e^{n\left(\frac{\pi}{4}+\frac{1}{2}\ln(2)\right)}$ pour des points équirépartis sur l'intervalle $[-1,1]$, i.e. $x_i = 1 - \frac{2i}{n-1}$, $i = 0, \ldots, n-1$;
- $\mathrm{cond}_\infty(V_n) \underset{n\to+\infty}{\sim} \frac{3^{\frac{3}{4}}}{4}(1+\sqrt{2})^n$ pour les racines du *polynôme de Chebyshev* [63] *de première espèce* de degré n, connues sous le nom de *points de Chebyshev*, $x_i = \cos\left(\frac{2i+1}{2n}\,\pi\right)$, $i = 0, \ldots, n-1$;
- $\mathrm{cond}_\infty(V_n) = n$ pour les racines de l'unité, $x_i = e^{\mathrm{i}\frac{2i\pi}{n}}$, $i = 0, \ldots, n-1$.

De fait, on peut montrer (voir [Bec00]) que pour tout choix arbitraire de points *réels* le conditionnement de V_n croît au moins exponentiellement avec n.

Évaluation numérique d'une intégrale par une formule de récurrence. On cherche à évaluer numériquement l'intégrale

$$I_n = \int_0^1 \frac{t^n}{6+t}\,\mathrm{d}t,$$

avec n un entier naturel fixé. En remarquant que l'on a d'une part

$$I_0 = \int_0^1 \frac{\mathrm{d}t}{6+t} = \ln(7) - \ln(6) = \ln\left(\frac{7}{6}\right),$$

et d'autre part

$$\forall n \geq 1,\ I_n = \int_0^1 \left(1 - \frac{6}{6+t}\right) t^{n-1}\,\mathrm{d}t = \int_0^1 t^{n-1}\,\mathrm{d}t - 10\int_0^1 \frac{t^{n-1}}{6+t}\,\mathrm{d}t = \frac{1}{n} - 6\,I_{n-1},$$

on déduit que l'on peut calculer I_n par la récurrence suivante

$$I_0 = \ln\left(\frac{7}{6}\right),\ I_k = \frac{1}{k} - 6\,I_{k-1},\ k = 1, \ldots, n. \tag{1.20}$$

Même en supposant que le calcul numérique de I_k à partir de I_{k-1}, $k \geq 1$, est exact (ce qui n'est pas le cas en pratique puisque l'évaluation de la fraction $\frac{1}{k}$ engendre inévitablement une erreur d'arrondi pour certaines valeurs de l'entier k), le fait que la valeur de I_0 n'est pas représentable en arithmétique en précision finie induit une perturbation et le résultat effectivement obtenu est donc une approximation de la valeur exacte de I_n, d'autant plus mauvaise que n est grand. En effet, en associant à la relation de récurrence (1.20) une application affine [64] φ_k liant I_k à I_0, $k = 1, \ldots, n$, on établit la relation suivante entre les erreurs relatives sur I_0 et I_n

$$\left|\frac{\delta I_n}{I_n}\right| = \kappa(\varphi_n, I_0)\left|\frac{\delta I_0}{I_0}\right|,$$

où, d'après (1.13),

$$\kappa(\varphi_n, I_0) = \left|{\varphi_n}'(I_0)\,\frac{I_0}{I_n}\right|.$$

Par la stricte décroissance de la suite $(t^k)_{k\in\mathbb{N}}$, $t \in]0,1[$, et la propriété de monotonie de l'intégrale, la suite $(I_k)_{k\in\mathbb{N}}$ est strictement décroissante. On a alors

$$\kappa(\varphi_n, I_0) = \left|(-6)^n \frac{I_0}{I_n}\right| > \left|(-6)^n \frac{I_0}{I_0}\right| = 6^n,$$

62. Le déterminant de la matrice V_n valant $\det(V_n) = \prod_{0\leq i<j\leq n-1}(x_j - x_i)$ (la preuve de ce résultat est laissée en exercice), cette dernière est inversible si et seulement si les valeurs x_i, $i = 0, \ldots, n-1$, sont deux à deux distinctes.

63. Pafnuti Lvovich Chebyshev (Пафну́тий Льво́вич Чебышёв en russe, 16 mai 1821 - 8 décembre 1894) était un mathématicien russe. Il est connu pour ses travaux dans le domaine des probabilités et des statistiques.

64. On a en effet $I_1 = 1 - 6\,I_0 = \varphi_1(I_0)$, $I_2 = \frac{1}{2} - 6\,I_1 = \frac{1}{2} - 6 + (-6)^2\,I_0 = \varphi_2(I_0)$, etc.

et le problème est donc extrêmement mal conditionné lorsque n est grand. Notons que l'on aurait pu s'en convaincre en s'intéressant à la relation (1.20), sur laquelle on voit que toute erreur sur la valeur de l'intégrale à une étape va être, *grosso modo*, multipliée par -6 à la suivante, ce qui résulte en son amplification au cours du processus.

On peut toutefois remédier à cette difficulté en « renversant » la relation de récurrence. Il vient alors, pour tout entier naturel p strictement plus grand que 1,

$$I_{k-1} = \frac{1}{6}\left(\frac{1}{k} - I_k\right),\ k = n+p, \ldots, n+1, \tag{1.21}$$

l'inconvénient, non négligeable, étant que l'on ne dispose pas de la valeur I_{n+p} permettant d'initier la récurrence. Mais, par un raisonnement similaire à celui conduit plus haut, on voit que l'erreur relative sur I_n satisfait maintenant

$$\left|\frac{\delta I_n}{I_n}\right| < \left(\frac{1}{6}\right)^p \left|\frac{\delta I_{n+p}}{I_{n+p}}\right|,$$

et, en approchant I_{n+p} par 0, c'est-à-dire en commettant une erreur relative de 100% sur cette valeur, on obtient la majoration

$$\left|\frac{\delta I_n}{I_n}\right| < \left(\frac{1}{6}\right)^p.$$

Pour approcher I_n avec une tolérance inférieure ou égale à une valeur $\varepsilon > 0$, il suffit alors de choisir un entier p vérifiant

$$p \geq -\frac{\ln(\varepsilon)}{\ln(6)}.$$

On observera pour finir que les erreurs d'arrondi, comme celles produites par l'évaluation des fractions dans la relation (1.21), ne constituent pas un problème, car elles sont, tout comme l'erreur sur la valeur « initiale » I_{n+p}, constamment attenuées au cours de la récurrence.

1.5 Analyse d'erreur et stabilité des méthodes numériques

Si le conditionnement d'un problème est très souvent la cause première du manque d'exactitude d'une solution calculée, la méthode numérique utilisée peut également contribuer à l'introduction d'importantes erreurs dans un résultat, même lorsque le problème est par ailleurs bien conditionné. On parle dans ce cas d'*instabilité* de la méthode. Dans cette section, nous donnons les bases de l'analyse d'erreur (*error analysis* en anglais), qui vise à l'appréciation de la précision de la solution calculée et à l'identification des contributions, de l'algorithme (qui génère et propage des erreurs d'arrondi) et du problème (au travers de son conditionnement), à l'erreur observée. C'est par le biais d'une telle analyse qu'est introduite l'importante notion de *stabilité numérique*[65] (*numerical stability* en anglais) d'un algorithme.

1.5.1 Analyse d'erreur directe et inverse

Considérons la résolution d'un problème bien posé par l'évaluation d'une application φ en un point $\boldsymbol{d}$ au moyen d'un algorithme exécuté dans une arithmétique en précision finie et notons $\hat{\boldsymbol{s}}$ le résultat obtenu. L'objectif de l'analyse d'erreur est d'estimer l'effet cumulé des erreurs d'arrondi sur la précision de $\hat{\boldsymbol{s}}$. Pour cela, on peut principalement procéder de deux façons (voir la figure 1.1).

Tout d'abord, on peut, de manière naturelle, chercher à « suivre » la propagation des erreurs d'arrondi à chaque étape intermédiaire de l'exécution de l'algorithme par des techniques similaires à celles de la sous-section 1.4.1. Cette technique porte le nom d'*analyse d'erreur directe* (*forward error analysis* en anglais) et répond à la question : *« Avec quelle précision le problème est-il résolu ? »*. Elle conduit à une majoration ou, plus rarement, une estimation de l'écart entre la solution attendue $\boldsymbol{s}$ et son approximation $\hat{\boldsymbol{s}}$, appelé l'*erreur directe* (*forward error* en anglais). Un de ses principaux inconvénients est que, dans de nombreux des cas, l'étude de la propagation des erreurs intermédiaires devient rapidement une tâche ardue.

65. On notera que cette notion est spécifique aux problèmes dans lesquels les erreurs d'arrondi sont la forme dominante d'erreur, le terme de stabilité pouvant avoir une signification différente dans d'autres domaines de l'analyse numérique, comme celui de la résolution numérique des équations différentielles par exemple (voir le chapitre 8).

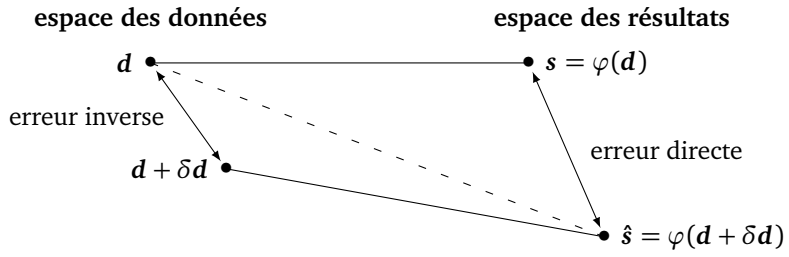


FIGURE 1.1 – Diagramme de représentation des erreurs directe et inverse pour l'évaluation de $\boldsymbol{s} = \varphi(\boldsymbol{d})$. Un trait plein représente une évaluation exacte, un trait discontinu le résultat d'un calcul en précision finie.

On voit par ailleurs que cette méthode prend en compte de façon indifférenciée l'influence de la sensibilité du problème (lorsque la donnée est entachée d'erreurs) et celle de l'algorithme de résolution (qui génère et propage des erreurs d'arrondi) dans l'erreur directe obtenue, ce qui rend son exploitation difficile en pratique.

Avec l'*analyse d'erreur inverse* ou *rétrograde* (*backward error analysis* en anglais), introduite par Wilkinson [Wil60] pour des problèmes d'algèbre linéaire, on contourne la difficulté à conduire l'analyse directe et à interpréter les résultats qu'elle fournit en montrant que la solution calculée est la solution « exacte » du problème que l'on cherche à résoudre *dans lequel la donnée a été perturbée*. En d'autres mots, on identifie la valeur $\hat{\boldsymbol{s}}$ à l'évaluation exacte de l'application φ en un point $\boldsymbol{d} + \delta\boldsymbol{d}$, correspondant à une donnée $\boldsymbol{d}$ perturbée par une *erreur inverse* [66] $\delta\boldsymbol{d}$. On répond alors à la question : *« Quel problème a-t-on effectivement résolu ? »*. Si l'on sait par ailleurs quelle est la sensibilité du problème à une perturbation de la donnée, on est en mesure d'estimer, ou d'au moins majorer, l'erreur directe sur le résultat.

Il est important de retenir que l'analyse inverse procède en deux temps : on cherche tout d'abord à estimer ou majorer (en norme ou bien par composante) l'erreur inverse et on réalise ensuite une analyse de sensibilité du problème à résoudre pour estimer ou majorer à son tour l'erreur directe. Elle présente l'avantage d'identifier clairement la contribution du problème dans la propagation des erreurs et de ramener sa détermination à une étude *générique*, car réalisée pour chaque *type* de problème (et non chaque problème) à résoudre, de conditionnement par les techniques de perturbation linéaire présentées dans la sous-section 1.4.2. Dans la pratique, cette démarche conduit à des estimations souvent plus simples et plus fines que celles fournies par l'analyse directe. Ainsi, si l'erreur inverse estimée n'est pas plus grande que l'incertitude sur la donnée, le résultat calculé sera considéré comme aussi bon que l'autorisent le problème et sa donnée. Cette considération est à la base de la notion de *stabilité inverse* d'un algorithme (voir la définition 1.8).

Indiquons qu'il n'est, dans certains cas, pas possible de conduire une analyse d'erreur inverse, c'est-à-dire d'établir une égalité de la forme $\hat{\boldsymbol{s}} = \varphi(\boldsymbol{d} + \delta\boldsymbol{d})$, mais seulement d'obtenir la relation suivante

$$\hat{\boldsymbol{s}} + \delta\boldsymbol{s} = \varphi(\boldsymbol{d} + \delta\boldsymbol{d}), \tag{1.22}$$

connue sous le nom de résultat d'*erreur mixte directe-inverse* (*mixed forward-backward error* en anglais). Lorsque les quantités $\frac{\|\delta\boldsymbol{d}\|}{\|\boldsymbol{d}\|}$ et $\frac{\|\delta\boldsymbol{s}\|}{\|\boldsymbol{s}\|}$ sont suffisamment petites, cette relation indique que la valeur calculée $\hat{\boldsymbol{s}}$ diffère légèrement de la solution $\hat{\boldsymbol{s}} + \delta\boldsymbol{s}$ produite par une donnée perturbée $\boldsymbol{d} + \delta\boldsymbol{d}$, elle-même légèrement différente de la « vraie » donnée du problème $\boldsymbol{d}$. Le diagramme de la figure 1.2 illustre ce principe.

Parlons pour finir de l'interprétation de l'erreur inverse comme un *résidu normalisé*. Le *résidu* associé à un résultat calculé $\hat{\boldsymbol{s}}$ est la quantité $\varphi(\boldsymbol{d}) - \hat{\boldsymbol{s}}$, qui peut être à valeurs scalaires (c'est le cas pour le problème de détermination des racines d'une équation algébrique), vectorielles (on peut par exemple considérer les problèmes de résolution d'un système linéaire ou de détermination d'éléments propres d'une matrice) ou matricielles (c'est le cas pour le problème de la détermination de l'inverse d'une matrice [67]). Le fait que le résidu associé à la solution d'un problème soit nul laisse à penser qu'un résidu « petit » en norme indique que la solution calculée est une

66. Plus précisément, l'erreur inverse associée à la solution calculée $\hat{\boldsymbol{s}}$ est la plus petite perturbation $\delta\boldsymbol{d}$ de la donnée $\boldsymbol{d}$ telle que la valeur (exacte) $\varphi(\boldsymbol{d} + \delta\boldsymbol{d})$ soit égale à $\hat{\boldsymbol{s}}$.

67. Dans ce problème, on observe que le résidu n'est pas défini de manière unique. Le problème s'énonçant comme *« étant donnée une matrice A d'ordre n, trouver la matrice X vérifiant* $AX = XA = I_n$ », on voit qu'il existe un résidu « à droite » $\hat{X}A - I_n$ et un résidu « à gauche » $A\hat{X} - I_n$, avec $\hat{X}$ la matrice obtenue par calcul numérique de l'inverse.

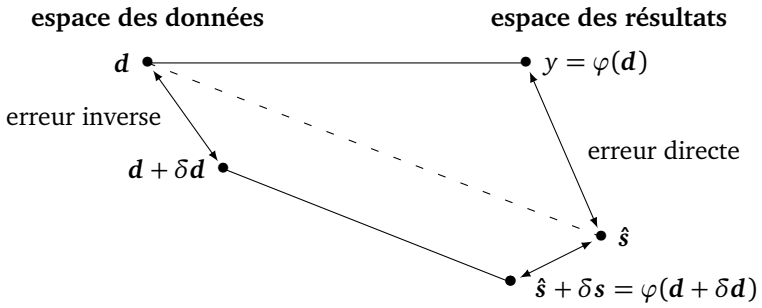


FIGURE 1.2 – Diagramme de représentation de l'erreur mixte directe-inverse pour l'évaluation de $\boldsymbol{s} = \varphi(\boldsymbol{d})$. Un trait plein représente une évaluation exacte, un trait discontinu le résultat d'un calcul en précision finie.

« bonne » approximation de la solution. On peut montrer que ceci est effectivement vrai pour la résolution de systèmes linéaires (voir [OP64, RG67]) ou la détermination des racines d'une équation algébrique, mais ce n'est pas toujours le cas [68]. Une telle propriété s'avère essentielle dans la pratique, car elle montre que le résidu permet, dans certaines situations, d'apprécier, simplement et à moindre coût, la qualité d'une solution calculée.

Quelques exemples (simples) d'analyse d'erreur

Notons pour commencer que la propriété (1.8) du modèle d'arithmétique à virgule flottante standard présenté dans la sous-section 1.3.3 fournit un premier exemple remarquable d'analyse d'erreur inverse. Elle implique en effet que le résultat $\mathrm{fl}(x \,\mathrm{op}\, y)$, où op désigne l'une des quatre opérations arithmétiques de base, correspond au résultat « exact » de l'opération considérée pour les opérandes perturbés $x(1+\delta)$ et/ou $y(1+\delta)$, avec $|\delta| \leq u$.

Pour être en mesure de faire l'analyse d'erreur d'opérations comportant plus de deux opérandes, nous aurons besoin du résultat suivant.

Lemme 1.7 *Soit n un entier naturel non nul tel que $nu < 1$, le nombre réel u désignant la précision machine. Si $|\delta_i| \leq u$ et $\rho_i = \pm 1$, $i = 1, \ldots, n$, alors on a*

$$\prod_{i=1}^{n} (1+\delta_i)^{\rho_i} = 1 + \theta_n,$$

avec $|\theta_n| \leq \gamma_n$, où γ_n est défini par (1.10).

DÉMONSTRATION. Pour démontrer le résultat, on raisonne par récurrence. Au rang un, on a $\theta_1 = \delta_1$ si $\rho_1 = 1$, et alors

$$|\theta_1| \leq u \leq \frac{u}{1-u} = \gamma_1,$$

ou bien $\theta_1 = \frac{1}{1+\delta_1} - 1$ si $\rho_1 = -1$, auquel cas

$$|\theta_1| \leq \frac{1}{1-u} - 1 = \frac{1-(1-u)}{1-u} = \frac{u}{1-u} = \gamma_1.$$

Soit à présent n un entier naturel strictement supérieur à 1. Pour $\rho_n = 1$, il vient

$$\prod_{i=1}^{n} (1+\delta_i)^{\rho_i} = (1+\theta_{n-1})(1+\delta_n) = 1 + \theta_n,$$

d'où

$$\theta_n = \delta_n + \theta_{n-1}(1+\delta_n)$$

68. Deux exemples de problèmes pour lesquels un « petit » résidu ne garantit pas une « petite » erreur inverse sont ceux de détermination de l'inverse d'une matrice et de la résolution de l'*équation de Sylvester* $AX - XB = C$ (voir [Hig93]), où $A \in M_m(\mathbb{R})$, $B \in M_n(\mathbb{R})$ et $C \in M_{m,n}(\mathbb{R})$ sont des matrices données et $X \in M_{m,n}(\mathbb{R})$ est à déterminer.

et

$$|\theta_n| \leq u + \frac{(n-1)u}{1-(n-1)u}(1+u) = \frac{u(1-(n-1)u)+(1+u)(n-1)u}{1-(n-1)u} = \frac{nu}{1-(n-1)u} \leq \gamma_n.$$

Si $\rho_n = -1$, on trouve de la même façon que

$$|\theta_n| \leq \frac{nu-(n-1)u^2}{1-nu+(n-1)u^2} \leq \gamma_n.$$

□

Considérons maintenant le calcul du produit de n nombres à virgule flottante x_i, $i = 1, \ldots, n$. En tenant compte de (1.8), on obtient

$$\mathrm{fl}(x_1 x_2 \ldots x_n) = x_1 x_2 (1+\delta_1) x_3 (1+\delta_2) \ldots x_n (1+\delta_{n-1}),\ |\delta_i| \leq u,\ i = 1, \ldots, n-1,$$

ce qui signifie que le produit calculé est égal au produit « exact » des nombres x_1 et $x_i(1+\delta_i)$, $i = 2, \ldots, n$. On en déduit, grâce au lemme 1.7, la majoration suivante de l'erreur directe

$$|x_1 x_2 \ldots x_n - \mathrm{fl}(x_1 x_2 \ldots x_n)| \leq \gamma_{n-1} |x_1 x_2 \ldots x_n|.$$

Envisageons ensuite le cas d'une somme de n nombres à virgule flottante x_i, $i = 1, \ldots, n$. En sommant dans l'ordre « naturel » des termes, on trouve

$$\mathrm{fl}((\ldots((x_1+x_2)+x_3)+\cdots+x_{n-1})+x_n) = x_1 \prod_{i=1}^{n-1}(1+\delta_i) + x_2 \prod_{i=1}^{n-1}(1+\delta_i) + x_3 \prod_{i=2}^{n-1}(1+\delta_i) + \cdots + x_{n-1}(1+\delta_{n-2})(1+\delta_{n-1}) + x_n(1+\delta_{n-1}),$$

avec $|\delta_i| \leq u$, $i = 1, \ldots, n-1$. Par une utilisation répétée du lemme 1.7, il vient alors

$$\mathrm{fl}((\ldots((x_1+x_2)+x_3)+\cdots+x_{n-1})+x_n) = (x_1+x_2)(1+\theta_{n-1}) + x_3(1+\theta_{n-2}) + \cdots + x_n(1+\theta_1),$$

où $|\theta_i| \leq \gamma_i$, $i = 1, \ldots, n-1$. On observe que la somme calculée est égale à la somme « exacte » des nombres $x_1(1+\theta_{n-1})$ et $x_i(1+\theta_{n+1-i})$, $i = 2, \ldots, n$. Cette analyse d'erreur inverse conduit aux majorations suivantes de l'erreur directe

$$|((\ldots((x_1+x_2)+x_3)+\cdots+x_{n-1})+x_n) - \mathrm{fl}((\ldots((x_1+x_2)+x_3)+\cdots+x_{n-1})+x_n)| \leq \sum_{i=1}^{n} \gamma_{n+1-i} |x_i| \leq \gamma_{n-1} \sum_{i=1}^{n} |x_i|, \quad (1.23)$$

la seconde étant indépendante de l'ordre de sommation. On voit cependant que l'on minimise *a priori* la première majoration en sommant les termes dans l'ordre *croissant* de leur valeur absolue, justifiant ainsi la règle selon laquelle l'erreur d'arrondi sur une somme a tendance [69] à être minimisée lorsque l'on additionne en premier les termes ayant la plus petite valeur absolue (voir l'exemple ci-dessous).

Exemple du calcul numérique d'une série infinie. Supposons que l'on souhaite approcher numériquement la valeur de la série $\sum_{k=1}^{+\infty} k^{-2}$, égale à $\frac{\pi^2}{6} = 1,6449340668482\ldots$. Pour cela, on réalise, avec le format simple précision de la norme IEEE, la somme de ses 10^9 premiers termes pour obtenir la valeur $1,6447253$. On observe alors que seuls quatre chiffres significatifs, sur les huit possibles, coïncident avec ceux de la valeur exacte. Ce manque de précision provient du fait que l'on a additionné les termes, strictement positifs, de la série dans l'ordre décroissant de leur valeur, les plus petits nombres ne contribuant plus à la somme une fois dépassé un certain rang en raison des erreurs d'arrondi. Le remède est d'effectuer la somme dans l'ordre inverse ; avec le même nombre de termes, on trouve alors la valeur $1,6449341$, qui est correcte pour la précision arithmétique considérée.

69. Cette règle est évidemment vraie si les quantités à sommer sont toutes de même signe, mais peut être contredite lorsque leur signe est arbitraire.

Des résultats d'analyse inverse pour les opérations couramment employées en algèbre linéaire peuvent être établis par des raisonnements similaires. Pour le produit scalaire entre deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ de $\mathbb{R}^n$, calculé dans l'ordre « naturel », on a ainsi

$$\mathrm{fl}(\boldsymbol{x}^\top \boldsymbol{y}) = x_1 y_1(1+\Theta_1) + x_2 y_2(1+\Theta_2) + \cdots + x_n y_n(1+\Theta_n), \tag{1.24}$$

avec $|\Theta_1| < \gamma_n$, $|\Theta_i| < \gamma_{n+2-i}$, $i = 2, \ldots, n$. Le produit scalaire calculé est donc égal au produit scalaire « exact » entre les vecteurs $\boldsymbol{x} + \delta\boldsymbol{x}$ et $\boldsymbol{y}$, ou encore $\boldsymbol{x}$ et $\boldsymbol{y} + \delta\boldsymbol{y}$, avec $\delta x_i = \delta y_i = \Theta_i$, $i = 1, \ldots, n$. Comme précédemment, ce résultat dépend de l'ordre dans lequel le produit scalaire est évalué. En observant cependant que chaque perturbation relative est majorée en valeur absolue par γ_n, on arrive à la majoration suivante, indépendante de l'ordre de sommation, de l'erreur directe

$$\left|\boldsymbol{x}^\top \boldsymbol{y} - \mathrm{fl}(\boldsymbol{x}^\top \boldsymbol{y})\right| \leq \sum_{i=1}^{n} \gamma_{n+2-i} \, |x_i| \, |y_i| \leq \gamma_n \sum_{i=1}^{n} |x_i| \, |y_i| = \gamma_n \, |\boldsymbol{x}|^\top \, |\boldsymbol{y}| \, ,$$

dans laquelle $|\boldsymbol{x}|$ et $|\boldsymbol{y}|$ désignent des vecteurs de composantes respectives $|x_i|$ et $|y_i|$, $i = 1, \ldots, n$. Ce résultat reste valable dans un système d'arithmétique à virgule flottante sans chiffre de garde et garantit que l'erreur relative sur le résultat sera petite lorsque, par exemple, $\boldsymbol{y} = \boldsymbol{x}$, car dans ce cas $|\boldsymbol{x}|^\top \, |\boldsymbol{y}| = \left|\boldsymbol{x}^\top \boldsymbol{y}\right|$. On ne peut en revanche rien affirmer si $\left|\boldsymbol{x}^\top \boldsymbol{y}\right| \ll |\boldsymbol{x}|^\top \, |\boldsymbol{y}|$.

L'analyse d'erreur du produit scalaire de deux vecteurs permet d'effectuer simplement celle du produit d'une matrice et d'un vecteur, que l'on peut voir comme une série de produits scalaires entre des vecteurs associés aux lignes de cette matrice et le vecteur en question. Soit A une matrice de $M_{m,n}(\mathbb{R})$ et $\boldsymbol{x}$ un vecteur de $\mathbb{R}^n$. D'après (1.24), on a pour les m produits scalaires entre les vecteurs $\boldsymbol{a}_i$, $i = 1, \ldots, m$, où ${\boldsymbol{a}_i}^\top$ désigne la i^{e} ligne de A, et le vecteur $\boldsymbol{x}$

$$\mathrm{fl}({\boldsymbol{a}_i}^\top \boldsymbol{x}) = (\boldsymbol{a}_i + \delta\boldsymbol{a}_i)^\top \boldsymbol{x}, \ |\delta\boldsymbol{a}_i| \leq \gamma_n \, |\boldsymbol{a}_i| \, , \ i = 1, \ldots, m,$$

l'inégalité entre les vecteurs $|\delta\boldsymbol{a}_i|$ et $|\boldsymbol{a}_i|$ étant entendue composante par composante. On en déduit le résultat suivant

$$\mathrm{fl}(A\boldsymbol{x}) = (A + \delta A)\boldsymbol{x}, \ |\delta A| \leq \gamma_n \, |A| \, , \tag{1.25}$$

où $|A|$ et $|\delta A|$ désignent les matrices d'éléments respectifs $\left|a_{ij}\right|$ et $\left|a_{ij}\right|$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, l'inégalité entre matrices étant comprise élément par élément. Ceci fournit la majoration élément par élément de l'erreur directe suivante

$$|A\boldsymbol{x} - \mathrm{fl}(A\boldsymbol{x})| \leq \gamma_n \, |A| \, |\boldsymbol{x}| \, ,$$

et, en ayant recours à des normes,

$$\|\mathrm{fl}(A\boldsymbol{x}) - A\boldsymbol{x}\|_p \leq \gamma_n \|A\|_p \|\boldsymbol{x}\|_p, \ p = 1, \infty,$$

ou encore [70]

$$\|A\boldsymbol{x} - \mathrm{fl}(A\boldsymbol{x})\|_2 \leq \sqrt{\min(m,n)} \, \gamma_n \|A\|_2 \|\boldsymbol{x}\|_2.$$

Pour traiter le produit de deux matrices $A \in M_{m,n}(\mathbb{R})$ et $B \in M_{n,p}(\mathbb{R})$, on observe que les mêmes erreurs d'arrondis sont produites quel que soit l'ordre des boucles nécessaires au calcul du produit (voir la section 1.2) ; il suffit donc d'en considérer un. En faisant choix de l'ordre « jik », pour lequel les colonnes $A\boldsymbol{b}_j$, $j = 1, \ldots, p$, de la matrice AB sont obtenues une à une, on obtient alors, en utilisant (1.25),

$$\mathrm{fl}(A\boldsymbol{b}_j) = (A + \delta A_j)\boldsymbol{b}_j, \ \left|\delta A_j\right| \leq \gamma_n \, |A| \, , \ j = 1, \ldots, p,$$

et, en notant $|B|$ la matrice d'éléments $\left|b_{ij}\right|$, $i = 1, \ldots, n$, $j = 1, \ldots, p$,

$$|AB - \mathrm{fl}(AB)| \leq \gamma_n \, |A| \, |B| \, ,$$

70. Pour établir cette dernière inégalité, on se sert du fait que si A et B sont deux matrices de $M_{m,n}(\mathrm{R})$ telles que $|A| \leq |B|$, alors $\|A\|_2 \leq \sqrt{\mathrm{rang}(B)}\|B\|_2$. En effet, l'hypothèse implique que $\left|a_{ij}\right| \leq \left|b_{ij}\right|$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, et donc que $\|\boldsymbol{a}_j\|_2 \leq \|\boldsymbol{b}_j\|_2$, $j = 1, \ldots, n$, où les vecteurs $\boldsymbol{a}_j$ et $\boldsymbol{b}_j$ sont les colonnes respectives des matrices A et B. On en déduit trivialement que $\|A\|_F \leq \|B\|_F$ et, en utilisant l'équivalence entre la norme spectrale et celle de Frobenius (voir le tableau A.1), on obtient que

$$\|A\|_2 \leq \|A\|_F \leq \|B\|_F \leq \sqrt{\mathrm{rang}(B)}\|B\|_2.$$

cette dernière majoration étant entendue élement par élément. Les majorations en norme correspondantes sont

$$\|AB - \mathrm{fl}(AB)\|_p \leq \gamma_n \|A\|_p \|B\|_p,\ p = 1, \infty, F,$$

et, pour $p = 2$ (sauf si les éléments de A et B sont positifs),

$$\|AB - \mathrm{fl}(AB)\|_2 \leq n\gamma_n \|A\|_2 \|B\|_2.$$

Terminons par un exemple d'analyse d'erreur mixte concernant la détermination des solutions réelles r_1 et r_2 d'une équation algébrique du second degré $a\,x^2 + b\,x + c = 0$ par l'algorithme 3. Dans ce cas, il a été démontré (voir [Kah72]) que les racines calculées $\hat{r}_1$ et $\hat{r}_2$ satisfont

$$|\tilde{r}_i - \hat{r}_i| \leq \gamma_5 \,|\tilde{r}_i|\,,\ i = 1, 2,$$

où l'on a noté $\tilde{r}_i$, $i = 1, 2$, les racines de l'équation perturbée $a\,x^2 + b\,x + \tilde{c} = 0$, avec $|\tilde{c} - c| \leq \gamma_2\,|\tilde{c}|$.

1.5.2 Stabilité numérique et précision d'un algorithme

L'effet des erreurs d'arrondi sur le résultat d'un algorithme exécuté en arithmétique en précision finie dépend *a priori* des opérations élémentaires composant ce dernier et de leur enchaînement. Ainsi, deux algorithmes associés à la résolution d'un même problème, et par conséquent mathématiquement équivalents, peuvent fournir des résultats numériquement différents. La notion de *stabilité numérique* que nous allons introduire sert à quantifier cet effet et à comparer entre eux des algorithmes. Comme pour l'analyse des erreurs d'arrondi, plusieurs définitions existent. Nous commençons par donner la plus couramment employée, basée sur l'erreur inverse.

Définition 1.8 (stabilité inverse d'un algorithme) *Un algorithme calculant une approximation $\hat{\boldsymbol{s}}$ de la solution d'un problème associée à une donnée $\boldsymbol{d}$ est dit **stable au sens inverse** en arithmétique en précision finie si $\hat{\boldsymbol{s}}$ est la solution exacte du problème pour une donnée $\hat{\boldsymbol{d}}$ telle que, pour une norme $\|\cdot\|$ choisie, on ait*

$$\|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq Cu\,\|\boldsymbol{d}\|,$$

où C est une constante « pas trop grande » et u est la précision machine.

En d'autres mots, un algorithme est stable au sens inverse (*backward stable* en anglais) si son erreur inverse relative est de l'ordre de grandeur de la précision machine u, ce qui signifie encore que le résultat qu'il fournit est la solution exacte du problème pour une donnée « légèrement » perturbée. De fait, on considère dans cette définition qu'un algorithme est stable dès que les erreurs d'arrondi qu'il propage sont du même ordre que les incertitudes présentes sur la donnée, ce qui les rend indiscernables avec la précision arithmétique disponible [71]. Ceci ne garantit évidemment pas que la précision de la solution calculée est bonne. Nous reviendrons sur ce point.

Exemples d'algorithmes stables au sens inverse. La propriété (1.8) du modèle d'arithmétique à virgule flottante standard garantit que les quatre opérations arithmétiques de base sont stables au sens inverse. Il en va de même pour le calcul du produit scalaire entre deux vecteurs de $\mathbb{R}^n$, en vertu de l'égalité (1.24).

Nous verrons que la plupart des méthodes directes « classiques » de résolution de systèmes linéaires (voir le chapitre 2) ou que l'évaluation d'un polynôme en un point par la méthode de Horner (voir la sous-section 5.7.2) sont stables au sens inverse.

Lorsque les données sont à valeurs vectorielles ou matricielles, on peut également mener l'étude de *stabilité inverse par composante* d'un algorithme en mesurant l'erreur par composante plutôt qu'en norme. Ceci à des conditions de stabilité plus restrictives, tout algorithme stable au sens inverse par composante étant en effet stable au sens inverse par rapport à toute norme sans que la réciproque soit forcément vraie.

La définition suivante de la stabilité fait usage de l'erreur directe et de ses liens avec l'erreur inverse.

Définition 1.9 (stabilité directe d'un algorithme) *Un algorithme calculant une approximation $\hat{\boldsymbol{s}}$ de la solution $\boldsymbol{s} = \varphi(\boldsymbol{d})$ d'un problème associée à une donnée $\boldsymbol{d}$ est dit **stable au sens direct** si l'erreur directe sur son résultat est d'un ordre de grandeur similaire à celle produite par d'un algorithme stable au sens inverse résolvant le même problème.*

71. On comprend ici que la notion de stabilité inverse, tout comme celle de bon ou de mauvais conditionnement, est *relative*, la « petitesse » de la constante apparaissant dans la majoration de l'erreur inverse étant en pratique fonction de la précision finie du calculateur.

En vertu de l'analyse de sensibilité réalisée dans la sous-section précédente et de la définition 1.8, cette définition signifie que l'erreur directe d'un algorithme stable au sens direct (*forward stable* en anglais) est telle qu'on ait [72], au premier ordre,

$$\|\hat{\boldsymbol{s}} - \boldsymbol{s}\| \leq \kappa(\varphi, \boldsymbol{d})\, Cu\, \|\boldsymbol{d}\|, \tag{1.26}$$

où $\kappa(\varphi, \boldsymbol{d})$ est le conditionnement relatif du problème en la donnée $\boldsymbol{d}$ et C est une constante « pas trop grande ». Comme le montre l'exemple ci-dessous, un algorithme stable au sens direct ne l'est pas forcément au sens inverse. En revanche, la stabilité inverse entraîne la stabilité directe par définition.

Exemple d'algorithme stable au sens direct. Considérons l'usage de la *règle de Cramer* [73] (voir la proposition A.163) pour la résolution d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ d'ordre 2. Dans ce cas, la solution $\boldsymbol{x}$ est donnée par

$$x_1 = \frac{b_1 a_{22} - b_2 a_{12}}{\det(A)},\ x_2 = \frac{a_{11} b_2 - a_{21} b_1}{\det(A)},\ \text{avec } \det(A) = a_{11}a_{22} - a_{21}a_{12}.$$

Supposons que le déterminant de A est évalué de manière exacte (ceci ne modifiera pas de manière substantielle les majorations que nous allons obtenir). En notant $C = (\mathrm{com}(A))^\top = \left(\begin{smallmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{smallmatrix}\right)$ la transposée de comatrice de A et $\hat{\boldsymbol{x}}$ la solution calculée, on trouve, en utilisant (1.25), que

$$\hat{\boldsymbol{x}} = \mathrm{fl}\left(\frac{1}{\det(A)}\, C\boldsymbol{b}\right) = \frac{1}{\det(A)}\,(C + \delta C)\boldsymbol{b} = \boldsymbol{x} + \frac{1}{\det(A)}\,\delta C\boldsymbol{b},\ \text{avec } |\delta C| \leq \gamma_2\, |C|\,.$$

Il vient alors, en vertu de (A.4),

$$\frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq \gamma_3 \frac{\left\| |A^{-1}|\, |\boldsymbol{b}| \right\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq \gamma_3 \left\| |A^{-1}|\, |A| \right\|_\infty,$$

ce qui implique la stabilité directe de la méthode par définition du conditionnement de Bauer–Skeel de la matrice A (voir (1.19)). Par ailleurs, on a

$$|\boldsymbol{b} - A\hat{\boldsymbol{x}}| \leq \gamma_3\, |A|\, |A^{-1}|\, |\boldsymbol{b}|\,,$$

d'où

$$\frac{\|\boldsymbol{b} - A\hat{\boldsymbol{x}}\|_\infty}{\|\boldsymbol{b}\|_\infty} \leq \gamma_3 \left\| |A|\, |A^{-1}| \right\|_\infty.$$

Le résidu normalisé dans le membre de gauche de cette dernière inégalité constituant une mesure de l'erreur inverse (voir [RG67]), on en déduit que la méthode n'est pas stable au sens inverse.

La définition la plus générale de la stabilité d'un algorithme fait appel à l'analyse d'erreur mixte directe-inverse.

Définition 1.10 (stabilité numérique d'un algorithme) *Un algorithme calculant une approximation $\hat{\boldsymbol{s}}$ de la solution $\boldsymbol{s} = \varphi(\boldsymbol{d})$ d'un problème associée à une donnée $\boldsymbol{d}$ est dit* ***numériquement stable*** *si les quantités $\delta\boldsymbol{d}$ et $\delta\boldsymbol{s}$ dans la relation* (1.22) *sont telles que*

$$\|\delta\boldsymbol{s}\| \leq C_1 u\, \|\boldsymbol{s}\| \text{ et } \|\delta\boldsymbol{d}\| \leq C_2 u\, \|\boldsymbol{d}\|,$$

où C_1 et C_2 sont des constantes « pas trop grandes » et u est la précision machine.

Il découle de cette dernière définition que la stabilité inverse d'un algorithme implique sa stabilité numérique, la réciproque n'étant pas vraie comme le montre l'exemple ci-après.

Exemple d'algorithme numériquement stable. On considère l'évaluation du produit tensoriel $\boldsymbol{x}\boldsymbol{y}^\top$ de deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ de $\mathbb{R}^m$ et $\mathbb{R}^n$ respectivement. Cet algorithme est numériquement stable (par composante), car on vérifie que

$$\mathrm{fl}(x_i y_j) = x_i y_j (1 + \delta_{ij}),\ \left|\delta_{ij}\right| \leq u, i = 1, \ldots, m,\ j = 1, \ldots, n,$$

en vertu de (1.8), mais il n'est pas stable au sens inverse. En effet, la matrice de $M_{m,n}(\mathbb{R})$ ayant pour éléments les réels δ_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, n'étant, en général, pas une matrice de rang 1, le résultat $\mathrm{fl}(\boldsymbol{x}\boldsymbol{y}^\top)$ ne peut s'écrire sous la forme d'un produit tensoriel de vecteurs.

72. Lorsque la donnée est à valeurs vectorielles ou matricielles, on peut également définir la stabilité directe par composante. Dans ce cas, c'est le conditionnement par composante $\kappa_c(\varphi, \boldsymbol{d})$ ou éventuellement mixte $\kappa_m(\varphi, \boldsymbol{d})$ qui intervient dans la majoration.

73. Gabriel Cramer (31 juillet 1704 - 4 janvier 1752) était un mathématicien suisse. Le travail par lequel il est le mieux connu est son traité, publié en 1750, d'*Introduction à l'analyse des lignes courbes algébriques* dans lequel il démontra qu'une courbe algébrique de degré n est déterminée par $\frac{n(n+3)}{2}$ de ses points en position générale.

Terminons par quelques mots sur la précision du résultat calculé par un algorithme. Un algorithme est considéré comme d'autant plus précis que l'erreur directe, mesurée en norme ou par composante, sur le résultat qu'il fournit est petite. Cette erreur satisfaisant la majoration (1.26) au premier ordre, on comprend que la précision du résultat dépend *a priori* à la fois de l'erreur inverse de l'algorithme et du conditionnement du problème que l'on résout de la manière (un peu grossière) suivante

précision $\lesssim$ conditionnement du problème $\times$ erreur inverse de l'algorithme.

Par conséquent, pour un problème mal conditionné, l'erreur directe peut être très importante même si la solution calculée possède une petite erreur inverse, car cette dernière peut se trouver amplifiée par un facteur aussi grand que l'est le conditionnement du problème. Cette remarque conduit à une séparation « pratique » entre problèmes bien et mal conditionnés relativement à la précision finie de l'arithmétique en virgule flottante employée. On peut ainsi considérer qu'un problème est mal conditionné si son conditionnement est d'un ordre de grandeur supérieur à l'inverse de la précision machine, c'est-à-dire

$$\kappa(\varphi, \boldsymbol{d})\, u \gtrsim 1.$$

Par exemple, si l'on cherche à résoudre un problème ayant un conditionnement de l'ordre 10^{10} par un algorithme stable au sens inverse exécuté en arithmétique en double précision de la norme IEEE 754. On a $u = \frac{1}{2}\, 2^{-53-1} \simeq 1,11\ 10^{-16}$ et l'on ne peut donc avoir raisonnablement confiance que dans les six premiers chiffres de la solution calculée.

On notera cependant que la majoration (1.26) est parfois extrêmement pessimiste, comme le montre l'exemple de l'algorithme introduit par Björck et Pereyra dans [BP70] pour résoudre efficacement [74] un système linéaire d'ordre n associé à une matrice de Vandermonde (voir la sous-section 6.2.2 pour une description). On a en effet vu dans la sous-section 1.4.2 qu'un tel problème était généralement très mal conditionné. Or, l'analyse directe montre que la majoration de l'erreur directe, et donc la précision de la méthode, est, dans ce cas, indépendante du conditionnement de la matrice de Vandermonde considérée (voir [Hig87]).

1.6 Notes sur le chapitre

Le mot « algorithme » dérive du nom du mathématicien al-Khwarizmi [75], latinisé au Moyen Âge en *Algoritmi*. Il existe une définition plus formelle de la notion d'algorithme que celle donnée en début de chapitre, basée sur les concepts de *calculabilité effective* et de *machine de Turing* introduits respectivement par Church [76] en 1936 et Turing en 1937 pour répondre négativement au *Entscheidungsproblem* formulé par Hilbert et Ackermann [77] en 1928.

Découvert en 1987, l'algorithme de Coppersmith [78]–Winograd [79] [CW90] permet d'effectuer le produit de deux matrices carrées de manière asymptotiquement plus rapide que l'algorithme de Strassen, sa complexité étant en $O(n^{2,376})$. S'il constitue une brique essentielle dans l'obtention de résultats théoriques de complexité pour d'autres algorithmes, il est cependant pratiquement inutilisable en raison de la présence d'énormes constantes dans les estimations de sa propre complexité.

Dans de nombreux cas, on peut montrer que le conditionnement d'un problème est proportionnel à l'inverse de la distance de ce problème au problème *mal posé* [80] le plus proche. Ce résultat est bien connu pour la résolution

74. Cette méthode ne requiert en effet que de l'ordre de $\frac{5}{2}\, n^2$ opérations arithmétiques.

75. Abu Abdullah Muhammad ibn Musa al-Khwarizmi (أبو عبد الله محمد بن موسى الخورزمي en arabe, v. 780 - v. 850) était un mathématicien, astronome et géographe perse, considéré comme l'un des fondateurs de l'algèbre. Il introduisit dans son aire culturelle les connaissances mathématiques indiennes (notamment le système de numération décimal) et traita de manière systématique la résolution des équations linéaires et quadratiques dans un ouvrage intitulé *« Abrégé du calcul par la restauration et la comparaison »* (كتلب المختصر في حسلب الجبر و المقلبلة en arabe).

76. Alonzo Church (14 juin 1903 - 11 août 1995) était un mathématicien américain, connu pour l'invention du λ-calcul. Il fit d'importantes contributions à la logique mathématique et aux fondements de l'informatique théorique.

77. Wilhelm Friedrich Ackermann (29 mars 1896 - 24 décembre 1962) était un mathématicien allemand. Il est célèbre pour avoir introduit la *fonction d'Ackermann*, qui est un exemple simple de fonction récursive non récursive primitive en théorie de la calculabilité.

78. Don Coppersmith est un mathématicien et cryptologue américain. Il est à l'origine d'algorithmes pour le calcul rapide de logarithmes discret et pour la cryptanalyse de l'algorithme de Rivest, Shamir et Adleman, ainsi que de méthodes pour la multiplication matricielle rapide et la factorisation. Il est aussi l'un des concepteurs des systèmes de chiffrement par bloc *Data Encryption Standard* (*DES*) et *MARS*.

79. Shmuel Winograd (né le 4 janvier 1936) est un informaticien américain. Il est connu pour ses travaux théoriques sur la complexité arithmétique des algorithmes.

80. Par opposition à la définition d'un problème bien posé, un problème mal posé, ou singulier, est un problème possédant plus d'une ou pas de solution ou bien dont la solution ne dépend pas continûment de la donnée.

de systèmes linéaires (voir le théorème A.170), mais il en existe de similaires pour le calcul d'éléments propres d'une matrice ou de racines d'un polynôme. Pour plus de détails sur le sujet, on pourra consulter l'article [Dem87].

Indiquons que la précision d'un résultat calculé peut être garantie par une approche reposant sur l'*arithmétique d'intervalles* [Moo66]. Dans celle-ci, tout nombre se trouve remplacé par un intervalle le contenant et dont les bornes sont représentables dans l'arithmétique en précision finie sous-jacente, ce qui permet de rendre compte à la fois de possibles incertitudes sur une donnée et des arrondis dûs à la représentation des nombres réels en machine. En définissant les opérations arithmétiques et les fonctions de base sur des intervalles élémentaires, on peut alors fournir un intervalle encadrant avec certitude le résultat de tout calcul effectué.

Il est possible de réduire l'effet des erreurs d'arrondi, sans pour autant avoir à augmenter la précision de l'arithmétique à virgule flottante utilisée, en faisant en sorte que celles-ci se compensent. L'exemple le plus connu d'un tel procédé est celui de l'*algorithme de somme compensée de Kahan* (*Kahan's compensated summation algorithm* en anglais) [Kah65], proposé pour le calcul de la somme de nombres à virgule flottante, qui consiste en l'estimation, à chaque addition effectuée, de l'erreur d'arrondi suivie d'une compensation par un terme correctif (voir l'algorithme 4 ci-dessous).

Algorithme 4 : Algorithme de somme compensée de Kahan pour le calcul de $s = \sum_{i=1}^{n} x_i$.

Entrée(s) : les nombres x_i, $i = 1, \ldots, n$.
Sortie(s) : la somme s.

```
s = x_1
c = 0
pour i = 2 à n faire
    y = x_i − c
    t = s + y
    c = (t − s) − y
    s = t
fin
```

En arithmétique à virgule flottante binaire avec chiffre de garde, on peut prouver (voir [Gol91]) que la somme de n nombres à virgule flottante x_i, $i = 1, \ldots, n$, ainsi calculée, ici notée $\hat{s}$, vérifie

$$\hat{s} = \sum_{i=1}^{n} x_i(1+\delta_i),\ |\delta_i| \leq 2u + O\left(nu^2\right),$$

ce qui est un résultat d'erreur inverse pratiquement idéal. La majoration de l'erreur directe correspondante est

$$\left|\sum_{i=1}^{n} x_i - \hat{s}\right| \leq \left(2u + O\left(nu^2\right)\right) \sum_{i=1}^{n} |x_i|,$$

ce dernier résultat étant indépendant de l'ordre de sommation si $nu < 1$. On constate une amélioration significative par rapport à (1.23).

Pour plus de détails sur la stabilité des méthodes numériques, on pourra consulter le très complet et excellent ouvrage de Higham [Hig02] sur le sujet, dont ce chapitre s'inspire en grande partie.

Références

[Bac94] P. BACHMANN. *Die analytische Zahlentheorie*, Band 2 der Reihe *Zahlentheorie*. B. G. Teubner, 1894.

[Bau66] F. L. BAUER. Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. *Z. Angew. Math. Mech.*, 46(7):409–421, 1966. DOI: 10.1002/zamm.19660460702.

[Bec00] B. BECKERMANN. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Math.*, 85(4):553–577, 2000. DOI: 10.1007/PL00005392.

[BP70] Å. BJÖRCK and V. PEREYRA. Solution of Vandermonde systems of equations. *Math. Comput.*, 24(112):893–903, 1970. DOI: 10.1090/S0025-5718-1970-0290541-1.

[BPZ07] R. BRENT, C. PERCIVAL, and P. ZIMMERMANN. Error bounds on complex floating-point multiplication. *Math. Comput.*, 76(259):1469–1481, 2007. DOI: 10.1090/S0025-5718-07-01931-X.

[CW90] D. COPPERSMITH and S. WINOGRAD. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.*, 9(3):251–280, 1990. DOI: 10.1016/S0747-7171(08)80013-2.

[Dem87] J. W. DEMMEL. On condition numbers and the distance to the nearest ill-posed problem. *Numer. Math.*, 51(3):251–289, 1987. DOI: 10.1007/BF01400115.

[Ede97] A. EDELMAN. The mathematics of the Pentium division bug. *SIAM Rev.*, 39(1):54–67, 1997. DOI: 10.1137/S0036144595293959.

[Gau73] W. GAUTSCHI. On the condition of algebraic equations. *Numer. Math.*, 21(5):405–424, 1973. DOI: 10.1007/BF01436491.

[Gau75] W. GAUTSCHI. Norm estimates for inverses of Vandermonde matrices. *Numer. Math.*, 23(4):337–347, 1975. DOI: 10.1007/BF01438260.

[GK93] I. GOHBERG and I. KOLTRACHT. Mixed, componentwise, and structured conditions numbers. *SIAM J. Matrix Anal. Appl.*, 14(3):688–704, 1993. DOI: 10.1137/0614049.

[Gol77] H. H. GOLDSTINE. *A history of numerical analysis from the 16th century through the 19th century*, volume 2 of *Studies in the history of mathematics and physical sciences*. Springer-Verlag, 1977. DOI: 10.1007/978-1-4684-9472-3.

[Gol91] D. GOLDBERG. What every computer scientist should know about floating-point arithmetic. *ACM Comput. Surveys*, 23(1):5–48, 1991. DOI: 10.1145/103162.103163.

[Had02] J. HADAMARD. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton Univ. Bull.*, 13 :49-52, 1902.

[Hig02] N. J. HIGHAM. *Accuracy and stability of numerical algorithms*. SIAM, second edition, 2002. DOI: 10.1137/1.9780898718027.

[Hig87] N. J. HIGHAM. Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems. *Numer. Math.*, 50(5):613–632, 1987. DOI: 10.1007/BF01408579.

[Hig90] N. J. HIGHAM. Exploiting fast matrix multiplication within the level 3 BLAS. *ACM Trans. Math. Software*, 16(4):352–368, 1990. DOI: 10.1145/98267.98290.

[Hig93] N. J. HIGHAM. Perturbation theory and backward error for $AX - XB = C$. *BIT*, 33(1):124–136, 1993. DOI: 10.1007/BF01990348.

[Hil94] D. HILBERT. Ein Beitrag zur Theorie des Legendre'schen Polynoms. *Acta Math.*, 18(1):155–159, 1894. DOI: 10.1007/BF02418278.

[HL14] G. H. HARDY and J. E. LITTLEWOOD. Some problems of diophantine approximation. *Acta Math.*, 37(1):193–239, 1914. DOI: 10.1007/BF02401834.

[Hoa61a] C. A. R. HOARE. Algorithm 63: partition. *Comm. ACM*, 4(7):321, 1961. DOI: 10.1145/366622.366642.

[Hoa61b] C. A. R. HOARE. Algorithm 63: Quicksort. *Comm. ACM*, 4(7):321, 1961. DOI: 10.1145/366622.366644.

[Hoa62] C. A. R. HOARE. Quicksort. *Comput. J.*, 5(1):10–16, 1962. DOI: 10.1093/comjnl/5.1.10.

[Kah65] W. KAHAN. Pracniques: further remarks on reducing truncation errors. *Comm. ACM*, 8(1):40, 1965. DOI: 10.1145/363707.363723.

[Kah72] W. KAHAN. A survey of error analysis. In *Proceedings IFIP Congress, Ljubljana, Information Processing 1971*, pages 1214–1239, 1972.

[Kah83] W. KAHAN. Mathematics written in sand – the hp-15C, Intel 8087, etc. In *Statistical computing section of the proceedings of the American Statistical Association*, pages 12–26, 1983.

[Knu76] D. E. KNUTH. Big Omicron and big Omega and big Theta. *SIGACT News*, 8(2):18–24, 1976. DOI: 10.1145/1008328.1008329.

[Lan09] E. LANDAU. *Handbuch der Lehre von der Verteilung der Primzahlen*. B. G. Teubner, 1909.

[Le 40] U. J. J. LE VERRIER. Sur les variations séculaires des éléments elliptiques des sept planètes principales : Mercure, Vénus, la Terre, Mars, Jupiter, Saturne et Uranus. *J. Math. Pures Appl. (1)*, 5 :220-254, 1840.

[Moo66] R. E. MOORE. *Interval analysis*. Prentice-Hall, 1966.

[OP64] W. OETTLI and W. PRAGER. Compatibility of approximate solution of linear equation with given error bounds for coefficients and right-hand sides. *Numer. Math.*, 6(1):405–409, 1964. DOI: 10.1007/BF01386090.

[RG67] J. L. RIGAL and J. GACHES. On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14(3):543–548, 1967. DOI: 10.1145/321406.321416.

[Ric66] J. R. RICE. A theory of condition. *SIAM J. Numer. Anal.*, 3(2):287–310, 1966. DOI: 10.1137/0703023.

[Ske79] R. D. SKEEL. Scaling for numerical stability in Gaussian elimination. *J. Assoc. Comput. Mach.*, 26(3):494–526, 1979. DOI: 10.1145/322139.322148.

[Ste74] P. H. STERBENZ. *Floating-point computation*. Prentice-Hall, 1974.

[Str69] V. STRASSEN. Gaussian elimination is not optimal. *Numer. Math.*, 13(4):354–356, 1969. DOI: 10.1007/BF02165411.

[Sze36] G. SZEGŐ. On some Hermitian forms associated with two given curves of the complex plane. *Trans. Amer. Math. Soc.*, 40(3):450–461, 1936. DOI: 10.1090/S0002-9947-1936-1501884-1.

[Tod61] J. TODD. Computational problems concerning the Hilbert matrix. *J. Res. Nat. Bur. Standards Sect. B*, 65B(1):19–22, 1961. DOI: 10.6028/jres.065B.005.

[Tre00] L. N. TREFETHEN. *Spectral methods in* MATLAB. SIAM, 2000. DOI: 10.1137/1.9780898719598.

[Tur48] A. M. TURING. Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.*, 1(1):287–308, 1948. DOI: 10.1093/qjmam/1.1.287.

[Wil60] J. H. WILKINSON. Error analysis of floating-point computation. *Numer. Math.*, 2(1):319–340, 1960. DOI: 10.1007/BF01386233.

[Wil64] J. W. J. WILLIAMS. Algorithm 232: Heapsort. *Comm. ACM*, 7(6):347–348, 1964. DOI: 10.1145/512274.512284.

[Wil94] J. H. WILKINSON. *Rounding errors in algebraic processes*. Dover, 1994.

Première partie

Algèbre linéaire numérique

L'*algèbre linéaire numérique* est une branche des mathématiques appliquées consacrée à l'étude de méthodes numériques pour la résolution de problèmes d'algèbre linéaire à l'aide de calculateurs.

Dans la plupart des applications du calcul scientifique issues de la physique, de la mécanique, de la biologie, de la chimie ou encore de la finance (cette liste n'étant évidemment pas exhaustive), l'algèbre linéaire, et plus particulièrement l'analyse matricielle, joue en effet un rôle remarquable. Par exemple, la simulation numérique d'un modèle mathématique se ramene très souvent à faire effectuer par un calculateur une série de calculs matriciels.

Si des questions théoriques fondamentales comme celles de l'existence et de l'unicité de la solution d'un système linéaire ou du fait qu'une matrice soit diagonalisable sont résolues depuis de nombreuses années, le développement de méthodes *robustes* (au sens de la notion de stabilité numérique introduite dans le chapitre 1) et *efficaces* (en termes du coût de calcul et d'occupation de l'espace mémoire requis) est toujours l'objet de recherches actives. De plus, ces méthodes doivent aussi être (ou pouvoir être) adaptées au caractère spécifique du problème à traiter. En effet, dans beaucoup d'applications [1], les matrices qui interviennent possèdent des propriétés [2] ou des structures [3] particulières, dont les algorithmes mis en œuvre doivent impérativement tirer profit.

Les trois prochains chapitres constituent une introduction aux techniques les plus couramment utilisées pour le traitement de deux types de problèmes apparaissant de manière récurrente dans le cadre que nous venons de définir, qui sont la *résolution de systèmes linéaires* et le *calcul des éléments propres d'une matrice*.

1. C'est, par exemple, le cas pour les systèmes linéaires provenant de méthodes de discrétisation utilisées pour la résolution approchées des équations différentielles.

2. On pense ici à des matrices hermitiennes ou symétriques, définies positives, à diagonale dominante, etc.

3. On pense ici à des matrices tridiagonales (par points ou par blocs), bandes ou, plus généralement, creuses, c'est-à-dire contenant beaucoup d'éléments nuls.

Chapitre 2

Méthodes directes de résolution des systèmes linéaires

On considère la résolution d'un système linéaire s'écrivant matriciellement

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{2.1}$$

avec A une matrice d'ordre n à coefficients réels inversible et $\boldsymbol{b}$ une matrice colonne de $M_{n,1}(\mathbb{R})$, par des méthodes dites *directes*, c'est-à-dire fournissant, en l'absence d'erreurs d'arrondi, la solution du système en un nombre *fini* [1] d'opérations élémentaires. Ces méthodes consistent en la construction d'une matrice inversible M telle que MA soit une matrice triangulaire, le système linéaire équivalent (au sens où il possède la même solution) obtenu,

$$MA\boldsymbol{x} = M\boldsymbol{b},$$

étant alors « facile » à résoudre (on verra ce que l'on entend précisément par là). Une telle idée est par exemple à la base de la célèbre *méthode d'élimination de Gauss* [2], qui permet de ramener la résolution d'un système linéaire quelconque à celle d'un système triangulaire supérieur.

Après avoir présenté quelques cas pratiques d'application de ces méthodes et donné des éléments sur la résolution numérique des systèmes triangulaires, nous introduisons dans le détail la méthode d'élimination de Gauss. Ce procédé est ensuite réinterprété en termes d'opérations matricielles, donnant lieu à une méthode de *factorisation* (*factorisation* ou *decomposition* en anglais) des matrices. Les propriétés d'une telle décomposition sont explorées, notamment dans le cas de matrices particulières. Le chapitre se conclut sur la présentation de quelques autres méthodes de factorisation.

2.1 Exemples d'application

Les méthodes de résolution de systèmes linéaires occupent une place centrale au sein des méthodes numériques. Bien entendu, de nombreux problèmes de mathématiques se posent en termes de résolution d'un système d'équations linéaires, comme pour la *méthode des moindres carrés* dans l'exemple qui suit, et le recours à un technique de résolution numérique est alors naturel. Mais il faut également souligner que de nombreuses méthodes numériques font intervenir la résolution de systèmes linéaires, de taille parfois conséquente, au sein d'étapes intermédiaires. C'est le cas pour les méthodes de résolution approchée des équations aux dérivées partielles, dont on donne un premier aperçu dans la sous-section 2.1.2 sur lesquelles nous reviendrons dans les derniers chapitres du cours.

1. On oppose ici ce type de méthodes avec les méthodes dites *itératives*, qui nécessitent *a priori* un nombre infini d'opérations pour fournir la solution. Celles-ci sont l'objet du chapitre 3.

2. Johann Carl Friedrich Gauß (30 avril 1777 - 23 février 1855) était un mathématicien, astronome et physicien allemand. Surnommé par ses pairs *« le prince des mathématiciens »*, il fit des contributions significatives dans de nombreux domaines des sciences de son époque, notamment en théorie des nombres, en statistique, en analyse, en géométrie différentielle, en électrostatique, en astronomie et en optique.

2.1.1 Estimation d'un modèle de régression linéaire en statistique *

Une des plus importantes applications pratiques de la statistique consiste en l'étude de la relation entre une variable observable, supposée aléatoire et dite *variable dépendante* [3], et une ou plusieurs autres variables observables, aléatoires ou non et que l'on qualifie de *variables indépendantes* [4]. Dans le cas de variables statistiques *quantitatives* et étant donné un échantillon de taille n, le *modèle de régression linéaire* suppose un lien de la forme

$$Y_i = \theta_0 + \sum_{j=1}^{p} \theta_j X_{ij} + U_i,\ i = 1, \dots, n, \tag{2.2}$$

entre la variable dépendante Y_i et les variables indépendantes X_{ij}, $j = 1, \dots, p$, $i = 1, \dots, n$; on parle de régression *simple* lorsque $p = 1$, de régression *multiple* sinon. Les coefficients θ_j, $j = 0, \dots, p$, sont les *paramètres* du modèle, et les variables aléatoires U_i, $i = 1, \dots, n$, sont des termes d'erreur résumant l'influence sur les variables dépendantes de facteurs autres que ceux modélisés par les variables indépendantes. Ces dernières quantités sont non observables et doivent par conséquent être estimées. Pour cela, on formule les hypothèses de base suivantes :

- les variables indépendantes sont *exogènes*, ce qui signifie qu'elles ne sont pas corrélées aux erreurs, ce qui se traduit par $E(U_i \,|\, X_{ij}) = 0$, $j = 1, \dots, p$, $i = 1, \dots, n$, si les variables indépendantes sont aléatoires, ou par $E(U_i) = 0$, $i = 1, \dots, n$, si elles sont déterministes, (condition plus faible : $E(U_i) = 0$ et $E(U_i X_{ij}) = 0$ dans le cas aléatoire)
- les erreurs U_i, $i = 1, \dots, n$, possèdent toutes la même variance, qui est indépendante des valeurs des variables X_j *homoscédasticité* (conditionnelle si les X_j sont aléatoires)
- indépendance/non corrélation des erreurs (conditionnellement aux X_j)
- les variables indépendantes (+ constante) ne sont pas colinéaires : la matrice des variables indépendantes est de rang p $(p+1)$ avec probabilité 1.

hypothèse de normalité des erreurs, implique les trois premières hypothèses. Noter que seule la dernière de ces hypothèses est nécessaire à l'estimation des paramètres du modèle, les autres permettent d'obtenir un estimateur non biaisé et/ou efficace (c'est-à-dire de variance minimale).

On considère une réalisation des variables observables dont les valeurs sont y_i et x_{ij}, il découle du modèle que

$$\boldsymbol{y} = X\boldsymbol{\theta} + \boldsymbol{u}$$

où l'on a introduit $\boldsymbol{y} \in \mathbb{R}^n$ le vecteur de composantes y_i, $X \in M_{n,k+1}(\mathbb{R})$ est la matrice

$$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$\boldsymbol{\theta} \in \mathbb{R}^k$ le vecteur de composantes θ_j et $\boldsymbol{u}$ le vecteur des erreurs u_i.

La méthode des moindres carrés consiste à estimer le vecteur $\boldsymbol{\theta}$ de façon à minimiser la somme des carrés des *résidus*, qui sont les différences entre les valeurs y_i observées et les valeurs estimées données par $\sum_j \hat{\theta}_j x_{ij}$, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\inf} \|\boldsymbol{y} - X\boldsymbol{\theta}\|^2.$$

Ce problème admet une unique solution. En effet, en développant la fonctionnelle à minimiser de la manière suivante

$$\|\boldsymbol{y} - X\boldsymbol{\theta}\|^2 = \boldsymbol{y}^T\boldsymbol{y} - 2\,\boldsymbol{\theta}^T X^T \boldsymbol{y} + \boldsymbol{\theta}^T X^T X \boldsymbol{\theta},$$

on obtient que l'estimateur recherché doit satisfaire le système linéaire, dit des *équations normales*

$$X^T X \boldsymbol{\theta} = X^T \boldsymbol{y},$$

la matrice $X^T X$ étant définie positive par construction et en vertu de l'hypothèse ... .

CONCLURE AVEC DES REMARQUES

3. On trouve encore, selon la discipline d'application, les noms de variable *réponse* ou *expliquée*.
4. On trouve aussi les noms de variables *prédictrices* ou *explicatives*.

A SUPPRIMER modèle linéaire gaussien : on ajoute au modèle précédent une hypothèse de normalité sur les erreurs, l'idée sous-jacente étant qu'il existe un vecteur $\boldsymbol{\theta}$ de vraies valeurs mais que l'estimation $\hat{\boldsymbol{\theta}}$ issue d'une série d'observations diffère selon les échantillons obtenus, mais, en vertu du théorème central limite, tend vers $\boldsymbol{\theta}$ en moyenne. Le vecteur $\boldsymbol{\theta}$ est donc une variable aléatoire dont on cherche la distribution. On cherche alors à déterminer des intervalles du type $[\hat{\theta}_j - \epsilon_j, \hat{\theta}_j + \epsilon_j]$ contenant très probablement θ_j.

ici, on suppose que les composantes $e_1, \ldots, e_n$ de $\boldsymbol{e}$ sont des observations indépendantes d'une variable aléatoires E distribuées selon une loi normale centrée de variance σ^2 inconnue ($\boldsymbol{y} \sim \mathcal{N}_n(X\boldsymbol{\theta}, \sigma^2 I_n)$. Cette hypothèse peut se justifier d'une part par un argument théorique/de modélisation, les déviations e_i étant interprétées comme des erreurs de mesure, d'autre part par un argument pratique, car elle est facile à vérifier *a posteriori*.

2.1.2 Résolution numérique d'un problème aux limites par la méthode des différences finies *

A REPRENDRE On considère le problème suivant : *étant donné deux fonctions c et f continues sur l'intervalle $[0,1]$ et deux constantes réelles α et β, trouver une fonction u de classe $\mathscr{C}^2$ sur $[0,1]$ vérifiant*

$$-u''(x) + c(x)u(x) = f(x),\ 0 < x < 1, \tag{2.3}$$

$$u(0) = \alpha,\ u(1) = \beta. \tag{2.4}$$

Un tel problème est appelé *problème aux limites*, car la fonction cherchée doit satisfaire des *conditions aux limites* (2.4) posées aux bornes de l'intervalle ouvert sur lequel l'équation différentielle (2.3) doit être vérifiée. Cette équation intervient dans la modélisation de divers phénomènes physiques (c'est en particulier une version indépendante du temps des équations linéaires *de la chaleur* ou *des ondes* en une dimension d'espace). Sous certaines conditions sur la fonction c, on peut montrer qu'il existe une unique solution au problème (2.3)-(2.4). Nous supposons que c est positive sur l'intervalle $[0,1]$, qui est une hypothèse suffisante, mais pas nécessaire, pour avoir existence et unicité. Sauf dans de rares cas, on ne connaît pas de solution explicite de (2.3)-(2.4) et on doit avoir recours à des méthodes d'approximation numérique de la solution. Nous faisons ici appel à l'une des plus simples d'entre elles, la *méthode des différences finies*.

Étant donné un entier $n \geq 1$, on commence par diviser l'intervalle $[0,1]$ en $n+1$ sous-intervalles de tailles égales, en posant

$$h = \frac{1}{n+1}$$

et en définissant un *maillage* uniforme de *pas* h comme étant l'ensemble des points $x_i = ih$, $0 \leq i \leq n+1$, appelés *nœuds* du maillage. La méthode des différences finies est alors un moyen d'obtenir une approximation de la solution de (2.3)-(2.4) aux nœuds du maillage. Plus précisément, on cherche un vecteur

$$\boldsymbol{u}_h = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathbb{R}^n,$$

tel que la valeur u_i soit « proche » de celle de $u(x_i)$, $i = 1, \ldots, n$, les valeurs de la solution aux points $x_0 = 0$ et $x_{n+1} = 1$ étant déjà connues.

Pour calculer le vecteur $\boldsymbol{u}_h$ des valeurs approchées de la solution u, le principe est de tout d'abord remplacer l'équation différentielle (2.3) par un système de n équations algébriques, obtenu en écrivant (2.3) en chaque nœud x_i, $1 \leq i \leq n$, du maillage et en substituant ensuite à chaque valeur $u''(x_i)$ une combinaison linéaire appropriée de valeurs de la fonction u en certains points du maillage. En effet, en supposant que u est quatre fois continûment dérivable sur l'intervalle $[0,1]$, on peut écrire, par la formule de Taylor–Lagrange (voir le théorème B.116), pour tout $i = 1, \ldots, n$,

$$u(x_{i+1}) = u(x_i) + h\,u'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(x_i - \theta_i^+ h),$$

et

$$u(x_{i-1}) = u(x_i) - h\,u'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)}(x_i) + \frac{h^4}{24}u^{(4)}(x_i + \theta_i^- h),$$

avec θ_i^+ et θ_i^- deux réels strictement compris entre 0 et 1. On en déduit, en sommant ces deux égalités et en utilisant le théorème des valeurs intermédiaires (voir le théorème B.89), que

$$u''(x_i) = \frac{u(x_{i-1}) - 2\,u(x_i) + u(x_{i+1})}{h^2} + \frac{h^2}{12}\,u^{(4)}(x_i + \theta_i h), \text{ avec } |\theta_i| \leq \max\{\theta_i^+, \theta_i^-\},\ 1 \leq i \leq n.$$

En négligeant le terme d'ordre deux en h dans cette dernière relation, on obtient une approximation de la dérivée seconde de la fonction u au nœud x_i, $1 \leq i \leq n$, du maillage correspondant au *schéma aux différences finies centrées* suivant

$$u''(x_i) \approx \frac{u(x_{i-1}) - 2\,u(x_i) + u(x_{i+1})}{h^2}, \tag{2.5}$$

en faisant le raisonnement, heuristique, que l'erreur commise sera d'autant plus « petite » que le pas h sera petit.

En notant alors $c_i = c(x_i)$ et $f_i = f(x_i)$, $1 \leq i \leq n$, pour alléger l'écriture, en substituant à $u''(x_i)$, $1 \leq i \leq n$, son approximation par le schéma aux différences finies (2.5) puis en remplaçant chaque valeur $u(x_i)$ par son approximation u_i, $1 \leq i \leq n$, on aboutit à un *problème discret*, associé à (2.3)-(2.4) et au maillage de pas h de l'intervalle $[0,1]$, qui prend la forme de la résolution d'un système linéaire : *trouver le vecteur* $\boldsymbol{u}_h$ *de* $\mathbb{R}^n$ *vérifiant*

$$A_h \boldsymbol{u}_h = \boldsymbol{b}_h, \tag{2.6}$$

avec

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2+c_1h^2 & -1 & & & \\ -1 & 2+c_2h^2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+c_{n-1}h^2 & -1 \\ & & & -1 & 2+c_nh^2 \end{pmatrix} \text{ et } \boldsymbol{b}_h = \begin{pmatrix} f_1 + \alpha\,h^{-2} \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n + \beta\,h^{-2} \end{pmatrix}.$$

La matrice A_h est dite *tridiagonale*, car elle ne possède des éléments non nuls que sur sa diagonale principale et les sous et sur-diagonales qui lui sont adjacentes. On remarque également que A_h et $\boldsymbol{b}_h$ sont directement calculables à partir des *données* du problème (2.3)-(2.4) que sont les fonctions c et f et les valeurs α et β.

On peut montrer que la matrice A_h est inversible (elle est symétrique et définie positive sous l'hypothèse que la fonction c est positive), c'est-à-dire que le système linéaire (2.6) admet une unique solution $\boldsymbol{u}_h$, qui fournit de plus une approximation convenable de la solution u au sens où la quantité

$$\max_{1 \leq i \leq n} |u(x_i) - u_i|$$

tend vers zéro quand l'entier n tend vers l'infini (on dit alors que la méthode des différences finies appliquée au problème (2.3)-(2.4) *converge*). Ce dernier résultat est relativement délicat à établir et nous renvoyons au chapitre 3 de [Cia98] (dont on s'est d'ailleurs inspiré pour cet exemple) pour une preuve.

On voit donc confirmée l'intuition, quelque peu heuristique, qui a conduit à l'établissement du problème discret et voulait que la qualité de l'approximation obtenue soit d'autant meilleure que le pas h est petit et, par voie de conséquence, l'entier n grand. On peut donc être amené à résoudre des systèmes linéaires de taille importante, la méthode des différences finies pouvant être appliquée à de nombreuses autres classes de problèmes aux limites en une, deux ou trois [5] dimensions d'espace. Cette résolution peut se faire au moyen des méthodes présentées dans le présent chapitre ou le suivant.

2.2 Stockage des matrices

Pour stocker une matrice carrée d'ordre n, réelle ou complexe, il suffit de placer en mémoire les n^2 éléments qui la définissent. Cependant, lorsque les matrices considérées possèdent une structure particulière, il peut être avantageux d'en tirer parti, notamment pour optimiser leur stockage. Ainsi, si la matrice du problème est symétrique (ou hermitienne dans le cas complexe), on a seulement besoin de stocker la partie triangulaire supérieure (ou inférieure) de la matrice, ce qui nécessite de garder $\frac{1}{2}\,n(n+1)$ nombres à virgule flottante en mémoire.

Dans les applications, il est courant que les matrices avec lesquelles on travaille possèdent de nombreux éléments nuls. Plus précisément, on parle de *matrice creuse* (*sparse matrix* en anglais) lorsque le nombre de

5. Dans ce dernier cas, la taille typique des systèmes linéaires couramment résolus est de plusieurs millions.

coefficients non nuls d'une matrice est petit devant le nombre total de coefficients qu'elle contient (typiquement de l'ordre de n pour une matrice carrée d'ordre n, avec n grand). Par exemple, une matrice tridiagonale est une matrice creuse et les matrices bandes (voir la définition A.105) produites par les méthodes de résolution approchée d'équations aux dérivées partielles sont, en général, creuses. Ce type de matrice intervient aussi dans de nombreux autres domaines, comme en analyse combinatoire, et plus particulièrement en théorie des réseaux et en *recherche opérationnelle*[6] (il semble que l'appellation soit due à Markowitz[7]).

Pour de telles matrices, on va chercher à ne stocker, dans la mesure du possible, que les coefficients non nuls, de manière à profiter d'un gain substantiel de place en espace mémoire par rapport à un stockage classique dès que l'on travaillera avec des matrices de grande taille, mais aussi d'un temps d'accès réduit aux données utiles, celles-ci étant en principe placées de manière contiguë en mémoire. Différentes techniques de stockage existent et conduisent à l'emploi d'algorithmes spécialisés, adaptés à la structure de données choisie et dont la complexité se trouve généralement réduite par rapport à celle des algorithmes classiques, pour manipuler ou effectuer des opérations sur les matrices creuses.

On va décrire ci-après quelques-uns des formats les plus couramment utilisés (la liste ne se voulant pas exhaustive). Dans toute la suite de cette section, on désigne par A une matrice creuse d'ordre n, *a priori* non symétrique, contenant un nombre `n_nonzero` d'éléments non nuls.

Tout d'abord, il est aisé de voir que, si A est une matrice bande de demi-largeur de bande inférieure valant p et de demi-largeur de bande supérieure valant q, le stockage peut se faire dans un tableau de nombres à virgule flottante de taille $(p+q+1)n$. On parle alors de *stockage bande* (*band storage* en anglais), en ligne ou en colonne selon que la matrice est parcourue ligne par ligne ou colonne par colonne lors du stockage. Dans le premier cas, si l'on cherche à déterminer l'indice k de l'élement du tableau contenant l'élément a_{ij} de la matrice A, on se sert du fait que le premier coefficient de la i^{e} ligne de A, c'est-à-dire a_{ii-p}, est stocké dans le tableau de nombres réels `matrix_elements` à la $(p+q+1)(i-1)+1^{\text{e}}$ position et on en déduit que $k=(p+q+1)(i-1)+j-i+p+1$. On notera que certains des éléments du tableau de stockage ne sont pas affectés, mais leur nombre, égal à $\frac{1}{2}(p(p-1)+q(q-1))$, reste négligeable. Ce stockage peut être utilisé sans modification si la matrice est triangulaire inférieure (on a alors $p=0$) ou supérieure (on a alors $q=0$). Il peut aussi être adapté pour une matrice symétrique ou hermitienne, en ne stockant que sa partie triangulaire supérieure ou inférieure et en précisant quelle est la partie stockée.

Exemple de stockage bande d'une matrice creuse. La matrice bande suivante, pour laquelle $p=2$ et $q=1$,

$$A=\begin{pmatrix}1&3&0&0&0\\2&5&0&0&0\\6&0&4&3&0\\0&0&2&7&9\\0&0&8&6&5\end{pmatrix}$$

a pour stockage bande en ligne le tableau

`matrix_elements(:)`	*	*	1	3	*	2	5	0	6	0	4	3	0	2	7	9	8	6	5	*

le symbole * désignant un élément du tableau non affecté avec la valeur d'un élément de A lors de son stockage.

Le *stockage profil* (*skyline storage (SKS)* en anglais) permet dans certains cas une réduction de l'espace requis par rapport au stockage bande. En effet, dans ce dernier, tous les termes situés à l'intérieur d'une bande fixée sont placés en mémoire. Avec le stockage profil, la largeur de cette bande varie d'une ligne (resp. colonne) à l'autre, commençant au premier élément non nul et finissant au dernier élément non nul, le *profil* d'une matrice étant défini comme la suite des indices de colonne (resp. de ligne) des premiers éléments non nuls de chaque ligne (resp. colonne) de cette matrice. Ce format est surtout utilisé avec les matrices symétriques, pour lesquelles on ne stocke que la partie triangulaire inférieure (stockage par lignes) ou supérieure (stockage par colonnes). Dans le premier

6. La recherche opérationnelle est la discipline regroupant le développement et l'application de méthodes, de modèles conceptuels et d'outils mathématiques et informatiques permettant de rationaliser le processus de prise de décision, généralement dans un but d'optimisation ou de contrôle.

7. Harry Max Markowitz (né le 24 août 1927) est un économiste américain, lauréat du prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel en 1990. Il est un des pionniers de la *théorie moderne du portefeuille*, ayant étudié dans sa thèse, soutenue en 1952, comment la diversification permettait d'améliorer le rendement d'un portefeuille d'actifs financiers tout en en réduisant le risque.

cas, on stocke ligne par ligne les éléments contenus dans le profil de la matrice jusqu'à la diagonale incluse dans un tableau de nombres à virgule flottante `matrix_elements`. Le tableau d'entiers `row_ptr` de longueur n indique alors la position du premier élément non nul de chaque ligne dans le tableau `matrix_elements`. Les indices de colonne des éléments stockés pouvant être facilement déduits, ils ne sont pas conservés. EXPLICATIONS ?

Une manière de stocker une matrice non symétrique consiste alors à utiliser le stockage profil par lignes décrit ci-dessus pour sa partie triangulaire inférieure et un stockage profil par colonnes pour sa partie triangulaire supérieure.

Ce mode de stockage est particulièrement adapté à la résolution de systèmes par factorisation LU, les profils inférieur et supérieur des matrices L et U de la factorisation étant les mêmes que ceux de A.

Exemple de stockage profil d'une matrice creuse. Le stockage profil de la matrice

$$A = \begin{pmatrix} 1 & -1 & -3 & 0 & 0 \\ -2 & 5 & 0 & 0 & 0 \\ 0 & 0 & 4 & 6 & 4 \\ -4 & 0 & 2 & 7 & 0 \\ 0 & 8 & 0 & 0 & 5 \end{pmatrix}$$

est donné par les tableaux suivants pour sa partie triangulaire inférieure

lower_matrix_elements(:)	1	-2	5	4	-4	0	2	7	8	0	0	5

row_ptr(:)	1	2	4	5	9

et par les tableaux suivants pour sa partie triangulaire supérieure

upper_matrix_elements(:)	1	-1	5	-3	0	4	6	7	4	0	5

col_ptr(:)	1	2	4	7	9

On observe sur les deux exemples précédents que les stockages bande ou profil peuvent dans certains cas conduire à placer en mémoire des éléments nuls de la matrice (tous ceux contenus dans la bande par exemple). Pour éviter ce stockage inutile, on a recours à d'autres techniques, comme celle, intuitive et flexible, du *stockage par coordonnées* (*coordinate list (COO) storage* en anglais) utilisé dans le format d'échange *Matrix Market* pour les matrices creuses. Cette dernière consiste en la création d'un tableau de nombres à virgule flottante et de deux tableaux d'entiers, ici respectivement nommés `values`, `row_ind` et `col_ind`, chacun de taille `n_nonzero` et contenant respectivement (dans un ordre arbitraire) les éléments non nuls de la matrice, leurs indices de ligne et leurs indices de colonne.

Exemple de stockage par coordonnées d'une matrice creuse. Un stockage par coordonnées possible de la matrice A de l'exemple précédent est

matrix_elements(:)	5	8	6	7	2	-1	-2	-4	1	-3	5	4	4

row_ind(:)	5	5	3	4	4	1	2	4	1	1	2	3	3

col_ind(:)	5	2	4	4	3	2	1	1	1	3	2	3	5

On remarquera que ce mode de stockage ne fait pas hypothèse sur la structure de la matrice creuse et ne stocke effectivement que ses éléments non nuls. Si l'ordre de parcours des éléments de la matrice à stocker est en théorie quelconque, on utilise généralement un balayage ligne par ligne ou colonne par colonne. Dans ces deux cas, l'un des tableaux `row_ind` et `col_ind` contient des informations redondantes et peut être remplacé de manière à réaliser des économies d'espace mémoire, au détriment cependant de la facilité d'accès aux éléments, qui doit se faire par un adressage indirect.

Ainsi, le *stockage morse* (*compressed row storage (CRS)* en anglais) utilise également un tableau de nombres à virgule flottante et deux tableaux d'entiers, ici respectivement nommés `matrix_elements`, `col_ind` et `row_ptr`. Le premier, de taille `n_nonzero`, contient les valeurs des éléments non nuls de la matrice, tels qu'on les trouve en parcourant la matrice ligne par ligne. Le deuxième tableau, également de taille `n_nonzero`, contient les indices de colonne des éléments non nuls stockés (on trouvera la valeur j dans `col_ind(k)` si `matrix_elements(k)` contient l'élément a_{ij}). Le troisième tableau, de taille n, contient des entiers qui sont les positions dans le

tableau `matrix_elements` du premier élément non nul de chaque ligne de la matrice. Au total, on a besoin de `2n_nonzero`+n cases en mémoire pour le stockage de la matrice. Notons qu'il existe une version de ce procédé de stockage dans laquelle la matrice est parcourue en colonne par colonne (on parle de *compressed column storage (CCS)* en anglais), par exemple utilisée dans le format de fichier de la collection Harwell–Boeing.

Exemple de stockage morse d'une matrice creuse. Le stockage morse de la matrice A de l'exemple précédent est donné par les tableaux

matrix_elements(:)	1	-1	-3	-2	5	4	6	4	-4	2	7	8	5

col_ind(:)	1	2	3	1	2	3	4	5	1	3	4	2	5

row_ptr(:)	1	4	6	9	12

Pour des matrices creuses dont la largeur de bande varie très peu d'une ligne à l'autre, on peut envisager un *stockage diagonal* (*compressed diagonal storage (CDS)* en anglais), c'est-à-dire de stocker consécutivement l'ensemble des diagonales de la matrice contenues à l'intérieur de la bande. Pour une matrice bande d'ordre n et de largeur de bande égale à $p+q+1$, on se sert pour cela d'un tableau de nombres à virgule flottante de taille $n \times (p+q+1)$, dont chaque colonne contient une diagonale. Comme pour le stockage bande précédemment introduit, ceci implique de stocker certains coefficients nuls de la matrice, mais aussi des coefficients n'existant pas dans la matrice (les diagonales ne contenant pas toutes n coefficients). En contrepartie, ce type de stockage permet un calcul efficace du produit de la matrice avec une matrice colonne. EXPLICATION ?

Exemple de stockage diagonal d'une matrice creuse. Le stockage diagonal de la matrice

$$A = \begin{pmatrix} 10 & -3 & 0 & 0 & 0 & 0 \\ 3 & 9 & 6 & 0 & 0 & 0 \\ 0 & 7 & 8 & 7 & 0 & 0 \\ 1 & 0 & 8 & 7 & 5 & 0 \\ 0 & 1 & 0 & 9 & 9 & 13 \\ 0 & 0 & 1 & 0 & 2 & -1 \end{pmatrix},$$

de largeur de bande égale à $5 (= 3+1+1)$, est donné par le tableau suivant

matrix_elements(:,-3)	*	*	*	1	1	1
matrix_elements(:,-2)	*	*	0	0	0	0
matrix_elements(:,-1)	*	3	7	8	9	2
matrix_elements(:,0)	10	9	8	7	9	-1
matrix_elements(:,1)	-3	6	7	5	13	*

le symbole * désignant un élément du tableau non affecté avec la valeur d'un élément de C lors de son stockage. Dans ce tableau, un indice négatif de colonne indique qu'il s'agit d'une sous-diagonale, un indice nul qu'il s'agit de la diagonale principale, un indice positif qu'il s'agit d'une sur-diagonale.

AJOUTER *stockage diagonal indenté* (*jagged diagonal storage (JDS)* en anglais)

On réordonne les lignes de la matrice suivant le nombre de zéros qu'elles contiennent respectivement, en conservant la trace de cette permutation dans le tableau `perm`. EXPLICATION Les diagonales sont ensuite stockées dans un tableau de nombres à virgule flottante à une dimension, le début de chaque diagonale étant repéré dans un tableau de pointeurs. Un dernier tableau d'entiers contient les indices de colonne des éléments stockés.

Exemple de stockage diagonal indenté d'une matrice creuse. Le stockage diagonal identé de la matrice

$$A = \begin{pmatrix} 10 & -3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 6 & 0 & -2 & 0 \\ 3 & 0 & 8 & 7 & 0 & 0 \\ 0 & 6 & 0 & 7 & 5 & 4 \\ 0 & 0 & 0 & 0 & 9 & 13 \\ 0 & 0 & 0 & 0 & 5 & -1 \end{pmatrix}$$

commence par un réordonnancement des lignes de la matrice par ordre croissant du nombre de coefficients nuls

$$PA = \begin{pmatrix} 0 & 6 & 0 & 7 & 5 & 4 \\ 0 & 9 & 6 & 0 & -2 & 0 \\ 3 & 0 & 8 & 7 & 0 & 0 \\ 10 & -3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9 & 13 \\ 0 & 0 & 0 & 0 & 5 & -1 \end{pmatrix}.$$

On décale ensuite tout les éléments de la matrice vers la gauche, en oubliant les zéros, et l'on stocke les colonnes contiguement

$$\begin{pmatrix} 6 & 7 & 5 & 4 \\ 9 & 6 & -2 & \\ 3 & 8 & 7 & \\ 10 & -3 & 1 & \\ 9 & 13 & & \\ 5 & -1 & & \end{pmatrix}.$$

matrix_elements(:)	6	9	3	10	9	5	7	6	8	-3	13	-1	5	-2	7	1	4

col_ind(:)	2	2	1	1	5	5	4	3	3	2	6	6	5	5	4	4	6

diag_ptr(:)	1	7	13	17

perm(:)	4	2	3	1	5	6

L'utilisation de tels formats de stockage nécessite des efforts de programmation conséquents pour la mise en œuvre des méthodes décrites dans le présent chapitre. On doit en effet manipuler des structures de données, optimisées mais complexes, qui rendent moins aisé l'accès aux données et tendent à particulariser les algorithmes. Des bibliothèques de programmes comme LAPACK (pour *Linear Algebra PACKage*) [ABB+99], qui met en œuvre un grand nombre d'algorithmes pour la résolution numérique de systèmes linéaires et de problèmes aux moindres carrés, aux valeurs propres ou de décomposition en valeurs singulières, proposent diverses implémentations d'une même méthode, chacune étant adaptée à un type de matrice, et donc au mode de stockage utilisé pour celui-ci.

2.3 Remarques sur la résolution des systèmes triangulaires

Observons tout d'abord que la solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec A une matrice inversible, *ne s'obtient pas*[8] en inversant A, puis en calculant le vecteur $A^{-1}\boldsymbol{b}$, mais en réalisant plutôt des combinaisons linéaires sur les lignes du système et des substitutions. En effet, on peut facilement voir que le calcul de la matrice A^{-1} équivaut à résoudre n systèmes linéaires[9], ce qui s'avère bien plus coûteux que la résolution du *seul* système dont on cherche la solution.

Considérons à présent un système linéaire (2.1) dont la matrice est triangulaire inférieure, c'est-à-dire de la forme

$$\begin{array}{ccccccccc} a_{11}x_1 & & & & & & & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & & & & & = & b_2 \\ \vdots & & \vdots & & \ddots & & & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \dots & + & a_{nn}x_n & = & b_n \end{array}.$$

Si la matrice A est inversible, ses termes diagonaux a_{ii}, $i = 1, \dots, n$, sont tous non nuls[10] et la résolution du système est alors extrêmement simple : on calcule x_1 par une division, que l'on substitue ensuite dans la deuxième équation pour obtenir x_2, et ainsi de suite... Cette méthode, dite *de « descente »* (*"forward substitution"* en anglais),

8. On doit sur ce sujet à Forsythe et Moler dans [FM67] la phrase particulièrement à propos : *"Almost anything you can do with A^{-1} can be done without it."*.

9. Ces systèmes sont

$$A\boldsymbol{x_i} = \boldsymbol{e_i},\ 1 \leq i \leq n,$$

où $\boldsymbol{e}_i$ désigne le i^{e} vecteur de la base canonique de $M_{n,1}(\mathbb{R})$.

10. On a en effet $a_{11}a_{22}\dots a_{nn} = \det(A) \neq 0$.

s'écrit

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j\right),\ i = 2, \ldots, n. \end{aligned} \tag{2.7}$$

L'algorithme mis en œuvre pour cette résolution effectue $\frac{1}{2}n(n-1)$ soustractions, $\frac{1}{2}n(n-1)$ multiplications et n divisions pour calculer la solution, soit un nombre d'opérations global de l'ordre de n^2. On notera que pour calculer la i^{e}, $2 \leq i \leq n$, composante du vecteur solution $\boldsymbol{x}$, on effectue un produit scalaire entre le vecteur constitué des $i-1$ premiers éléments de la i^{e} ligne de la matrice A et le vecteur contenant les $i-1$ premières composantes de $\boldsymbol{x}$. L'accès aux éléments de A se fait donc ligne par ligne et on parle pour cette raison d'algorithme est *orienté ligne* (voir l'algorithme 5).

Algorithme 5 : Algorithme de la méthode de descente (version orientée ligne).

Entrée(s) : les tableaux contenant la matrice triangulaire inférieure A et le vecteur $\boldsymbol{b}$.
Sortie(s) : le tableau contenant le vecteur $\boldsymbol{x}$ solution du système $A\boldsymbol{x} = \boldsymbol{b}$.

```
x(1) = b(1)/a(1,1)
pour i = 2 à n faire
    x(i) = b(i)
    pour j = 1 à i-1 faire
        x(i) = x(i) - A(i,j) * x(j)
    fin
    x(i) = x(i)/A(i,i)
fin
```

On peut obtenir un algorithme *orienté colonne* pour la méthode en tirant parti du fait que la i^{e} composante du vecteur $\boldsymbol{x}$, une fois calculée, peut être éliminée du système. L'ordre des boucles d'indices i et j est alors inversé (voir l'algorithme 6, dans lequel la solution $\boldsymbol{x}$ calculée est commodément stockée dans le tableau contenant initialement le second membre du système linéaire).

Exemple de résolution d'un système triangulaire inférieur. Appliquons une approche orientée colonne pour la résolution du système

$$\begin{pmatrix} 2 & 0 & 0 \\ 1 & 5 & 0 \\ 7 & 9 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ 5 \end{pmatrix}.$$

On trouve que $x_1 = 3$ et l'on considère ensuite le système à deux équations et deux inconnues

$$\begin{pmatrix} 5 & 0 \\ 9 & 8 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix} - 3 \begin{pmatrix} 1 \\ 7 \end{pmatrix},$$

pour lequel on trouve $x_2 = -\dfrac{1}{5}$. On a enfin

$$8\, x_3 = -16 + \frac{9}{5},$$

soit $x_3 = -\dfrac{71}{40}$.

Le choix d'une approche orientée ligne ou colonne dans l'écriture d'un même algorithme peut considérablement modifier ses performances en fonction de l'architecture du calculateur utilisé.

Le cas d'un système linéaire dont la matrice est inversible et triangulaire supérieure se traite de manière analogue, par la méthode dite *de « remontée »* (*"back substitution"* en anglais) suivante

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j\right),\ i = n-1, \ldots, 1, \end{aligned} \tag{2.8}$$

Algorithme 6 : Algorithme de la méthode de descente (version orientée colonne).

Entrée(s) : les tableaux contenant la matrice triangulaire inférieure A et le vecteur $\boldsymbol{b}$.
Sortie(s) : le vecteur $\boldsymbol{x}$ solution du système $A\boldsymbol{x} = \boldsymbol{b}$, contenu dans le tableau dans lequel était stocké le vecteur $\boldsymbol{b}$ en entrée.

```
pour j = 1 à n−1 faire
    b(j) = b(j)/A(j,j)
    pour i = j+1 à n faire
        b(i) = b(i) − A(i,j) ∗ b(j)
    fin
fin
b(n) = b(n)/A(n,n)
```

et dont la complexité est également de l'ordre de n^2 opérations. Là encore, on peut produire des algorithmes orientés ligne ou colonne pour le codage de la méthode.

Dans la pratique, il est par ailleurs utile de remarquer que seule la partie *a priori* non nulle de la matrice nécessite d'être stockée [11] pour la résolution d'un système triangulaire, d'où une économie de mémoire conséquente dans le cas de grands systèmes.

2.4 Méthode d'élimination de Gauss

Une technique de choix pour ramener la résolution d'un système linéaire quelconque à celle d'un système triangulaire est la *méthode d'élimination de Gauss* (*Gaussian elimination* en anglais). Celle-ci consiste en premier lieu à transformer, par des opérations simples sur les équations, le système en un système équivalent, c'est-à-dire ayant la (ou les) même(s) solution(s), $MA\boldsymbol{x} = M\boldsymbol{b}$, dans lequel MA est une matrice triangulaire supérieure [12] (on dit encore que la matrice du système est sous forme *échelonnée*). Cette étape de mise à zéro d'une partie des coefficients de la matrice est qualifiée d'*élimination* et utilise de manière essentielle le fait qu'on ne modifie pas la solution d'un système linéaire en ajoutant à une équation donnée une combinaison linéaire des autres équations. Lorsque la matrice du système est inversible, la solution du système peut ensuite être obtenue par une méthode de remontée, mais le procédé d'élimination est très général et amène toute matrice sous forme échelonnée.

2.4.1 Élimination sans échange

Commençons par décrire étape par étape la méthode dans sa forme de base, dite *sans échange*, en considérant le système linéaire (2.1), avec A étant une matrice inversible d'ordre n. Supposons de plus que le terme a_{11} de la matrice A est non nul. Nous pouvons alors éliminer l'inconnue x_1 de la deuxième à la n^{e} ligne du système en leur retranchant respectivement la première ligne multipliée par le coefficient $\frac{a_{i1}}{a_{11}}$, $i = 2, \ldots, n$. En notant $A^{(1)}$ et $\boldsymbol{b}^{(1)}$ la matrice et le vecteur second membre résultant de ces opérations [13], on a alors

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}}\, a_{1j} \text{ et } b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}}\, b_1,\ i = 2, \ldots, n, j = 1, \ldots, n,$$

les coefficients de la première ligne restant les mêmes, et le système $A^{(1)}\boldsymbol{x} = \boldsymbol{b}^{(1)}$ est équivalent au système de départ. En supposant non nul le coefficient diagonal $a_{22}^{(1)}$ de $A^{(1)}$, on peut ensuite procéder à l'élimination de l'inconnue x_2 de la troisième à la n^{e} ligne de ce nouveau système, et ainsi de suite. On obtient, sous l'hypothèse

11. Les éléments de la matrice triangulaire sont généralement stockés dans un tableau à une seule entrée de dimension $\frac{1}{2}(n+1)n$ en gérant la correspondance entre les indices i et j d'un élément de la matrice et l'indice $k(= k(i,j))$ de l'élément le représentant dans le tableau. Par exemple, pour une matrice triangulaire inférieure stockée ligne par ligne, on vérifie facilement que $k(i,j) = j + \frac{1}{2}i(i-1)$.
12. Il faut bien noter qu'on ne calcule en pratique jamais explicitement la matrice d'élimination M, mais seulement les produits MA et $M\boldsymbol{b}$.
13. On pose $A^{(0)} = A$ et $\boldsymbol{b}^{(0)} = \boldsymbol{b}$ pour être consistant.

$a_{k+1\,k+1}^{(k)} \neq 0$, $k = 0, \ldots, n-2$, une suite finie de matrices $A^{(k)}$, $k = 1, \ldots, n-1$, de la forme

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & \cdots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & & & a_{2n}^{(k)} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & a_{k+1\,k+1}^{(k)} & \cdots & a_{k+1\,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n\,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix},$$

telle que le système $A^{(n-1)}\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ est triangulaire supérieur. Les quantités $a_{k+1\,k+1}^{(k)}$, $k = 0, \ldots, n-2$, sont appelées les *pivots* et l'on a supposé qu'elles étaient non nulles à chaque étape, les formules permettant de passer du système linéaire de matrice $A^{(k)}$ et de second membre $\boldsymbol{b}^{(k)}$ au suivant se résumant à

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik+1}^{(k)}}{a_{k+1\,k+1}^{(k)}}\, a_{k+1\,j}^{(k)} \text{ et } b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik+1}^{(k)}}{a_{k+1\,k+1}^{(k)}}\, b_{k+1}^{(k)},\ i = k+2, \ldots, n, j = k+1, \ldots, n,$$

les autres coefficients étant inchangés. En pratique, pour une résolution « à la main » d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ par cette méthode, il est commode pour les calculs d'appliquer l'élimination à la matrice *« augmentée »* $\begin{pmatrix} A & \boldsymbol{b} \end{pmatrix}$.

Exemple d'application de la méthode d'élimination de Gauss sans échange. Considérons la résolution par la méthode d'élimination de Gauss sans échange du système linéaire suivant

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ 2x_1 & + & 3x_2 & + & 4x_3 & + & x_4 & = & 12 \\ 3x_1 & + & 4x_2 & + & x_3 & + & 2x_4 & = & 13 \\ 4x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 14 \end{array} \right. .$$

À la première étape, le pivot vaut 1 et on soustrait de la deuxième (resp. troisième (resp. quatrième)) équation la première équation multipliée par 2 (resp. 3 (resp. 4)) pour obtenir

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & - & 2x_2 & - & 8x_3 & - & 10x_4 & = & -20 \\ & - & 7x_2 & - & 10x_3 & - & 13x_4 & = & -3 \end{array} \right. .$$

Le pivot vaut -1 à la deuxième étape. On retranche alors à la troisième (resp. quatrième) équation la deuxième équation multipliée par -2 (resp. -7), d'où le système

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & & & - & 4x_3 & + & 4x_4 & = & 0 \\ & & & & 4x_3 & + & 36x_4 & = & 40 \end{array} \right. .$$

À la dernière étape, le pivot est égal à -4 et on soustrait à la dernière équation l'avant-derniére multipliée par -1 pour arriver à

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & & & - & 4x_3 & + & 4x_4 & = & 0 \\ & & & & & & 40x_4 & = & 40 \end{array} \right. .$$

Ce système triangulaire, équivalent au système d'origine, est enfin résolu par remontée :

$$\left\{ \begin{array}{l} x_4 = 1 \\ x_3 = x_4 = 1 \\ x_2 = 10 - 2 - 7 = 1 \\ x_1 = 11 - 2 - 3 - 4 = 2 \end{array} \right. .$$

Comme on l'a vu, la méthode d'élimination de Gauss, dans sa forme sans échange, ne peut s'appliquer que si tous les pivots $a_{k+1\,k+1}^{(k)}$, $k=0,\ldots,n-2$, sont non nuls, ce qui rejette de fait des matrices inversibles aussi simples que

$$A=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}.$$

De plus, le fait que la matrice soit inversible n'empêche aucunement l'apparition d'un pivot nul durant l'élimination, comme le montre l'exemple ci-dessous.

Exemple de mise en échec de la méthode d'élimination de Gauss sans échange. Considérons la matrice inversible

$$A=\begin{pmatrix}1 & 2 & 3\\ 2 & 4 & 5\\ 7 & 8 & 9\end{pmatrix}=A^{(0)}.$$

On a alors

$$A^{(1)}=\begin{pmatrix}1 & 2 & 3\\ 0 & 0 & -1\\ 0 & -6 & -12\end{pmatrix},$$

et l'élimination s'interrompt à l'issue de la première étape, le pivot $a_{22}^{(1)}$ étant nul.

Il apparaît donc que des conditions plus restrictives que l'inversibilité de la matrice sont nécessaires pour assurer la bonne exécution de cette méthode. Celles-ci sont fournies par le théorème 2.2. Indiquons qu'il existe des catégories de matrices pour lesquelles la méthode de Gauss sans échange peut-être utilisée sans aucun risque. Parmi celles-ci, on trouve les matrices *à diagonale dominante par lignes ou par colonnes* (voir à ce titre le théorème 2.5) et les matrices *symétriques définies positives* (voir le théorème 2.12).

2.4.2 Élimination de Gauss avec échange

Dans sa forme générale, la méthode d'élimination de Gauss permet de transformer un système linéaire dont la matrice est carrée (inversible ou non) ou même rectangulaire en un système échelonné équivalent. En considérant le cas d'une matrice A carrée inversible, nous allons maintenant décrire les modifications à apporter à la méthode déjà présentée pour mener l'élimination à son terme. Dans tout ce qui suit, les notations de la section 2.4.1 sont conservées.

À la première étape, au moins l'un des coefficients de la première colonne de la matrice $A^{(0)}(=A)$ est non nul, faute de quoi la matrice A ne serait pas inversible. On choisit [14] un de ces éléments comme premier pivot d'élimination et l'on échange alors la première ligne du système avec celle du pivot avant de procéder à l'élimination de la première colonne de la matrice résultante, c'est-à-dire l'annulation de tous les éléments de la première colonne de la matrice (permutée) du système situés sous la diagonale. On note $A^{(1)}$ et $\boldsymbol{b}^{(1)}$ la matrice et le second membre du système obtenu et l'on réitère ce procédé. Par la suite, à l'étape $k+1$, $1\leq k\leq n-2$, la matrice $A^{(k)}$ est inversible [15], et donc l'un au moins des éléments $a_{ik+1}^{(k)}$, $k+1\leq i\leq n$, est différent de zéro. Après avoir choisi comme pivot l'un de ces coefficients non nuls, on effectue l'échange de la ligne de ce pivot avec la $k+1^{\text{e}}$ ligne de la matrice $A^{(k)}$, puis l'élimination conduisant à la matrice $A^{(k+1)}$. Ainsi, on arrive après $n-1$ étapes à la matrice $A^{(n-1)}$, dont le coefficient $a_{nn}^{(n-1)}$ est non nul.

En raison de l'échange de lignes qui a éventuellement lieu avant chaque étape d'élimination, on parle de méthode d'élimination de Gauss *avec échange*.

Exemple d'application de la méthode d'élimination de Gauss avec échange. Considérons la résolution du système linéaire $A\boldsymbol{x}=\boldsymbol{b}$, avec

$$A=\begin{pmatrix}2 & 4 & -4 & 1\\ 3 & 6 & 1 & -2\\ -1 & 1 & 2 & 3\\ 1 & 1 & -4 & 1\end{pmatrix}\text{ et }\boldsymbol{b}=\begin{pmatrix}0\\ -7\\ 4\\ 2\end{pmatrix},$$

14. Pour l'instant, on ne s'intéresse pas au choix *effectif* du pivot, qui est cependant d'une importance cruciale pour la stabilité numérique de la méthode. Ce point est abordé dans la section 2.4.4.

15. On a en effet que $\det(A^{(k)})=\pm\det(A)$, $k=0,\ldots,n-1$. On renvoie à la section 2.5.1 pour une justification de ce fait.

par application de la méthode d'élimination de Gauss avec échange. On trouve successivement

$$A^{(1)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & -1 & -2 & \frac{1}{2} \end{pmatrix} \text{ et } \boldsymbol{b}^{(1)} = \begin{pmatrix} 0 \\ -7 \\ 4 \\ 2 \end{pmatrix},$$

$$A^{(2)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & -2 & \frac{5}{3} \end{pmatrix} \text{ et } \boldsymbol{b}^{(2)} = \begin{pmatrix} 0 \\ 4 \\ -7 \\ \frac{10}{3} \end{pmatrix},$$

et

$$A^{(3)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix} \text{ et } \boldsymbol{b}^{(3)} = \begin{pmatrix} 0 \\ 4 \\ -7 \\ \frac{4}{3} \end{pmatrix},$$

d'où la solution $\boldsymbol{x} = \begin{pmatrix} 1 & -1 & 0 & 2 \end{pmatrix}^\top$. On note que l'on a procédé au cours de la deuxième étape à l'échange des deuxième et troisième lignes.

On pourra remarquer que si la matrice A est non inversible, alors tous les éléments $a_{ik+1}^{(k)}$, $k+1 \leq i \leq n$, seront nuls pour au moins une valeur de k entre 0 et $n-1$. Si cela vient à se produire alors que $k < n-1$, on n'a pas besoin de réaliser l'élimination dans la $k+1^{\text{e}}$ colonne (puisque celle-ci est déjà nulle) et l'on passe simplement à l'étape suivante en posant $A^{(k+1)} = A^{(k)}$ et $\boldsymbol{b}^{(k+1)} = \boldsymbol{b}^{(k)}$. L'élimination est donc bien possible pour une matrice carrée non inversible et l'on a démontré le résultat suivant.

Théorème 2.1 *Soit A une matrice carrée, inversible ou non. Il existe au moins une matrice inversible M telle que la matrice MA soit triangulaire supérieure.*

Il reste à compter le nombre d'opérations élémentaires que requiert l'application de la méthode d'élimination de Gauss pour la résolution d'un système linéaire de n équations à n inconnues. Tout d'abord, pour passer de la matrice $A^{(k)}$ à la matrice $A^{(k+1)}$, $0 \leq k \leq n-2$, on effectue $(n-k-1)^2$ soustractions, $(n-k-1)^2$ multiplications et $n-k-1$ divisions, ce qui correspond à un total de $\frac{1}{6}(2n-1)n(n-1)$ soustractions, $\frac{1}{6}(2n-1)n(n-1)$ multiplications et $\frac{1}{2}n(n-1)$ divisions pour l'élimination complète. Pour la mise à jour du second membre à l'étape $k+1$, on a besoin de $n-k-1$ soustractions et autant de multiplications, soit en tout $\frac{1}{2}n(n-1)$ soustractions et $\frac{1}{2}n(n-1)$ multiplications. Enfin, il faut faire $\frac{1}{2}n(n-1)$ soustractions, autant de multiplications et n divisions pour résoudre le système final par une méthode de remontée.

En tout, la résolution du système par la méthode d'élimination de Gauss nécessite donc environ $\frac{n^3}{3}$ soustractions, $\frac{n^3}{3}$ multiplications et $\frac{n^2}{2}$ divisions. À titre de comparaison, le calcul de la solution du système par la règle de Cramer (voir la proposition A.163) requiert, en utilisant un développement « brutal » par ligne ou colonne pour le calcul des déterminants, environ $(n+1)!$ additions ou soustractions, $(n+2)!$ multiplications et n divisions. Ainsi, pour $n = 10$ par exemple, on obtient un compte d'approximativement 700 opérations pour la méthode d'élimination de Gauss contre près de 479000000 opérations pour la règle de Cramer !

2.4.3 Résolution de systèmes rectangulaires par élimination

Nous n'avons jusqu'à présent considéré que des systèmes linéaires de n équations à n inconnues, mais la méthode d'élimination avec échange peut être utilisée pour la résolution de tout système à m équations et n inconnues, avec $m \neq n$. Ce procédé ramène en effet toute matrice rectangulaire sous forme échelonnée (voir la définition A.166), et l'on peut alors résoudre le système associé comme expliqué dans la section A.5. La méthode d'élimination de Gauss constitue à ce titre un moyen simple de détermination du rang d'une matrice quelconque.

2.4.4 Choix du pivot

Revenons à présent sur le choix des pivots lors de l'élimination. À la $k+1^{\text{e}}$ étape du procédé, si l'élément $a_{k+1\,k+1}^{(k)}$ est non nul, il semble naturel de l'utiliser comme pivot (c'est d'ailleurs ce que l'on fait dans la méthode de Gauss sans échange). Cependant, à cause de la présence d'erreurs d'arrondi en pratique, cette manière de procéder est en général à proscrire, comme l'illustre l'exemple d'instabilité numérique suivant.

Exemple d'application numérique (tiré de [FM67]). Supposons que les calculs soient effectués en virgule flottante dans le système décimal, avec une mantisse à trois chiffres, et considérons le système

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

dont la solution « exacte » est $x_1 = 1,\overline{0001}$ et $x_2 = 0,\overline{9998}$. En choisissant le nombre 10^{-4} comme pivot à la première étape de l'élimination de Gauss, on obtient le système triangulaire

$$\begin{pmatrix} 10^{-4} & 1 \\ 0 & -9990 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -9990 \end{pmatrix},$$

car les nombres $-10^4 + 1 = -9999$ et $-10^4 + 2 = -9998$ sont tous deux arrondis au nombre -9990. La solution numérique calculée est alors

$$x_1 = 0 \text{ et } x_2 = 1,$$

ce qui très différent de la véritable solution du système. Si, par contre, on commence par échanger les deux équations du système pour utiliser le nombre 1 comme pivot, on trouve

$$\begin{pmatrix} 1 & 1 \\ 0 & 0,999 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0,999 \end{pmatrix},$$

puisque les nombres $-10^{-4} + 1 = 0,9999$ et $-2\ 10^{-4} + 1 = 0,9998$ sont chacun arrondis au nombre $0,999$. La solution calculée vaut alors

$$x_1 = 1 \text{ et } x_2 = 1,$$

ce qui est cette fois très satisfaisant.

En général, le changement de pivot n'a pas un effet aussi spectaculaire que dans cet exemple, mais il n'en demeure pas moins essentiel lorsque les calculs sont effectués en arithmétique à virgule flottante. De fait, pour éviter la propagation d'erreurs et obtenir une meilleure stabilité numérique de la méthode, il faut chercher, même lorsque le pivot « naturel » est non nul, à choisir le plus grand pivot en valeur absolue. On peut pour cela suivre, au début de la $k+1^{\text{e}}$ étape, $0 \leq k \leq n-2$, de l'élimination,

- soit une stratégie de *pivot partiel* (*partial pivoting* en anglais) dans laquelle le pivot est l'élément $a^{(k)}_{rk+1}$ de la $k+1^{\text{e}}$ colonne de la matrice $A^{(k)}$ situé sous la diagonale ayant la plus grande valeur absolue,
$$|a^{(k)}_{rk+1}| = \max_{k+1 \leq i \leq n} |a^{(k)}_{ik+1}|,$$
- soit une stratégie de *pivot total* (*complete pivoting* en anglais), plus coûteuse, dans laquelle le pivot est l'élément $a^{(k)}_{rs}$ de la sous-matrice $(a^{(k)}_{ij})_{k+1 \leq i,j \leq n}$ le plus grand en valeur absolue,
$$|a^{(k)}_{rs}| = \max_{k+1 \leq i,j \leq n} |a^{(k)}_{ij}|,$$
- soit encore une stratégie intermédiaire aux précédentes, portant en anglais le nom de *rook pivoting* et introduite dans [NP92], qui consiste à prendre pour pivot l'élément $a^{(k)}_{rs}$ de la sous-matrice $(a^{(k)}_{ij})_{k+1 \leq i,j \leq n}$ ayant la plus grande valeur absolue dans la colonne *et* la ligne dans auxquelles il appartient [16], c'est-à-dire
$$|a^{(k)}_{rs}| = \max_{k+1 \leq i \leq n} |a^{(k)}_{is}| = \max_{k+1 \leq j \leq n} |a^{(k)}_{rj}|.$$

Dans les deux derniers cas, si le pivot n'est pas dans la $k+1^{\text{e}}$ colonne, il faut procéder à un échange de colonnes en plus d'un éventuel échange de lignes. L'effet de la stratégie de choix de pivot sur la stabilité numérique de la factorisation est analysé dans la sous-section 2.7.2.

Quelle que soit la stratégie adoptée, la recherche des pivots doit également être prise en compte dans l'évaluation du coût global de la méthode d'élimination de Gauss. Celle-ci demande de l'ordre de n^2 comparaisons au total pour la stratégie de pivot partiel et de l'ordre de n^3 comparaisons pour celle de pivot total, la première étant privilégiée en raison de sa complexité algorithmique moindre et de performances généralement [17] très bonnes. Pour la technique de *rook pivoting*, cette recherche nécessite de l'ordre de n^3 comparaisons dans le pire des cas, mais une complexité en $O(n^2)$ est souvent observée en pratique et démontrée (voir [Fos97]) moyennant une hypothèse statistique sur les coefficients des matrices $A^{(k)}$, $k = 0, \ldots, n-2$.

16. Cette technique de choix de pivot tire son nom du fait que la recherche effective du pivot parmi les éléments de la matrice rappelle les déplacements de la tour (*rook* en anglais) dans le jeu d'échecs.

17. On fait néanmoins appel à la recherche d'un pivot total dans quelques situations particulières, comme la résolution d'équations de Sylvester, le calcul d'une décomposition en valeurs singulières ou d'une décomposition de Schur généralisée.

2.4.5 Méthode d'élimination de Gauss–Jordan

Introduite indépendamment [18] par Jordan [19] [Jor88] et Clasen [Cla88], la *méthode d'élimination de Gauss–Jordan* est une variante de la méthode d'élimination de Gauss ramenant toute matrice sous forme échelonnée *réduite* (voir la définition A.166). Dans le cas d'une matrice A inversible, cette méthode revient à chercher une matrice M telle que la matrice MA soit non pas triangulaire supérieure mais *diagonale*, égale à la matrice identité. Pour cela, on procède comme pour l'élimination de Gauss, mais en annulant à chaque étape tous les éléments de la colonne considérée situés au dessous *et au dessus* de la diagonale, ainsi qu'en normalisant les éléments diagonaux.

Si elle est bien moins efficace [20] que la méthode d'élimination de Gauss pour la résolution de systèmes linéaires, la méthode d'élimination de Gauss–Jordan est utile pour le calcul de l'inverse d'une matrice A carrée d'ordre n. Il suffit pour cela de résoudre simultanément les n systèmes linéaires

$$A\boldsymbol{x}_j = \boldsymbol{e}_j,\ 1 \leq j \leq n,$$

en appliquant à chaque second membre $\boldsymbol{e}_j$ les transformations nécessaires à l'élimination de Gauss–Jordan. D'un point de vue pratique, on a coutume d'« augmenter » la matrice A à inverser avec la matrice identité d'ordre n (les n seconds membres « élémentaires ») et d'appliquer la méthode de Gauss–Jordan à la matrice écrite par blocs $\begin{pmatrix} A & I_n \end{pmatrix}$. À l'issue du processus d'élimination, dans lequel des échanges de lignes peuvent s'avérer nécessaires, le premier bloc contient la matrice identité et le second l'inverse de la matrice A.

Exemple d'application de la méthode d'élimination de Gauss–Jordan sans échange pour l'inversion d'une matrice. On considère la matrice inversible

$$A = \begin{pmatrix} 2 & 1 & -4 \\ 3 & 3 & -5 \\ 4 & 5 & -2 \end{pmatrix},$$

dont on veut calculer l'inverse par la méthode d'élimination de Gauss–Jordan sans échange. On forme tout d'aord la matrice augmentée

$$\begin{pmatrix} A & I_n \end{pmatrix} = \begin{pmatrix} 2 & 1 & -4 & 1 & 0 & 0 \\ 3 & 3 & -5 & 0 & 1 & 0 \\ 4 & 5 & -2 & 0 & 0 & 1 \end{pmatrix},$$

et l'on trouve alors successivement, l'entier k indiquant le numéro de l'étape effectuée,

$$k=1,\quad \begin{pmatrix} 1 & \frac{1}{2} & -2 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{3}{2} & 1 & -\frac{3}{2} & 1 & 0 \\ 0 & 3 & 6 & -2 & 0 & 1 \end{pmatrix},$$

$$k=2,\quad \begin{pmatrix} 1 & 0 & -\frac{7}{3} & 1 & -\frac{1}{3} & 0 \\ 0 & 1 & \frac{2}{3} & -1 & \frac{2}{3} & 0 \\ 0 & 0 & 4 & 1 & -2 & 1 \end{pmatrix},$$

$$k=3,\quad \begin{pmatrix} 1 & 0 & 0 & \frac{19}{12} & -\frac{3}{2} & \frac{7}{12} \\ 0 & 1 & 0 & -\frac{7}{6} & 1 & -\frac{1}{6} \\ 0 & 0 & 1 & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \end{pmatrix},$$

d'où $A^{-1} = \dfrac{1}{12}\begin{pmatrix} 19 & -18 & 7 \\ -14 & 12 & -2 \\ 3 & -6 & 3 \end{pmatrix}$.

18. On pourra consulter l'article [AM87] pour des détails sur l'histoire cette méthode.

19. Wilhelm Jordan (1er mars 1842 - 17 avril 1899) était un géodésiste allemand. Il est connu parmi les mathématiciens pour le procédé d'élimination portant son nom, publié en 1888 dans son *Handbuch der Vermessungskunde*, qu'il appliqua à la résolution de problèmes aux moindres carrés en géodésie.

20. Effectuons un compte des opérations effectuées pour la résolution d'un système de n équations à n inconnues. À l'étape $k+1$, $0 \leq k \leq n-1$, il faut faire $(n-k+1)(n-1)$ soustractions, $(n-k+1)(n-1)$ multiplications et $(n-k+1)$ divisions pour mettre à jour la matrice et le second membre du système. En revanche, la résolution du système (diagonal) final ne nécessite aucune opération supplémentaire. Une résolution par la méthode d'élimination de Gauss–Jordan nécessite donc de l'ordre de n^3 opérations.

2.5 Interprétation matricielle de l'élimination de Gauss : la factorisation LU

Nous allons maintenant montrer que la méthode de Gauss dans sa forme sans échange est équivalente à la décomposition de la matrice A sous la forme d'un produit de deux matrices, $A = LU$, avec L une matrice triangulaire inférieure (*lower triangular* en anglais), qui est l'inverse de la matrice M des transformations successives appliquées à la matrice A lors de l'élimination de Gauss sans échange, et U une matrice triangulaire supérieure (*upper triangular* en anglais), avec $U = A^{(n-1)}$ en reprenant la notation utilisée dans la section 2.4.1.

2.5.1 Formalisme matriciel

Chacune des opérations que nous avons effectuées pour transformer le système linéaire lors de l'élimination de Gauss, que ce soit l'échange de deux lignes ou l'annulation d'une partie des coefficients d'une colonne de la matrice $A^{(k)}$, $0 \leq k \leq n-2$, peut se traduire matriciellement par la multiplication de la matrice et du second membre du système linéaire courant par une matrice inversible particulière. L'introduction de ces matrices va permettre de traduire le procédé d'élimination dans un formalisme matriciel débouchant sur une factorisation remarquable de la matrice A.

Matrices des transformations élémentaires

Soit (m,n) un couple d'entiers de $(\mathbb{N}\backslash\{0,1\})^2$ et $A = (a_{ij})_{1\leq i\leq m,\ 1\leq j\leq n}$ une matrice de $M_{m,n}(\mathbb{R})$. On appelle *opérations élémentaires sur les lignes de A* les transformations suivantes :

- l'échange (entre elles) des i^{e} et j^{e} lignes de A ;
- la multiplication de la i^{e} ligne de A par un scalaire $\lambda \in \mathbb{R}\backslash\{0\}$;
- le remplacement de la i^{e} ligne de A par la somme de cette même ligne avec la j^{e} ligne de A multipliée par un réel non nul λ.

Explicitons à présent les opérations matricielles correspondant à chacunes de ces opérations. Tout d'abord, échanger les i^{e} et j^{e} lignes, $i \neq j$, de la matrice A revient à multiplier à gauche cette matrice par la *matrice de permutation* (*permutation matrix* en anglais)

$$P_{ij} = \begin{pmatrix}
1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & \ddots & \ddots & & & & & & & & \vdots \\
\vdots & \ddots & 1 & \ddots & & & & & & & \vdots \\
\vdots & & \ddots & 0 & 0 & \cdots & 0 & 1 & & & \vdots \\
\vdots & & & 0 & 1 & \ddots & & 0 & & & \vdots \\
\vdots & & & \vdots & \ddots & \ddots & \ddots & \vdots & & & \vdots \\
\vdots & & & 0 & & \ddots & 1 & 0 & & & \vdots \\
\vdots & & & 1 & 0 & \dots & 0 & 0 & & & \vdots \\
\vdots & & & & & & & & 1 & & \vdots \\
\vdots & & & & & & & & & \ddots & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1
\end{pmatrix} = I_m + (E_{ij} + E_{ji} - E_{ii} - E_{jj}) \in M_m(\mathbb{R}).$$

Cette matrice est symétrique et orthogonale, de déterminant valant -1.

La multiplication de la i^{e} ligne de la matrice A par un scalaire non nul λ s'effectue en multipliant à gauche

cette matrice par la *matrice de dilatation* (*directional scaling matrix* en anglais)

$$D_i(\lambda) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ \vdots & & \ddots & \lambda & \ddots & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_m + (\lambda - 1)E_{ii} \in M_m(\mathbb{R}).$$

Cette matrice est inversible et $D_i(\lambda)^{-1} = D_i\left(\frac{1}{\lambda}\right)$.

Enfin, le remplacement de la i^{e} ligne de A par la somme de la i^{e} ligne et de la j^{e} ligne, $i \neq j$, multipliée par un scalaire non nul λ est obtenu en multipliant à gauche la matrice A par la *matrice de transvection* (*shear matrix* en anglais)

$$T_{ij}(\lambda) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \cdots & 0 & & \vdots \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ \vdots & & \lambda & \cdots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_m + \lambda\, E_{ij} \in M_m(\mathbb{R})$$

(on a supposé dans l'exemple que $j < i$). Cette matrice a pour inverse[21] $T_{ij}(-\lambda)$. Pour la suite, il est important de remarquer que, étant donné trois entiers naturels i, k et l, tels que $1 \leq k < i < l \leq m$, et deux scalaires non nuls λ et μ, le produit des matrices de tranvection $T_{ik}(\lambda)$ et $T_{lk}(\mu)$ de $M_m(\mathbb{R})$ est commutatif et vaut

$$T_{ik}(\lambda)T_{lk}(\mu) = I_m + \lambda\, E_{ik} + \mu\, E_{lk}.$$

On effectue de manière analogue des opérations élémentaires *sur les colonnes* de la matrice A en multipliant *à droite* cette dernière par les matrices d'ordre n correspondantes.

Exemple d'action d'une matrice de permutation. Soit les matrices $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ et $P_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. On a

$$P_{23}A = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 4 & 5 & 6 \end{pmatrix} \text{ et } AP_{23} = \begin{pmatrix} 1 & 3 & 2 \\ 4 & 6 & 5 \\ 7 & 9 & 8 \end{pmatrix}.$$

Factorisation LU

Si l'élimination arrive à son terme sans qu'il y ait besoin d'échanger des lignes du système linéaire, la matrice inversible M du théorème 2.1 est unique et égale au produit

$$M = E^{(n-1)} \dots E^{(2)} E^{(1)}$$

21. L'ensemble des matrices de transvection de $M_m(\mathbb{R})$ engendre le *groupe spécial linéaire* (*special linear group* en anglais) $\mathrm{SL}_m(\mathbb{R})$, constitué des matrices d'ordre m dont le déterminant vaut 1, et, avec l'ensemble des matrices de dilatation de $M_m(\mathbb{R})$, le groupe général linéaire $\mathrm{GL}_m(\mathbb{R})$.

de $n-1$ *matrices d'élimination* définies par

$$E^{(k)} = \prod_{i=k+1}^{n} T_{ik}\left(-\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+2k}^{(k-1)}}{a_{kk}^{(k-1)}} & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & 0 & \dots & 0 & 1 \end{pmatrix}, \ 1 \leq k \leq n-1. \tag{2.9}$$

Par construction, la matrice M est triangulaire inférieure et son inverse est donc également une matrice triangulaire inférieure. Il en résulte que la matrice A s'écrit comme le produit

$$A = LU, \tag{2.10}$$

dans lequel $L = M^{-1}$ et $U = MA = A^{(n-1)}$ est une matrice triangulaire supérieure. Fait remarquable, la matrice L se calcule de manière immédiate à partir des matrices $E^{(k)}$, $1 \leq k \leq n-1$, alors qu'il n'existe pas d'expression simple pour M. En effet, chacune des matrices d'élimination définies par (2.9) étant produit de matrices de transvection, il est facile de vérifier que son inverse vaut

$$(E^{(k)})^{-1} = \prod_{i=k+1}^{n} T_{ik}\left(\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & \frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \frac{a_{k+2k}^{(k-1)}}{a_{kk}^{(k-1)}} & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & 0 & \dots & 0 & 1 \end{pmatrix}, \ 1 \leq k \leq n-1,$$

et l'on a [22]

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1}\dots(E^{(n-1)})^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \frac{a_{21}^{(0)}}{a_{11}^{(0)}} & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}} & 1 & \ddots & \vdots \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ \frac{a_{n1}^{(0)}}{a_{11}^{(0)}} & \dots & \frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & \dots & \frac{a_{nn-1}^{(n-2)}}{a_{n-1n-1}^{(n-2)}} & 1 \end{pmatrix}.$$

Si des échanges de lignes ont eu lieu lors de l'élimination, la [23] matrice M s'écrit

$$M = E^{(n-1)}P^{(n-1)}\dots E^{(2)}P^{(2)}E^{(1)}P^{(1)},$$

22. La vérification est laissée en exercice.
23. Il n'y a pas forcément unicité de la matrice dans ce cas, en raison de possibles multiples choix de pivots.

où la matrice $P^{(k)}$, $1 \leq k \leq n-1$, est soit la matrice de permutation correspondant à l'échange de lignes effectué à la k^{e} étape, soit la matrice identité si le pivot « naturel » est utilisé. En écrivant que

$$M = E^{(n-1)}(P^{(n-1)}E^{(n-2)}P^{(n-1)})\dots(P^{(n-1)}\dots P^{(2)}E^{(1)}P^{(2)}\dots P^{(n-1)})(P^{(n-1)}\dots P^{(2)}P^{(1)}),$$

et en posant $P = P^{(n-1)}\dots P^{(2)}P^{(1)}$, on obtient $L = PM^{-1}$ et $U = (MP^{-1})PA$, d'où

$$PA = LU.$$

Enfin, si la stratégie de choix de pivot employée conduit à des échanges de colonnes en plus des éventuels échanges de lignes, la factorisation prend la forme

$$PAQ = LU,$$

les matrices P et Q rendant respectivement compte des échanges de lignes et de colonnes.

Exemple d'application de la factorisation $PA = LU$. Revenons à l'exemple de mise en échec de la méthode d'élimination de Gauss, pour lequel le pivot « naturel » est nul à la seconde étape. La recherche d'un pivot partiel conduit à l'échange de la deuxième ligne avec la troisième et l'on arrive à

$$A^{(2)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix} = U.$$

Les matrices d'élimination aux deux étapes effectuées sont respectivement

$$E^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -7 & 0 & 1 \end{pmatrix} \text{ et } E^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

et les matrices d'échange sont respectivement

$$P^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } P^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

d'où

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 7 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}.$$

et la matrice de permutation

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

On a

$$LU = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} = PA.$$

La résolution du système linéaire (2.1) connaissant la factorisation LU de la matrice A se ramène à celle du système triangulaire inférieur

$$L\boldsymbol{y} = \boldsymbol{b} \tag{2.11}$$

par une méthode de descente, suivie de celle du système triangulaire supérieur

$$U\boldsymbol{x} = \boldsymbol{y} \tag{2.12}$$

par une méthode de remontée, ce que l'on accomplit en effectuant $n(n-1)$ additions, $n(n-1)$ multiplications et n divisions. Dans le cas d'une factorisation de type $PA = LU$ (resp. $PAQ = LU$), il faudra appliquer la matrice de permutation P au vecteur $\boldsymbol{b}$ de manière à tout d'abord résoudre le système $L\boldsymbol{y} = P\boldsymbol{b}$ avant de considérer le système $U\boldsymbol{x} = \boldsymbol{y}$ (resp. les systèmes $U\boldsymbol{z} = \boldsymbol{y}$ et $Q^{-1}\boldsymbol{x} = \boldsymbol{z}$).

Exemple d'application de la factorisation LU pour la résolution d'un système linéaire. Considérons la matrice

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{pmatrix}.$$

En appliquant de l'algorithme de factorisation, on arrive à

$$A = LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{pmatrix}.$$

Si $\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, la solution de $L\boldsymbol{y} = \boldsymbol{b}$ est $\boldsymbol{y} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ et celle de $U\boldsymbol{x} = \boldsymbol{y}$ est $\boldsymbol{x} = \frac{1}{3}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$.

On observera que ce type de factorisation est particulièrement avantageux lorsque l'on doit résoudre plusieurs systèmes linéaires ayant tous la même matrice et des seconds membres différents. En effet, une fois obtenue la factorisation LU de la matrice, la résolution de chaque système ne nécessite que de l'ordre de n^2 opérations. Il en est de même si l'on souhaite résoudre une *modification de rang un* (*rank-one modification* en anglais) d'un système linéaire dont la factorisation LU de la matrice est connue. Plus explicitement, étant donné une matrice A d'ordre n inversible et deux vecteurs non nuls $\boldsymbol{u}$ et $\boldsymbol{v}$ de $M_{n,1}(\mathbb{R})$ tels que $1 + \boldsymbol{v}^\top A^{-1}\boldsymbol{u} \neq 0$, supposons que l'on souhaite calculer la solution du système linéaire

$$\left(A + \boldsymbol{u}\boldsymbol{v}^\top\right)\boldsymbol{x} = \boldsymbol{b}, \tag{2.13}$$

le second membre $\boldsymbol{b}$ dans $M_{n,1}(\mathbb{R})$ étant lui aussi donné, sachant que l'on a disposition la factorisation LU de la matrice A. On peut dans ce cas se passer de la factorisation de la matrice $A + \boldsymbol{u}\boldsymbol{v}^\top$ en faisant appel à la *formule de Sherman–Morrison* [SM50] (dont la vérification est laissée en exercice),

$$(A + \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^\top A^{-1}}{1 + \boldsymbol{v}^\top A^{-1}\boldsymbol{u}}.$$

Le calcul de la solution du système (2.13) se résume alors à l'évaluation de la formule

$$A^{-1}\boldsymbol{b} - \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^\top A^{-1}\boldsymbol{b}}{1 + \boldsymbol{v}^\top A^{-1}\boldsymbol{u}},$$

ce qui requiert $2n(n-1)$ soustractions, $2n(n-1)$ multiplications et $2n$ divisions pour le calcul des vecteurs $A^{-1}\boldsymbol{b}$ et $A^{-1}\boldsymbol{u}$ à partir de la factorisation de la matrice A, $2(n-1)$ additions et $2n$ multiplications pour le calcul des scalaires $\boldsymbol{v}^\top A^{-1}\boldsymbol{b}$ et $\boldsymbol{v}^\top A^{-1}\boldsymbol{u}$ connaissant $A^{-1}\boldsymbol{b}$ et $A^{-1}\boldsymbol{u}$, et enfin une addition, n soustractions, n multiplications et une division pour finir de former le vecteur solution, soit de l'ordre de $4n^2$ opérations au total.

Terminons cette section en indiquant que la méthode de factorisation LU fournit une manière rapide de calculer le déterminant de la matrice A, qui n'est autre, au signe près, que le produit des pivots, puisque [24]

$$\det(PA) = \det(LU) = \det(L)\det(U) = \det(U) = \left(\prod_{i=1}^{n} u_{ii}\right),$$

et

$$\det(A) = \frac{\det(PA)}{\det(P)} = \begin{cases} \det(PA) & \text{si on a effectué un nombre pair d'échanges de lignes,} \\ -\det(PA) & \text{si on a effectué un nombre impair d'échanges de lignes,} \end{cases}$$

le déterminant d'une matrice de permutation étant égal à -1.

2.5.2 Condition d'existence de la factorisation LU

Commençons par donner une condition suffisante assurant qu'il n'y aura pas d'échange de lignes durant l'élimination de Gauss, ce qui conduira bien à une factorisation de la forme (2.10) de la matrice. On va à cette occasion aussi établir que cette décomposition est unique si l'on impose la valeur 1 aux éléments diagonaux de L (c'est précisément la valeur obtenue avec la construction par élimination de Gauss).

24. On laisse le lecteur traiter le cas d'une factorisation de la forme $PAQ = LU$.

Théorème 2.2 (condition suffisante d'existence et d'unicité de la factorisation LU) *Soit A une matrice d'ordre n. La factorisation LU de A, avec $l_{ii} = 1$ pour $i = 1, \ldots, n$, existe et est unique si toutes les sous-matrices principales*

$$A_k = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}, \ 1 \leq k \leq n, \tag{2.14}$$

extraites de A sont inversibles.

DÉMONSTRATION. Il est possible de montrer l'existence de la factorisation LU de manière constructive, en utilisant le procédé d'élimination de Gauss. En supposant que les n sous-matrices principales extraites de A sont inversibles, on va ici prouver en même temps l'existence et l'unicité par un raisonnement par récurrence [25].

Pour $k = 1$, on a

$$A_1 = a_{11} \neq 0,$$

et il suffit de poser $L_1 = 1$ et $U_1 = a_{11}$. Montrons à présent que s'il existe une unique factorisation de la sous-matrice A_{k-1}, $2 \leq k \leq n$, de la forme $A_{k-1} = L_{k-1}U_{k-1}$, avec $(L_{k-1})_{ii} = 1$, $i = 1, \ldots, k-1$, alors il existe une unique factorisation de ce type pour A_k. Pour cela, décomposons A_k en blocs

$$A_k = \begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{pmatrix},$$

avec $\boldsymbol{b}$ et $\boldsymbol{c}$ des matrices de $M_{k-1,1}(\mathbb{R})$ et d une matrice d'ordre 1, et cherchons une factorisation de A_k de la forme

$$\begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{pmatrix} = \begin{pmatrix} L_{k-1} & \boldsymbol{0} \\ \boldsymbol{l}^\top & 1 \end{pmatrix} \begin{pmatrix} U_{k-1} & \boldsymbol{u} \\ \boldsymbol{0}^\top & \mu \end{pmatrix}$$

où $\boldsymbol{0}$ désigne le vecteur nul de $M_{k-1,1}(\mathbb{R})$, $\boldsymbol{l}$ et $\boldsymbol{u}$ sont des matrices de $M_{k-1,1}(\mathbb{R})$ et μ est une matrice d'ordre 1. En effectuant le produit de matrices et en identifiant par blocs avec A_k, on obtient

$$L_{k-1}U_{k-1} = A_{k-1}, \ L_{k-1}\boldsymbol{u} = \boldsymbol{b}, \ \boldsymbol{l}^\top U_{k-1} = \boldsymbol{c}^\top \text{ et } \boldsymbol{l}^\top \boldsymbol{u} + \mu = d.$$

Si la première de ces égalités n'apporte aucune nouvelle information, les trois suivantes permettent de déterminer les matrices $\boldsymbol{l}$, $\boldsymbol{u}$ et μ. En effet, on a par hypothèse $0 \neq \det(A_{k-1}) = \det(L_{k-1})\det(U_{k-1})$, les matrices L_{k-1} et U_{k-1} sont donc inversibles. Par conséquent, les matrices $\boldsymbol{l}$ et $\boldsymbol{u}$ existent et sont uniques et $\mu = d - \boldsymbol{l}^\top \boldsymbol{u}$. Ceci achève la preuve par récurrence. □

Dans cette preuve, on utilise de manière fondamentale le fait les termes diagonaux de la matrice L sont tous égaux à 1. On aurait tout aussi bien pu choisir d'imposer d'autres valeurs (non nulles) ou encore décider de fixer les valeurs des éléments diagonaux de la matrice U. Ceci implique que plusieurs factorisations LU existent, chacune pouvant être déduite d'une autre par multiplication par une matrice diagonale convenable (voir la section 2.6.1).

On remarque également que la condition du théorème n'est que *suffisante*. Il n'est en effet pas nécessaire que la matrice A soit inversible pour que sa factorisation LU existe et soit unique (ce cas étant cependant le seul ayant vraiment un intérêt pratique). Nous laissons au lecteur le soin d'adapter et de compléter la démonstration précédente pour obtenir le résultat ci-après.

Théorème 2.3 (condition nécessaire et suffisante d'existence et d'unicité de la factorisation LU) *Soit A une matrice d'ordre n et de rang non nul r. Une factorisation LU de A, avec $l_{ii} = 1$ pour $i = 1, \ldots, n$, existe si et seulement si les sous-matrices principales A_k d'ordre $k = 1, \ldots, r$ extraites de A sont inversibles. Cette factorisation est de plus unique sous la même condition si et seulement si r est supérieur ou égal à $n-1$.*

Pour toute matrice A inversible, il est possible de se ramener à la condition suffisante du théorème 2.2 après des échanges préalable de lignes de la matrice (comme on l'a vu lors de la description de l'élimination de Gauss avec échange). En ce sens, la factorisation LU des matrices inversibles est toujours possible. Si une stratégie de pivot partiel, de pivot total ou de *rook pivoting* est appliquée à l'élimination de Gauss, on a plus précisément le résultat suivant.

Théorème 2.4 *Soit A une matrice d'ordre n inversible. Alors, il existe une matrice P (resp. des matrices P et Q) tenant compte d'une stratégie de pivot partiel (resp. de pivot total ou de rook pivoting), une matrice triangulaire inférieure L, dont les éléments sont inférieurs ou égaux à 1 en valeur absolue, et une matrice triangulaire supérieure U telles que*

$$PA = LU \quad (\textit{resp. } PAQ = LU).$$

25. Notons que ce procédé de démonstration permet aussi de prouver *directement* (c'est-à-dire sans faire appel à un résultat sur la factorisation LU) l'existence et l'unicité de la factorisation de Cholesky d'une matrice symétrique définie positive (voir le théorème 2.12).

2.5.3 Mise en œuvre

La matrice L étant triangulaire inférieure à diagonale ne contenant que des 1 et la matrice U triangulaire supérieure, celles-ci peuvent être commodément stockées dans le tableau contenant initialement A, les éléments non triviaux de la matrice U étant stockés dans la partie triangulaire supérieure et ceux de L occupant la partie triangulaire inférieure stricte (puisque sa diagonale est connue *a priori*). L'algorithme 7, écrit en pseudo-code, présente une première version de la factorisation LU.

Algorithme 7 : Algorithme de factorisation LU (version « kji »).

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U, stockées dans le tableau contenant la matrice A en entrée.

```
pour k = 1 à n−1 faire
    si A(k,k) ≠ 0 alors
        pour i = k+1 à n faire
            A(i,k) = A(i,k)/A(k,k)
            pour j = k+1 à n faire
                A(i,j) = A(i,j) − A(i,k) ∗ A(k,j)
            fin
        fin
    sinon
        arrêt
    fin
fin
```

Cet algorithme contient trois boucles imbriquées, portant respectivement sur les indices k, j et i. Il peut être réécrit de six manières distinctes en modifiant l'ordre des boucles et la nature des opérations sous-jacentes. Lorsque la boucle sur l'indice i précède celle sur j, on dit que l'algorithme est *orienté ligne* ; il est dit *orienté colonne* lorsque c'est l'inverse. Dans LAPACK [ABB+99], on dit que la version « kji », orientée colonne, de l'algorithme de factorisation fait appel à des opérations *saxpy* (acronyme pour *"scalar a x plus y"*), car l'opération de base de l'algorithme consiste à effectuer le produit d'un scalaire par un vecteur puis à additionner le résultat avec un vecteur. La version « jki », également orientée colonne, de l'algorithme 8 utilise des opérations *gaxpy* (acronyme pour *"generalized saxpy"*), l'opération de base étant cette fois-ci le produit d'une matrice par un vecteur, suivi de l'addition du résultat avec un vecteur.

Algorithme 8 : Algorithme de factorisation LU (version « jki »).

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U, stockées dans le tableau contenant la matrice A en entrée.

```
pour j = 1 à n faire
    pour k = 1 à j−1 faire
        pour i = k+1 à n faire
            A(i,j) = A(i,j) − A(i,k) ∗ A(k,j)
        fin
    fin
    si A(j,j) ≠ 0 alors
        pour i = j+1 à n faire
            A(i,j) = A(i,j)/A(j,j)
        fin
    sinon
        arrêt
    fin
fin
```

Terminons cette sous-section sur une variante de l'algorithme d'élimination nécessitant moins de résultats

intermédiaires (et donc d'écritures en mémoire [26]) que la méthode de Gauss « classique » pour produire la factorisation LU d'une matrice. Il s'agit de la *méthode de Doolittle* [27] (la *méthode de Crout* [28] [Cro41], également remarquable, ne diffère de cette dernière que par le choix d'avoir les éléments diagonaux de U, et non de L, tous égaux à 1). On l'obtient en remarquant que, si aucun échange de lignes n'est requis, la factorisation LU de la matrice A est formellement équivalente à la résolution du système linéaire de n^2 équations suivant

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir} u_{rj},$$

les inconnues étant les $n^2 + n$ coefficients des matrices L et U. Étant donné que les termes diagonaux de L sont fixés et égaux à 1 et en supposant les $k-1$, $2 \leq k \leq n$, colonnes de L et U sont connues, la relation ci-dessus conduit à

$$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj},\ j = k, \ldots, n,$$
$$l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rj} \right),\ i = k+1, \ldots, n,$$

ce qui permet de calculer les coefficients de manière séquentielle. Cette façon de procéder correspond à la version « ijk » de l'algorithme de factorisation. On peut remarquer que l'opération principale est à présent un produit scalaire. Une mise en œuvre de la méthode de Doolittle est proposée ci-après avec l'algorithme 9.

Algorithme 9 : Algorithme de factorisation LU (version « ijk », dite méthode de Doolittle).

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U, stockées dans le tableau contenant la matrice A en entrée.

```
pour i = 1 à n faire
    pour j = 2 à i faire
        si A(j−1, j−1) ≠ 0 alors
            A(i, j−1) = A(i, j−1)/A(j−1, j−1)
        sinon
            arrêt
        fin
        pour k = 1 à j−1 faire
            A(i, j) = A(i, j) − A(i, k) ∗ A(k, j)
        fin
    fin
    pour j = i+1 à n faire
        pour k = 1 à i−1 faire
            A(i, j) = A(i, j) − A(i, k) ∗ A(k, j)
        fin
    fin
fin
```

Les algorithmes 10 à 12 proposent des mises en œuvre de la version « kji » de la factorisation LU utilisant respectivement une stratégie de pivot partiel, de pivot total et de *rook pivoting*.

26. Ceci était particulièrement avantageux à l'époque de l'usage de calculateurs mécaniques possédant un registre dédié à l'accumulation des résultats d'opérations élémentaires.

27. Myrick Hascall Doolittle (17 mars 1830 - 27 juin 1913) était un mathématicien américain qui travailla pour la *United States coast and geodetic survey*. Il proposa en 1878 une modification de la méthode d'élimination de Gauss pour la résolution d'équations normales provenant de problèmes de triangulation.

28. Prescott Durand Crout (28 juillet 1907 - 25 septembre 1984) était un mathématicien américain. Il inventa une version du procédé d'élimination de Gauss dans laquelle une réorganisation de l'ordre des opérations arithmétiques élémentaires en permet une implémentation efficace sur un calculateur mécanique.

Algorithme 10 : Algorithme de factorisation LU avec une stratégie de pivot partiel.

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : le tableau P contenant les échanges de lignes, la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U de la factorisation $PA = LU$ étant stockées dans le tableau contenant la matrice A en entrée.

pour $k = 1$ *à* $n-1$ **faire**
  trouver un entier r tel que $k \leq r \leq n$ et $|A(r,k)| \geq |A(i,k)|$, $\forall i \in \{k,\ldots,n\}$
  $P(k) = r$
  pour $j = 1$ *à* n **faire**
    échanger $A(k,j)$ et $A(r,j)$
  fin
  si $A(k,k) \neq 0$ **alors**
    pour $i = k+1$ *à* n **faire**
      $A(i,k) = A(i,k)/A(k,k)$
      pour $j = k+1$ *à* n **faire**
        $A(i,j) = A(i,j) - A(i,k) * A(k,j)$
      fin
    fin
  sinon
    arrêt
  fin
fin

Algorithme 11 : Algorithme de factorisation LU avec une stratégie de pivot total.

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : les tableaux P et Q contenant respectivement les échanges de lignes et de colonnes, la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U de la factorisation $PAQ = LU$ étant stockées dans le tableau contenant la matrice A en entrée.

pour $k = 1$ *à* $n-1$ **faire**
  trouver des entiers r et s tels que $k \leq r \leq n$, $k \leq s \leq n$ et $|A(r,s)| \geq |A(i,j)|$, $\forall (i,j) \in \{k,\ldots,n\}^2$
  $P(k) = r$
  $Q(k) = s$
  pour $j = 1$ *à* n **faire**
    échanger $A(k,j)$ et $A(r,j)$
  fin
  pour $i = 1$ *à* n **faire**
    échanger $A(i,k)$ et $A(i,s)$
  fin
  si $A(k,k) \neq 0$ **alors**
    pour $i = k+1$ *à* n **faire**
      $A(i,k) = A(i,k)/A(k,k)$
      pour $j = k+1$ *à* n **faire**
        $A(i,j) = A(i,j) - A(i,k) * A(k,j)$
      fin
    fin
  sinon
    arrêt
  fin
fin

Algorithme 12 : Algorithme de factorisation LU avec une stratégie de *rook pivoting*.

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : les tableaux P et Q contenant respectivement les échanges de lignes et de colonnes, la partie triangulaire inférieure stricte de la matrice L et la partie triangulaire supérieure de la matrice U de la factorisation $PAQ = LU$ étant stockées dans le tableau contenant la matrice A en entrée.

pour $k = 1$ *à* $n-1$ **faire**
  répéter
    trouver un entier r tel que $k \leq r \leq n$ et $|A(r,k)| \geq |A(i,k)|$, $\forall i \in \{k,\ldots,n\}$
    trouver un entier s tel que $k \leq s \leq n$ et $|A(r,s)| \geq |A(r,j)|$, $\forall j \in \{k,\ldots,n\}$
    trouver un entier r tel que $k \leq r \leq n$ et $|A(r,s)| \geq |A(i,s)|$, $\forall i \in \{k,\ldots,n\}$
    trouver un entier s tel que $k \leq s \leq n$ et $|A(r,s)| \geq |A(r,j)|$, $\forall j \in \{k,\ldots,n\}$
  jusqu'à *ce que r et s ne soient plus modifiés*
  $P(k) = r$
  $Q(k) = s$
  pour $j = 1$ *à* n **faire**
    échanger $A(k,j)$ et $A(r,j)$
  fin
  pour $i = 1$ *à* n **faire**
    échanger $A(i,k)$ et $A(i,s)$
  fin
  si $A(k,k) \neq 0$ **alors**
    pour $i = k+1$ *à* n **faire**
      $A(i,k) = A(i,k)/A(k,k)$
      pour $j = k+1$ *à* n **faire**
        $A(i,j) = A(i,j) - A(i,k) * A(k,j)$
      fin
    fin
  sinon
    arrêt
  fin
fin

Indiquons que le choix d'un ordonnancement de boucles particulier n'affecte pas la stabilité numérique de la méthode. En revanche, l'ordre et la mise en œuvre de la stratégie de pivot à employer préférentiellement dépendent de manière cruciale de l'architecture du calculateur utilisé et de son efficacité à effectuer des opérations algébriques sur des tableaux à une ou plusieurs dimensions, ainsi que de la façon dont ces tableaux sont stockés en mémoire. Cette problématique est abordée dans de nombreuses publications, comme l'article [GPS90].

2.5.4 Factorisation LU de matrices particulières

Nous examinons dans cette section l'application de la factorisation LU à plusieurs types de matrices fréquemment rencontrées en pratique. Exploiter la structure spécifique d'une matrice peut en effet conduire à un renforcement des résultats théoriques établis dans le cas général et/ou à une réduction considérable du coût des algorithmes utilisés. À ces premiers cas particuliers, il faut ajouter ceux des matrices symétriques et symétriques définies positives, abordés respectivement dans les sous-sections 2.6.1 et 2.6.2.

Cas des matrices à diagonale strictement dominante

Certaines matrices, comme celles produites par des méthodes de discrétisation des équations aux dérivées partielles, possèdent la particularité d'être à diagonale dominante (voir la définition A.106). Le résultat suivant montre qu'une matrice à diagonale strictement dominante admet toujours une factorisation LU.

Théorème 2.5 *Si A est une matrice d'ordre n à diagonale strictement dominante (par lignes ou par colonnes) alors elle admet une unique factorisation LU. En particulier, si A est une matrice d'ordre n à diagonale strictement dominante par colonnes, on a*

$$\left|l_{ij}\right| \leq 1,\ 1 \leq i,j \leq n.$$

DÉMONSTRATION. Nous reprenons un argument provenant de [Wil61]. Supposons que A est une matrice à diagonale strictement dominante par colonnes. Posons $A^{(0)} = A$. On sait par hypothèse que

$$\left|a_{11}^{(0)}\right| > \sum_{j=2}^{n} \left|a_{j1}^{(0)}\right|,$$

et $a_{11}^{(0)}$ est donc non nul. L'application du procédé d'élimination sans échange donne

$$a_{ij}^{(1)} = a_{ij}^{(0)} - \frac{a_{ij}^{(0)}}{a_{11}^{(0)}}\, a_{1j}^{(0)},\ 2 \leq i,j \leq n,$$

d'où, $\forall j \in \{2,\ldots,n\}$,

$$\begin{aligned}
\sum_{i=2}^{n} \left|a_{ij}^{(1)}\right| &\leq \sum_{i=2}^{n} \left(\left|a_{ij}^{(0)}\right| + \left|\frac{a_{ij}^{(0)}}{a_{11}^{(0)}}\right| \left|a_{1j}^{(0)}\right| \right) \\
&\leq \sum_{i=2}^{n} \left|a_{ij}^{(0)}\right| + \left|a_{1j}^{(0)}\right| \sum_{i=2}^{n} \left|\frac{a_{ij}^{(0)}}{a_{11}^{(0)}}\right| \\
&< \sum_{i=1}^{n} \left|a_{ij}^{(0)}\right|.
\end{aligned}$$

De plus, on a

$$\begin{aligned}\left|a_{ii}^{(1)}\right| &\geq \left|a_{ii}^{(0)}\right| - \left|\frac{a_{i1}^{(0)}}{a_{11}^{(0)}}\right| \left|a_{1i}^{(0)}\right| \\ &> \sum_{\substack{j=1\\ j\neq i}}^{n} \left|a_{ji}^{(0)}\right| - \left(1 - \sum_{\substack{j=2\\ j\neq i}}^{n} \left|\frac{a_{j1}^{(0)}}{a_{11}^{(0)}}\right|\right) \left|a_{1i}^{(0)}\right| \\ &= \sum_{\substack{j=2\\ j\neq i}}^{n} \left(\left|a_{ji}^{(0)}\right| + \left|\frac{a_{j1}^{(0)}}{a_{11}^{(0)}}\right| \left|a_{1i}^{(0)}\right|\right) \\ &\geq \sum_{\substack{j=1\\ j\neq i}}^{n} \left|a_{ji}^{(1)}\right|,\end{aligned}$$

et $A^{(1)}$ est donc une matrice à diagonale strictement dominante par colonnes. Par un calcul analogue, on montre que si la matrice $A^{(k)}$, $1 \leq k \leq n-2$, est à diagonale strictement dominante par colonnes, alors $A^{(k+1)}$ l'est aussi, ce qui permet de prouver le résultat par récurrence sur k.

Dans le cas d'une matrice A à diagonale strictement dominante par lignes, on utilise que sa transposée $A^{\top}$ est à diagonale strictement dominante par colonnes et admet donc une factorisation LU. On conclut alors en utilisant la proposition 2.11. □

Cas des matrices bandes

Il est remarquable que les matrices L et U issues de la factorisation LU d'une matrice bande A sont elles-mêmes des matrices bandes, de largeurs de bande (respectivement inférieure pour L et supérieure pour U) identiques à celles de A. La zone d'espace mémoire allouée par le stockage bande, décrit dans la section 2.2, pour contenir une matrice bande est par conséquent de taille suffisante pour que l'on puisse y stocker sa factorisation.

Proposition 2.6 *La factorisation LU conserve la structure des matrices bandes.*

DÉMONSTRATION. Soit A une matrice bande d'ordre n et de largeur de bande valant $p+q+1$, admettant une factorisation LU telle que

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir} u_{rj},\ 1 \leq i, j \leq n.$$

Pour prouver l'assertion, on raisonne par récurrence sur l'indice $k = \min(i, j)$. Pour $k = 1$, on obtient d'une part

$$a_{1j} = l_{11} u_{1j} = u_{1j},\ 1 \leq j \leq n,$$

d'où $u_{1j} = 0$ si $j > q+1$, et d'autre part

$$a_{i1} = l_{i1} u_{11},\ 1 \leq i \leq n.$$

En particulier, on a $a_{11} = l_{11}u_{11} = u_{11}$ et donc $u_{11} \neq 0$. Par conséquent, on trouve que

$$l_{i1} = \frac{a_{i1}}{u_{11}},\ 1 \leq i \leq n,$$

d'où $l_{i1} = 0$ si $i > p+1$.

Supposons à présent que, pour tout $k = 1, \ldots, K-1$ avec $2 \leq K \leq n$, on ait

$$u_{kj} = 0 \text{ si } j > q+k \text{ et } l_{ik} = 0 \text{ si } i > p+k.$$

Soit $j > q+K$. Pour tout $r = 1, \ldots, K-1$, on a dans ce cas $j > q+K \geq q+r+1 > q+r$ et, par hypothèse de récurrence, le coefficient u_{rj}. Ceci implique alors

$$0 = a_{Kj} = \sum_{r=1}^{K} l_{ir} u_{rj} = l_{KK} u_{Kj} + \sum_{r=1}^{K-1} l_{Kr} u_{rj} = u_{Kj},\ j > q+K.$$

De la même manière, on prouve que

$$0 = a_{iK} = l_{iK} u_{KK} + \sum_{r=1}^{K-1} l_{ir} u_{rK} = l_{iK} u_{KK},\ i > p+K,$$

et on conclut en utilisant que u_{KK} est non nul, ce qui achève la démonstration par récurrence. □

Cas des matrices tridiagonales

On considère dans cette section un cas particulier de matrice bandes : les matrices *tridiagonales*, dont seules la diagonale principale et les deux diagonales qui lui sont adjacentes possèdent des éléments non nuls.

Définition 2.7 (matrice tridiagonale) *Soit n un entier supérieur ou égal à 3. On dit qu'une matrice A de $M_n(\mathbb{R})$ est* ***tridiagonale*** *si*

$$\forall (i,j) \in \{1,\dots,n\}^2,\ a_{ij} = 0 \ si\ |i-j| > 1.$$

Considérons [29] la matrice réelle tridiagonale d'ordre n

$$A = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ \gamma_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & 0 & \gamma_n & \alpha_n \end{pmatrix}.$$

en supposant qu'elle soit inversible et admette, sans qu'il y ait besoin d'échange de lignes, une factorisation LU (c'est le cas par exemple si elle est à diagonale strictement dominante par lignes, *i.e.*, $|\alpha_1| > |\beta_1|$, $|\alpha_i| > |\beta_i| + |\gamma_i|$, $i = 2, dots, n-1$, et $|\alpha_n| > |\gamma_n|$). Dans ce cas, les matrices L et U de la factorisation sont de la forme

$$L = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & l_n & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_1 & v_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & v_{n-1} \\ 0 & \dots & \dots & 0 & u_n \end{pmatrix},$$

et une identification terme à terme entre A et le produit LU conduit aux relations suivantes

$$v_i = \beta_i,\ i = 1,\dots,n-1,\ u_1 = \alpha_1,\ l_j = \frac{\gamma_j}{u_{j-1}} \text{ et } u_j = \alpha_j - l_j\,\beta_{j-1},\ j = 2,\dots,n.$$

Cette méthode spécifique de factorisation LU d'une matrice tridiagonale est connue sous le nom d'*algorithme de Thomas* [30] [Tho49] et constitue un cas particulier de factorisation de Doolittle sans changement de pivot. Elle requiert $n-1$ soustractions et autant de multiplications et de divisions pour factoriser une matrice d'ordre n.

Si l'on souhaite résoudre le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec $\boldsymbol{b}$ dans $M_{n,1}(\mathbb{R})$, dans lequel A est une matrice tridiagonale factorisable, on doit de plus, une fois la factorisation obtenue, résoudre les systèmes $L\boldsymbol{y} = \boldsymbol{b}$ et $U\boldsymbol{x} = \boldsymbol{y}$. Les méthodes de descente et de remontée se résument dans ce cas particulier aux formules

$$y_1 = b_1,\ y_i = b_i - l_i\,y_{i-1},\ i = 2,\dots,n,$$

et

$$x_n = \frac{y_n}{u_n},\ x_j = \frac{1}{u_j}\left(y_j - v_j\,x_{j+1}\right),\ j = n-1,\dots,1,$$

ce qui revient à effectuer $2(n-1)$ soustractions, $2(n-1)$ multiplications et n divisions. La résolution d'un système linéaire tridiagonal nécessite donc un total de $8n-7$ opérations, ce qui représente une diminution importante de coût par rapport au cas général.

29. Par commodité, on a posé $\alpha_i = a_{ii}$, $i = 1,\dots,n$, $\beta_i = a_{ii+1}$, $i = 1,\dots,n-1$, et $\gamma_i = a_{ii-1}$, $i = 2,\dots,n$.

30. Llewellyn Hilleth Thomas (21 octobre 1903 - 20 avril 1992) était un physicien et mathématicien britannique. Il est connu pour ses contributions en physique atomique, et plus particulièrement la *précession de Thomas* (une correction relativiste qui s'applique au spin d'une particule possédant une trajectoire accélérée) et le *modèle de Thomas–Fermi* (un modèle statistique d'approximation de la distribution des électrons dans un atome à l'origine de la théorie de la fonctionnelle de densité).

Cas des matrices de Hessenberg

Un autre type de matrice apparaissant de manière récurrente dans les applications est celui des *matrices de Hessenberg* [31].

Définition 2.8 (matrice de Hessenberg) *Une matrice carrée d'ordre n est dite de Hessenberg supérieure (**upper Hessenberg matrix** en anglais) si tous ses coefficients situés au dessous de sa première sous-diagonale sont nuls :* $\forall (i,j) \in \{1,\dots,n\}^2,\ i > j+1 \implies a_{ij} = 0$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & & & \vdots \\ 0 & a_{32} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn-1} & a_{nn} \end{pmatrix}.$$

*Elle est dite de Hessenberg inférieure (**lower Hessenberg matrix** en anglais) si la matrice* $A^\top$ *est de Hessenberg supérieure.*

On remarque qu'une matrice tridiagonale est à la fois de Hessenberg supérieure et inférieure.

Considérons la factorisation LU d'une matrice de Hessenberg supérieure d'ordre n notée A, en supposant dans un premier temps qu'il n'y a pas besoin de procéder à des échanges de lignes pour y parvenir. Comme précédemment, on pose $A^{(0)} = A$ et l'on construit une suite de matrices $(A^{(k)})_{0 \le k \le n-1}$ en effectuant des opérations sur les lignes de ces matrices de manière à obtenir une matrice $A^{(n-1)}$ triangulaire supérieure. Compte tenu de la structure particulière de la matrice A, il n'y a qu'un coefficient à annuler à chaque étape et seuls les coefficients de la même ligne et situés à droite de la colonne à laquelle appartient le pivot sont modifiés. L'algorithme d'élimination prend ainsi la forme suivante :

$$a_{k+2\,j}^{(k+1)} = a_{k+2\,j}^{(k)} - \frac{a_{k+2\,k+1}^{(k)}}{a_{k+1\,k+1}^{(k)}}\, a_{k+1\,j}^{(k)},\ k = 0,\dots,n-2,\ j = k+1,\dots,n,$$

les autres coefficients étant inchangés. Pour factoriser la matrice, on a ainsi besoin de $\frac{1}{2}n(n-1)$ soustractions, $\frac{1}{2}n(n-1)$ multiplications et $n-1$ divisions. Si le coût de la résolution du système triangulaire supérieur (2.12) obtenu est *a priori* le même que pour une matrice quelconque, celui de la résolution du système triangulaire inférieur (2.11) est en revanche ramené dans ce cas à $n-1$ soustractions et $n-1$ multiplications.

Si l'on est dans l'obligation de changer de pivot (parce que le pivot « naturel » est nul à une étape donnée ou bien si l'on emploie une stratégie de choix de pivot pour assurer une meilleure stabilité numérique), il faut remarquer que le choix d'un pivot total ne préserve généralement pas la structure de Hessenberg de la matrice. En revanche, la stratégie de pivot partiel consiste dans ce cas à choisir le pivot parmi deux seuls candidats à chaque étape (les éléments $a_{k+1\,k+1}^{(k)}$ et $a_{k+2\,k+1}^{(k)}$), ce qui n'introduit aucune modification structurelle en cas d'échange de lignes.

Cas des matrices de Toeplitz

Les *matrices de Toeplitz* [32] jouent un rôle prépondérant dans de nombreuses applications en physique mathématique, en statistique ou en algèbre, pour des problèmes de convolution, d'équations intégrales, d'approximation au sens des moindres carrés par des polynômes ou de séries temporelles stationnaires.

Définition 2.9 (matrice de Toeplitz) *Soit n un entier naturel non nul. Une matrice carrée A d'ordre n est dite de*

31. Karl Adolf Hessenberg (8 septembre 1904 - 22 février 1959) était un mathématicien et ingénieur allemand. Il s'intéressa à la résolution de problèmes aux valeurs propres et introduisit à cette occasion les matrices particulières portant aujourd'hui son nom.

32. Otto Toeplitz (1er août 1881 - 15 février 1940) était un mathématicien allemand, principalement connu pour ses travaux en analyse fonctionnelle.

Toeplitz s'il existe au plus $2n-1$ *scalaires distincts* $a_{-n+1},\dots,a_0,\dots,a_{n-1}$ *tels que* $a_{ij}=a_{j-i}$, $1\le i,j\le n$, *c'est-à-dire*

$$A=\begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{n-1}\\ a_{-1} & a_0 & a_1 & \dots & a_{n-2}\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ a_{-n+2} & \dots & a_{-1} & a_0 & a_1\\ a_{-n+1} & \dots & a_{-2} & a_{-1} & a_0 \end{pmatrix}.$$

Toute matrice de Toeplitz A est dite *persymétrique* (*persymmetric* en anglais), c'est-à-dire que ses coefficients sont symétriques par rapport à l'*antidiagonale* et qu'elle vérifie par conséquent l'égalité

$$A=J_nA^\top J_n,$$

où J_n est la matrice de permutation d'ordre n telle que

$$J_n=\begin{pmatrix} 0 & \dots & 0 & 1\\ \vdots & \iddots & \iddots & 0\\ 0 & \iddots & \iddots & \vdots\\ 1 & 0 & \dots & 0 \end{pmatrix}.$$

La *méthode de Bareiss* [Bar69] permet la résolution [33] d'un système linéaire dont la matrice est de Toeplitz en la ramenant, à la manière du procédé d'élimination de Gauss, à celle d'un système triangulaire équivalent. Pour une matrice de Toeplitz A d'ordre n, elle construit, à partir de $A^{(0)}=A$, deux familles de matrices $(A^{(-k)})_{0\le k\le n-1}$ et $(A^{(k)})_{0\le k\le n-1}$, telles que la matrice $A^{(-k)}$ (resp. $A^{(k)}$), $1\le k\le n-1$, possède des éléments nuls sur les k premières diagonales situées au-dessous (resp. au-dessus) de la diagonale principale, les dernières (resp. premières) $n-k$ lignes de cette matrice formant une matrice rectangulaire ayant encore une structure de Toeplitz. Plus explicitement, on a

$$A^{(-k)}=\begin{pmatrix} a_0^{(0)} & a_1^{(0)} & \dots & \dots & \dots & \dots & \dots & a_{n-1}^{(0)}\\ 0 & a_0^{(-1)} & & & & & & \vdots\\ \vdots & \ddots & \ddots & & & & & \vdots\\ \vdots & & \ddots & a_0^{(-k+1)} & a_1^{(-k+1)} & \dots & \dots & a_{n-k+1}^{(-k+1)}\\ 0 & & & \ddots & a_0^{(-k)} & a_1^{(-k)} & \dots & a_{n-k}^{(-k)}\\ a_{-k-1}^{(-k)} & \ddots & & & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & & & \ddots & \ddots & a_1^{(-k)}\\ a_{-n+1}^{(-k)} & \dots & a_{-k-1}^{(-k)} & 0 & \dots & \dots & 0 & a_0^{(-k)} \end{pmatrix},$$

et

$$A^{(k)}=\begin{pmatrix} a_0^{(0)} & 0 & \dots & \dots & 0 & a_{k+1}^{(k)} & \dots & a_{n+1}^{(k)}\\ a_{-1}^{(k)} & \ddots & \ddots & & & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & & & \ddots & a_{k+1}^{(k)}\\ a_{-n+k}^{(k)} & \dots & a_{-1}^{(k)} & a_0^{(0)} & 0 & \dots & \dots & 0\\ a_{-n+k-1}^{(k-1)} & \dots & \dots & a_{-1}^{(k-1)} & a_0^{(0)} & \ddots & & \vdots\\ \vdots & & & & \ddots & \ddots & \ddots & \vdots\\ \vdots & & & & & \ddots & \ddots & 0\\ a_{-n+1}^{(0)} & \dots & \dots & \dots & \dots & \dots & a_{-1}^{(0)} & a_0^{(0)} \end{pmatrix}.$$

33. D'autres méthodes directes, d'efficacité comparable, existent, comme celles de Levinson [Lev47] pour la résolution d'un système linéaire de Toeplitz général, de Durbin [Dur60] pour la résolution d'un système d'*équations de Yule–Walker* ou de Trench [Tre64, Zoh69] pour le calcul de l'inverse d'une matrice de Toeplitz. Celles-ci possèdent toutes, lorsque la matrice de Toeplitz considérée est de plus symétrique ou hermitienne, un lien fort avec la théorie des polynômes orthogonaux relativement à une mesure sur le cercle unité formulée par Szegő [Sze39].

Pour obtenir la matrice $A^{(-k)}$ (resp. $A^{(k)}$), $1 \leq k \leq n-1$, on modifie les lignes $k+1$ à n (resp. 1 à $n-k$) de la matrice $A^{(-k+1)}$ (resp. $A^{(k-1)}$), ce que l'on peut résumer [34] dans les algorithmes de transformation suivants

$$a_i^{(0)} = a_i,\ i = -n+1, \ldots, n-1,$$

$$a_i^{(-k)} = a_i^{(-k+1)} - m^{(-k)}\, a_{i+k}^{(k-1)},\ m^{(-k)} = \frac{a_{-k}^{(-k+1)}}{a_0^{(k-1)}},\ i = -n+1, \ldots, -k-1, 0, \ldots, n-1-k,$$
$$k = 1, \ldots, n-1, \quad (2.15)$$

$$a_i^{(k)} = a_i^{(k-1)} - m^{(k)}\, a_{i-k}^{(-k)},\ m^{(k)} = \frac{a_k^{(k-1)}}{a_0^{(-k)}},\ i = -n+1+k, \ldots, -1, k+1, \ldots, n-1,$$
$$k = 1, \ldots, n-1. \quad (2.16)$$

Les relations (2.15) demandent d'effectuer, pour chaque valeur considérée de l'entier k, $2n-2k-1$ soustractions, $2n-2k-1$ multiplications et une division, tandis que les relations (2.16) nécessitent $2(n-k-1)$ soustractions, $2(n-k-1)$ multiplications et une division. L'application de la méthode entraîne donc un coût de $2n^2-5n-3$ soustractions, autant de multiplications et $2(n-1)$ divisions au total.

Interprétées matriciellement, les relations (2.15) et (2.16) sur les coefficients s'écrivent comme

$$A^{(-k)} = A^{(-k+1)} - m^{(-k)}\, Z^{(-k)} A^{(k-1)},\ A^{(k)} = A^{(k-1)} - m^{(k)}\, Z^{(k)} A^{(-k)},\ k = 1, \ldots, n-1, \quad (2.17)$$

où l'on a introduit les matrices $Z^{(\pm k)}$ d'ordre n, définies [35] par $z_{ij}^{(\pm k)} = \delta_{i\pm k, j}$, $1 \leq i, j \leq n$. En posant alors

$$A^{(\pm k)} = M^{(\pm k)} A,\ k = 1, \ldots, n-1, \quad (2.18)$$

avec $M^{(0)} = I_n$. Si le procédé arrive à son terme, ce qui est le cas si la matrice A vérifie l'hypothèse [36] du théorème 2.2, on montre en utilisant les relations de récurrence, déduites de (2.17), satisfaites par les matrices $M^{(\pm k)}$, $k = 1, \ldots, n-1$ que la matrice $M^{(-n+1)}$ est triangulaire inférieure et que ses éléments diagonaux sont tous égaux à 1. La matrice $A^{(-n+1)}$ étant triangulaire supérieure, il découle de l'unicité de la factorisation LU et de la définition (2.18) que

$$L = \left(M^{(-n+1)}\right)^{-1} \text{ et } U = A^{(-n+1)}.$$

D'autre part, en considérant la factorisation UL de A de matrices $\tilde{U}$ et $\tilde{L}$, on obtient, par des observations et un raisonnement similaires à ceux faits précédemment, que

$$\tilde{U} = a_0 \left(M^{(n-1)}\right)^{-1} \text{ et } \tilde{L} = {a_0}^{-1} A^{(n-1)}.$$

En se servant de la propriété de persymétrie de la matrice A, il vient alors que

$$L = {a_0}^{-1} J_n A^{(n-1)} J_n.$$

La méthode de Bareiss fournit par conséquent la factorisation LU effective d'une matrice de Toeplitz, en un nombre d'opérations arithmétiques bien inférieur à celui de l'élimination de Gauss.

34. On observera avec ces formules que la mise en œuvre de la méthode demande un effort particulier, en raison de la structure de stockage spécifique utilisée pour les matrices de Toeplitz et du fait que les matrices $A^{(k)}$ et $A^{(-k)}$, $k = 1, \ldots, n-1$, ne sont *a priori* pas des matrices de Toeplitz.

35. On remarquera que le produit par la gauche avec $Z^{(k)}$ (resp. $Z^{(-k)}$), $1 \leq k \leq n-1$, d'une matrice d'ordre n décale les lignes de cette matrice de k positions vers le haut (resp. le bas).

36. On peut en effet montrer (voir [Bar69]) que le mineur principal dominant d'ordre k, $k = 1, \ldots, n$, de la matrice A vaut

$$\begin{vmatrix} a_0 & \cdots & a_{k-1} \\ \vdots & \ddots & \vdots \\ a_{-k+1} & \cdots & a_0 \end{vmatrix} = a_0\, a_0^{(-1)} \ldots a_0^{(-k+1)}.$$

Phénomène de remplissage lors de la factorisation de matrices creuses

Un des inconvénients de la factorisation LU appliquée à une matrice creuse est qu'elle entraîne l'apparition de termes non nuls dans les matrices triangulaires L et U à des endroits où les éléments de la matrice initiale sont nuls. Ce phénomène, connu sous le nom de *remplissage* (*fill-in* en anglais), pose problème, puisque certains des modes de stockage utilisés [37] pour la matrice à factoriser ne peuvent alors pas contenir sa factorisation.

Exemple de remplissage lors de la factorisation d'une matrice creuse. Considérons la matrice

$$A = \begin{pmatrix} 4 & -1 & -1 & -1 \\ -1 & 2 & 0 & 0 \\ -1 & 0 & 2 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix},$$

dont seules la première ligne, la première colonne et la diagonale principale contiennent des éléments non nuls et qui est, pour cette raison, dite *en pointe de flèche* (*arrowhead matrix* en anglais). La factorisation LU de cette matrice sans stratégie de choix de pivot [38], c'est-à-dire en utilisant à chaque étape de l'élimination le pivot « naturel », conduit aux matrices

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{4} & -\frac{1}{7} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{7} & -\frac{1}{6} & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} 4 & -1 & -1 & -1 \\ 0 & \frac{7}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & \frac{12}{7} & -\frac{2}{7} \\ 0 & 0 & 0 & \frac{5}{3} \end{pmatrix},$$

dont le stockage ne peut être fait dans une structure de donnée spécialement adaptée aux matrices en pointe de flèche (en ne conservant en mémoire que les éléments de la première ligne, de la première colonne et de la diagonale principale). En revanche, la factorisation de la matrice permutée $PAP^\top$, avec

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

donne

$$PAP^\top = \begin{pmatrix} 2 & 0 & 0 & -1 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 2 & -1 \\ -1 & -1 & -1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 & -1 \\ 0 & 2 & 0 & -1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & \frac{5}{2} \end{pmatrix},$$

et ne pose donc aucun problème de stockage.

Si le cas des matrices en pointe de flèche est particulièrement symptomatique [39] du phénomène de remplissage, il est cependant loin d'être isolé.

Notons que l'on peut anticiper ce remplissage en réalisant préalablement une *factorisation symbolique* (*symbolic factorisation* en anglais) de la matrice, qui consiste à déterminer le nombre et la position des coefficients non nuls créés au cours de la factorisation effective, ce qui permet d'adapter le stockage en conséquence. Une autre approche est de chercher à le limiter en procédant à une renumérotation des inconnues et des équations du système linéaire en question. Si le problème de détermination de la renumérotation minimisant le remplissage d'une matrice creuse est en général NP-complet [Yan81], des méthodes heuristiques ont été proposées.

Une première approche est de modifier la façon de choisir le pivot pour tenter de minimiser le remplissage, comme dans la stratégie élaborée par Markowitz [Mar57]. Cette dernière est basée sur une vision « locale » du problème, au sens où le remplissage à chaque étape est minimisé sans envisager le remplissage effectif après $n-1$ étapes d'élimination. Pour cela, on introduit à l'étape $k+1$, $k = 0, \ldots, n-2$, la quantité, appelée *compte de Markowitz* (*Markowitz count* en anglais), définie par le produit

$$(r_i^{(k)} - 1)(c_j^{(k)} - 1),$$

37. Ce n'est pas le cas des stockages bande ou profil, comme on l'a vu plus haut, mais de tout stockage ne gardant en mémoire que les éléments non nuls de la matrice, comme le stockage Morse.

38. La matrice admet en effet une factorisation LU en vertu du théorème 2.2, puisque ses mineurs principaux dominants valent respectivement 4, 7, 12 et 20, et sont donc non nuls (notons que l'on peut aussi invoquer le théorème 2.5 en remarquant que la matrice est à diagonale strictement dominante par lignes).

39. Il faut en effet ajouter que le coût de la factorisation d'une matrice en pointe de flèche d'ordre n sous la forme permutée proposée dans l'exemple possède une complexité en $O(n)$, contre $O(n^3)$ sous sa forme originelle.

dans lequel les entiers $r_i^{(k)}$ et $c_j^{(k)}$ sont respectivement le nombre d'éléments non nuls dans la ligne i et le nombre d'éléments non nuls dans la colonne j de la sous-matrice d'ordre $n-k$ de la matrice courante $A^{(k)}$ dans laquelle l'élimination doit être faite. Le pivot de Markowitz est alors choisi parmi les éléments non nuls dont les indices de ligne et de colonne minimisent le compte. Ce procédé peut être modifié afin d'y introduite davantage de stabilité numérique, en rejetant notamment les éléments candidats considérés comme trop petits en valeur absolue. Quand c'est le cas, on est conduit à décider d'un compromis entre la préservation du caractère creux d'une part et la stabilité des calculs d'autre part (voir [DR79]).

Une autre technique consiste à réduire la largeur de bande de la matrice avant la factorisation, sachant que cette dernière préserve la structure bande (voir la proposition 2.6). Ceci peut se faire en voyant la matrice creuse à factoriser comme une *matrice d'adjacence* [40] pour un graphe. Dans le cas d'une matrice symétrique, un exemple d'un tel procédé est donné par l'*algorithme de Cuthill–McKee* [CM69]. Celui-ci vise à renuméroter les sommets du graphe sans changer la structure de ce dernier, de manière à réduire la largeur de bande de la matrice d'adjacence associée.

Pour produire ce réordonnancement, on procède de la manière suivante. On choisit tout d'abord un sommet du graphe de plus petit degré, c'est-à-dire dont le nombre de voisins dans le graphe est le plus petit, comme le premier numéroté. On numérote ensuite chacun de ses voisins, par ordre de degré croissant (si deux voisins possèdent le même degré, on les numérote dans un ordre arbitraire). On poursuit le processus en numérotant les voisins des sommets déjà numérotés, par ordre de degré croissant dans le graphe ne contenant plus les sommets déjà numérotés, jusqu'à exhaustion des sommets encore non numérotés.

Il a été observé par George [Geo71] qu'une meilleure réduction était généralement atteinte en renversant la numérotation obtenue avec cet algorithme, donnant lieu à l'*algorithme de Cuthill–McKee inverse* (*reverse Cuthill–McKee algorithm* en anglais) utilisé en pratique (voir la figure 2.1 pour un exemple).

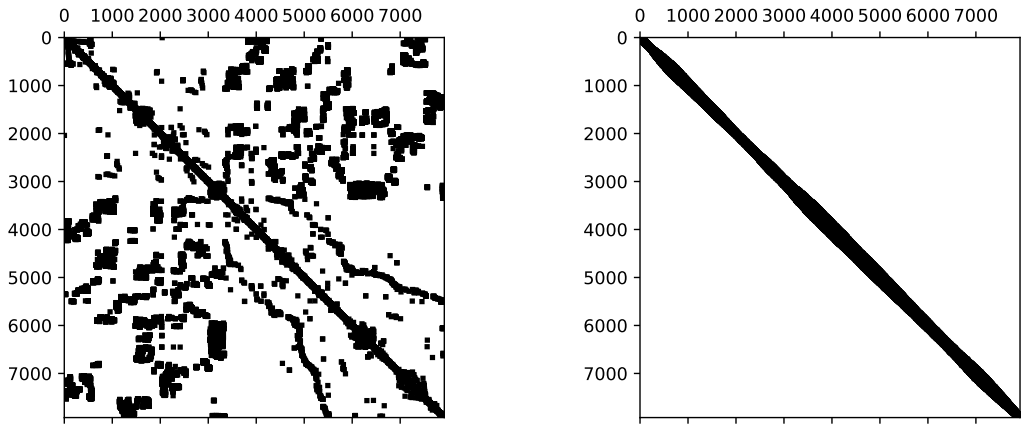


FIGURE 2.1 – Exemple d'application de l'algorithme de Cuthill–McKee inverse pour la réduction de la largeur de bande de la matrice creuse symétrique d'ordre 7920 `commanche_dual` du groupe Pothen dans la SuiteSparse Matrix Collection. On a représenté à gauche les positions des éléments non nuls de la matrice (au nombre de 31680) et à droite celles de la permutation de la matrice, obtenue après renumérotation par l'algorithme.

On pourra consulter l'article [GPS76] pour la présentation et la comparaison d'autres méthodes de réduction de largeur de bande et de profil.

2.6 Autres méthodes de factorisation

Nous présentons dans cette dernière section d'autres types de factorisation, adaptés à des matrices particulières. Il s'agit de la *factorisation* $LDM^\top$ d'une matrice carrée, qui devient la *factorisation* $LDL^\top$ lorsque cette matrice est

40. Usuellement, une matrice d'adjacence pour un graphe fini à n sommets est une matrice d'ordre n dont l'élément non diagonal situé à la i^{e} ligne et à la j^{e} colonne est le nombre d'arêtes liant le sommet i au sommet j et l'élément diagonal situé à la i^{e} ligne est le nombre de boucles au sommet i. Dans l'interprétation de la matrice faite par l'algorithme de Cuthill-McKee, on considère simplement que les sommets i et j du graphe sont reliés par une arête si l'élément situé à la i^{e} ligne et à la j^{e} colonne est non nul.

symétrique, de la *factorisation de Cholesky* [41], pour une matrice réelle *symétrique définie positive*, et de la *factorisation QR*, que l'on peut généraliser aux matrices rectangulaires (dans le cadre de la résolution d'un problème aux moindres carrés par exemple) ou bien carrées, mais non inversibles.

2.6.1 Factorisation LDM$^\top$

Cette méthode considère une décomposition sous la forme d'un produit d'une matrice triangulaire inférieure, d'une matrice diagonale et d'une matrice triangulaire supérieure. Une fois obtenue la factorisation de la matrice A (d'un coût identique à celui de la factorisation LU), la résolution du système linéaire (2.1) fait intervenir la résolution d'un système triangulaire inférieur (par une méthode de descente), puis celle (triviale) d'un système diagonal et enfin la résolution d'un système triangulaire supérieur (par une méthode de remontée), ce qui représente un coût de $n^2 + n$ opérations.

Proposition 2.10 *Sous les hypothèses du théorème 2.2, il existe une unique matrice triangulaire inférieure L, une unique matrice diagonale D et une unique matrice triangulaire supérieure $M^\top$, les éléments diagonaux de L et M étant tous égaux à 1, telles que*

$$A = LDM^\top.$$

DÉMONSTRATION. Les hypothèses du théorème 2.2 étant satisfaites, on sait qu'il existe une unique factorisation LU de la matrice A. En choisissant les éléments diagonaux de la matrice D égaux à u_{ii}, $1 \leq i \leq n$, (tous non nuls puisque la matrice U est inversible), on a

$$A = LU = LDD^{-1}U.$$

Il suffit alors de poser $M^\top = D^{-1}U$ pour obtenir l'existence de la factorisation. Son unicité est une conséquence de l'unicité de la factorisation LU. □

Lorsque la matrice A considérée est inversible, la factorisation LDM$^\top$ permet également de démontrer simplement le résultat suivant, sans qu'il y ait besoin d'avoir recours au théorème 2.2.

Proposition 2.11 *Soit A une matrice carrée d'ordre n inversible admettant une factorisation LU. Alors, sa transposée $A^\top$ admet une factorisation LU.*

DÉMONSTRATION. Puisque A admet une factorisation LU, elle admet aussi une factorisation LDM$^\top$ et l'on a

$$A^\top = (LDM^\top)^\top = (M^\top)^\top D^\top L^\top = MDL^\top.$$

La matrice $A^\top$ admet donc elle aussi une factorisation LDM$^\top$ et, par suite, une factorisation LU. □

L'intérêt de la factorisation LDM$^\top$ devient clair lorsque la matrice A est symétrique, puisque $M = L$ dans ce cas. La factorisation résultante peut alors être calculée avec un coût et un stockage environ deux fois moindres que ceux d'une factorisation LU classique. Cependant, comme pour cette dernière méthode, il n'est pas conseillé [42], pour des questions de stabilité numérique, d'utiliser cette factorisation si la matrice A n'est pas réelle symétrique définie positive ou à diagonale dominante. De manière générale, tout système linéaire pouvant être résolu au moyen de la factorisation de Cholesky (introduite dans la section 2.6.2 ci-après) peut également l'être par la factorisation LDL$^\top$ et, lorsque la matrice de ce système est une matrice bande (par exemple tridiagonale), il s'avère plus avantageux de préférer la seconde méthode, les extractions de racines carrées requises par la première représentant, dans ce cas particulier, une fraction importante du nombre d'opérations arithmétiques effectuées.

2.6.2 Factorisation de Cholesky

Une matrice réelle symétrique définie positive vérifiant les hypothèses du théorème 2.2 en vertu du critère de Sylvester (voir le théorème A.151), elle admet une factorisation LDL$^\top$, dont la matrice D est de plus à termes diagonaux *strictement positifs*. Cette observation conduit à une factorisation ne faisant intervenir qu'une seule matrice triangulaire inférieure, appelée *factorisation de Cholesky* [Cho, Ben24]. Plus précisément, on a le résultat suivant.

41. André-Louis Cholesky (15 octobre 1875 - 31 août 1918) était un mathématicien et officier français. Il inventa, alors qu'il effectuait une carrière dans les services géographiques et topographiques de l'armée, une méthode pour la résolution de systèmes d'équations linéaires dont la matrice est symétrique définie positive.

42. Dans les autres situations, on se doit de faire appel à des stratégies de choix de pivot conservant le caractère symétrique de la matrice à factoriser, c'est-à-dire trouver une matrice de permutation P telle que la factorisation LDL$^\top$ de $PAP^\top$ soit stable. Nous renvoyons aux notes de fin de chapitre pour plus de détails sur les approches possibles.

Théorème 2.12 (« factorisation de Cholesky ») *Soit A une matrice réelle symétrique définie positive. Alors, il existe une unique matrice triangulaire inférieure B, dont les éléments diagonaux sont strictement positifs, telle que*

$$A = BB^\top.$$

DÉMONSTRATION. Supposons que la matrice A est d'ordre n. On sait, par le théorème A.151, que les déterminants des sous-matrices principales extraites A_k, $1 \leq k \leq n$, de A (définies par (2.14)), sont strictement positifs et les conditions du théorème 2.2 sont vérifiées. La matrice A admet donc une unique factorisation LU. Les éléments diagonaux de la matrice U sont de plus strictement positifs, car on a

$$\prod_{i=1}^{k} u_{ii} = \det(A_k) > 0,\ 1 \leq k \leq n.$$

En introduisant la matrice diagonale Δ définie par $(\Delta)_{ii} = \sqrt{u_{ii}}$, $1 \leq i \leq n$, la factorisation se réécrit

$$A = L\Delta\Delta^{-1}U.$$

En posant $B = L\Delta$ et $C = \Delta^{-1}U$, la symétrie de A entraîne que $BC = C^\top B^\top$, d'où $C(B^\top)^{-1} = B^{-1}C^\top = I_n$ (une matrice étant triangulaire supérieure, l'autre triangulaire inférieure et toutes deux à coefficients diagonaux égaux à 1) et donc $C = B^\top$. On a donc montré l'existence d'au moins une factorisation de Cholesky. Pour montrer l'unicité de cette décomposition, on suppose qu'il existe deux matrices triangulaires inférieures B_1 et B_2 telles que

$$A = B_1{B_1}^\top = B_2{B_2}^\top,$$

d'où ${B_2}^{-1}B_1 = {B_2}^\top(B_1^\top)^{-1}$. Il existe donc une matrice diagonale D telle que ${B_2}^{-1}B_1 = D$ et, par conséquent, $B_1 = B_2D$. Finalement, on a

$$B_2{B_2}^\top = B_1{B_1}^\top = B_2DD^\top{B_2}^\top,$$

et donc $D^2 = I_n$. Les coefficients diagonaux d'une matrice de factorisation de Cholesky étant par hypothèse positifs, on a nécessairement $D = I_n$ et donc $B_1 = B_2$. □

Pour la mise en œuvre de cette factorisation, on procède de la manière suivante. On pose $B = (b_{ij})_{1\leq i,j\leq n}$ avec $b_{ij} = 0$ si $i < j$ et l'on déduit alors de l'égalité $A = BB^\top$ que

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{\min(i,j)} b_{ik}b_{jk},\ 1 \leq i,j \leq n.$$

La matrice A étant réelle symétrique, il suffit, par exemple, que les relations ci-dessus soient vérifiées pour $j \leq i$ et l'on construit alors les colonnes de la matrice B à partir de celles de A. En fixant l'indice j à 1 et en faisant varier l'indice i de 1 à n, on trouve

$$\begin{array}{rcl@{\quad}rcl}
a_{11} &=& (b_{11})^2, & \text{d'où} \quad b_{11} &=& \sqrt{a_{11}}, \\
a_{21} &=& b_{11}b_{21}, & \text{d'où} \quad b_{21} &=& \dfrac{a_{21}}{b_{11}}, \\
\vdots & & & \vdots & & \\
a_{n1} &=& b_{11}b_{n1}, & \text{d'où} \quad b_{n1} &=& \dfrac{a_{n1}}{b_{11}},
\end{array}$$

ce qui permet la détermination de la première colonne de B. Les coefficients de la j^{e} colonne de B, $2 \leq j \leq n$, s'obtiennent en utilisant les relations

$$\begin{array}{rcl@{\quad}rcl}
a_{jj} &=& (b_{j1})^2 + (b_{j2})^2 + \cdots + (b_{jj})^2, & \text{d'où} \quad b_{jj} &=& \sqrt{a_{jj} - \sum_{k=1}^{j-1}(b_{jk})^2}, \\
a_{j+1j} &=& b_{j1}b_{j+11} + b_{j2}b_{j+12} + \cdots + b_{jj}b_{j+1j}, & \text{d'où} \quad b_{j+1j} &=& \dfrac{a_{j+1j} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}}, \\
\vdots & & & \vdots & & \\
a_{nj} &=& b_{j1}b_{n1} + b_{j2}b_{n2} + \cdots + b_{jj}b_{nj}, & \text{d'où} \quad b_{nj} &=& \dfrac{a_{nj} - \sum_{k=1}^{j-1} b_{jk}b_{nk}}{b_{jj}},
\end{array}$$

après avoir préalablement déterminé les $j-1$ premières colonnes, le théorème 2.12 assurant que les quantités sous les racines carrées sont strictement positives. En pratique, on ne vérifie d'ailleurs pas que la matrice réelle A

est définie positive, mais simplement qu'elle est symétrique, avant de débuter la factorisation. En effet, si la valeur trouvée à la k^{e} étape, $1 \leq k \leq n$, pour la quantité $(b_{kk})^2$ est négative ou nulle, c'est que A n'est pas définie positive. Au contraire, si l'algorithme de factorisation arrive à son terme, cela prouve que A est définie positive, car, pour toute matrice inversible B et tout vecteur $\boldsymbol{v}$ non nul, on a

$$\left(BB^\top \boldsymbol{v}, \boldsymbol{v}\right) = \|B^\top \boldsymbol{v}\|_2^{\,2} > 0.$$

Il est à noter que le calcul du déterminant d'une matrice dont on connaît la factorisation de Cholesky est immédiat, puisque

$$\det(A) = \det(BB^\top) = (\det(B))^2 = \left(\prod_{i=1}^{n} b_{ii}\right)^2.$$

Le nombre d'opérations nécessaires pour effectuer la factorisation de Cholesky d'une matrice réelle A symétrique définie positive d'ordre n par les formules ci-dessus est de $\frac{1}{6}(n^2-1)n$ additions et soustractions, $\frac{1}{6}(n^2-1)n$ multiplications, $\frac{1}{2}n(n-1)$ divisions et n extractions de racines carrées, soit une complexité favorable par rapport à la factorisation LU de la même matrice. Si l'on souhaite résoudre un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ associé, il faut ajouter $n(n-1)$ additions et soustractions, $n(n-1)$ multiplications et $2n$ divisions pour la résolution des systèmes triangulaires $B\boldsymbol{y} = \boldsymbol{b}$ et $B^\top \boldsymbol{x} = \boldsymbol{y}$.

Exemple d'application de la factorisation de Cholesky. Considérons la matrice réelle symétrique définie positive

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 10 \\ 3 & 10 & 26 \end{pmatrix}.$$

En appliquant de l'algorithme de factorisation de Cholesky, on obtient

$$A = BB^\top = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}.$$

2.6.3 Factorisation QR

Le principe de cette méthode n'est plus d'écrire la matrice A comme le produit de deux matrices triangulaires, mais comme le produit d'une matrice *orthogonale* (*unitaire* dans le cas complexe) Q, qu'il est facile d'inverser puisque $Q^{-1} = Q^\top$, et d'une matrice *triangulaire supérieure* R. Pour résoudre le système linéaire (2.1), on effectue donc tout d'abord la factorisation de la matrice A, on procède ensuite au calcul du second membre du système $R\boldsymbol{x} = Q^\top \boldsymbol{b}$, qui est enfin résolu par une méthode de remontée.

Commençons par donner un résultat d'existence et d'unicité de cette factorisation lorsque que la matrice A est carrée et inversible, que l'on va prouver [43] en s'appuyant sur le fameux *procédé (d'orthonormalisation) de Gram–Schmidt* [44] (*Gram–Schmidt (orthogonalization) process* en anglais).

Théorème 2.13 (factorisation QR) *Soit A une matrice réelle d'ordre n inversible. Alors il existe une matrice orthogonale Q et une matrice triangulaire supérieure R, dont les éléments diagonaux sont strictement positifs, telles que*

$$A = QR.$$

Cette factorisation est unique.

43. On pourra noter que l'existence et l'unicité de la factorisation QR d'une matrice inversible découlent de celles de la factorisation de Cholesky. En effet, si A est une matrice inversible d'ordre n, alors la matrice $A^\top A$ est symétrique définie positive et il existe donc une unique matrice B triangulaire inférieure, dont les éléments diagonaux sont strictement positifs, telle que $A^\top A = BB^\top$. En posant alors $Q = \left(A^\top\right)^{-1} B$ et $R = B^\top$, on vérifie que $Q^\top Q = I_n$, que la matrice R est triangulaire supérieure, que ses éléments diagonaux sont strictement positifs et que $A = QR$.

44. Erhard Schmidt (13 janvier 1876 - 6 décembre 1959) était un mathématicien allemand. Il est considéré comme l'un des fondateurs de l'analyse fonctionnelle abstraite moderne.

DÉMONSTRATION. La matrice A étant inversible, ses colonnes, notées $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ forment une base de $M_{n,1}(\mathbb{R})$. On peut alors obtenir une base orthonormée $\{\boldsymbol{q}_j\}_{1\leq j\leq n}$ de $M_{n,1}(\mathbb{R})$ à partir de la famille $\{\boldsymbol{a}_j\}_{1\leq j\leq n}$ en appliquant le procédé de Gram–Schmidt, *i.e.*

$$\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|_2},$$

$$\widetilde{\boldsymbol{q}}_{j+1} = \boldsymbol{a}_{j+1} - \sum_{k=1}^{j} \big(\boldsymbol{q}_k, \boldsymbol{a}_{j+1}\big)\boldsymbol{q}_k,\ \boldsymbol{q}_{j+1} = \frac{\widetilde{\boldsymbol{q}}_{j+1}}{\|\widetilde{\boldsymbol{q}}_{j+1}\|_2},\ j = 1, \ldots, n-1.$$

On en déduit alors que

$$\boldsymbol{a}_j = \sum_{i=1}^{j} r_{ij}\boldsymbol{q}_i,$$

avec $r_{jj} = \|\boldsymbol{a}_j - \sum_{k=1}^{j-1}\big(\boldsymbol{q}_k, \boldsymbol{a}_j\big)\boldsymbol{q}_k\|_2 > 0$, $r_{ij} = \big(\boldsymbol{q}_i, \boldsymbol{a}_j\big)$ pour $1 \leq i \leq j-1$, et $r_{ij} = 0$ pour $j < i \leq n$, $1 \leq j \leq n$. En notant R la matrice triangulaire supérieure (inversible) de coefficients r_{ij}, $1 \leq i, j \leq n$, et Q la matrice orthogonale dont les colonnes sont les vecteurs $\boldsymbol{q}_j$, $1 \leq j \leq n$, on vient d'établir que $A = QR$.
Pour montrer l'unicité de la factorisation, on suppose que

$$A = Q_1R_1 = Q_2R_2,$$

d'où

$$Q_2^\top Q_1 = R_2{R_1}^{-1}.$$

En posant $T = R_2{R_1}^{-1}$, on a $T^\top T = (Q_2^\top Q_1)^\top Q_2^\top Q_1 = I_n$, qui est une factorisation de Cholesky de la matrice identité. Ceci entraîne que $T = I_n$, par unicité de cette dernière factorisation (établie dans le théorème 2.12). □

Le caractère constructif de la démonstration ci-dessus fournit directement une méthode de calcul de la factorisation QR utilisant le procédé de Gram–Schmidt. L'algorithme 13 propose une mise en œuvre de cette méthode pour le calcul de la factorisation QR d'une matrice A d'ordre n inversible. Cette approche nécessite d'effectuer $n^2(n-1)$ additions et soustractions, n^3 multiplications, n^2 divisions et n extractions de racines carrées pour le calcul de la matrice Q, soit de l'ordre de n^3 opérations.

Algorithme 13 : Algorithme de factorisation QR d'une matrice inversible par le procédé d'orthonormalisation de Gram–Schmidt.

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : les tableaux contenant la matrice orthogonale Q et la matrice triangulaire supérieure R.
pour $j = 1$ *à* n **faire**
  $Q(1:n, j) = A(1:n, j)$
  pour $i = 1$ *à* $j-1$ **faire**
    $R(i, j) = Q(1:n, i)^\top * Q(1:n, j)$
    $Q(1:n, j) = Q(1:n, j) - R(i, j) * Q(1:n, i)$
  fin
  $R(j, j) = \|Q(1:n, j)\|_2$
  $Q(1:n, j) = Q(1:n, j)/R(j, j)$
fin

En pratique cependant, et plus particulièrement pour les problèmes de grande taille, la propagation des erreurs d'arrondi entraîne une perte d'orthogonalité entre les vecteurs calculés par le procédé, ce qui fait que la matrice Q obtenue n'est pas « numériquement » orthogonale. Ces instabilités numériques sont dues au fait que la procédure d'orthonormalisation produit des valeurs très petites, ce qui peut entraîner de possibles *annulations* en arithmétique en précision finie (voir la sous-section 1.4.1). Il convient alors de recourir à une version plus stable de l'algorithme, appelée *procédé de Gram–Schmidt modifié* (*modified Gram–Schmidt process* en anglais).

Cette variante consiste en un réordonnancement des calculs de façon à ce que, dès qu'un vecteur de la base orthonormée est obtenu, tous les vecteurs restants à orthonormaliser lui soient rendus orthogonaux (voir l'algorithme 14). Une différence majeure concerne alors le calcul des coefficients r_{ij}, puisque là où la méthode classique fait intervenir une colonne de la matrice à factoriser, sa variante utilise un vecteur déjà partiellement orthogonalisé. Pour cette raison, et malgré l'équivalence mathématique entre les deux versions du procédé, cette seconde version du procédé est préférable à la première si les calculs sont effectués en arithmétique à virgule flottante (voir [Ric66, Bjö67]). Pour la factorisation d'une matrice inversible d'ordre n, celle-ci requiert

Algorithme 14 : Algorithme de factorisation QR d'une matrice inversible par le procédé d'orthonormalisation de Gram–Schmidt modifié.

Entrée(s) : le tableau contenant la matrice A.
Sortie(s) : les tableaux contenant la matrice orthogonale Q et la matrice triangulaire supérieure R.
pour $j = 1$ *à* n **faire**
  $Q(1:n,j) = A(1:n,j)$
fin
pour $i = 1$ *à* n **faire**
  $R(i,i) = \|Q(1:n,i)\|_2$
  $Q(1:n,i) = Q(1:n,i)/R(i,i)$
  pour $j = i+1$ *à* n **faire**
    $R(i,j) = Q(1:n,i)^\top * Q(1:n,j)$
    $Q(1:n,j) = Q(1:n,j) - R(i,j) * Q(1:n,i)$
  fin
fin

$\frac{1}{2}(2n+1)n(n-1)$ additions et soustractions, n^3 multiplications, n^2 divisions et n extractions de racines carrées, soit de l'ordre de n^3 opérations au total.

Indiquons à présent comment réaliser la factorisation QR d'une matrice non inversible ou rectangulaire. Supposons pour commencer que la matrice A est d'ordre n et non inversible. L'ensemble $\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_n\}$ des colonnes de A forment alors une famille liée de vecteurs de $M_{n,1}(\mathbb{R})$ et il existe un entier k, $1 < k \leq n$, tel que la famille $\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_k\}$ est libre et engendre $\boldsymbol{a}_{k+1}$. Le procédé de Gram–Schmidt utilisé pour la factorisation de cette matrice va donc s'arrêter à l'étape $k+1$, puisque l'on aura

$$\|\boldsymbol{a}_{k+1} - \sum_{l=1}^{k} (\boldsymbol{q}_l, \boldsymbol{a}_{k+1})\,\boldsymbol{q}_l\|_2 = 0.$$

On commence donc par échanger les colonnes de A pour amener les colonnes libres aux premières positions. Ceci revient à multiplier A par une matrice de permutation P telle que les $\mathrm{rang}(A)$ premières colonnes de $\tilde{A} = AP$ sont libres, les $n - \mathrm{rang}(A)$ colonnes restantes étant engendrées par les $\mathrm{rang}(A)$ premières (cette permutation peut d'ailleurs se faire au fur et à mesure du procédé d'orthornormalisation, en effectuant une permutation circulaire de la k^{e} à la n^{e} colonne dès que l'on trouve une norme nulle). On applique alors le procédé de Gram–Schmidt jusqu'à l'étape $\mathrm{rang}(A)$ pour construire une famille orthonormée $\{\boldsymbol{q}_1,\dots,\boldsymbol{q}_{\mathrm{rang}(A)}\}$ que l'on complète ensuite par des vecteurs $\boldsymbol{q}_{\mathrm{rang}(A)+1},\dots,\boldsymbol{q}_n$ pour obtenir une base de orthonormée de $M_{n,1}(\mathbb{R})$. On note Q la matrice carrée d'ordre n ayant ces vecteurs pour colonnes. On en déduit qu'il existe des scalaires r_{ij} tels que

$$\tilde{\boldsymbol{a}}_i = \begin{cases} \sum_{j=1}^{i} r_{ij}\,\boldsymbol{q}_j & \text{si } 1 \leq i \leq \mathrm{rang}(A), \\ \sum_{j=1}^{\mathrm{rang}(A)} r_{ij}\,\boldsymbol{q}_j & \text{si } \mathrm{rang}(A)+1 \leq i \leq n, \end{cases}$$

avec $r_{ii} > 0$, $1 \leq i \leq \mathrm{rang}(A)$, et on note R la matrice carrée d'ordre n telle que

$$R = \begin{pmatrix} r_{11} & \dots & \dots & \dots & r_{1n} \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & r_{\mathrm{rang}(A)\,\mathrm{rang}(A)} & \dots & r_{\mathrm{rang}(A)\,n} \\ \vdots & & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}.$$

Considérons ensuite une matrice A rectangulaire de taille $m \times n$ et supposons que $m < n$. Dans ce cas, on a toujours $\ker(A) \neq \{\mathbf{0}\}$ et tout système linéaire associé à A admet une infinité de solutions. On suppose de plus que A est de rang maximal, sinon il faut légèrement modifier l'argumentaire qui suit. Puisque les colonnes de A sont

des vecteurs de $M_{m,1}(\mathbb{R})$ et que $\text{rang}(A) = m$, les m premières colonnes de A sont, à d'éventuelles permutations de colonnes près, libres. On peut donc construire une matrice orthogonale Q d'ordre m à partir de $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\}$ par le procédé de Gram–Schmidt. D'autre part, les colonnes $\boldsymbol{a}_{m+1}, \ldots, \boldsymbol{a}_n$ de A sont engendrées par les colonnes de Q et il existe donc des coefficients r_{ij} tels que

$$\boldsymbol{a}_i = \begin{cases} \sum_{j=1}^{i} r_{ij}\, \boldsymbol{q}_j & \text{si } 1 \leq i \leq m, \\ \sum_{j=1}^{m} r_{ij}\, \boldsymbol{q}_j & \text{si } m+1 \leq i \leq n, \end{cases}$$

avec $r_{ii} > 0$, $1 \leq i \leq m$. On note alors R la matrice de taille $m \times n$ définie par

$$R = \begin{pmatrix} r_{11} & \cdots & \cdots & \cdots & \cdots & r_{1n} \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \ldots & 0 & r_{mm} & \ldots & r_{mn} \end{pmatrix}.$$

Faisons maintenant l'hypothèse que $m > n$, qui est le cas le plus répandu en pratique. Pour simplifier, on va supposer que $\ker(A) = \{\mathbf{0}\}$, c'est-à-dire que $\text{rang}(A) = n$ (si ce n'est pas le cas, il faut procéder comme dans le cas d'une matrice carrée non inversible). On commence par appliquer le procédé de Gram–Schmidt aux colonnes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ de la matrice A pour obtenir la famille de vecteurs $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$, que l'on complète par des vecteurs $\boldsymbol{q}_{n+1}, \ldots, \boldsymbol{q}_m$ pour arriver à une base orthonormée de $M_{m,1}(\mathbb{R})$. On note alors Q la matrice carrée d'ordre m ayant pour colonnes les vecteurs $\boldsymbol{q}_j$, $j = 1, \ldots, m$. On a par ailleurs

$$\boldsymbol{a}_j = \sum_{i=1}^{j} r_{ij}\, \boldsymbol{q}_j,\ 1 \leq j \leq n,$$

avec $r_{ii} > 0$, $1 \leq i \leq n$. On pose alors

$$R = \begin{pmatrix} r_{11} & \ldots & r_{1n} \\ 0 & \ddots & \vdots \\ \vdots & \ddots & r_{nn} \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix},$$

qui est une matrice de taille $m \times n$.

Malgré l'amélioration apportée par le procédé de Gram–Schmidt modifié, cette méthode reste relativement peu utilisée en pratique pour le calcul d'une factorisation QR, car on lui préfère la *méthode de Householder* [45] [Hou58], dont le principe est de multiplier la matrice A par une suite de matrices de transformation très simples, dites *de Householder*, pour l'amener progressivement sous forme triangulaire supérieure.

Définition 2.14 (matrice de Householder) *Soit $\boldsymbol{v}$ une vecteur non nul de $M_{n,1}(\mathbb{R})$. On appelle* ***matrice de Householder*** *(**Householder matrix** en anglais) associée au vecteur $\boldsymbol{v}$, et on note $H(\boldsymbol{v})$, la matrice définie par*

$$H(\boldsymbol{v}) = I_n - 2\,\frac{\boldsymbol{v}\boldsymbol{v}^\top}{\boldsymbol{v}^\top \boldsymbol{v}}. \tag{2.19}$$

On pose de plus $H(\mathbf{0}) = I_n$, ce qui permet de considérer la matrice identité comme une matrice de Householder.

Les matrices de Householder possèdent des propriétés intéressantes, que l'on résume dans le résultat suivant.

45. Alston Scott Householder (5 mai 1904 - 4 juillet 1993) était un mathématicien américain. Il s'intéressa aux applications des mathématiques, notamment en biomathématiques et en analyse numérique.

Lemme 2.15 *Soit $\boldsymbol{v}$ un vecteur non nul de $M_{n,1}(\mathbb{R})$ et $H(\boldsymbol{v})$ la matrice de Householder qui lui est associée. Alors, $H(\boldsymbol{v})$ est symétrique et orthogonale. De plus, si $\boldsymbol{x}$ est un vecteur de $M_{n,1}(\mathbb{R})$ et $\boldsymbol{e}$ est un vecteur unitaire tels que $\boldsymbol{x} \neq \pm\|\boldsymbol{x}\|_2 \boldsymbol{e}$, on a*

$$H(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e})\boldsymbol{x} = \mp\|\boldsymbol{x}\|_2 \boldsymbol{e}.$$

DÉMONSTRATION. Il est facile de voir que $H(\boldsymbol{v}) = H(\boldsymbol{v})^\top$. Par ailleurs, on vérifie que

$$H(\boldsymbol{v})^2 = I_n - 4\frac{\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|_2^{\,2}} + 4\frac{\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|_2^{\,4}} = I_n - 4\frac{\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|_2^{\,2}} + 4\frac{\|\boldsymbol{v}\|_2^{\,2}\boldsymbol{v}\boldsymbol{v}^\top}{\|\boldsymbol{v}\|_2^{\,4}} = I_n.$$

Sans perte de généralité, on peut ensuite supposer que $\boldsymbol{e}$ est le premier vecteur de la base canonique $\{\boldsymbol{e}_i\}_{1\leq i\leq n}$ de $M_{n,1}(\mathbb{R})$ et l'on a

$$\begin{aligned} H(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)\boldsymbol{x} &= \boldsymbol{x} - 2\frac{(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)^\top}{(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)^\top(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)}\boldsymbol{x} \\ &= \boldsymbol{x} - 2\frac{(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2 \boldsymbol{e}_1)\left(\|\boldsymbol{x}\|_2^{\,2} \pm \|\boldsymbol{x}\|_2 x_1\right)}{2\|\boldsymbol{x}\|_2^{\,2} \pm 2\|\boldsymbol{x}\|_2 x_1} \\ &= \mp\|\boldsymbol{x}\|_2 \boldsymbol{e}_1. \end{aligned}$$

□

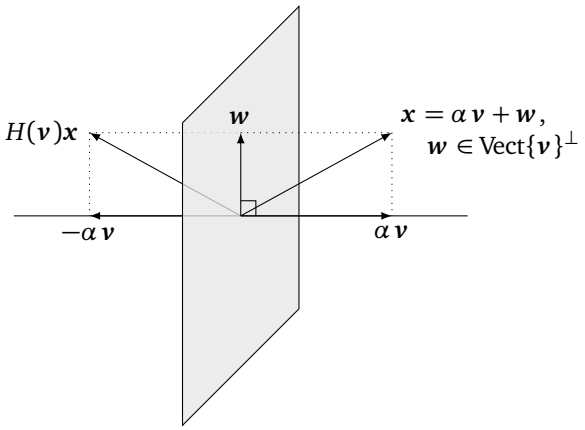


FIGURE 2.2 – Transformation d'un vecteur $\boldsymbol{x}$ de l'espace par la matrice de Householder $H(\boldsymbol{v})$.

En considérant les matrices de $M_{n,1}(\mathbb{R})$ comme des représentations matricielles de vecteurs de $\mathbb{R}^n$, la matrice de Householder $H(\boldsymbol{v})$ s'interprète comme la représentation de la *symétrie orthogonale (ou réflexion) par rapport à l'hyperplan orthogonal au vecteur représenté par $\boldsymbol{v}$* (voir la figure 2.2 pour une illustration en dimention trois). Les matrices de Householder peuvent par conséquent être utilisées pour annuler certaines composantes d'un vecteur de $M_{n,1}(\mathbb{R})$ donné, comme le montre l'exemple suivant.

Exemple de transformation d'un vecteur par une matrice de Householder. Considérons les choix de vecteurs $\boldsymbol{x} = \begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix}^\top$ et $\boldsymbol{e} = \boldsymbol{e}_3$. On a $\|\boldsymbol{x}\|_2 = 2$, d'où

$$\boldsymbol{v} = \boldsymbol{x} + \|\boldsymbol{x}\|_2 \boldsymbol{e}_3 = \begin{pmatrix}1\\1\\3\\1\end{pmatrix},\ H(\boldsymbol{v}) = \frac{1}{6}\begin{pmatrix}5 & -1 & -3 & -1\\ -1 & 5 & -3 & -1\\ -3 & -3 & -3 & -3\\ -1 & -1 & -3 & 5\end{pmatrix} \text{ et } H(\boldsymbol{v})\boldsymbol{x} = \begin{pmatrix}0\\0\\-2\\0\end{pmatrix}.$$

Décrivons à présent la méthode de Householder pour la factorisation d'une matrice réelle A d'ordre n (on peut l'étendre sans grande difficulté aux matrices complexes). Dans ce cas, celle-ci revient à trouver $n-1$ matrices $H^{(k)}$, $1 \leq k \leq n-1$, d'ordre n telles que $H^{(n-1)}\ldots H^{(2)}H^{(1)}A$ soit triangulaire supérieure.

On procède pour cela de la manière suivante. On commence par poser $A^{(1)} = A$. À la k^{e} étape, $1 \leq k \leq n-2$, de la méthode, la répartition des zéros dans la matrice $A^{(k)}$ est identique à celle obtenue au même stade de

l'élimination de Gauss avec échange. On doit donc mettre à zéro des coefficients sous-diagonaux de la k^{e} colonne de $A^{(k)}$.

Soit $\tilde{\boldsymbol{a}}^{(k)}$ la matrice de $M_{n-k+1,1}(\mathbb{R})$ contenant les éléments $a_{ik}^{(k)}$, $k \leq i \leq n$, de $A^{(k)}$. Si $\sum_{i=k+1}^{n} \left|a_{ik}^{(k)}\right| = 0$, alors $A^{(k)}$ est déjà de la « forme » de $A^{(k+1)}$ et on pose $H^{(k)} = I_n$. Si $\sum_{i=k+1}^{n} \left|a_{ik}^{(k)}\right| > 0$, alors il existe, en vertu du lemme 2.15, une matrice $\tilde{\boldsymbol{v}}^{(k)}$ de $M_{n-k+1,1}(\mathbb{R})$, donné par

$$\tilde{\boldsymbol{v}}^{(k)} = \tilde{\boldsymbol{a}}^{(k)} \pm \|\tilde{\boldsymbol{a}}^{(k)}\|_2 \, \tilde{\boldsymbol{e}}_1^{(n-k+1)}, \tag{2.20}$$

où $\tilde{\boldsymbol{e}}_1^{(n-k+1)}$ désigne le premier vecteur de la base canonique de $M_{n-k+1,1}(\mathbb{R})$, tel que le vecteur $H(\tilde{\boldsymbol{v}}^{(k)})\tilde{\boldsymbol{a}}^{(k)}$ ait toutes ses composantes nulles à l'exception de la première. On pose alors

$$H^{(k)} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H(\tilde{\boldsymbol{v}}^{(k)}) \end{pmatrix} = H(\boldsymbol{v}^{(k)}), \text{ avec } \boldsymbol{v}^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\boldsymbol{v}}^{(k)} \end{pmatrix} \in M_{n,1}(\mathbb{R}). \tag{2.21}$$

On réitère ces opérations jusqu'à obtenir la matrice triangulaire supérieure

$$A^{(n)} = H^{(n-1)} \ldots H^{(1)} A^{(1)},$$

et alors $A^{(n)} = R$ et $Q = \left(H^{(n-1)} \ldots H^{(1)}\right)^{\top} = H^{(1)} \ldots H^{(n-1)}$. Notons au passage que nous n'avons supposé A inversible et qu'aucun n'échange de colonne n'a été nécessaire comme avec le procédé d'orthonormalisation de Gram–Schmidt.

Revenons sur le choix du signe dans (2.20) lors de la construction du vecteur de Householder à la k^{e} étape. Dans le cas réel, il est commode de choisir le vecteur de telle manière à ce que le coefficient $a_{kk}^{(k+1)}$ soit positif. Ceci peut néanmoins conduire à d'importantes erreurs d'annulation si le vecteur $\tilde{\boldsymbol{a}}^{(k)}$ est « proche » d'un multiple positif de $\tilde{\boldsymbol{e}}_1^{(n-k+1)}$, mais ceci peut s'éviter en ayant recours à la formule suivante dans le calcul de $\tilde{\boldsymbol{v}}^{(k)}$

$$\tilde{v}_1^{(k)} = \frac{\left(\tilde{a}_1^{(k)}\right)^2 - \|\tilde{\boldsymbol{a}}^{(k)}\|_2^2}{\tilde{a}_1^{(k)} + \|\tilde{\boldsymbol{a}}^{(k)}\|_2} = \frac{-\sum_{i=k+1}^{n} \left(a_{ik}^{(k)}\right)^2}{\tilde{a}_1^{(k)} + \|\tilde{\boldsymbol{a}}^{(k)}\|_2}.$$

Cette méthode s'applique de la même manière aux matrices rectangulaires, à quelques modifications évidentes près. Par exemple, dans le cas d'une matrice de taille $m \times n$ avec $m > n$, la méthode construit n matrices $H^{(k)}$, $1 \leq k \leq n$, d'ordre m telles que la matrice $A^{(n+1)}$ est de la forme

$$A^{(n+1)} = \begin{pmatrix} a_{11}^{(n+1)} & \ldots & a_{1n}^{(n+1)} \\ 0 & \ddots & \vdots \\ \vdots & \ddots & a_{nn}^{(n+1)} \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix}.$$

Une des raisons du succès de la méthode de Householder est sa grande stabilité numérique. Elle ne modifie en effet pas le conditionnement du problème, puisque

$$\text{cond}_2(A^{(n)}) = \text{cond}_2(A), \; A \in M_n(\mathbb{R}),$$

en vertu de la proposition A.156. De plus, la base contenue dans la matrice Q est *numériquement* orthonormale et ne dépend pas du degré d'indépendance des colonnes de la matrice A, comme c'est le cas pour le procédé de Gram–Schmidt. Ces avantages sont cependant tempérés par un coût sensiblement supérieur.

Abordons pour finir quelques aspects de la mise en œuvre de la méthode de Householder. Dans cette dernière, il faut absolument tenir compte de la structure particulière des matrices $H^{(k)}$, $1 \leq k \leq n-1$ intervenant dans la

factorisation. En particulier, il s'avère qu'il n'est pas nécessaire d'assembler une matrice de Householder pour en effectuer le produit avec une autre matrice. Prenons en effet l'exemple d'une matrice M d'ordre m quelconque que l'on veut multiplier par la matrice $H(\boldsymbol{v})$ avec $\boldsymbol{v}$ un vecteur de $M_{m,1}(\mathbb{R})$. En utilisant (2.19), on obtient que [46]

$$H(v)M = M - \frac{2}{\|\boldsymbol{v}\|_2^2}\,\boldsymbol{v}(M^\top \boldsymbol{v})^\top.$$

Ainsi, calculer le produit $H(v)M$ se ramène *grosso modo* à effectuer un produit scalaire (le coefficient $\beta = \frac{2}{\|\boldsymbol{v}\|_2^2}$), un produit matrice-vecteur (le produit $\boldsymbol{w} = M^\top \boldsymbol{v}$), un produit tensoriel de deux vecteurs (la matrice $\boldsymbol{v}(\beta\,\boldsymbol{w})^\top$), suivis de la différence de deux matrices, ce qui nécessite au total $2m^2-1$ additions et soustractions, $2(m+1)m$ multiplications et une division. Ce résultat est à comparer aux $2m-1$ additions et soustractions, $m(m+2)$ multiplications et une division requises pour la construction de $H(\boldsymbol{v})$, ajoutées aux $m^2(m-1)$ additions et soustractions et m^3 multiplications nécessaires au produit de deux matrices quelconques.

Une conséquence de cette remarque est que l'on n'a, *a priori*, pas à stocker, ni même à calculer, la matrice Q lors de la résolution d'un système linéaire via une factorisation QR, puisque l'on a seulement besoin à chaque étape k, $k = 1,\ldots,n$ dans le cas d'une matrice A d'ordre n, d'effectuer le produit de la matrice $H^{(k)}$ avec $A^{(k)}$ et de mettre à jour le second membre du système considéré. Le coût total de ces opérations est de $\frac{1}{3}(2n^2+7n+3)(n+1)$ additions et soustractions, $\frac{2}{3}(n+3)(n+2)(n+1)$ multiplications et $n+1$ divisions, soit environ le double de celui de l'élimination de Gauss.

Si l'on a besoin de connaître explicitement la matrice Q, il est possible de l'obtenir par un procédé consistant, à partir de la matrice $Q^{(1)} = I_n$, à utiliser soit la formule de récurrence

$$Q^{(k+1)} = Q^{(k)}H^{(k)},\ k = 1,\ldots,n-1,$$

et l'on parle alors d'*accumulation directe*, soit la formule

$$Q^{(k+1)} = H^{(n-k)}Q^{(k)},\ k = 1,\ldots,n-1,$$

correspondant à une *accumulation rétrograde*. En se rappelant qu'une sous-matrice principale d'ordre $k-1$ correspond à l'identité dans chaque matrice $H^{(k)}$ (voir (2.21)), $1 \le k \le n-1$, on constate que les matrices $Q^{(k)}$ se « remplissent » graduellement au cours des itérations de l'accumulation rétrograde, ce qui peut être exploité pour diminuer le nombre d'opérations requises pour effectuer le calcul, alors que la matrice $Q^{(2)}$ est, au contraire, pleine à l'issue de la première étape de l'accumulation directe. Pour cette raison, la version rétrograde du procédé d'accumulation est la solution la moins onéreuse et donc celle à privilégier pour le calcul effectif de Q.

Notons que l'on peut encore parvenir à la factorisation QR d'une matrice par la *méthode de Givens* [47] [Giv58], qui consiste à utiliser des *matrices de rotation de Givens* (voir la sous-section 4.5.1) pour annuler les coefficients sous-diagonaux de la matrice à factoriser, en la parcourant ligne par ligne ou colonne par colonne et avec un coût voisin du double de celui de la méthode de Householder. Cette approche peut néanmoins s'avérer avantageuse dans le cas d'une matrice creuse.

2.7 Stabilité numérique des méthodes directes *

Pour que les méthodes de factorisation présentées puissent servir à la résolution numérique de systèmes linéaires, il faut bien sûr qu'elles soient stables lorsque les calculs sont effectués par un calculateur fonctionnant en arithmétique en précision finie. Il en va de même pour les méthodes de descente et de remontée utilisées pour la résolution des systèmes triangulaires obtenus.

explications sur les notations (γ_n) et le sens des inégalités entre matrices

46. Par des considérations analogues, on trouve que

$$MH(v) = M - \frac{2}{\|\boldsymbol{v}\|_2^2}\,(M\boldsymbol{v})\boldsymbol{v}^\top.$$

47. James Wallace Givens, Jr. (14 décembre 1910 - 5 mars 1993) était un mathématicien américain et l'un des pionnier de l'informatique et du calcul scientifiques. Il reste connu pour les matrices de rotation portant son nom.

2.7.1 Résolution des systèmes triangulaires *

l'analyse d'erreur est très simple et l'on conclut que les algorithmes de substitution sont extrèmement stables

en pratique, l'erreur directe est souvent bien plus petite que ne le laissent espérer les majorations faisant intervenir le conditionnement

sources d'explications :

- la précision du résultat dépend fortement du second membre du système à résoudre
- une matrice triangulaire peut-être beaucoup mieux ou beaucoup moins bien conditionnée que sa transposée
- l'utilisation d'un choix de pivot pour lors de la factorisation LU, de Cholesky ou QR d'une matrice peut fortement améliorer le conditionnement des systèmes triangulaires résultants

On donne un résultat pour la méthode de remontée seulement, analogue pour la méthode de descente.

Théorème 2.16 *La solution $\hat{\boldsymbol{x}}$, calculée par l'algorithme implémentant les formules 2.8 exécutée en arithmétique en précision finie, d'un système linéaire triangulaire supérieur $U\boldsymbol{x} = \boldsymbol{b}$ satisfait*

$$(U + \delta U)\hat{\boldsymbol{x}} = \boldsymbol{b},\ avec\ \left|\delta u_{ij}\right| \leq \begin{cases} \gamma_{n-i+1}\left|u_{ij}\right| & si\ i = j \\ \gamma_{|i-j|}\left|u_{ij}\right| & si\ i \neq j \end{cases},\ 1 \leq i \leq j \leq n.$$

Théorème 2.17 *Soit le système linéaire $T\boldsymbol{x} = \boldsymbol{b}$, avec T une matrice triangulaire d'ordre n inversible, que l'on résout par une méthode de substitution. La solution calculée $\hat{\boldsymbol{x}}$ vérifie*

$$(T + \delta T)\hat{\boldsymbol{x}} = \boldsymbol{b},\ |\delta T| \leq \gamma_n |T|.$$

2.7.2 Élimination de Gauss et factorisation LU *

REPRENDRE l'analyse d'erreur de l'élimination de Gauss combine celles des calculs de produits scalaires et des méthodes de substitution pour la résolution des systèmes triangulaires.

Toutes les variantes de la méthode conduisent aux mêmes bornes d'erreur, on réalise celle de la méthode de Doolittle sans choix de pivot (car ceci revient à appliquer l'élimination sur une matrice dans laquelle des lignes et/ou des colonnes ont été permutées)

Les matrices $\widehat{L}$ et $\widehat{U}$ calculées satisfont ($\hat{l}_{ii} = 1$)

$$\left| a_{kj} - \sum_{r=1}^{k-1} \hat{l}_{kr}\hat{u}_{rj} - \hat{u}_{kj} \right| \leq \gamma_k \sum_{r=1}^{k} \left|\hat{l}_{kr}\right| \left|\hat{u}_{rj}\right|,\ j = k, \ldots, n,$$

$$\left| a_{ik} - \sum_{r=1}^{k-1} \hat{l}_{ir}\hat{u}_{rj} \right| \leq \gamma_k \sum_{r=1}^{k} \left|\hat{l}_{ir}\right| \left|\hat{u}_{rj}\right|,\ i = k+1, \ldots, n.$$

On déduit des ces inégalités le résultat d'analyse inverse suivant pour la factorisation LU.

Théorème 2.18 *Soit A une matrice d'ordre n pour laquelle le procédé d'élimination de Gauss peut être mené à son terme. Alors les matrices $\widehat{L}$ et $\widehat{U}$ de la factorisation obtenues vérifient*

$$\widehat{L}\widehat{U} = A + \delta A,\ |\delta A| \leq \gamma_n \left|\widehat{L}\right|\left|\widehat{U}\right|.$$

Avec quelques efforts supplémentaires, on arrive au résultat suivant pour l'erreur inverse sur la solution de $Ax = b$.

Théorème 2.19 *Soit A une matrice d'ordre n pour laquelle le procédé d'élimination de Gauss conduit à une factorisation LU avec les matrices $\widehat{L}$ et $\widehat{U}$. La solution calculée $\widehat{\boldsymbol{x}}$ est telle que*

$$(A + \delta A)\widehat{\boldsymbol{x}} = \boldsymbol{b},\ |\delta A| \leq 2\gamma_n \left|\widehat{L}\right|\left|\widehat{U}\right|$$

A VOIR : dans [Hig02], th. 9.3 p 193, on a γ_{3n} en place de $2\gamma_n$.

DÉMONSTRATION. A ECRIRE □

INTERPRETATION DU RESULTAT : on voudrait $|\delta A| \leq u\,|A|$, mais comme les éléments de la matrice subissent jusqu'à n opérations arithmétiques, on ne peut espérer mieux qu'une estimation de la forme $|\delta A| \leq c_n u\,|A|$, avec c_n une constante dépendant de n. Une telle borne est obtenue si $\widehat{L}$ et $\widehat{U}$ sont telles que $\left|\widehat{L}\right|\left|\widehat{U}\right| = \left|\widehat{L}\widehat{U}\right|$ car alors

$$\left|\widehat{L}\right|\left|\widehat{U}\right| = \left|\widehat{L}\widehat{U}\right| = |A+\delta A| \leq |A| + \gamma_n \left|\widehat{L}\right|\left|\widehat{U}\right|,$$

d'où $\left|\widehat{L}\right|\left|\widehat{U}\right| \leq \frac{1}{1-\gamma_n}\,|A|$ et donc

$$|\delta A| \leq \frac{2\gamma_n}{1-\gamma_n}\,|A|\,.$$

Une classe de matrices pour lesquelles ceci est vérifié est celle des *matrices totalement positives* (matrices dont tous les mineurs sont (strictement ?) positifs [48]). Si A est totalement positive, on peut montrer [dBP76] que les coefficients de $\widehat{L}$ et de $\widehat{U}$ sont positifs si u est suffisamment petit.

La stabilité de l'élimination est donc déterminée par la taille des éléments de la matrice $\left|\widehat{L}\right|\left|\widehat{U}\right|$. Sans choix de pivot durant l'élimination, le rapport $\|\left|\widehat{L}\right|\left|\widehat{U}\right|\|/\|A\|$ peut devenir arbitrairement grand.

Avec une stratégie de pivot partiel, on peut montrer que L est petite et U est bornée relativement à A.

On souhaiterait par conséquent obtenir un contrôle de la quantité $|\delta A|$ en fonction de $|A|$. Ce type d'analyse d'erreur fait traditionnellement intervenir le *facteur de croissance* de la méthode, qui est la quantité

$$\rho_n = \frac{\max\limits_{1\leq i,j,k\leq n}\left|a_{ij}^{(k)}\right|}{\max\limits_{1\leq i,j\leq n}\left|a_{ij}\right|}. \tag{2.22}$$

En utilisant la majoration

$$\left|u_{ij}\right| = \left|a_{ij}^{(i)}\right| \leq \rho_n \max_{i,j}\left|a_{ij}\right|,$$

on obtient le résultat classique, dü à Wilkinson [Wil61] (A VOIR car pas trouvé dans l'article), suivant.

Théorème 2.20 *Soit A une matrice d'ordre n pour laquelle on suppose que l'élimination de Gauss avec une stratégie de pivot partiel calcule une approximation $\hat{\boldsymbol{x}}$ de la solution du problème $A\boldsymbol{x} = \boldsymbol{b}$. Alors on a*

$$(A+\delta A)\hat{\boldsymbol{x}} = \boldsymbol{b},\ \|\delta A\|_\infty \leq 2n^2\gamma_n\rho_n\,\|A\|_\infty.$$

A VOIR : dans [Hig02], th. 9.5 p 194, on a γ_{3n} en place de $2\gamma_n$.

Ajouter des remarques/résultats sur les classes de matrices pour lesquelles il n'est pas utile de chercher un pivot (le facteur de croissance de la méthode d'élimination sans échange est un $O(1)$) : symétriques définies positives, à diagonale dominante (par lignes ou colonnes), totalement positives [dBP76]

On peut par ailleurs montrer que l'on a $\rho_n \leq 2^{n-1}$ pour la stratégie de pivot partiel [49] (malgré ce résultat, on constate en pratique que le facteur de croissance est généralement petit), $\rho_n \leq n^{\frac{1}{2}}(2\,3^{\frac{1}{2}}\ldots n^{\frac{1}{n-1}} \sim c\,n^{\frac{1}{4}\log n+\frac{1}{2}}$ pour celle de pivot total (majorant pas atteignable, voir Wilkinson), $\rho_n \leq \frac{1}{2}\,n^{\frac{3}{4}\log n}$ pour le rook pivoting [Fos97].

2.7.3 Factorisation de Cholesky *

REPRENDRE L'analyse d'erreur pour cette factorisation est similaire à celle pour la factorisation LU. En utilisant le lemme 8.4 dans [Hig02], on obtient

$$\left|a_{ij} - \sum_{k=1}^{i}\hat{b}_{ik}\hat{b}_{jk}\right| \leq \gamma_i \sum_{k=1}^{i}\left|\hat{b}_{ik}\right|\left|\hat{b}_{jk}\right|,$$

48. VERIFIER la ref : Si A est une matrice totalement positive dont les mineurs principaux sont tous strictement positifs, alors on montre facilement (voir [Cry73]) qu'elle admet une factorisation LU avec $L \geq 0$ et $U \geq 0$.

49. Cette borne est atteinte pour les matrices de la forme $\begin{pmatrix} 1 & 0 & \ldots & 0 & 1 \\ -1 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ -1 & \ldots & \ldots & -1 & 1 \end{pmatrix}$.

et, par une modification,

$$\left| a_{ii} - \sum_{k=1}^{i} \hat{b}_{ik}^2 \right| \leq \gamma_{i+1} \sum_{k=1}^{i} \hat{b}_{ik}^2.$$

On a alors les résultats de stabilité inverse suivants, analogues aux théorèmes ref et ref :
- la matrice $\hat{B}$ effectivement calculée est telle que $\hat{B}^T\hat{B} = A + \delta A$, avec $|\delta A| \leq \gamma_{n+1} \left|\widehat{B}\right| \left|\widehat{B}^T\right|$
- la solution $\widehat{\boldsymbol{x}}$ calculée est telle que $(A+\delta A)\widehat{\boldsymbol{x}} = \boldsymbol{b}$, avec $|\delta A| \leq \gamma_{3n+1} \left|\widehat{B}\right| \left|\widehat{B}^T\right|$

Ces résultats impliquent la parfaite stabilité inverse en norme de la factorisation de Cholesky. On a en effet (A VOIR)

$$\| |B| \left|B^T\right| \|_2 = \| |B| \|_2^{\,2} \leq n \, \|B\|_2^{\,2} = n \, \|A\|_2,$$

dont l'analogue pour $\widehat{B}$ est

$$\| \left|\widehat{B}\right| \left|\widehat{B}^T\right| \|_2 \leq \frac{n}{1 - n\gamma_{n+1}} \, \|A\|_2,$$

et donc

$$\|\delta A\|_2 \leq 8n(n+1)u \, \|A\|_2 \text{ ou bien } 4n(3n+1)u \, \|A\|_2$$

(la première en supposant que $2n(n+1)u \leq 1-(n+1)u$, voir Wilkinson 1968)

Note : la dernière matrice δA n'est généralement pas symétrique, car les matrices de l'analyse d'erreur inverse pour la résolution des deux systèmes triangulaires ne sont pas les transposées l'une de l'autre.

2.7.4 Factorisation QR **

stabilité inverse avec une meilleure erreur inverse que celle de la factorisation LU (REF ?)

Remarque sur la stabilité du procédé d'orthogonalisation de Gram–Schmidt : analyse des différentes versions du procédé dans [Bjö94]

2.8 Notes sur le chapitre

Si la méthode d'élimination est attribuée à Gauss qui l'utilisa en 1809 dans son ouvrage *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* pour la résolution de problèmes aux moindres carrés, celle-ci apparaît déjà dans le huitième chapitre d'un livre anonyme chinois de mathématiques intitulé *« Les neuf chapitres sur l'art mathématique »* (九章算術 en sinogrammes traditionnels, *Jiǔzhāng Suànshù* en pinyin) et compilé entre le deuxième et le premier siècle avant J.-C., avec un exemple de résolution d'un système de cinq équations à cinq inconnues. On la retrouve plus tard dans des notes rédigées par Newton [50] vers 1670 et publiées en 1707 sous le titre *Arithmetica universalis*. Le long article [Grc11a] relate en détail l'histoire de ce procédé. Enfin, on doit à Turing son interprétation comme procédé de factorisation de matrice [Tur48]. Sur ce point, on pourra plus particulièrement consulter la référence [Dop13].

Cette méthode fut par ailleurs le premier algorithme sujet à l'analyse de sa sensibilité aux erreurs d'arrondi. Après une tentative initiale concluant de manière pessimiste à l'instabilité du procédé [Hot43], plusieurs articles [vNG47, FHW48, Tur48], aujourd'hui considérés comme les actes de naissance de l'analyse numérique « moderne » (voir [Grc11b]), apportèrent des preuves théoriques et empiriques de son bon comportement dans la plupart des cas. L'analyse d'erreur inverse de la méthode telle qu'on la connaît aujourd'hui est l'œuvre de Wilkinson [Wil61].

Enfin, si des stratégies de choix de pivot étaient déjà couramment utilisées dans les années quarante (voir par exemple [vNG47]), les dénominations de « pivot total » et « pivot partiel » semblent dues à Wilkinson [Wil61].

Appliquée à une matrice symétrique définie positive, la méthode de Bareiss est un avatar d'un algorithme développé par Schur [51] [Sch17], motivé par des travaux de Toeplitz et Carathéodory [52] sur le *problème des moments*

50. Sir Isaac Newton (4 janvier 1643 - 31 mars 1727) était un philosophe, mathématicien, physicien et astronome anglais. Figure emblématique des sciences, il est surtout reconnu pour sa théorie de la gravitation universelle et l'invention du calcul infinitésimal.

51. Issai Schur (Исáй Шур en russe, 10 janvier 1875 – 10 janvier 1941) était un mathématicien russe qui travailla surtout en Allemagne. Il s'intéressa à la combinatoire et à la représentation des groupes et a donné son nom à différents concepts et résultats mathématiques.

52. Constantin Carathéodory (Κωνσταντίνος Καραθεοδωρή en grec, 13 septembre 1873 - 2 février 1950) était un mathématicien grec. Il apporta des contributions significatives à la théorie des fonctions d'une variable réelle, au calcul des variations et à la théorie de la mesure. Il fit aussi œuvre de pionnier en développant la formulation axiomatique de la thermodynamique selon une approche purement géométrique.

trigonométriques, pour vérifier qu'une fonction définie par une série entière est analytique et bornée sur le disque unité (voir [KS95] pour beaucoup plus de détails).

Il existe principalement deux approches pour obtenir une factorisation symétrique numériquement stable d'une matrice symétrique quelconque. L'une, due à Parlett et Reid [PR70], conduit à une décomposition de la forme

$$PAP^\top = LTL^\top,$$

où L est une matrice triangulaire inférieure, dont les éléments vérifient $l_{ii} = 1$ et $\left|l_{ij}\right| \leq 1$, $j < i$, et T est une matrice tridiagonale. Une version améliorée de l'algorithme de factorisation initialement proposé, introduite par Aasen [Aas71], requiert environ $\frac{n^3}{6}$ additions et multiplications pour y parvenir. Une fois la matrice symétrique A factorisée, la solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ est obtenue en résolvant successivement

$$L\boldsymbol{z} = P\boldsymbol{b},\ T\boldsymbol{y} = \boldsymbol{z},\ L^\top \boldsymbol{w} = \boldsymbol{y},\ \boldsymbol{x} = P\boldsymbol{w},$$

pour un coût d'environ $2n^2$ opérations [53]. Une autre manière de procéder consiste en la construction, par une stratégie de pivot *diagonal* total [BP71], d'une matrice de permutation P telle que

$$PAP^\top = LDL^\top,$$

où D est une matrice diagonale par blocs [54]. L'analyse de cette dernière méthode de factorisation (voir [Bun71]) montre que sa stabilité est très satisfaisante, mais qu'elle nécessite d'effectuer jusqu'à $\frac{n^3}{6}$ comparaisons pour le choix des pivots, ce qui la rend coûteuse. Se basant sur un principe analogue à la stratégie de pivot partiel pour la factorisation LU, Bunch et Kaufman [BK77] inventèrent un algorithme permettant d'arriver à une telle factorisation de manière stable en réalisant seulement de l'ordre de n^2 comparaisons.

Schmidt introduisit en 1907, dans un article sur les équations intégrales [Sch07], le procédé aujourd'hui dit de Gram–Schmidt pour construire, de manière théorique, une famille orthonormée de fonctions à partir d'une famille libre infinie. Faisant explicitement référence à des travaux de Gram [Gra83] sur le développement en série de fonctions par des méthodes de moindres carrés, il lui en attribua l'idée [55]. Il apparaît cependant que cette méthode est bien plus ancienne. Laplace [56] se servit en effet dès 1812 d'une version orientée ligne du procédé de Gram–Schmidt modifié, sans faire de lien avec l'orthogonalisation, pour calculer les masses des planètes Jupiter et Saturne à partir d'un système d'équations normales issu de données fournies par Bouvard [57] et estimer l'écart type de la distribution de l'erreur sur la solution en supposant que le bruit sur les observations astronomiques suit une loi normale (voir *« Sur l'application du calcul des probabilités à la philosophie naturelle »* dans [Lap20], premier supplément). On le retrouve par la suite dans un compte-rendu de Bienaymé [58] [Bie53], dans lequel une technique proposée par Cauchy [59] pour interpoler des séries convergentes [Cau37] est interprétée comme une méthode d'élimination de Gauss pour la résolution de systèmes linéaires de la forme $Z^\top A\boldsymbol{x} = Z^\top \boldsymbol{b}$, avec A et Z des matrices de $M_{m,n}(\mathbb{R})$, $\boldsymbol{x}$ et $\boldsymbol{b}$ des vecteurs de $M_{n,1}(\mathbb{R})$ et $M_{m,1}(\mathbb{R})$ respectivement, et ensuite améliorée au moyen d'un ajustement par la méthode des moindres carrés.

L'application de la factorisation QR par la méthode de Householder pour la résolution de problèmes aux moindres carrés est détaillée dans l'article [Gol65] et porte pour cette raison parfois le nom de *méthode de Golub* [60].

53. Rappelons qu'un système linéaire dont la matrice est tridiagonale se résout de manière très efficace (en $O(n)$ opérations) par l'algorithme de Thomas présenté dans la sous-section 2.5.4.

54. Dans ce cas, la résolution du système linéaire $D\boldsymbol{y} = \boldsymbol{z}$ se ramène à celle d'un ensemble de systèmes linéaires carrés d'ordre 1 ou 2.

55. Dans l'article en question, c'est le procédé de Gram–Schmidt modifié qu'utilise Gram.

56. Pierre-Simon Laplace (23 mars 1749 - 5 mars 1827) était un mathématicien, astronome et physicien français. Son œuvre la plus importante concerne le calcul des probabilités et la mécanique céleste.

57. Alexis Bouvard (27 juin 1767 - 7 juin 1843) était un astronome français. Parmi ses travaux les plus significatifs figurent la découverte de huit comètes et la compilation de tables astronomiques pour les planètes Jupiter, Saturne et Uranus.

58. Irénée-Jules Bienaymé (28 août 1796 - 19 octobre 1878) était un mathématicien français. Il généralisa la méthode des moindres carrés introduite par Laplace et contribua à la théorie des probabilités, au développement des statistiques ainsi qu'à leur application à la finance, à la démographie et aux sciences sociales.

59. Augustin-Louis Cauchy (21 août 1789 – 23 mai 1857) était un mathématicien français. Très prolifique, ses recherches couvrent l'ensemble des domaines mathématiques de son époque. On lui doit notamment en analyse l'introduction des fonctions holomorphes et des critères de convergence des séries. Ses travaux sur les permutations furent précurseurs de la théorie des groupes. Il fit aussi d'importantes contributions à l'étude de la propagation des ondes en optique et en mécanique.

60. Gene Howard Golub (29 février 1932 - 16 novembre 2007) était un mathématicien américain. Il fut une figure marquante dans le domaine de l'analyse numérique, que soit par ses travaux sur les algorithmes de décomposition et de factorisation de matrices ou par son rôle dans la création du congrès international de mathématiques appliquées et industrielles et de journaux consacrés au calcul scientifique et à l'analyse matricielle.

Si la stabilité numérique du procédé de Gram–Schmidt modifié s'avère suffisante pour un grand nombre d'applications, on doit néanmoins avoir recours à d'autres variantes de la technique lorsque l'on demande [61] que les relations d'orthogonalité entre les vecteurs soient satisfaites à la précision machine. On peut alors choisir de « ré-orthogonaliser » la famille obtenue, en lui appliquant simplement une seconde [62] fois le procédé, ou bien directement employer la correction du procédé proposée dans l'article [DGK+76].

Références

[Aas71] J. O. AASENN. On the reduction of a symmetric matrix to tridiagonal form. *BIT*, 11(3):233–242, 1971. DOI: 10.1007/BF01931804.

[ABB+99] E. ANDERSON, Z. BAI, C. BISCHOF, S. BLACKFORD, J. DEMMEL, J. DONGARRA, J. DU CROZ, A. GREENBAUM, S. HAMMARLING, A. MCKENNEY, and D. SORENSEN. *LAPACK users' guide*. SIAM, third edition, 1999. DOI: 10.1137/1.9780898719604.

[AM87] S. C. ALTHOEN and R. MCLAUGHLIN. Gauss–Jordan reduction: a brief history. *Amer. Math. Monthly*, 94(2):130–142, 1987. DOI: 10.2307/2322413.

[Arn51] W. E. ARNOLDI. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9(1):17–29, 1951. DOI: 10.1090/qam/42792.

[Bar69] E. H. BAREISS. Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices. *Numer. Math.*, 13(5):204–224, 1969. DOI: 10.1007/BF02163269.

[Ben24] Commandant BENOÎT. Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues. – Application de la méthode à la résolution d'un système défini d'équations linéaires (procédé du Commandant Cholesky). *Bull. Géodésique*, 2(1) :67-77, 1924. DOI : 10.1007/BF03031308.

[Bie53] J. BIENAYMÉ. Remarques sur les différences qui distinguent l'interpolation de M. Cauchy de la méthode des moindres carrés, et qui assurent la supériorité de cette méthode. *C. R. Acad. Sci. Paris*, 37 :5-13, 1853.

[Bjö67] Å. BJÖRCK. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7(1):1–21, 1967. DOI: 10.1007/BF01934122.

[Bjö94] Å. BJÖRCK. Numerics of Gram-Schmidt orthogonalization. *Linear Algebra and Appl.*, 197-198:297–316, 1994. DOI: 10.1016/0024-3795(94)90493-6.

[BK77] J. R. BUNCH and L. KAUFMAN. Some stable methods for calculating inertia and solving symmetric linear systems. *Math. Comput.*, 31(137):163–179, 1977. DOI: 10.1090/S0025-5718-1977-0428694-0.

[BP71] J. R. BUNCH and B. N. PARLETT. Direct methods for solving symmetric indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 8(4):639–655, 1971. DOI: 10.1137/0708060.

[Bun71] J. R. BUNCH. Analysis of the diagonal pivoting method. *SIAM J. Numer. Anal.*, 8(4):656–680, 1971. DOI: 10.1137/0708061.

[Cau37] A. CAUCHY. Mémoire sur l'interpolation. *J. Math. Pures Appl. (1)*, 2 :193-205, 1837.

[Cho] A.-L. CHOLESKY. Sur la résolution numérique des systèmes d'équations linéaires. Manuscrit daté du 2 décembre 1910.

[Cia98] P. G. CIARLET. *Introduction à l'analyse numérique matricielle et à l'optimisation – cours et exercices corrigés*, *Mathématiques appliquées pour la maîtrise*. Dunod, 1998.

[Cla88] B.-I. CLASEN. Sur une nouvelle méthode de résolution des équations linéaires et sur l'application de cette méthode au calcul des déterminants. *Ann. Soc. Sci. Bruxelles*, 12(2) :251-281, 1888.

[CM69] E. CUTHILL and J. MCKEE. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 24th ACM national conference*, pages 157–172, 1969. DOI: 10.1145/800195.805928.

[Cro41] P. D. CROUT. A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients. *Trans. Amer. Inst. Elec. Eng.*, 60(12):1235–1241, 1941. DOI: 10.1109/T-AIEE.1941.5058258.

61. C'est par exemple le cas avec le *procédé d'Arnoldi* [Arn51], intervenant dans des méthodes de résolution de systèmes linéaires ou de problèmes aux valeurs propres.

62. Dans [GLR+05], il est montré que, si la famille de vecteurs initiale est « numériquement » de rang maximal, deux applications successives du procédé de Gram–Schmidt classique suffisent pour obtenir une famille « numériquement » orthogonale.

[Cry73] C. W. CRYER. The *LU*-factorization of totally positive matrices. *Linear Algebra and Appl.*, 7(1):83–92, 1973. DOI: 10.1016/0024-3795(73)90039-6.

[dBP76] C. de BOOR and A. PINKUS. Backward error analysis for totally positive linear systems. *Numer. Math.*, 27(4):485–490, 1976. DOI: 10.1007/BF01399609.

[DGK+76] J. W. DANIEL, W. B. GRAGG, L. KAUFMAN, and G. W. STEWART. Reorthogonalization and stable algorithms for updating the Gram-Schmidt *QR* factorization. *Math. Comput.*, 30(136):772–795, 1976. DOI: 10.1090/S0025-5718-1976-0431641-8.

[Dop13] F. M. DOPICO. Alan Turing and the origins of modern Gaussian elimination. *Arbor*, 189(764):a084, 2013. DOI: 10.3989/arbor.2013.764n6007.

[DR79] I. S. DUFF and J. K. REID. Some design features of a sparse matrix code. *ACM Trans. Math. Software*, 5(1):18–35, 1979. DOI: 10.1145/355815.355817.

[Dur60] J. DURBIN. The fitting of time-series models. *Rev. Inst. Internat. Statist.*, 28(3):233–244, 1960. DOI: 10.2307/1401322.

[FHW48] L. FOX, H. D. HUSKEY, and J. H. WILKINSON. Notes on the solution of algebraic linear simultaneous equations. *Quart. J. Mech. Appl. Math.*, 1(1):149–173, 1948. DOI: 10.1093/qjmam/1.1.149.

[FM67] G. E. FORSYTHE and C. B. MOLER. *Computer solution of linear systems*, *Series in automatic computation*. Prentice-Hall, 1967.

[Fos97] L. V. FOSTER. The growth factor and efficiency of Gaussian elimination with rook pivoting. *J. Comput. Appl. Math.*, 86(1):177–194, 1997. DOI: 10.1016/S0377-0427(97)00154-4. Corrigendum in J. Comput. Appl. Math., 98(1):177, 1998.

[Geo71] J. A. GEORGE. *Computer implementation of the finite element method*. PhD thesis, Stanford University, 1971.

[Giv58] W. GIVENS. Computation of plane unitary rotations transforming a general matrix to triangular form. *J. Soc. Ind. Appl. Math.*, 6(1):26–50, 1958. DOI: 10.1137/0106004.

[GLR+05] L. GIRAUD, J. LANGOU, M. ROZLOŽNÍK, and J. van den ESHOF. Rounding error analysis of the classical Gram-Schmidt orthogonalization process. *Numer. Math.*, 101(1):87–100, 2005. DOI: 10.1007/s00211-005-0615-4.

[Gol65] G. GOLUB. Numerical methods for solving linear least squares problems. *Numer. Math.*, 7(3):206–216, 1965. DOI: 10.1007/BF01436075.

[GPS76] N. E. GIBBS, W. G. POOLE, JR., and P. K. STOCKMEYER. A comparison of several bandwidth and profile reduction algorithms. *ACM Trans. Math. Software*, 2(4):322–330, 1976. DOI: 10.1145/355705.355707.

[GPS90] K. A. GALLIVAN, R. J. PLEMMONS, and A. H. SAMEH. Parallel algorithms for dense linear algebra computations. *SIAM Rev.*, 32(1):54–135, 1990. DOI: 10.1137/1032002.

[Gra83] J. P. GRAM. Ueber die Entwicklung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate. *J. Reine Angew. Math.*, 1883(94):41–73, 1883. DOI: 10.1515/crll.1883.94.41.

[Grc11a] J. F. GRCAR. How ordinary elimination became Gaussian elimination. *Historia Math.*, 38(2):163–218, 2011. DOI: 10.1016/j.hm.2010.06.003.

[Grc11b] J. F. GRCAR. John von Neumann's analysis of Gaussian elimination and the origins of modern numerical analysis. *SIAM Rev.*, 53(4):607–682, 2011. DOI: 10.1137/080734716.

[Hig02] N. J. HIGHAM. *Accuracy and stability of numerical algorithms*. SIAM, second edition, 2002. DOI: 10.1137/1.9780898718027.

[Hot43] H. HOTELLING. Some new methods in matrix calculation. *Ann. Math. Statist.*, 14(1):1–34, 1943. DOI: 10.1214/aoms/1177731489.

[Hou58] A. S. HOUSEHOLDER. Unitary triangularization of a nonsymmetric matrix. *J. Assoc. Comput. Mach.*, 5(4):339–342, 1958. DOI: 10.1145/320941.320947.

[Jor88] W. JORDAN. *Handbuch der Vermessungskunde, Erster Band*. Metzler, dritte verbesserte Auflage, 1888.

[KS95] T. KAILATH and A. H. SAYED. Displacement structure: theory and applications. *SIAM Rev.*, 37(3):297–386, 1995. DOI: 10.1137/1037082.

[Lap20] P-S. LAPLACE. *Théorie analytique des probabilités*. Courcier, troisième édition, 1820.

[Lev47] N. LEVINSON. The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25(4):261–278, 1947. DOI: 10.1002/sapm1946251261.

[Mar57] H. M. MARKOWITZ. The elimination form of the inverse and its application to linear programming. *Management Sci.*, 3(3):255–269, 1957. DOI: 10.1287/mnsc.3.3.255.

[NP92] L. NEAL and G. POOLE. A geometric analysis of Gaussian elimination. II. *Linear Algebra and Appl.*, 173:239–264, 1992. DOI: 10.1016/0024-3795(92)90432-A.

[PR70] B. N. PARLETT and J. K. REID. On the solution of a system of linear equations whose matrix is symmetric but not definite. *BIT*, 10(3):386–397, 1970. DOI: 10.1007/BF01934207.

[Ric66] J. R. RICE. Experiments on Gram-Schmidt orthogonalization. *Math. Comput.*, 20(94):325–328, 1966. DOI: 10.1090/S0025-5718-1966-0192673-4.

[Sch07] E. SCHMIDT. Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Math. Ann.*, 63(4):433–476, 1907. DOI: 10.1007/BF01449770.

[Sch17] J. SCHUR. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *J. Reine Angew. Math.*, 1917(147):205–232, 1917. DOI: 10.1515/crll.1917.147.205.

[SM50] J. SHERMAN and W. J. MORRISON. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.*, 21(1):124–127, 1950. DOI: 10.1214/aoms/1177729893.

[Sze39] G. SZEGŐ. *Orthogonal polynomials*, volume 23 of *Colloquium publications*. American Mathematical Society, 1939.

[Tho49] L. H. THOMAS. Elliptic problems in linear difference equations over a network. Technical report, Watson Scientific Computing Laboratory, Columbia University, New York, 1949.

[Tre64] W. F. TRENCH. An algorithm for the inversion of finite Toeplitz matrices. *J. Soc. Indust. Appl. Math.*, 12(3):515–522, 1964. DOI: 10.1137/0112045.

[Tur48] A. M. TURING. Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.*, 1(1):287–308, 1948. DOI: 10.1093/qjmam/1.1.287.

[vNG47] J. von NEUMANN and H. H. GOLDSTINE. Numerical inverting of matrices of high order. *Bull. Amer. Math. Soc.*, 53(11):1021–1099, 1947. DOI: 10.1090/S0002-9904-1947-08909-6.

[Wil61] J. H. WILKINSON. Error analysis of direct methods of matrix inversion. *J. Assoc. Comput. Mach.*, 8(3):281–330, 1961. DOI: 10.1145/321075.321076.

[Yan81] M. YANNAKAKIS. Computing the minimum fill-in is NP-complete. *SIAM. J. Algebraic Discrete Methods*, 2(1):77–79, 1981. DOI: 10.1137/0602010.

[Zoh69] S. ZOHAR. Toeplitz matrix inversion: the algorithm of W. F. Trench. *J. Assoc. Comput. Mach.*, 16(4):592–601, 1969. DOI: 10.1145/321541.321549.

Chapitre 3

Méthodes itératives de résolution des systèmes linéaires

L'idée des méthodes itératives (*iterative methods* en anglais) de résolution de systèmes linéaires est de construire une suite convergente $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ de vecteurs vérifiant

$$\lim_{k\to+\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}, \tag{3.1}$$

où $\boldsymbol{x}$ est solution du système (2.1). Dans ce chapitre, nous présentons des méthodes itératives parmi les plus simples à mettre en œuvre, à savoir les *méthodes de Jacobi* [1], *de Gauss–Seidel* [2] et *de Richardson* [3], ainsi que leurs variantes. Dans celles-ci, partant d'un vecteur initial arbitraire $\boldsymbol{x}^{(0)}$, la suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ vérifie une relation de récurrence de la forme

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \tag{3.2}$$

la matrice carrée B, appelée *matrice d'itération* (*iteration matrix* en anglais) de la méthode, et le vecteur $\boldsymbol{c}$ dépendant de la matrice A et du second membre $\boldsymbol{b}$ du système à résoudre, et possiblement de l'entier k. Ces méthodes suivent ainsi le même principe que les méthodes de point fixe, dont l'application à la résolution d'équations non linéaires est présentée dans le chapitre 5.

Pour une matrice d'ordre n pleine, le coût de calcul induit par une méthode dite *linéaire stationnaire du premier degré* [4] (*linear stationary method of first degree* en anglais) [For53] est de l'ordre de n^2 opérations à *chaque* itération. On a vu au chapitre 2 que le coût *total* d'une méthode directe pour la résolution d'un système linéaire à n équations et n inconnues est de l'ordre de n^3 opérations. Ainsi, une méthode itérative de ce type ne sera compétitive que si elle est en mesure de fournir une solution approchée [5] suffisamment précise en un nombre d'itérations indépendant de l'entier n, ou bien croissant de manière sous-linéaire avec n. Les méthodes directes pouvant cependant s'avérer coûteuses, notamment en termes d'allocation d'espace mémoire, dans certains cas particuliers (un exemple est celui des grandes matrices creuses [6], en raison du phénomène de remplissage évoqué dans le précédent chapitre), les méthodes itératives constituent souvent une alternative intéressante pour la résolution des systèmes linéaires.

Avant d'aborder la description des méthodes mentionnées plus haut, on donne quelques résultats généraux de convergence et de stabilité, ainsi que des principes de comparaison (en terme de *« vitesse » de convergence*), relatifs

1. Carl Gustav Jacob Jacobi (10 décembre 1804 - 18 février 1851) était un mathématicien allemand. Ses travaux portèrent essentiellement sur l'étude des fonctions elliptiques, les équations différentielles et aux dérivées partielles, les systèmes d'équations linéaires et la théorie des déterminants. Un grand nombre de résultats d'algèbre et d'analyse portent ou utilisent son nom.

2. Philipp Ludwig von Seidel (24 octobre 1821 - 13 août 1896) était un mathématicien, physicien de l'optique et astronome allemand. Il a étudié l'aberration optique en astronomie en la décomposant en cinq phénomènes constitutifs, appelés *« les cinq aberrations de Seidel »*, et reste aussi connu pour la méthode de résolution numérique de systèmes linéaires portant son nom.

3. Lewis Fry Richardson (11 octobre 1881 - 30 septembre 1953) était un mathématicien, météorologiste et psychologue britannique. Il imagina de prévoir le temps à partir des équations primitives atmosphériques, les lois de la mécanique des fluides qui régissent les mouvements de l'air.

4. La méthode itérative est *linéaire* si ni la matrice B ni le vecteur $\boldsymbol{c}$ ne dépendent du vecteur $\boldsymbol{x}^{(k)}$ et *stationnaire* si ni B ni $\boldsymbol{c}$ ne dépendent de l'entier k. Enfin, elle est *du premier degré* si le vecteur $\boldsymbol{x}^{(k+1)}$ ne dépend explicitement que du vecteur $\boldsymbol{x}^{(k)}$.

5. On comprend en effet que, à la différence d'une méthode directe, une méthode itérative ne conduit à la solution exacte du problème qu'après avoir, en théorie, effectué un nombre *infini* d'opérations.

6. Il existe néanmoins des solveurs efficaces basés sur des méthodes directes pour ces cas particuliers (voir par exemple [DER86]).

à la classe de méthodes itératives de la forme (3.2). Des résultats plus précis, traitant l'utilisation d'une méthode donnée pour la résolution d'un système linéaire possédant des propriétés ou une structure particulières, sont ensuite établis. Un autre type de méthodes itératives, qui se servent de projections orthogonales pour construire la relation de récurrence (3.2), est présenté en fin de chapitre.

3.1 Méthodes linéaires stationnaires du premier degré

3.1.1 Généralités

Dans cette section, nous abordons quelques aspects généraux des méthodes itératives linéaires stationnaires du premier degré. Dans toute la suite, on désigne par n un entier naturel non nul et l'on considère la résolution d'un système linéaire dont la matrice est d'ordre n à coefficients complexes, les résultats énoncés restant valables *mutatis mutandis* dans le cas réel.

Nous commençons par introduire la notion fondamentale de *convergence* (*convergence* en anglais) d'une méthode itérative.

Définition 3.1 (convergence d'une méthode itérative linéaire stationnaire du premier degré) *On dit qu'une méthode itérative définie par* (3.2) *est* ***convergente*** *(****convergent*** *en anglais) si, pour tout choix d'initialisation* $\boldsymbol{x}^{(0)}$*, la suite* $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ *possède une limite indépendante de* $\boldsymbol{x}^{(0)}$.

Cette première définition ne fait aucunement intervenir le système linéaire (2.1) que l'on cherche à résoudre. Il convient par conséquent de préciser en quel sens la matrice d'itération B et le vecteur $\boldsymbol{c}$, intervenant dans la relation de récurrence définissant la méthode, sont liés à la matrice A et au second membre $\boldsymbol{b}$ de ce système. On utilise pour cela la notion de *consistance* (*consistency* en anglais) d'une méthode itérative [You72].

Définitions 3.2 (consistance d'une méthode itérative linéaire stationnaire du premier degré) *Une méthode itérative définie par* (3.2) *est dite* ***consistante*** *(****consistent*** *en anglais) avec le système linéaire* (2.1) *si toute solution* $\boldsymbol{x}$ *de ce dernier satisfait l'égalité*

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}. \tag{3.3}$$

Elle est dite ***réciproquement consistante*** *(****reciprocally consistent*** *en anglais) avec* (2.1) *si toute solution de* (3.3) *satisfait* (2.1) *et* ***complètement consistante*** *(****completely consistent*** *en anglais) avec* (2.1) *si elle est à la fois consistante et réciproquement consistante avec* (2.1).

Les précédentes définitions n'exigent pas que les matrices A et $I_n - B$ soient inversibles, mais que les systèmes linéaires qui leur sont respectivement associés possèdent au moins une solution. On peut montrer (voir le théorème 1 dans [You72]) qu'une méthode itérative définie par (3.2) est complètement consistante avec le système (2.1) si et seulement s'il existe une matrice M inversible telle que

$$B = I_n - M^{-1}A \text{ et } \boldsymbol{c} = M^{-1}\boldsymbol{b}.$$

Comme nous l'avons fait dans le précédent chapitre, nous supposons dorénavant que la matrice du système linéaire (2.1) est inversible. On a dans ce cas le résultat suivant.

Théorème 3.3 *Soit A une matrice d'ordre n inversible. Une méthode itérative définie par* (3.2) *est*

- *consistante avec le système* (2.1) *si et seulement si*

$$\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b},$$

- *complètement consistante avec le système* (2.1) *si et seulement si elle est consistante et la matrice* $I_n - B$ *est inversible.*

DÉMONSTRATION. La matrice A étant inversible, on a $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Si la méthode itérative est consistante, ce vecteur est solution du système (3.3), d'où $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$. Réciproquement, si $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ et $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$, alors $\boldsymbol{c} = (I_n - B)\boldsymbol{x}$ et la méthode est consistante.

Supposons à présent que la méthode itérative est consistante et que la matrice $I_n - B$ est inversible, la matrice A étant inversible. D'après la première assertion du théorème, on a alors $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$, d'où $\boldsymbol{b} = A(I_n - B)^{-1}\boldsymbol{c}$. On en déduit que la

méthode est réciproquement consistante (il suffit pour cela de procéder comme dans la preuve de la première assertion en échangeant A et $\boldsymbol{b}$ avec $I_n - B$ et $\boldsymbol{c}$) et donc bien complètement consistante. Réciproquement, supposons que la méthode est complètement consistante. Alors, le système linéaire (2.1) admettant une unique solution, il en va de même pour le système (3.3), ce qui implique que la matrice $I_n - B$ est inversible. □

La consistance complète d'une méthode itérative est une propriété essentielle si l'on veut que celle-ci converge vers la solution du système linéaire (2.1), comme le souligne le résultat suivant.

Théorème 3.4 *Soit A une matrice d'ordre n inversible, $\boldsymbol{x}$ l'unique vecteur solution du système linéaire* (2.1) *et une méthode itérative définie par* (3.2)*. Si la suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode converge vers $\boldsymbol{x}$ pour toute initialisation $\boldsymbol{x}^{(0)}$, alors la méthode est complètement consistante. Réciproquement, si la méthode est complètement consistante et convergente, alors la suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ a pour limite $\boldsymbol{x}$.*

DÉMONSTRATION. Si la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{x}$, alors, par passage à la limite dans la relation (3.2), on obtient que $\boldsymbol{x}$ est solution du système (3.3) et la méthode itérative est consistante. Supposons maintenant qu'il existe une solution $\boldsymbol{z}$ du système (3.3), différente de $\boldsymbol{x}$. Si $\boldsymbol{x}^{(0)} = \boldsymbol{z}$, on a $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(2)} = \cdots = \boldsymbol{z}$, ce qui vient contredire le fait que la méthode est convergente. Le vecteur $\boldsymbol{x}$ est donc l'unique solution de (3.3) et la méthode est complètement consistante.

D'autre part, si la méthode est convergente, la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ a pour limite une solution du système (3.3). Si la méthode est de plus complètement consistante, elle est en particulier réciproquement consistante et $\boldsymbol{x}$ est alors l'unique solution de (3.3). □

Pour caractériser la convergence d'une méthode itérative vers la solution du système linéaire (2.1), on utilise deux quantités particulières.

Définitions 3.5 (erreur et résidu d'une méthode itérative de résolution de système linéaire) *Soit k un entier naturel. On appelle **erreur** (**error** en anglais) à l'itération k d'une méthode itérative le vecteur $\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$, où $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ est la solution de* (2.1)*. Le **résidu** (**residual** en anglais) à l'itération k d'une méthode itérative est le vecteur $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$.*

On déduit des précédentes définitions qu'une méthode itérative converge vers la solution du système linéaire (2.1) si et seulement si $\lim_{k\to+\infty} \boldsymbol{e}^{(k)} = \boldsymbol{0}$ (soit encore si $\lim_{k\to+\infty} \boldsymbol{r}^{(k)} = -\lim_{k\to+\infty} A\boldsymbol{e}^{(k)} = \boldsymbol{0}$) pour tout choix d'initialisation $\boldsymbol{x}^{(0)}$.

La seule propriété de consistance complète d'une méthode itérative ne suffisant pas à assurer qu'elle converge [7], le résultat ci-après fournit un critère fondamental de convergence.

Théorème 3.6 *Soit A une matrice inversible. Une méthode itérative linéaire stationnaire du premier ordre, définie par* (3.2) *et complètement consistante avec le système* (2.1)*, est convergente si et seulement si $\rho(B) < 1$, où $\rho(B)$ désigne le rayon spectral de la matrice d'itération de la méthode.*

DÉMONSTRATION. Soit k un entier naturel. La méthode étant supposée complètement consistante avec (2.1), l'erreur à la $k+1^{\text{e}}$ itération vérifie la relation

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x} = B\boldsymbol{x}^{(k)} - \boldsymbol{c} - (B\boldsymbol{x} - \boldsymbol{c}) = B\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}\right) = B\boldsymbol{e}^{(k)}.$$

Le résultat se déduit alors du théorème A.159. □

En pratique, le rayon spectral d'une matrice peut être difficile à calculer, mais on déduit du théorème A.157 que le rayon spectral d'une matrice B est strictement inférieur à 1 s'il existe au moins une norme matricielle pour laquelle $\|B\| < 1$. L'étude de convergence des méthodes itératives de résolution de systèmes linéaires de la forme (3.2) repose donc sur la détermination de $\rho(B)$ ou, de manière équivalente, la recherche d'une norme matricielle pour laquelle $\|B\| < 1$.

7. Un exemple particulièrement simple de ce fait est celui donné par Young dans [You71]. Soit le système linéaire

$$2I_n\boldsymbol{x} = \boldsymbol{b}$$

et la méthode itérative ayant pour relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = -\boldsymbol{x}^{(k)} + \boldsymbol{b}.$$

On a $A = 2I_n$, $B = -I_n$ et $\boldsymbol{c} = \boldsymbol{b} = (I_n - B)A^{-1}\boldsymbol{b}$, la méthode est donc complètement consistante d'après le théorème 3.3. On trouve cependant que

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(2k)} = \boldsymbol{x}^{(0)} \text{ et } \boldsymbol{x}^{(2k+1)} = -\boldsymbol{x}^{(0)} + \boldsymbol{b}.$$

La suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ n'est donc convergente que si $\boldsymbol{x}^{(0)} = \boldsymbol{x} = \frac{1}{2}\boldsymbol{b}$.

Une autre question se posant lorsque l'on est en présence de méthodes itératives convergentes est de savoir laquelle converge le plus rapidement. Une réponse est fournie par le résultat suivant, que l'on peut résumer ainsi : la méthode la plus « rapide » est celle dont la matrice d'itération a le plus petit rayon spectral.

Théorème 3.7 *Soit $\|\cdot\|$ une norme vectorielle quelconque. On considère deux méthodes itératives complètement consistantes avec le système* (2.1)*,*

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c} \text{ et } \tilde{\boldsymbol{x}}^{(k+1)} = \tilde{B}\tilde{\boldsymbol{x}}^{(k)} + \tilde{\boldsymbol{c}},$$

avec $\boldsymbol{x}^{(0)} = \tilde{\boldsymbol{x}}^{(0)}$ et $\rho(B) < \rho(\tilde{B})$. Alors, pour tout réel strictement positif ε, il existe un entier N tel que

$$k \geq N \implies \sup_{\substack{\boldsymbol{x}^{(0)} \in M_{n,1}(\mathbb{C}) \\ \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| = 1}} \left(\frac{\|\tilde{\boldsymbol{x}}^{(k)} - \boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|} \right)^{1/k} \geq \frac{\rho(\tilde{B})}{\rho(B) + \varepsilon},$$

où $\boldsymbol{x}$ désigne la solution de (2.1).

DÉMONSTRATION. D'après la formule de Gelfand (voir le théorème A.161), étant donné un réel ε strictement positif, il existe un entier naturel N, dépendant de ε, tel que

$$\forall k \in \mathbb{N},\ k \geq N \implies \sup_{\substack{\boldsymbol{e}^{(0)} \in M_{n,1}(\mathbb{C}) \\ \|\boldsymbol{e}^{(0)}\| = 1}} \|B^k \boldsymbol{e}^{(0)}\|^{1/k} \leq (\rho(B) + \varepsilon).$$

Par ailleurs, pour tout entier naturel k supérieur ou égal à N, il existe un vecteur $\boldsymbol{e}^{(0)}$, dépendant de k, tel que

$$\|\boldsymbol{e}^{(0)}\| = 1 \text{ et } \|\tilde{B}^k \boldsymbol{e}^{(0)}\|^{1/k} = \|\tilde{B}^k\|^{1/k} \geq \rho(\tilde{B}),$$

en vertu du théorème A.157 et en notant également $\|\cdot\|$ la norme matricielle subordonnée à la norme vectorielle considérée. Ceci achève de démontrer l'assertion. □

On mesure la vitesse de convergence d'une méthode itérative convergente de matrice d'itération B en introduisant le *taux asymptotique de convergence* (*asymptotic rate of convergence* en anglais) [You54], défini par

$$R_\infty(B) = -\ln(\rho(B)).$$

Un autre taux utilisé est le *taux moyen de convergence après k itérations* (*average rate of convergence after k iterations* en anglais), qui, pour tout entier naturel k non nul, est égal à

$$R_k(B) = -\ln\left(\|B^k\|^{1/k}\right) = -\frac{\ln(\|B^k\|)}{k},$$

où $\|\cdot\|$ désigne une norme de matrice compatible avec la norme vectorielle dans laquelle on mesure l'erreur de la méthode itérative (le taux $R_k(B)$ dépend par conséquent d'un choix de norme, ce qui n'est pas le cas pour $R_\infty(B)$). Pour toute méthode convergente, on peut facilement montrer que, pour tout entier naturel k non nul tel que $\|B^k\| < 1$, on a

$$R_k(B) \leq R_\infty(B)$$

et, en utilisant la formule de Gelfand (voir le théorème A.161),

$$\lim_{k \to +\infty} R_k(B) = R_\infty(B),$$

justifiant ainsi le qualificatif d'« asymptotique » pour $R_\infty(B)$.

Si le taux asymptotique de convergence est de loin le plus commode [8] d'emploi, il fournit parfois une estimation trop optimiste du taux moyen de convergence (qui est celui réellement observé en pratique), la convergence de la suite $(\|B^k\|)_{k \in \mathbb{N}^*}$ n'étant pas nécessairement monotone (voir l'exemple 5.2 dans [Axe94]). Néanmoins, l'inverse de

8. L'utilisation du taux moyen de convergence implique en effet l'évaluation de puissances de la matrice d'itération, ce qui est en général coûteux, voire impossible si la matrice d'itération n'est pas explicitement disponible (comme dans le cas de certaines méthodes itératives définies par (3.5)).

l'un ou l'autre de ces taux fournit une estimation du nombre d'itérations à effectuer pour diviser l'erreur par un facteur e, qui est la *constante de Napier*[9] ($e = 2,718281828459\ldots$).

On peut voir la méthode itérative (3.2) comme un cas particulier de méthode de point fixe, à l'instar des méthodes introduites dans le chapitre 5 pour la résolution d'équations non linéaires (on pourra à ce titre comparer les relations (3.2) et (5.10)). Il a aussi été remarqué (voir [Wit36] pour l'une des plus anciennes références) que les méthodes itératives inventées avant les années 1950 appartenaient à une même famille (voir le tableau 3.1), obtenue en considérant une « décomposition » (on parle en anglais de *"splitting"*) de la matrice du système à résoudre de la forme

$$A = M - N, \tag{3.4}$$

avec M une matrice inversible, appelée dans certains contextes le *préconditionneur* (*preconditioner* en anglais) de la méthode, et en posant

$$\forall k \in \mathbb{N},\ M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b}. \tag{3.5}$$

Il découle des résultats donnés plus haut qu'une méthode itérative basée sur une telle décomposition est, par construction, complètement consistante avec le système (2.1) dès que ce dernier possède au moins une solution.

Pour que la dernière formule de récurrence soit d'utilité en pratique, il convient que tout système linéaire ayant M pour matrice puisse être résolu simplement et à faible coût. Les méthodes de Jacobi, de Gauss–Seidel et de Richardson stationnaire, basées sur la décomposition (3.4) et présentées ci-après, correspondent ainsi au choix d'une matrice inversible M soit diagonale, soit triangulaire inférieure. Par ailleurs, pour qu'une méthode définie par (3.5) converge, il faut (et il suffit), en vertu du théorème 3.6, que la valeur du rayon spectral de sa matrice d'itération, donnée par $M^{-1}N$, soit strictement inférieure à 1 et il est même souhaitable, compte tenu du théorème 3.7, qu'elle soit la plus petite possible.

nom de la méthode	M	N
Jacobi	D	$E+F$
sur-relaxation simultanée	$\frac{1}{\omega}D$	$\frac{1-\omega}{\omega}D+E+F$
Gauss–Seidel	$D-E$	F
sur-relaxation successive	$\frac{1}{\omega}D-E$	$\frac{1-\omega}{\omega}D+F$
sur-relax. succes. symétrique	$\frac{1}{\omega(2-\omega)}(D-\omega E)D^{-1}(D-\omega F)$	$\frac{1}{\omega(2-\omega)}((1-\omega)D+\omega E)D^{-1}((1-\omega)D+\omega F)$
Richardson stationnaire	$\frac{1}{\alpha}I_n$	$\frac{1}{\alpha}I_n - A$

TABLE 3.1 – Choix des matrices M et N apparaissant dans la relation de récurrence (3.5) pour les méthodes itératives linéaires stationnaires du premier degré présentées dans ce chapitre (les matrices D, E et F apparaissant dans les expressions sont celles de la décomposition (3.8) de A et les réels ω et α sont supposés non nuls).

Plusieurs résultats de convergence, propres aux méthodes de Jacobi et de Gauss–Seidel (ainsi que leurs variantes relaxées) ou à la méthode de Richardson stationnaire, sont donnés dans la section 3.1.6. De manière plus générale, le résultat ci-après fournit une condition nécessaire et suffisante de convergence d'une méthode itérative associée à un choix de décomposition (3.4) d'une matrice A hermitienne[10] définie positive.

Théorème 3.8 (« théorème de Householder–John »[11]) *Soit A une matrice hermitienne inversible, dont la décomposition sous la forme* (3.4)*, avec M une matrice inversible, est telle que la matrice hermitienne $M^* + N$ est définie positive. On a alors $\rho(M^{-1}N) < 1$ si et seulement si A est définie positive.*

DÉMONSTRATION. La matrice A (supposée d'ordre n) étant hermitienne, la matrice $M^* + N$ est effectivement hermitienne puisque

$$M^* + N = M^* + M - A = M + M^* - A^* = M + N^*.$$

9. John Napier (1550 - 4 avril 1617) était un physicien, mathématicien, astronome et astrologue écossais. Il établit quelques formules de trigonométrie sphérique, popularisa l'usage anglo-saxon du point comme séparateur entre les parties entière et fractionnaire d'un nombre et inventa les logarithmes.

10. Comme on l'a déjà mentionné, tous les résultats énoncés le sont pour une matrice à coefficients complexes, mais restent vrais dans le cas réel. Il suffit de remplacer le mot « hermitienne » par l'expression « réelle symétrique » et la transconjugaison par la transposition dans les énoncés.

11. La preuve de suffisance de la condition est attribuée à John [Joh66], celle de sa nécessité à Householder [Hou58] (ce dernier créditant Reich [Rei49]).

Supposons la matrice A définie positive. L'application $\|\cdot\|$ de $M_{n,1}(\mathbb{C})$ dans $\mathbb{R}$ définie par

$$\forall \boldsymbol{v} \in M_{n,1}(\mathbb{C}),\ \|\boldsymbol{v}\| = (\boldsymbol{v}^* A \boldsymbol{v})^{1/2},$$

est alors une norme vectorielle euclidienne. On désigne par $\|\cdot\|$ la norme matricielle qui lui est subordonnée.

Nous allons établir que $\|M^{-1}N\| < 1$. Par définition, on a

$$\|M^{-1}N\| = \|I_n - M^{-1}A\| = \sup_{\substack{\boldsymbol{v} \in M_{n,1}(\mathbb{C}) \\ \|\boldsymbol{v}\|=1}} \|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|.$$

D'autre part, pour tout vecteur $\boldsymbol{v}$ de $M_{n,1}(\mathbb{C})$ tel que $\|\boldsymbol{v}\| = 1$, on vérifie que

$$\begin{aligned}
\|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|^2 &= (\boldsymbol{v} - M^{-1}A\boldsymbol{v})^* A (\boldsymbol{v} - M^{-1}A\boldsymbol{v}) \\
&= \boldsymbol{v}^* A \boldsymbol{v} - \boldsymbol{v}^* A (M^{-1}A\boldsymbol{v}) - (M^{-1}A\boldsymbol{v})^* A \boldsymbol{v} + (M^{-1}A\boldsymbol{v})^* A (M^{-1}A\boldsymbol{v}) \\
&= \|\boldsymbol{v}\|^2 - (M^{-1}A\boldsymbol{v})^* M^* (M^{-1}A\boldsymbol{v}) - (M^{-1}A\boldsymbol{v})^* M (M^{-1}A\boldsymbol{v}) + (M^{-1}A\boldsymbol{v})^* A (M^{-1}A\boldsymbol{v}) \\
&= 1 - (M^{-1}A\boldsymbol{v})^* (M^* + N)(M^{-1}A\boldsymbol{v}) < 1,
\end{aligned}$$

puisque la matrice $M^* + N$ est définie positive par hypothèse. La fonction de $M_{n,1}(\mathbb{C})$ dans $\mathbb{R}$ qui à $\boldsymbol{v}$ associe $\|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|$ étant continue sur le compact $\{\boldsymbol{v} \in M_{n,1}(\mathbb{C}) \mid \|\boldsymbol{v}\| = 1\}$, elle y atteint sa borne supérieure en vertu d'une généralisation du théorème B.88. Ceci achève la première partie de la démonstration.

Supposons à présent que $\rho(M^{-1}N) < 1$. En vertu du théorème 3.6, la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$, définie par

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = M^{-1}N\boldsymbol{x}^{(k)},$$

converge vers le vecteur nul pour toute initialisation $\boldsymbol{x}^{(0)}$.

Raisonnons par l'absurde et faisons l'hypothèse que la matrice A n'est pas définie positive. Il existe alors un vecteur $\boldsymbol{x}^{(0)}$ non nul tel que $(\boldsymbol{x}^{(0)})^* A \boldsymbol{x}^{(0)} \leq 0$. Le vecteur $M^{-1}A\boldsymbol{x}^{(0)}$ étant non nul, on déduit des calculs effectués plus haut et de l'hypothèse sur $M^* + N$ que

$$(\boldsymbol{x}^{(0)})^* (A - (M^{-1}N)^* A (M^{-1}N))\boldsymbol{x}^{(0)} = (M^{-1}A\boldsymbol{x}^{(0)})^* (M^* + N)(M^{-1}A\boldsymbol{x}^{(0)}) > 0.$$

La matrice $A - (M^{-1}N)^* A (M^{-1}N)$ étant définie positive (elle est en effet congruente à $M^* + N$ qui est définie positive), on a par ailleurs

$$\forall k \in \mathbb{N},\ 0 \leq (\boldsymbol{x}^{(k)})^* (A - (M^{-1}N)^* A (M^{-1}N))\boldsymbol{x}^{(k)} = (\boldsymbol{x}^{(k)})^* A \boldsymbol{x}^{(k)} - (\boldsymbol{x}^{(k+1)})^* A \boldsymbol{x}^{(k+1)},$$

l'inégalité étant stricte pour $k = 0$, d'où

$$\forall k \in \mathbb{N}^*,\ (\boldsymbol{x}^{(k+1)})^* A \boldsymbol{x}^{(k+1)} \leq (\boldsymbol{x}^{(k)})^* A \boldsymbol{x}^{(k)} \leq \cdots \leq (\boldsymbol{x}^{(1)})^* A \boldsymbol{x}^{(1)} < (\boldsymbol{x}^{(0)})^* A \boldsymbol{x}^{(0)} \leq 0.$$

Ceci contredit le fait que $\boldsymbol{x}^{(k)}$ tend vers $\mathbf{0}$ lorsque l'entier k tend vers l'infini ; la matrice A est donc définie positive. □

Terminons cette section en abordant la question de la stabilité numérique des méthodes itératives de la forme (3.5). Celles-ci sont en effet destinées à être mises en œuvre sur des calculateurs pour lesquels le résultat de toute opération arithmétique peut être affecté par une erreur d'arrondi. Il convient donc de s'assurer que la convergence des méthodes ne se trouve pas altérée en pratique. Le résultat suivant montre qu'il n'en est rien.

Théorème 3.9 *Soit A une matrice d'ordre n inversible, décomposée sous la forme* (3.4)*, avec M une matrice inversible et $\rho(M^{-1}N) < 1$, $\boldsymbol{b}$ une matrice de $M_{n,1}(\mathbb{C})$ et $\boldsymbol{x}$ l'unique solution de* (2.1)*. On suppose de plus qu'à chaque étape la méthode itérative est affectée d'une erreur, au sens où*

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = M^{-1}N\boldsymbol{x}^{(k)} + M^{-1}\boldsymbol{b} + \boldsymbol{\epsilon}^{(k)} \tag{3.6}$$

avec $\boldsymbol{\epsilon}^{(k)}$ un vecteur de $M_{n,1}(\mathbb{C})$, et qu'il existe une norme vectorielle $\|\cdot\|$ et une constante positive ϵ telles que, pour tout entier naturel k,

$$\|\boldsymbol{\epsilon}^{(k)}\| \leq \epsilon.$$

Alors, il existe une constante positive K, ne dépendant que de $M^{-1}N$ telle que

$$\limsup_{k\to+\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq K\,\epsilon.$$

DÉMONSTRATION. Compte tenu de (3.6), l'erreur à l'étape $k+1$ vérifie la relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k+1)} = M^{-1}N\boldsymbol{e}^{(k)} + \boldsymbol{\epsilon}^{(k)},$$

dont on déduit que

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k)} = (M^{-1}N)^k\boldsymbol{e}^{(0)} + \sum_{i=0}^{k-1}(M^{-1}N)^i\boldsymbol{\epsilon}^{(k-i-1)}.$$

Puisque $\rho(M^{-1}N) < 1$, il existe, par application du théorème A.157, une norme matricielle subordonnée $\|\cdot\|_s$ telle que $\|M^{-1}N\| < 1$; on note également $\|\cdot\|_s$ la norme vectorielle qui lui est associée. Les normes vectorielles sur $M_{n,1}(\mathbb{C})$ étant équivalentes, il existe une constante C, strictement plus grande que 1 et ne dépendant que de $M^{-1}N$, telle que

$$\forall \boldsymbol{v} \in M_{n,1}(\mathbb{C}),\ C^{-1}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_s \leq C\,\|\boldsymbol{v}\|.$$

Par majoration, il vient alors

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k)}\|_s \leq \|M^{-1}N\|_s{}^k\|\boldsymbol{e}^{(0)}\|_s + C\,\epsilon\sum_{i=0}^{k-1}\|M^{-1}N\|_s{}^i \leq \|M^{-1}N\|_s{}^k\|\boldsymbol{e}^{(0)}\|_s + \frac{C\,\epsilon}{1+\|M^{-1}N\|_s},$$

d'où on tire le résultat en posant $K = \dfrac{C^2}{1+\|M^{-1}N\|_s}$. □

3.1.2 Méthode de Jacobi

Observons tout d'abord que, si les coefficients diagonaux de la matrice A sont non nuls, il est possible d'isoler la i^{e} inconnue dans la i^{e} équation du système linéaire (2.1), $1 \leq i \leq n$ et l'on obtient alors le système équivalent

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\,x_j\right),\ i = 1,\ldots,n.$$

La méthode de Jacobi [Jac45] se base sur ces relations pour construire, à partir d'un vecteur initial $\boldsymbol{x}^{(0)}$ donné, une suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ par récurrence

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\,x_j^{(k)}\right),\ i = 1,\ldots,n, \tag{3.7}$$

ce qui implique que $M = D$ et $N = E+F$ dans la décomposition (3.4) de la matrice A, où D est la matrice diagonale contenant les coefficients diagonaux de A, $d_{ij} = a_{ij}\,\delta_{ij}$, $1 \leq i \leq j \leq n$, E est la matrice triangulaire inférieure de coefficients $e_{ij} = -a_{ij}$ si $1 \leq j < i \leq n$ et 0 sinon, et F est la matrice triangulaire supérieure telle que $f_{ij} = -a_{ij}$ si $1 \leq i < j \leq n$ et 0 sinon. Autrement dit, on a

$$A = D - (E+F), \tag{3.8}$$

la matrice d'itération de la méthode de Jacobi étant donnée par

$$B_J = D^{-1}(E+F).$$

On remarquera que la matrice diagonale D doit ici être inversible pour que la méthode soit bien définie. Cette condition n'est cependant pas très restrictive dans la mesure où l'ordre des équations et des inconnues peut être modifié. On observe par ailleurs que le calcul des composantes de la nouvelle approximation de la solution à chaque itération peut se faire de manière concomitante (on parle de calculs pouvant être effectués *en parallèle*). Pour cette raison, la méthode est aussi connue sous le nom de *méthode des déplacements simultanés* (*method of simultaneous displacements* en anglais).

Une généralisation de la méthode de Jacobi est la *méthode de sur-relaxation simultanée* ou *de sur-relaxation de Jacobi* (*simultaneous over-relaxation* ou *Jacobi over-relaxation (JOR) method* en anglais), dans laquelle un paramètre de relaxation réel non nul, noté ω, est introduit. Dans ce cas, les relations de récurrence deviennent

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} \right) + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n,$$

ce qui correspond aux choix

$$M = \frac{1}{\omega}\, D,\ N = \frac{1-\omega}{\omega}\, D + E + F$$

et à la matrice d'itération

$$B_{JOR}(\omega) = \omega\, D^{-1}(E+F) + (1-\omega)\, I_n. \tag{3.9}$$

Cette méthode est consistante pour toute valeur non nulle de ω et coïncide avec la méthode de Jacobi pour $\omega = 1$. L'idée de relaxer la méthode repose sur le fait que, si l'efficacité de la méthode se mesure par la valeur du rayon spectral de la matrice d'itération, alors, en utilisant le fait que la fonction $\omega \mapsto \rho(B_{JOR}(\omega))$ est continue, on peut trouver une valeur du paramètre ω pour laquelle ce rayon est le plus petit possible, donnant ainsi une méthode itérative potentiellement plus efficace que la méthode de Jacobi. Ce type de technique s'applique également à la méthode de Gauss–Seidel (voir la prochaine section), pour laquelle il est d'ailleurs bien plus couramment employé.

L'étude des méthodes de relaxation pour un type de matrice donné consiste en général à déterminer, s'ils existent, un intervalle I de $\mathbb{R}$ ne contenant pas l'origine, tel que la méthode converge pour toute valeur de ω choisie dans I, et une valeur optimale ω_o du paramètre de relaxation telle que (dans le cas présent)

$$\rho(B_{JOR}(\omega_o)) = \inf_{\omega \in I} \rho(B_{JOR}(\omega)).$$

3.1.3 Méthodes de Gauss–Seidel et de sur-relaxation successive

Remarquons à présent que, lors d'un calcul *séquentiel* des composantes du vecteur $\boldsymbol{x}^{(k+1)}$ par les formules de récurrence (3.7), les premières $i-1^{\text{es}}$ composantes sont connues au moment de la détermination de i^{e} composante, si l'entier i est compris entre 2 et n. La méthode de Gauss–Seidel, parfois appelée [12] *méthode des déplacements successifs* (*method of successive displacements* en anglais), utilise ce fait, en se servant des composantes du vecteur $\boldsymbol{x}^{(k+1)}$ déjà obtenues pour le calcul des suivantes. Ceci conduit aux relations

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right),\ i = 1, \ldots, n, \tag{3.10}$$

ce qui équivaut, en utilisant la décomposition (3.8), à poser $M = D - E$ et $N = F$ dans la relation (3.4), la matrice d'itération associée à la méthode étant alors

$$B_{GS} = (D-E)^{-1} F.$$

Pour que cette méthode soit bien définie, il faut que la matrice $D-E$ soit inversible, ce qui équivaut une nouvelle fois à ce que la matrice D soit inversible. Comme on l'a vu précédemment, une condition de ce type n'est pas très restrictive en pratique si A est inversible. Par rapport à la méthode de Jacobi, la méthode de Gauss–Seidel présente l'avantage de ne pas nécessiter le stockage simultané de deux approximations successives de la solution.

Comme pour la méthode de Jacobi, on peut utiliser une version relaxée de cette méthode. On parle alors de *méthode de sur-relaxation successive* (*successive over-relaxation (SOR) method* en anglais) [Fra50, You54],

12. On trouve encore dans la littérature (voir [Fra50]) le nom de *méthode de Liebmann* [Lie18] lorsque la méthode est appliquée à la résolution d'un système linéaire issu de la discrétisation par différences finies d'une équation de Laplace.

définie par [13]

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n,$$

ce qui revient à poser

$$M = \frac{1}{\omega} D - E \text{ et } N = \frac{1-\omega}{\omega} D + F.$$

la matrice d'itération étant

$$B_{SOR}(\omega) = \left(\frac{1}{\omega} D - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right) = (D - \omega E)^{-1}((1-\omega) D + \omega F) = (I_n - \omega D^{-1} E)^{-1} \left((1-\omega) I_n + \omega D^{-1} F \right).$$

Cette dernière méthode est consistante pour toute valeur de ω non nulle et coïncide avec la méthode de Gauss–Seidel pour $\omega = 1$. Si $\omega > 1$, on parle de *sur-relaxation* (*over-relaxation* en anglais) et de *sous-relaxation* (*under-relaxation* en anglais) si $\omega < 1$. Dans les applications, il s'avère que la valeur optimale ω_o du paramètre est généralement plus grande que 1, d'où le nom de la méthode.

Notons que l'on peut en quelque sorte « symétriser » la méthode de sur-relaxation successive en combinant à chaque itération la relation de récurrence de cette méthode, dans laquelle l'approximation est mise à jour dans l'ordre de la numération des inconnues du système, ce qui s'écrit encore

$$\forall k \in \mathbb{N},\ (D - \omega E)\, \boldsymbol{x}^{(k+\frac{1}{2})} = (\omega F + (1-\omega) D)\, \boldsymbol{x}^{(k)} + \omega\, \boldsymbol{b}, \tag{3.11}$$

avec celle d'une méthode de sur-relaxation successive *rétrograde* (*backward SOR method* en anglais), consistant à mettre à jour l'approximation dans l'ordre inverse de la numérotation des inconnues, c'est-à-dire

$$\forall k \in \mathbb{N},\ (D - \omega F)\, \boldsymbol{x}^{(k+1)} = (\omega E + (1-\omega) D)\, \boldsymbol{x}^{(k+\frac{1}{2})} + \omega\, \boldsymbol{b}. \tag{3.12}$$

Cette formulation à deux étapes définit la méthode dite de *sur-relaxation successive symétrique* (*symmetric successive over-relaxation (SSOR) method* en anglais), introduite pour la résolution numérique de problèmes aux limites elliptiques discrétisés par la méthode des différences finies par Sheldon [She55] et généralisant une méthode proposée par Aitken [14] [Ait50], de matrice d'itération

$$B_{SSOR}(\omega) = (D - \omega F)^{-1}(\omega E + (1-\omega) D)(D - \omega E)^{-1}(\omega F + (1-\omega) D),$$

et de vecteur [15]

$$\boldsymbol{c} = \omega(2-\omega)(D - \omega F)^{-1} D (D - \omega E)^{-1} \boldsymbol{b}.$$

On peut vérifier que cette méthode correspond au choix de préconditionneur

$$M = \frac{1}{\omega(2-\omega)} (D - \omega E) D^{-1} (D - \omega F).$$

13. Du fait du caractère nécessairement séquentiel de la méthode de Gauss–Seidel, la relaxation doit s'interpréter dans le sens suivant : en introduisant le vecteur « auxiliaire » $\boldsymbol{x}^{(k+\frac{1}{2})}$ dont les composantes sont définies par

$$\forall k \in \mathbb{N},\ a_{ii}\, x_i^{(k+\frac{1}{2})} = b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)},\ i = 1, \ldots, n,$$

on voit que les composantes de l'approximation $\boldsymbol{x}^{(k+1)}$ de la méthode SOR sont données par les moyennes pondérées

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \omega\, x_i^{(k+\frac{1}{2})} + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n.$$

14. Alexander Craig Aitken (1er avril 1895 - 3 novembre 1967) était un mathématicien néo-zélandais et l'un des meilleurs calculateurs mentaux connus. Il fût élu à la *Royal Society of London* en 1936 pour ses travaux dans le domaine des statistiques, de l'algèbre et de l'analyse numérique.

15. On obtient cette expression en remarquant que $I_n + (\omega E + (1-\omega) D)(D - \omega E)^{-1} = I_n + ((2-\omega) D - (D - \omega E))(D - \omega E)^{-1} = (2-\omega) D (D - \omega E)^{-1}$.

Si elle converge sous les mêmes hypothèses que la méthode de sur-relaxation successive, la méthode de sur-relaxation successive symétrique est moins sensible au choix de la valeur du paramètre ω, ce qui rend d'autant moins cruciale la connaissance précise de la valeur optimale de ce paramètre. Il est aussi remarquable que la matrice d'itération de cette méthode est semblable [16] à une matrice hermitienne (resp. symétrique) si la matrice A est hermitienne (resp. symétrique), ceci n'étant généralement pas le cas avec les autres méthodes présentées. Elle est ainsi particulièrement adaptée à l'utilisation du procédé d'accélération connu sous le nom de *méthode semi-itérative de Tchebychev*.

Indiquons enfin que le coût de cette dernière méthode n'est pas nécessairement le double de celui de la méthode de sur-relaxation successive, comme on pourrait naïvement le penser. En effet, il a été observé par Niethammer [Nie64] que les relations de récurrence (3.11) et (3.12) pouvaient respectivement être réécrites sous la forme

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+\frac{1}{2})} = \boldsymbol{x}^{(k)} + \omega\,(D^{-1}E\boldsymbol{x}^{(k+\frac{1}{2})} - \boldsymbol{x}^{(k)} + D^{-1}F\boldsymbol{x}^{(k)} + D^{-1}\boldsymbol{b}), \tag{3.13}$$

et

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k+\frac{1}{2})} + \omega\,(D^{-1}F\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k+\frac{1}{2})} + D^{-1}E\boldsymbol{x}^{(k+\frac{1}{2})} + D^{-1}\boldsymbol{b}). \tag{3.14}$$

Une fois passée la première étape, on constate que le vecteur $D^{-1}E\boldsymbol{x}^{(k+\frac{1}{2})}$ apparaissant dans la relation (3.14) a été calculé lors de l'application de la relation (3.13) pour obtenir le vecteur $\boldsymbol{x}^{(k+\frac{1}{2})}$. De manière analogue, le vecteur $D^{-1}F\boldsymbol{x}^{(k+1)}$, calculé lors l'évaluation de (3.14), est utilisé à l'itération suivante dans la relation

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+\frac{3}{2})} = \boldsymbol{x}^{(k+1)} + \omega\,(D^{-1}E\boldsymbol{x}^{(k+\frac{3}{2})} - \boldsymbol{x}^{(k+1)} + D^{-1}F\boldsymbol{x}^{(k+1)} + D^{-1}\boldsymbol{b}).$$

Ainsi, moyennant le stockage d'un vecteur additionnel en mémoire, le coût de la méthode s'avère en fait voisin de celui de la méthode de sur-relaxation successive.

3.1.4 Méthode de Richardson stationnaire

On peut remarquer que la relation de récurrence (3.5) définissant les méthodes itératives vues jusqu'à présent peut encore s'écrire sous la forme « corrective » suivante

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + M^{-1}\boldsymbol{r}^{(k)}, \tag{3.15}$$

faisant intervenir le vecteur $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ qui est le résidu à l'étape k de la méthode. Dans cette dernière, le choix $M^{-1} = \alpha I_n$ avec α un réel non nul, correspondant à la décomposition $M = \frac{1}{\alpha} I_n$ et $N = \frac{1}{\alpha} I_n - A$, conduit à la méthode de Richardson *stationnaire* [17]

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\,(\boldsymbol{b} - A\boldsymbol{x}^{(k)}), \tag{3.16}$$

de matrice d'itération

$$B_R(\alpha) = I_n - \alpha A$$

et de vecteur

$$\boldsymbol{c} = \alpha\boldsymbol{b}.$$

16. On peut effectivement montrer que la matrice $(I_n - \omega D^{-1}E^*)B_{SSOR}(\omega)(I_n - \omega D^{-1}E^*)^{-1}$ (resp. $(I_n - \omega D^{-1}E^\top)B_{SSOR}(\omega)(I_n - \omega D^{-1}E^\top)^{-1}$) est hermitienne (resp. réelle symétrique) si la matrice A est hermitienne (resp. réelle symétrique), c'est-à-dire que $F = E^*$ (resp. $F = E^\top$). Il suffit pour cela de remarquer que la matrice d'itération de la méthode peut, dans le cas où A est hermitienne, encore s'écrire

$$B_{SSOR}(\omega) = (I_n - \omega D^{-1}E^*)^{-1}(\omega D^{-1}E + (1-\omega)I_n)(I_n - \omega D^{-1}E)^{-1}(\omega D^{-1}E^* + (1-\omega)I_n),$$

et que les matrices $(I_n - \omega D^{-1}E)^{-1}$ et $\omega D^{-1}E + (1-\omega)I_n$ commutent pour le produit matriciel. On trouve alors que

$$(I_n - \omega D^{-1}E^*)B_{SSOR}(\omega)(I_n - \omega D^{-1}E^*)^{-1} = (D - \omega E)^{-1}(\omega E + (1-\omega)D)(\omega E^* + (1-\omega)D)(D - \omega E^*)^{-1}.$$

17. Dans l'article original de Richardson [Ric11], il s'avère que la valeur du paramètre α dépend de l'itération, c'est-à-dire que l'on a la relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\,(\boldsymbol{b} - A\boldsymbol{x}^{(k)}),$$

qui donne lieu à une méthode itérative *instationnaire* (*non-stationary iterative method* en anglais).

3.1.5 Méthode des directions alternées

Cette méthode a été introduite par Peaceman et Rachford pour la résolution de systèmes linéaires provenant de la discrétisation par la méthode des différences finies d'équations aux dérivées partielles de type elliptique ou parabolique posées dans un domaine bidimensionnel [PR55]. Elle considère que la matrice A du système linéaire à résoudre peut s'écrire comme la somme $A = A_1 + A_2$, dans laquelle les matrices A_1 et A_2 sont telles que la résolution des systèmes linéaires dont les matrices respective sont $\gamma I_n + A_1$ et $\gamma I_n + A_2$, avec γ un paramètre réel, est facile et peu coûteuse.

À partir de l'égalité $A\boldsymbol{x} = (A_1 + A_2)\boldsymbol{x} = \boldsymbol{b}$, on peut écrire

$$\begin{aligned}(\gamma I_n + A_1)\boldsymbol{x} &= (\gamma I_n - A_2)\boldsymbol{x} + \boldsymbol{b},\\ (\gamma I_n + A_2)\boldsymbol{x} &= (\gamma I_n - A_1)\boldsymbol{x} + \boldsymbol{b},\end{aligned}$$

dont on tire une méthode dans laquelle chaque itération contient deux étapes

$$\begin{aligned}\forall k \in \mathbb{N},\ (\gamma I_n + A_1)\boldsymbol{x}^{(k+\frac{1}{2})} &= (\gamma I_n - A_2)\boldsymbol{x}^{(k)} + \boldsymbol{b},\\ (\gamma I_n + A_2)\boldsymbol{x}^{(k+1)} &= (\gamma I_n - A_1)\boldsymbol{x}^{(k+\frac{1}{2})} + \boldsymbol{b},\end{aligned}$$

le nombre réel γ, que l'on peut d'ailleurs faire dépendre de l'itération, étant à choisir de manière à obtenir la convergence la plus rapide possible. Le nom de la méthode (*alternating direction implicit (ADI) method* en anglais) vient du fait que les matrices A_1 et A_2 de la décomposition utilisée représentent chacune un opérateur aux différences finies dans une direction d'espace distincte.

Tout comme la méthode de sur-relaxation successive symétrique précédemment vue, la méthode des directions alternées entre dans le cadre des méthodes linéaires stationnaires du premier degré caractérisées par la relation de récurrence (3.2). On montre en effet par une substitution que celle-ci a pour matrice d'itération

$$B_{ADI}(\gamma) = (\gamma I_n + A_2)^{-1}(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)$$

et vecteur

$$\boldsymbol{c} = (\gamma I_n + A_2)^{-1}(I_n + (\gamma I_n - A_1)(\gamma I_n + A_1)^{-1})\boldsymbol{b}.$$

3.1.6 Résultats de convergence

Dans toute la suite, nous faisons l'hypothèse que les coefficients diagonaux de la matrice A sont non nuls. Avant de considérer la résolution de systèmes linéaires dont les matrices possèdent des propriétés ou une structure particulières, nous commençons par donner un résultat général pour la méthode de sur-relaxation successive.

Théorème 3.10 (condition nécessaire de convergence pour la méthode SOR [Kah58]) *Le rayon spectral de la matrice de la méthode de sur-relaxation successive vérifie toujours l'inégalité*

$$\forall \omega \in]0, +\infty[,\ \rho(B_{SOR}(\omega)) \geq |\omega - 1|\,.$$

Cette méthode ne peut donc converger que si ω appartient à l'intervalle $]0, 2[$.

DÉMONSTRATION. On remarque que le déterminant de la matrice $B_{SOR}(\omega)$ vaut

$$\det(B_{SOR}(\omega)) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = (1-\omega)^n,$$

compte tenu des structures, respectivement diagonale et triangulaires, des matrices D, E et F de la décomposition (3.8). En notant $\lambda_1, \ldots, \lambda_n$ les valeurs propres de la matrice, on en déduit alors que

$$\rho(B_{SOR}(\omega))^n \geq \prod_{i=1}^{n} |\lambda_i| = |\det(B_{SOR}(\omega))| = |1-\omega|^n\,,$$

l'égalité n'ayant lieu que lorsque toutes les valeurs propres de $B_{SOR}(\omega)$ sont de module égal à $|1-\omega|$. □

Nous mentionnons ensuite un résultat concernant la méthode de sur-relaxation de Jacobi.

Théorème 3.11 *Si la méthode de Jacobi converge, alors la méthode de sur-relaxation de Jacobi converge pour ω appartenant à l'intervalle* $]0,1]$.

DÉMONSTRATION. D'après l'identité (3.9), les valeur propres de la matrice $B_{JOR}(\omega)$ sont

$$\eta_i = \omega\,\lambda_i + 1 - \omega,\ i = 1,\dots,n,$$

où les nombres $\lambda_1,\dots,\lambda_n$ sont les valeurs propres de la matrice B_J. En posant $\lambda_i = \rho_i\, e^{\mathrm{i}\theta_i}$, $i = 1,\dots,n$, on a alors

$$|\eta_i| = \omega^2 {\rho_i}^2 + 2\omega\rho_i\cos(\theta_i)(1-\omega) + (1-\omega)^2 \leq (\omega\rho_i + 1 - \omega)^2,\ i = 1,\dots,n.$$

On en déduit que le rayon spectral est strictement inférieur à 1 si ω appartient à $]0,1]$. □

On peut enfin énoncer le résultat général suivant concernant la méthode de Richardson stationnaire.

Théorème 3.12 (condition nécessaire et suffisante de convergence pour la méthode de Richardson stationnaire) *La méthode de Richardson stationnaire est convergente si et seulement si*

$$\forall i \in \{1,\dots,n\},\ \frac{2\,\mathrm{Re}(\lambda_i)}{\alpha\,|\lambda_i|^2} > 1,$$

où les scalaires $\lambda_1,\dots,\lambda_n$, sont les valeurs propres de la matrice A du système linéaire à résoudre.

DÉMONSTRATION. Les valeurs propres de la matrice d'itération $B_R(\alpha) = I_n - \alpha A$ étant données par $1 - \alpha\,\lambda_i$, $i = 1,\dots,n$. La condition nécessaire et suffisante de convergence équivaut alors à satisfaire les inégalités

$$\forall i \in \{1,\dots,n\},\ |1 - \alpha\,\lambda_i| < 1,$$

soit encore

$$\forall i \in \{1,\dots,n\},\ (1 - \alpha\,\mathrm{Re}(\lambda_i))^2 + (\alpha\,\mathrm{Im}(\lambda_i))^2 < 1,$$

dont se déduit immédiatement la condition donnée dans l'énoncé. □

Cas des matrices à diagonale strictement dominante

Nous avons déjà abordé le cas particulier des matrices à diagonale strictement dominante dans le cadre de leur factorisation au chapitre précédent. Pour de telles matrices, on est en mesure d'affirmer *a priori* que les méthodes itératives que l'on a introduites convergent, comme initialement établi par von Mises [18] et Geiringer [19] [vMP29] et par Collatz [20] [Col42].

Théorème 3.13 *Si A est une matrice à diagonale strictement dominante par lignes, alors les méthodes de Jacobi et de Gauss–Seidel sont convergentes.*

DÉMONSTRATION. Soit A une matrice d'ordre n à diagonale strictement dominante par lignes, c'est-à-dire que $|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$ pour $i = 1,\dots,n$. En posant

$$r = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|,$$

et en observant alors que $\|B_J\|_\infty = r < 1$, on en déduit que la méthode de Jacobi est convergente.

Soit k un entier naturel. On considère l'erreur à l'itération $k+1$ de la méthode de Gauss–Seidel. Celle-ci vérifie

$$e_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}\, e_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}\, e_j^{(k)},\ i = 1,\dots,n.$$

18. Richard Edler von Mises (19 avril 1883 - 14 juillet 1953) était un scientifique américain d'origine autrichienne. Ses contributions couvrent des disciplines aussi variées que la mécanique des solides et des fluides, l'aérodynamique et l'aéronautique, la géométrie de construction, la statistique, la théorie des probabilités ou encore la philosophie des sciences.

19. Hilda Geiringer (aussi connue sous le nom de Hilda von Mises, 28 septembre 1893 - 22 mars 1973) était une mathématicienne autrichienne qui travailla dans les domaines des statistiques, des probabilités et de la théorie mathématique de la plasticité.

20. Lothar Collatz (6 juillet 1910 - 26 septembre 1990) était un mathématicien allemand. Il est connu pour la conjecture portant son nom, relative aux suites de Syracuse, énoncée en 1937 et non résolue à ce jour. Il fût un promoteur des mathématiques appliquées au calcul numérique, avec des contributions importantes à la résolution approchée des équations différentielles et des problèmes aux valeurs propres.

On va établir que

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\|_\infty \leq r\,\|\boldsymbol{e}^{(k)}\|_\infty,$$

en raisonnant par récurrence sur les composantes du vecteur d'erreur. Pour $i = 1$, on a

$$\forall k \in \mathbb{N},\ e_1^{(k+1)} = -\sum_{j=2}^{n} \frac{a_{1j}}{a_{11}}\, e_j^{(k)},$$

d'où $|e_1^{(k+1)}| \leq r\,\|\boldsymbol{e}^{(k)}\|_\infty$. Supposons maintenant que $|e_j^{(k+1)}| \leq r\,\|\boldsymbol{e}^{(k)}\|_\infty$ pour $j = 1, \ldots, i-1$. On a

$$\forall k \in \mathbb{N},\ |e_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k+1)}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k)}| \leq \|\boldsymbol{e}^{(k)}\|_\infty \left(r \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \right) < \|\boldsymbol{e}^{(k)}\|_\infty \sum_{\substack{i=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|,$$

d'où $|e_i^{(k+1)}| \leq r\,\|\boldsymbol{e}^{(k)}\|_\infty$, ce qui achève la preuve par récurrence. On a par conséquent

$$\forall k \in \mathbb{N}^*,\ \|\boldsymbol{e}^{(k)}\|_\infty \leq r\,\|\boldsymbol{e}^{(k-1)}\|_\infty \leq \cdots \leq r^k\,\|\boldsymbol{e}^{(0)}\|_\infty,$$

et, par suite,

$$\lim_{k\to+\infty} \|\boldsymbol{e}^{(k)}\|_\infty = 0,$$

ce qui prouve la convergence de la méthode de Gauss–Seidel. □

Cas des matrices hermitiennes définies positives

Lorsque la matrice du système à résoudre est hermitienne définie positive, on peut établir que la condition nécessaire de convergence de la méthode de sur-relaxation successive du théorème 3.10 est suffisante.

Théorème 3.14 (« théorème d'Ostrowski[21]–Reich »[22]) *Si la matrice A est hermitienne à coefficients diagonaux strictement positifs, alors la méthode de sur-relaxation successive converge pour tout ω appartenant à $]0, 2[$ si et seulement si A est définie positive.*

DÉMONSTRATION. On peut voir ce résultat comme un corollaire du théorème de Householder–John. En effet, la matrice A étant hermitienne, on a, à partir de (3.8), $D - E - F = D^* - E^* - F^*$, et donc $D = D^*$ et $F = E^*$, compte tenu de la définition de ces matrices. Le paramètre ω étant un réel non nul, il vient alors

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega} D + F = \frac{2-\omega}{\omega}\, D.$$

La matrice D étant définie positive par hypothèse, la matrice $M^* + N$ est définie positive pour ω appartenant à l'intervalle $]0, 2[$ et il suffit alors d'appliquer le théorème 3.8 pour conclure. □

On a aussi le résultat suivant pour la méthode de sur-relaxation de Jacobi.

Théorème 3.15 (condition suffisante de convergence de la méthode JOR) *Si la matrice A est hermitienne définie positive, alors la méthode de sur-relaxation de Jacobi converge si le paramètre ω appartient à l'intervalle $\left]0, \frac{2}{\rho(D^{-1}A)}\right[$, où D est la matrice diagonale de la décomposition* (3.8).

DÉMONSTRATION. Puisque la matrice A est hermitienne, on peut utiliser le théorème 3.8 pour conclure, à condition que la matrice hermitienne $\frac{2}{\omega} D - A$ soit définie positive. Ses valeurs propres étant données par $\frac{2}{\omega} a_{ii} - \lambda_i$, $i = 1, \ldots, n$, où les réels $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de la matrice A, ceci implique que

$$0 < \omega < \frac{2\,a_{ii}}{\lambda_i},\ i = 1, \ldots, n,$$

d'où le résultat. □

Enfin, le théorème 3.12, traitant de la convergence de la méthode de Richardson stationnaire, peut également être complété dans ce cas.

21. Alexander Markowich Ostrowski (Александр Маркович Островский en russe, 25 septembre 1893 - 20 novembre 1986) était un mathématicien suisse d'origine russe. Ses contributions, extrêmement variées, portent sur divers domaines des mathématiques, au nombre desquels l'analyse numérique.

22. Ce résultat a été établi par Reich [Rei49] pour la méthode de Gauss–Seidel ($\omega = 1$) et par Ostrowski [Ost54] dans le cas général ($0 < \omega < 2$).

Théorème 3.16 *Si la matrice A est hermitienne définie positive, alors la méthode de Richardson stationnaire converge si et seulement si*

$$0 < \alpha < \frac{2}{\rho(A)}.$$

De plus, la valeur du rayon spectral de la matrice d'itération de la méthode est minimale pour le choix

$$\alpha_o = \frac{2}{\lambda_n + \lambda_1}$$

les réels strictement positifs λ_1 et λ_n étant respectivement la plus petite et la plus grande des valeurs propres de la matrice A, et vaut

$$\rho(B_R(\alpha_o)) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1},$$

où le réel $\mathrm{cond}_2(A)$ est le conditionnement de la matrice A relativement à la norme spectrale.

DÉMONSTRATION. La matrice A étant hermitienne définie positive, ses valeurs propres sont des réels strictement positifs et l'on a $\rho(B_R(\alpha)) = \max\{|1-\alpha\lambda_1|, |1-\alpha\lambda_n|\}$ (voir la figure 3.1). La condition nécessaire et suffisante du théorème 3.12 se résume alors à

$$1 - \alpha\lambda_1 < 1 \text{ et } 1 - \alpha\lambda_n > -1,$$

d'où

$$0 < \alpha < \frac{2}{\lambda_n}.$$

On conclut alors en remarquant que $\rho(A) = \lambda_n$.

Pour minimiser la valeur du rayon spectral $\rho(B_R(\alpha))$, il suffit de voir, comme observé sur la figure 3.1, que la valeur optimale α_o du paramètre α est celle l'abscisse du point d'intersection des graphes des fonctions $\alpha \mapsto |1-\alpha\lambda_1|$ et $\alpha \mapsto |1-\alpha\lambda_n|$, ce qui signifie encore que

$$\alpha_o\lambda_n - 1 = 1 - \alpha_o\lambda_1.$$

On en déduit par conséquent que

$$\alpha_o = \frac{2}{\lambda_n + \lambda_1}$$

et alors

$$\rho(B_R(\alpha_o)) = 1 - \frac{2}{\lambda_n + \lambda_1}\lambda_1 = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Enfin, la dernière égalité de l'énoncé est une conséquence du théorème A.168. □

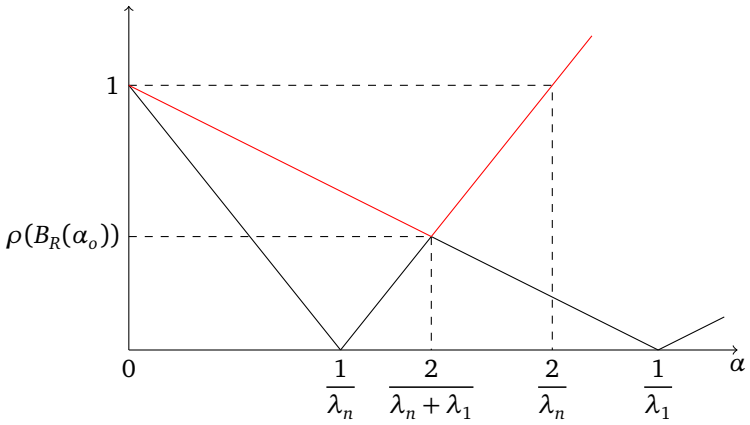


FIGURE 3.1 – Graphes des fonctions $\alpha \mapsto |1-\alpha\lambda_1|$, $\alpha \mapsto |1-\alpha\lambda_n|$ et de la valeur du rayon spectral de la matrice d'itération $B_R(\alpha)$ en fonction de la valeur du paramètre α (le dernier étant tracé en rouge) dans le cas d'une matrice A hermitienne définie positive.

On conclut en donnant un résultat de convergence pour la méthode des directions alternées.

Théorème 3.17 (condition suffisante de convergence de la méthode des directions alternées) *Si la matrice A est hermitienne définie positive et que les matrices A_1 et A_2 de la décomposition $A = A_1 + A_2$ sont hermitiennes positives, alors la méthode des directions alternées converge si le paramètre γ est strictement positif.*

DÉMONSTRATION. Observons tout d'abord que, la matrice A_1 étant hermitienne positive, ses valeurs propres sont réelles positives et la matrice $\gamma I_n + A_1$ est donc inversible si γ est strictement positif. De plus, les matrices $\gamma I_n + A_1$ et $\gamma I_n - A_1$ étant hermitiennes et commutant entre elles, on en déduit que la matrice $(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}$ est hermitienne. Ainsi, on montre que

$$\|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}\|_2 = \rho((\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}) = \max_{\lambda \in \sigma(A_1)} \left| \frac{\gamma - \lambda}{\gamma + \lambda} \right| \leq 1,$$

avec égalité si et seulement si 1 est valeur propre de la matrice $(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}$. Ces résultats restent vrais si l'on remplace A_1 par A_2.

La matrice d'itération $B_{ADI}(\gamma)$ étant semblable à la matrice $(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}$, on a alors

$$\rho(B_{ADI}(\gamma)) \leq \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\|_2 \leq \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}\|_2 \|(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\|_2 \leq 1.$$

Il reste à montrer que l'on a $\rho(B_{ADI}(\gamma)) < 1$. On raisonne pour cela par l'absurde. Si $\rho(B_{ADI}(\gamma)) = 1$, alors $\|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\|_2 = 1$. Ainsi, par définition de d'une norme subordonnée, il existe $\boldsymbol{y}$ appartenant à $M_{n,1}(\mathbb{C})$ tel que $\|\boldsymbol{y}\|_2 = 1$ et

$$\begin{aligned} 1 = \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\boldsymbol{y}\|_2 &\leq \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}\|_2 \|(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\boldsymbol{y}\|_2 \\ &\leq \|(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\boldsymbol{y}\|_2 \\ &\leq \|(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\|_2 \|\boldsymbol{y}\|_2 \\ &\leq 1, \end{aligned}$$

d'où $\|(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\boldsymbol{y}\|_2 = 1$ La décomposition de $\boldsymbol{y}$ dans une base orthonormée de vecteurs propres de la matrice $(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}$ et la relation de Pythagore permettent alors de montrer que ce vecteur est un vecteur propre de $(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}$ associé à 1, d'où $(\gamma I_n - A_2)\boldsymbol{y} = (\gamma I_n + A_2)\boldsymbol{y}$, soit encore $A_2\boldsymbol{y} = \boldsymbol{0}$. Ces dernières conclusions restent vraies en remplaçant A_2 par A_1, puisque l'on a à présent

$$1 = \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}(\gamma I_n - A_2)(\gamma I_n + A_2)^{-1}\boldsymbol{y}\|_2 = \|(\gamma I_n - A_1)(\gamma I_n + A_1)^{-1}\boldsymbol{y}\|_2.$$

On a ainsi obtenu que $\boldsymbol{y}$ est un vecteur non nul appartenant à $\ker(A_1) \cap \ker(A_2) \subset \ker(A)$, ce qui contredit le fait que A est définie positive (et par conséquent inversible). □

Cas des matrices tridiagonales

On en mesure comparer la convergence des méthodes de Jacobi, de Gauss–Seidel et de sur-relaxation successive dans le cas particulier des matrices tridiagonales.

Théorème 3.18 *Si la matrice A est tridiagonale, alors les rayons spectraux des matrices d'itération des méthodes de Jacobi et de Gauss–Seidel sont liés par la relation*

$$\rho(B_{GS}) = \rho(B_J)^2$$

de sorte que les deux méthodes convergent ou divergent simultanément. En cas de convergence, la méthode de Gauss–Seidel converge plus rapidement que celle de Jacobi.

Pour prouver ce résultat, on a besoin du lemme technique suivant.

Lemme 3.19 *Le déterminant de la matrice $A(\alpha)$, tridiagonale d'ordre n et définie pour tout scalaire α non nul par*

$$A(\alpha) = \begin{pmatrix} a_{11} & \alpha^{-1}a_{12} & 0 & \dots & 0 \\ \alpha\, a_{21} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha^{-1}a_{n-1\,n} \\ 0 & \dots & 0 & \alpha\, a_{n\,n-1} & a_{nn} \end{pmatrix}, \tag{3.17}$$

ne dépend pas de α.

DÉMONSTRATION. On remarque, en introduisant la matrice diagonale d'ordre n inversible (α étant non nul)

$$Q(\alpha) = \begin{pmatrix} \alpha & & & \\ & \alpha^2 & & \\ & & \ddots & \\ & & & \alpha^n \end{pmatrix},$$

que les matrices $A(\alpha)$ et $A(1)$ sont semblables, car on a $A(\alpha) = Q(\alpha)A(1)Q(\alpha)^{-1}$. □

On est à présent en mesure de démontrer le théorème 3.18.

DÉMONSTRATION DU THÉORÈME 3.18. Les valeurs propres de la matrice d'itération de la méthode de Jacobi $B_J = D^{-1}(E+F)$ sont les racines du polynôme caractéristique

$$\chi_{B_J}(\lambda) = \det(\lambda I_n - B_J) = \det(D^{-1})\det(\lambda D - E - F).$$

les matrices D, E et F étant celles de la décomposition (3.8). De même, les valeurs propres de la matrice d'itération de la méthode de Gauss–Seidel $B_{GS} = (D-E)^{-1}F$ sont les zéros du polynôme

$$\chi_{B_{GS}}(\lambda) = \det(\lambda I_n - B_{GS}) = \det((D-E)^{-1})\det(\lambda D - \lambda E - F).$$

Compte tenu de la structure tridiagonale de A, la matrice $\lambda^2 D - \alpha\lambda^2 E - \alpha^{-1}F$ est de la forme (3.17) et l'application du lemme 3.19 avec le choix $\alpha = \lambda^{-1}$ conduit alors à

$$\det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n \det(\lambda D - E - F),$$

d'où

$$\chi_{B_{GS}}(\lambda^2) = \frac{\det(D)}{\det(D-E)}\,\lambda^n \chi_J(\lambda) = \lambda^n \chi_J(\lambda).$$

De cette dernière relation, on déduit que, pour tout scalaire λ non nul,

$$\lambda^2 \in \sigma(B_{GS}) \iff \pm\lambda \in \sigma(B_J),$$

et donc $\rho(B_{GS}) = \rho(B_J)^2$. □

On remarque que, dans la démonstration ci-dessus, on a établi une bijection entre les valeurs propres non nulles de la matrice B_{GS} et les paires de valeurs propres non nulles de signes opposés de la matrice B_J.

Si la matrice tridiagonale est de plus hermitienne définie positive, le théorème 3.14 assure que la méthode de sur-relaxation successive converge pour ω appartenant à l'intervalle $]0,2[$. La méthode de Gauss–Seidel (qui correspond au choix $\omega = 1$) est par conséquent convergente, ainsi que la méthode de Jacobi en vertu du théorème 3.18. De plus, on est en mesure de déterminer une valeur explicite du paramètre de relaxation optimal de la méthode de sur-relaxation successive. Ceci est l'objet du résultat suivant.

Théorème 3.20 *Si la matrice A est tridiagonale, hermitienne et définie positive, alors la méthode de sur-relaxation successive converge pour tout réel ω appartenant à l'intervalle $]0,2[$ et il existe un unique paramètre optimal, valant*

$$\omega_o = \frac{2}{1+\sqrt{1-\rho(B_J)^2}},$$

minimisant le rayon spectral de la matrice d'itération de cette méthode.

DÉMONSTRATION. La matrice A étant hermitienne définie positive, on sait en vertu du théorème 3.14 que la méthode de sur-relaxation successive est convergente si et seulement si le paramètre ω appartenant à l'intervalle $]0,2[$. Il reste donc à déterminer la valeur du paramètre optimal ω_o. Pour cela, on commence par définir, pour tout scalaire α non nul, la matrice

$$A(\alpha) = \frac{\lambda^2+\omega-1}{\omega}\,D - \alpha\lambda^2 E - \frac{1}{\alpha}\,F,$$

les matrices D, E et F étant issues de la décomposition (3.8) de A. Par une application du lemme 3.19, on obtient alors que

$$\det\left(\frac{\lambda^2+\omega-1}{\omega}\,D - \lambda E - \lambda F\right) = \det(A(\lambda^{-1})) = \det(A(1)) = \det\left(\frac{\lambda^2+\omega-1}{\omega}\,D - \lambda^2 E - F\right).$$

En remarquant alors que

$$\chi_{B_{SOR}(\omega)}(\lambda^2) = \det\left(\left(E-\frac{D}{\omega}\right)^{-1}\right)\det\left(\frac{\lambda^2+\omega-1}{\omega}D-\lambda^2E-F\right),$$

il vient

$$\chi_{B_{SOR}(\omega)}(\lambda^2) = \frac{\det\left(E-\frac{D}{\omega}\right)}{\det(-D)}\lambda^n\chi_{B_J}\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right).$$

On déduit que, pour tout scalaire λ non nul,

$$\lambda^2 \in \sigma(B_{SOR}(\omega)) \iff \pm\frac{\lambda^2+\omega-1}{\lambda\omega} \in \sigma(B_J).$$

Ainsi, si le réel μ est valeur propre de la matrice d'itération B_J, son opposé $-\mu$ l'est également les carrés des racines

$$\lambda_\pm(\mu,\omega) = \frac{\mu\omega \pm \sqrt{\mu^2\omega^2-4(\omega-1)}}{2}$$

de l'équation du second degré (en λ)

$$\frac{\lambda^2+\omega-1}{\lambda\omega} = \mu,$$

sont des valeurs propres de la matrice $B_{SOR}(\omega)$. Par conséquent, on a la caractérisation suivante

$$\rho(B_{SOR}(\omega)) = \max_{\mu\in\sigma(B_J)} \max\left\{\left|\lambda_+(\mu,\omega)^2\right|, \left|\lambda_-(\mu,\omega)^2\right|\right\}.$$

On va maintenant montrer que les valeurs propres de la matrice B_J sont réelles. On a

$$B_J\boldsymbol{v} = \mu\,\boldsymbol{v} \iff (E+F)\boldsymbol{v} = \mu D\boldsymbol{v} \iff A\boldsymbol{v} = (1-\mu)D\boldsymbol{v} \implies \boldsymbol{v}^*A\boldsymbol{v} = (1-\mu)\boldsymbol{v}^*D\boldsymbol{v},$$

et le scalaire $1-\mu$ est donc un réel positif, les matrices A et D étant toutes deux hermitiennes définies positives. Pour déterminer le rayon spectral $\rho(B_{SOR}(\omega))$, il suffit alors d'étudier la fonction

$$\begin{array}{rccl} m: & [0,1[\times]0,2[& \mapsto & \mathbb{R} \\ & (\mu,\omega) & \to & \max\left\{\left|\lambda_+(\mu,\omega)^2\right|, \left|\lambda_-(\mu,\omega)^2\right|\right\}. \end{array}$$

En effet, on a $|\mu|<1$ et il est par ailleurs inutile de considérer la fonction pour des valeurs strictement négatives de μ, puisque [23] $\lambda_\pm(-\mu,\omega)^2 = \lambda_\mp(\mu,\omega)^2$. D'autre part, on sait que la méthode ne peut converger si le paramètre ω n'est pas contenu dans $]0,2[$.
Pour $\mu=0$, on vérifie que

$$m(0,\omega) = |\omega-1|.$$

Pour $0<\mu<1$, le trinôme $\omega\to\mu^2\omega^2-4(\omega-1)$ possède deux racines réelles $\omega_\pm(\mu)$ vérifiant

$$1<\omega_+(\mu) = \frac{2}{1+\sqrt{1-\mu^2}} < 2 < \omega_-(\mu) = \frac{2}{1-\sqrt{1-\mu^2}}.$$

Si $\frac{2}{1+\sqrt{1-\mu^2}} < \omega < 2$, les nombres complexes $\lambda_+(\mu,\omega)^2$ et $\lambda_-(\mu,\omega)^2$ sont conjugués et un calcul simple montre alors que

$$m(\mu,\omega) = \left|\lambda_+(\mu,\omega)^2\right| = \left|\lambda_-(\mu,\omega)^2\right| = \omega-1.$$

Si $0<\omega<\frac{2}{1+\sqrt{1-\mu^2}}$, on voit facilement que

$$m(\mu,\omega) = \lambda_+(\mu,\omega)^2.$$

On a ainsi, pour $0<\mu<1$ et $0<\omega<\frac{2}{1+\sqrt{1-\mu^2}}$,

$$\frac{\partial m}{\partial\mu}(\mu,\omega) = 2\,\lambda_+(\mu,\omega)\frac{\partial\lambda_+}{\partial\mu}(\mu,\omega) = \lambda_+(\mu,\omega)\left(\omega+\frac{\mu\omega^2}{\sqrt{\mu^2\omega^2-4(\omega-1)}}\right) > 0,$$

et donc, à valeur fixée de ω, il vient

$$\max_{\mu\in\sigma(B_J)}\left|\lambda_+(\mu,\omega)^2\right| = \left|\lambda_+(\rho(B_J),\omega)^2\right|.$$

23. On a en effet $\lambda_\pm(\mu,\omega)^2 = \frac{1}{2}(\mu^2\omega^2-2(\omega-1)) \pm \frac{\mu\omega}{2}\sqrt{\mu^2\omega^2-4(\omega-1)}$.

On peut à présent minimiser le rayon spectral $\rho(B_{SOR}(\omega))$ par rapport à ω. Pour $0 < \omega < \frac{2}{1+\sqrt{1-\rho(B_J)^2}}$, il vient

$$\frac{\partial}{\partial \mu}\left|\lambda_+(\rho(B_J),\omega)^2\right| = 2\,\lambda_+(\rho(B_J),\omega)\,\frac{\partial \lambda_+}{\partial \omega}(\rho(B_J),\omega) = \lambda_+(\rho(B_J),\omega)\left(\rho(B_J) + \frac{\rho(B_J)\omega - 2}{2\sqrt{\rho(B_J)^2\omega^2 - 4(\omega-1)}}\right)$$
$$= 2\,\lambda_+(\rho(B_J),\omega)\,\frac{\rho(B_J)\lambda_+(\rho(B_J),\omega) - 1}{\sqrt{\rho(B_J)^2\omega^2 - 4(\omega-1)}}.$$

Sachant que $0 < \rho(B_J) < 1$, on trouve que le minimum de $\left|\lambda_+(\rho(B_J),\omega)^2\right|$ sur $\left]0, \frac{2}{1+\sqrt{1-\rho(B_J)^2}}\right]$ est atteint en $\frac{2}{1+\sqrt{1-\rho(B_J)^2}}$. D'autre part, le minimum de la fonction $\omega - 1$ sur $\left[\frac{2}{1+\sqrt{1-\rho(B_J)^2}}, 2\right]$ est également atteint en ce point. On en déduit que, lorsque ω varie dans l'intervalle $]0,2[$, le minimum de $\rho(B_{SOR}(\omega))$ est atteint en

$$\omega_o = \frac{2}{1+\sqrt{1-\rho(B_J)^2}},$$

et l'on a alors $\rho(B_{SOR}(\omega_o)) = \omega_o - 1$ (voir la figure 3.2). □

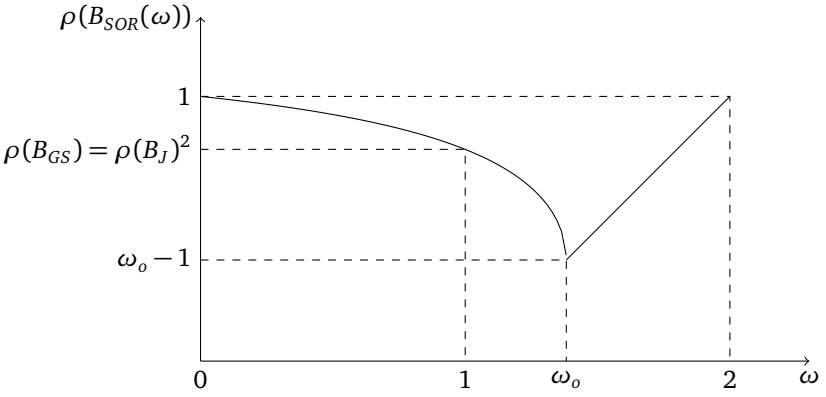


FIGURE 3.2 – Graphe de la valeur du rayon spectral de la matrice d'itération $B_{SOR}(\omega)$ en fonction de la valeur du paramètre de relaxation ω dans le cas d'une matrice A tridiagonale, hermitienne et définie positive.

Remarque 3.21 *Les précédents résultats restent plus généralement vrais pour toute matrice A possédant la* « propriété (A) »*, c'est-à-dire pour laquelle il existe une matrice de permutation P telle qu'on a la structure par blocs suivante*

$$PAP^\top = \begin{pmatrix} D_1 & C_1 \\ C_2 & D_2 \end{pmatrix},$$

où D_1 et D_2 sont des matrices diagonales, pas nécessairement de même ordre (cela signifie encore que A peut être vue comme la matrice d'adjacence d'un graphe biparti [24]). Pour de telles matrices, comme les matrices tridiagonales ou tridiagonales par blocs, on peut en effet montrer (voir [You54]) que :

- *si μ est une valeur propre de la matrice d'itération B_J, alors $-\mu$ est aussi une valeur propre de B_J, de même ordre de multiplicité algébrique,*
- *si le réel ω est non nul et si le scalaire λ est une valeur propre non nulle de la matrice d'itération $B_{SOR}(\omega)$, alors il existe une valeur propre μ de B_J telle que*

$$(\lambda + \omega - 1)^2 = \omega^2\mu^2\lambda, \tag{3.18}$$

- *réciproquement, si μ est une valeur propre de B_J et si λ est un scalaire satisfaisant* (3.18)*, alors λ est une valeur propre de $B_{SOR}(\omega)$.*

24. Un graphe est dit *biparti* si l'ensemble de ses sommets peut être divisé en deux sous-ensembles disjoints U et V de manière à ce que chaque arête du graphe relie l'un des sommets de U à l'un de ceux de V.

3.1.7 Remarques sur la mise en œuvre des méthodes

En pratique, une méthode itérative de résolution de système linéaire ne fournit qu'une *approximation* [25] de la solution du système, puisque le nombre d'itérations qu'il est possible d'effectuer est nécessairement fini. Idéalement, il conviendrait de mettre fin aux calculs à la première itération pour laquelle l'erreur est « suffisamment petite », c'est-à-dire pour le premier entier naturel k tel que

$$\|\boldsymbol{e}^{(k)}\| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \varepsilon,$$

où ε est une tolérance fixée et $\|\cdot\|$ est une norme donnée. Cependant, on ne sait généralement pas évaluer l'erreur, puisque la solution $\boldsymbol{x}$ n'est pas connue, et il faut donc avoir recours à un autre critère d'arrêt. Deux choix naturels s'imposent alors.

Tout d'abord, le résidu $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ étant facile à calculer, on peut tester si $\|\boldsymbol{r}^{(k)}\| \leq \delta$, avec δ une tolérance fixée. Puisque l'on a

$$\|\boldsymbol{e}^{(k)}\| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = \|\boldsymbol{x}^{(k)} - A^{-1}\boldsymbol{b}\| = \|A^{-1}\boldsymbol{r}^{(k)}\| \leq \|A^{-1}\|\,\|\boldsymbol{r}^{(k)}\|,$$

on voit que doit alors choisir δ tel que $\delta \leq \frac{\varepsilon}{\|A^{-1}\|}$. Ce critère peut être trompeur si la norme de A^{-1} est grande et qu'on ne dispose pas d'une bonne estimation de cette dernière. Il est en général plus judicieux de considérer dans le test d'arrêt un résidu *normalisé*,

$$\frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{r}^{(0)}\|} \leq \delta, \text{ ou encore } \frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{b}\|} \leq \delta,$$

la seconde possibilité correspondant au choix de l'initialisation $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. Dans ce dernier cas, on obtient le contrôle suivant de l'erreur *relative*

$$\frac{\|\boldsymbol{e}^{(k)}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}\|\,\frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\,\delta,$$

où $\mathrm{cond}(A)$ est le conditionnement de la matrice A relativement à la norme subordonnée à la norme $\|\cdot\|$ considérée dans le test (voir la sous-section A.5.4 de l'annexe A).

Un autre critère parfois utilisé est basé sur l'*incrément* $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$. La suite des erreurs d'une méthode itérative de la forme (3.2) vérifiant la relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)},$$

on obtient, par utilisation de l'inégalité triangulaire,

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\| \leq \|B\|\,\|\boldsymbol{e}^{(k)}\| \leq \|B\|\left(\|\boldsymbol{e}^{(k+1)}\| + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|\right),$$

d'où

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\| \leq \frac{\|B\|}{1 - \|B\|}\,\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|.$$

Parlons maintenant de la mise en œuvre proprement dite des méthodes présentées. Supposant qu'un test d'arrêt basé sur le résidu est employé, on considère chacune des méthodes décrites dans ce chapitre au travers de la relation de récurrence (3.5) écrite sous la forme « corrective » (3.15).

Pour l'initialisation de la méthode, on choisit habituellement, sauf si l'on dispose *a priori* d'informations sur la solution, le vecteur nul, c'est-à-dire $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. Ensuite, à chaque étape de la boucle de l'algorithme, on devra réaliser les opérations suivantes :

- le calcul du résidu,
- la résolution du système linéaire ayant M pour matrice et le résidu comme second membre,
- la mise à jour de l'approximation de la solution,

25. Ceci n'est pas nécessairement rédhibitoire : l'emploi d'une méthode directe (voir le chapitre 2) conduit aussi, sauf cas particulier, à une approximation de la solution dès lors que les calculs sont faits en précision finie.

jusqu'à [26] ce que la norme du résidu soit plus petite qu'une tolérance prescrite.

Le nombre d'opérations élémentaires requises à chaque itération pour un système linéaire d'ordre n se décompose en n^2 additions et soustractions et n^2 multiplications pour le calcul du résidu, n divisions (pour la méthode de Jacobi) ou $\frac{n(n-1)}{2}$ additions et soustractions, $\frac{n(n-1)}{2}$ multiplications et n divisions (pour la méthode de Gauss–Seidel) ou n multiplications (pour la méthode de Richardson stationnaire) pour la résolution du système linéaire associé à la matrice M, n additions pour la mise à jour de la solution approchée, $n-1$ additions, n multiplications et une extraction de racine carrée pour le calcul de la norme euclidienne du résidu servant au critère d'arrêt (on peut également réaliser le test directement sur la norme du résidu au carré, ce qui évite d'extraire une racine carrée). Ce compte d'opérations, de l'ordre de $\frac{1}{2}n^2$ additions et soustractions, $\frac{3}{2}n^2$ multiplications et n divisions, montre que l'utilisation d'une méthode itérative s'avère très favorable par rapport à celle d'une des méthodes directes du chapitre 2 si le nombre d'itérations à effectuer reste petit devant n.

Terminons en répétant que chaque composante de l'approximation courante de la solution peut être calculée indépendamment des autres dans la méthode de Jacobi (ou de sur-relaxation simultanée). Cette méthode est donc facilement parallélisable. Au contraire, pour la méthode de Gauss–Seidel (ou de sur-relaxation successive), ce calcul ne peut se faire que séquentiellement, mais sans qu'on ait toutefois besoin de conserver l'approximation de la solution à l'étape précédente au cours de celui-ci, ce qui constitue un (léger) gain en termes d'espace mémoire alloué à la méthode.

3.2 Méthodes de projection

A REDIGER autre classe de méthodes itératives qui nécessitent le calcul de projections orthogonales de l'approximation courante sur les sous-espace vectoriels engendrés par différents blocs de lignes issus d'une partition de la matrice A du système à résoudre (on parle pour cette raison en anglais de *row projection methods*).

3.2.1 Méthode de Kaczmarz

La *méthode de Kaczmarz* [27] [Kac37], caractère séquentiel de la méthode
analyse de convergence [Tan71]

3.2.2 Méthode de Cimmino

La *méthode de Cimmino* [28] [Cim38] est une méthode itérative de résolution de système linéaire d'un type bien différent de celui des méthodes précédemment étudiées. Elle utilise le fait que la solution du système linéaire (2.1) est l'unique point d'intersection de n hyperplans d'équations respectives

$$\boldsymbol{a_i}^T \boldsymbol{x} = b_i,\ i = 1, \ldots, n,$$

le vecteur $\boldsymbol{a_i}^T$ désignant la i^{e} ligne de la matrice A.

Étant donné une approximation $\boldsymbol{x}^{(0)}$ de $\boldsymbol{x}$, la réflexion $\boldsymbol{x_i}^{(0)}$ de $\boldsymbol{x}^{(0)}$ par rapport au i^{e} hyperplan est donnée par

$$\boldsymbol{x_i}^{(0)} = \boldsymbol{x}^{(0)} + 2\,\frac{b_i - \boldsymbol{a_i}^T \boldsymbol{x}^{(0)}}{\boldsymbol{a_i}^T \boldsymbol{a_i}}\,\boldsymbol{a_i},\ i = 1, \ldots, n.$$

Étant alors donné n nombres strictement positifs $m_1, \ldots, m_n$, on définit le vecteur $\boldsymbol{x}^{(1)}$ comme le barycentre du système formé des points $\boldsymbol{x_i}^{(0)}$, $i = 1, \ldots, n$, respectivement pondérés par les masses m_i. En remarquant que le point $\boldsymbol{x}^{(0)}$ et ses réflexions se trouvent sur le bord d'une hypersphère centrée en $\boldsymbol{x}$, le barycentre se trouve nécessairement à l'intérieur de cette hypersphère et le vecteur $\boldsymbol{x}^{(1)}$ est donc une meilleure approximation de la solution que ne l'était $\boldsymbol{x}^{(0)}$, au sens où l'on a

$$\|\boldsymbol{x}^{(1)} - \boldsymbol{x}\|_2 < \|\boldsymbol{x}^{(0)} - \boldsymbol{x}\|_2.$$

26. En pratique, il est aussi nécessaire de limiter le nombre d'itérations, afin d'éliminer tout problème lié à l'absence de convergence d'une méthode.

27. Stefan Kaczmarz (20 mars 1895 - 1939) était un mathématicien polonais. Il inventa une méthode itérative de résolution de systèmes linéaires à la base d'une importante technique de reconstruction d'image utilisée en tomographie axiale.

28. Gianfranco Luigi Giuseppe Cimmino (12 mars 1908 - 30 mai 1989) était un mathématicien italien, connu pour ses contributions à l'étude des équations aux dérivées partielles de type elliptique et à l'analyse numérique.

On répète alors le procédé pour construire une suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ convergeant nécessairement vers $\boldsymbol{x}$.

Matriciellement, la méthode repose sur la relation de récurrence suivante :

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + A^T D(\boldsymbol{b} - A\boldsymbol{x}^{(k)}),$$

avec D une matrice diagonale ayant pour coefficients diagonaux

$$d_{ii} = \frac{2}{\sum_{j=1}^{n} m_j} \frac{m_i}{\|\boldsymbol{a_i}^T\|_2^{\,2}},\ i = 1, \ldots, n.$$

(choix courant : $m_i = 1$?)

caractère parallèle par construction de la méthode

3.3 Notes sur le chapitre

Les méthodes itératives de résolution des systèmes linéaires trouvent leur origine dans une lettre [29] de Gauss adressée en 1823 à son ancien élève Gerling [30] [Gau23]. Gauss y décrit par l'exemple une nouvelle méthode, permettant le calcul précis d'angles intervenant dans un problème de géodésie, qui l'enthousiasme au point qu'il indique à son correspondant que cette « procédure indirecte [31] peut être effectuée alors qu'on est à moitié endormi, ou en train de penser à d'autres choses » [32]. Bien que la notation matricielle n'existât pas à l'époque et que l'ordre dans lequel Gauss effectuait ses calculs fût naturellement déterminé par le choix à chaque étape de l'inconnue permettant la plus grande réduction du résidu, on y reconnaît l'algorithme de la méthode développée plus tard par Seidel [vSei74] et aujourd'hui connue, dans sa forme moderne (dans laquelle les inconnues sont traitées de manière cyclique et non dans un ordre visant à faire décroître au mieux le résidu à chaque itération), sous le nom de « méthode de Gauss–Seidel ».

Les plus anciens résultats de convergence de la méthode de Gauss–Seidel utilisée pour la résolution d'équations normales (c'est-à-dire de systèmes linéaires dont les matrices sont réelles symétriques et définies positives) apparaissent indépendamment dans les travaux de Nekrasov [Nek85] et de Pizzetti [Piz87], qui établissent tous deux des conditions en termes du rayon spectral de la matrice d'itération de la méthode. Le premier traitement systématique et rigoureux de l'étude de convergence des méthodes de Gauss–Seidel et de Jacobi semble cependant dû à von Mises et Geiringer dans [vMP29]. Enfin, une grande partie des résultats obtenus pour la méthode de sur-relaxation successive et le choix d'un paramètre optimal sont l'œuvre de Young [33]. On en trouvera une présentation complète dans le livre [You71].

Un autre résultat remarquable de comparaison de la convergence des méthodes de Jacobi et de Gauss–Seidel, démontré à l'aide de la *théorie de Perron* [34]*–Frobenius* [35], s'énonce comme suit.

Théorème 3.22 (« théorème de Stein–Rosenberg » [SR48]) *Si la matrice A est telle que la matrice d'itération B_J de la méthode de Jacobi a tous ses éléments positifs, alors seule l'une des assertions suivantes est vraie :*

(i) $\rho(B_J) = \rho(B_{GS}) = 0$,

(ii) $0 < \rho(B_{GS}) < \rho(B_J) < 1$,

(iii) $\rho(B_J) = \rho(B_{GS}) = 1$,

(iv) $1 < \rho(B_J) < \rho(B_{GS})$,

29. Le lecteur non germanophone intéressé pourra prendre connaissance de l'intégralité du contenu de cette lettre via la traduction anglaise proposée par Forsythe [For51].

30. Christian Ludwig Gerling (10 juillet 1788 - 15 janvier 1864) était un astronome, mathématicien et physicien allemand. Il est connu pour ses travaux en géodésie et pour l'importante correspondance qu'il eut avec son directeur de thèse, Carl Friedrich Gauss, sur le sujet.

31. Gauss compare ici sa nouvelle méthode avec le procédé d'élimination qui porte aujourd'hui son nom, qu'il qualifie de méthode « directe » (voir le précédent chapitre).

32. *Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken.*

33. David M. Young, Jr. (20 octobre 1923 - 21 décembre 2008) était un mathématicien et informaticien américain. Il est connu pour avoir posé en grande partie le cadre d'analyse de méthodes itératives de résolution de systèmes linéaires.

34. Oskar Perron (7 mai 1880 - 22 février 1975) était un mathématicien allemand à qui l'on doit notamment des contributions concernant les équations aux dérivées partielles.

35. Ferdinand Georg Frobenius (26 octobre 1849 - 3 août 1917) était un mathématicien allemand. Il s'intéressa principalement à la théorie des groupes et à l'algèbre linéaire, mais travailla également en analyse et en théorie des nombres.

où B_{GS} désigne la matrice d'itération de la méthode de Gauss–Seidel.

D'un point de vue pratique, les méthodes itératives présentées dans ce chapitre furent très populaires jusque dans les années 1970. Elles sont encore utilisées de nos jours, notamment pour le préconditionnement des systèmes linéaires, mais elles se sont généralement trouvées supplantées par des méthodes plus récentes, comme la *méthode du gradient conjugué* (*conjugate gradient method* en anglais) [HS52]. Cette dernière fait partie des méthodes dites *à direction de descente* [Tem39], dont le point de départ est la minimisation de la fonction

$$\forall \boldsymbol{x} \in M_{n,1}(\mathbb{C}),\ J(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^* A\boldsymbol{x} - \boldsymbol{x}^*\boldsymbol{b},$$

avec A une matrice hermitienne définie positive d'ordre n et $\boldsymbol{b}$ une matrice de $M_{n,1}(\mathbb{C})$. Dans ce cas, la fonction J atteint son minimum en $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ et la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$ équivaut bien à celle du problème de minimisation. Pour la résolution numérique de ce problème par une méthode itérative, l'idée est de se servir d'une suite minimisante de la forme

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{p}^{(k)},$$

où le vecteur $\boldsymbol{p}^{(k)}$ et le scalaire $\alpha^{(k)}$ sont respectivement la *direction* (*search direction* en anglais) et la *longueur du pas* (*step length* en anglais) de descente à l'étape k, à partir d'une initialisation $\boldsymbol{x}^{(0)}$ donnée. On remarque que le choix du résidu $\boldsymbol{r}^{(k)}$ comme direction de descente, comme proposé par Cauchy [Cau47], ainsi que d'un pas suffisamment petit et indépendant de l'itération conduit à la méthode de Richardson stationnaire introduite dans la section 3.1.4, dans ce cas appelée *méthode du gradient à pas fixe* (le résidu $\boldsymbol{r}^{(k)}$ étant l'opposé du gradient de J en $\boldsymbol{x}^{(k)}$). La *méthode du gradient à pas optimal* est obtenue en déterminant le pas de descente $\alpha^{(k)}$ à chaque étape (c'est donc une méthode instationnaire) de manière à ce que

$$\alpha^{(k)} = \underset{\alpha\in\mathbb{R}}{\arg\min}\, J(\boldsymbol{x}^{(k)} + \alpha\,\boldsymbol{p}^{(k)}).$$

Dans la méthode du gradient conjugué, la direction de descente fait intervenir le résidu à l'étape courante, mais également la direction de descente à l'étape précédente (de manière à « garder une mémoire » des itérations précédentes et d'éviter ainsi des phénomènes d'oscillations en tirant partie d'une certaine propriété d'orthogonalité) et un pas optimal est utilisé.

Cette dernière méthode est en fait une méthode directe employée comme une méthode itérative, puisque l'on peut montrer qu'elle converge en au plus n itérations. C'est une *méthode de Krylov* [36], une propriété fondamentale étant que le vecteur $\boldsymbol{x}^{(k)}$ minimise la fonction J sur l'espace affine $\boldsymbol{x}^{(0)}+\mathscr{K}_k$, avec $\mathscr{K}_k = \mathrm{Vect}\{\boldsymbol{r}^{(0)}, A\boldsymbol{r}^{(0)}, \ldots, A^{k-1}\boldsymbol{r}^{(0)}\}$ est le *sous-espace de Krylov d'ordre k* généré par la matrice A et le vecteur $\boldsymbol{r}^{(0)}$.

Si la matrice A n'est pas hermitienne définie positive, on ne peut plus appliquer la méthode du gradient conjugué car la matrice n'étant plus celle d'un produit scalaire (hermitien), ce point intervenant de manière critique dans les propriétés de la fonction J. Cependant, le cadre des sous-espaces de Krylov reste propice à la construction de méthodes itératives consistant à minimiser la norme euclidienne du résidu. Parmi les nombreuses méthodes existantes, on peut citer la *méthode du gradient biconjugué* (*biconjugate gradient (BiCG) method* en anglais), introduite par Lanczos [37] [Lan52] et popularisée par Fletcher [Fle76], et ses variantes comme la *méthode du gradient conjugué au carré* (*conjugate gradient squared (CGS) method* en anglais) [Son89] ou la *méthode du gradient biconjugué stabilisée* (*biconjugate gradient stabilized (BiCGSTAB) method* en anglais) [vdVor92], la *méthode orthomin* [Vin76], la *méthode d'orthogonalisation complète* (*full orthogonalization method (FOM)* en anglais) [Saa81] ou encore la *méthode du résidu minimal* (*minimal residual (MINRES) method* en anglais) [PS75] et sa généralisation (*generalized minimal residual (GMRES) method* en anglais) [SS86].

On trouvera de nombreuses références sur les méthodes itératives de résolution de systèmes linéaires, ainsi qu'une perspective historique sur leur important développement au cours du vingtième siècle avec l'essor des calculateurs automatiques, dans l'article [SvdV00].

36. Alexei Nikolaevich Krylov (Алексе́й Никола́евич Крыло́в en russe, 15 août 1863 - 26 octobre 1945) était un ingénieur naval, mathématicien et mémorialiste russe. Il est célèbre pour ses travaux en mathématiques appliquées, et plus particulièrement un article consacré aux problèmes aux valeurs propres paru en 1931, dans lequel il introduisit ce que l'on appelle aujourd'hui les *sous-espaces de Krylov*.

37. Cornelius Lanczos (Lánczos Kornél en hongrois, né Lőwy Kornél, 2 février 1893 - 25 juin 1974) était un mathématicien et physicien hongrois. Il développa plusieurs méthodes numériques, pour la recherche de valeurs propres, la résolution de systèmes linéaires, l'approximation de la fonction gamma ou encore le ré-échantillonage de signaux numériques.

Références

[Ait50] A. C. AITKEN. Studies in practical mathematics. V. On the iterative solution of a system of linear equations. *Proc. Roy. Soc. Edinburgh Sect. A*, 63(1):52–60, 1950. DOI: 10.1017/S008045410000697X.

[Axe94] O. AXELSSON. *Iterative solution methods*. Cambridge University Press, 1994. DOI: 10 . 1017 / CBO9780511624100.

[Cau47] A. CAUCHY. Méthode générale pour la résolution des systèmes d'équations simultanées. *C. R. Acad. Sci. Paris*, 25 :536-538, 1847.

[Cim38] G. CIMMINO. Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari. *La Ricerca Scientifica*, 9:326–333, 1938.

[Col42] L. COLLATZ. Fehlerabschätzung für das Iterationsverfahren zur Auflösung linearer Gleichungssysteme. *Z. Angew. Math. Mech.*, 22(6):357–361, 1942. DOI: 10.1002/zamm.19420220607.

[DER86] I. DUFF, A. ERISMAN, and J. REID. *Direct methods for sparse matrices*. Oxford University Press, 1986.

[Fle76] R. FLETCHER. Conjugate gradient methods for indefinite systems. In G. A. WATSON, editor, *Numerical analysis - proceedings of the Dundee conference on numerical analysis, 1975*. Volume 506, Lecture Notes in Mathematics, pages 73–89. Springer, 1976. DOI: 10.1007/BFb0080116.

[For51] G. E. FORSYTHE. Gauss to Gerling on relaxation. *Math. Tables Aids Comput.*, 5(36):255–258, 1951. DOI: 10.1090/S0025-5718-51-99414-8.

[For53] G. E. FORSYTHE. Solving linear algebraic equations can be interesting. *Bull. Amer. Math. Soc.*, 59(4):299–329, 1953. DOI: 10.1090/S0002-9904-1953-09718-X.

[Fra50] S. P. FRANKEL. Convergence rates of iterative treatments of partial differential equations. *Math. Tables Aids Comput.*, 4(30):65–75, 1950. DOI: 10.1090/S0025-5718-1950-0046149-3.

[Gau23] C. F. GAUSS. Lettre du 26 décembre adressée à Christian Ludwig Gerling, 1823.

[Hou58] A. S. HOUSEHOLDER. The approximate solution of matrix problems. *J. Assoc. Comput. Mach.*, 5(3):205–243, 1958. DOI: 10.1145/320932.320933.

[HS52] M. R. HESTENES and E. STIEFEL. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49(6):409–436, 1952. DOI: 10.6028/jres.049.044.

[Jac45] C. G. J. JACOBI. Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden lineären Gleichungen. *Astronom. Nachr.*, 22(20):297–306, 1845. DOI: 10.1002/asna.18450222002.

[Joh66] F. JOHN. *Lectures on advanced numerical analysis*. Gordon and Breach, 1966.

[Kac37] S. KACZMARZ. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bull. Int. Acad. Sci. Cracovie, Cl. Sci. Math. Nat., Sér. A Sci. Math.*, 35:355–357, 1937.

[Kah58] W. KAHAN. *Gauss-Seidel methods of solving large systems of linear equations*. PhD thesis, university of Toronto, 1958.

[Lan52] C. LANCZOS. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952. DOI: 10.6028/jres.049.006.

[Lie18] H. LIEBMANN. Die angenäherte Ermittelung harmonischer Funktionen und konformer Abbildungen (nach Ideen von Boltzmann und Jacobi). *Bayer. Akad. Wiss. Math.-Phys. Kl. Sitzungsber.*:385–416, 1918.

[Nek85] P. A. NEKRASOV. Détermination des inconnues par la méthode de moindres carrés dans le cas où le nombre d'inconnues est considérable (russe). *Rec. Math. [Mat. Sbornik] N.S.*, 12(1) :189-204, 1885.

[Nie64] W. NIETHAMMER. Relaxation bei komplexen Matrizen. *Math. Z.*, 86(1):34–40, 1964. DOI: 10.1007/BF01111275.

[Ost54] A. M. OSTROWSKI. On the linear iteration procedures for symmetric matrices. *Rend. Mat. Appl. (5)*, 14:140–163, 1954.

[Piz87] P. PIZZETTI. Sulla compensazione delle osservazioni secondo il metodo dei minimi quadrati, nota 1, nota 2. *Atti Accad. Lincei Rend. (4)*, 3(2):230–235, 288–293, 1887.

[PR55] D. W. PEACEMAN and H. H. RACHFORD, JR. The numerical solution of parabolic and elliptic differential equations. *SIAM J.*, 3(1):28–41, 1955. DOI: 10.1137/0103003.

[PS75] C. C. PAIGE and M. A. SAUNDERS. Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975. DOI: 10.1137/0712047.

[Rei49] E. REICH. On the convergence of the classical iterative method of solving linear simultaneous equations. *Ann. Math. Statist.*, 20(3):448–451, 1949. DOI: 10.1214/aoms/1177729998.

[Ric11] L. F. RICHARDSON. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philos. Trans. Roy. Soc. London Ser. A*, 210(459-470):307–357, 1911. DOI: 10.1098/rsta.1911.0009.

[Saa81] Y. SAAD. Krylov subspace methods for solving large unsymmetric linear systems. *Math. Comput.*, 37(155):105–126, 1981. DOI: 10.1090/S0025-5718-1981-0616364-6.

[She55] J. W. SHELDON. On the numerical solution of elliptic difference equations. *Math. Tables Aids Comput.*, 9(51):101–112, 1955. DOI: 10.1090/S0025-5718-1955-0074929-1.

[Son89] P. SONNEVELD. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 10(1):36–52, 1989. DOI: 10.1137/0910004.

[SR48] P. STEIN and R. L. ROSENBERG. On the solution of linear simultaneous equations by iteration. *J. London Math. Soc. (1)*, 23(2):111–118, 1948. DOI: 10.1112/jlms/s1-23.2.111.

[SS86] Y. SAAD and M. H. SCHULTZ. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986. DOI: 10.1137/0907058.

[SvdV00] Y. SAAD and H. A. van der VORST. Iterative solution of linear systems in the 20th century. *J. Comput. Appl. Math.*, 123(1-2):1–33, 2000. DOI: 10.1016/S0377-0427(00)00412-X.

[Tan71] K. TANABE. Projection method for solving a singular system of linear equations and its applications. *Numer. Math.*, 17(3):203–214, 1971. DOI: 10.1007/BF01436376.

[Tem39] G. TEMPLE. The general theory of relaxation methods applied to linear systems. *Proc. Roy. Soc. London Ser. A*, 169(939):476–500, 1939. DOI: 10.1098/rspa.1939.0012.

[vdVor92] H. A. van der VORST. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 13(2):631–644, 1992. DOI: 10.1137/0913035.

[Vin76] P. K. W. VINSOME. Orthomin, an iterative method for solving sparse sets of simultaneous linear equations. In *Proceedings of the fourth symposium on numerical simulation of reservoir performance*, pages 49–59. Society of Petroleum Engineers of AIME, 1976. DOI: 10.2118/5729-MS.

[vMP29] R. von MISES und H. POLLACZEK-GEIRINGER. Praktische Verfahren der Gleichungsauflösung. *Z. Angew. Math. Mech.*, 9(1):58–77, 1929. DOI: 10.1002/zamm.19290090105.

[vSei74] P. L. von SEIDEL. Über ein Verfahren die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineare Gleichungen überhaupt, durch successive Annäherung aufzulösen. *Abh. Kgl. Bayer Akad. Wiss. Math. Phys. Kl.*, 11(3):81–108, 1874.

[Wit36] H. WITTMEYER. Über die Lösung von linearen Gleichungssystemen durch Iteration. *Z. Angew. Math. Mech.*, 16(5):301–310, 1936. DOI: 10.1002/zamm.19360160505.

[You54] D. YOUNG. Iterative methods for solving partial difference equations of elliptic type. *Trans. Amer. Math. Soc.*, 76(1):92–111, 1954. DOI: 10.1090/S0002-9947-1954-0059635-7.

[You71] D. M. YOUNG. *Iterative solution of large linear systems*. Academic Press, 1971.

[You72] D. M. YOUNG. On the consistency of linear stationary iterative methods. *SIAM J. Numer. Anal.*, 9(1):89–96, 1972. DOI: 10.1137/0709010.

Chapitre 4

Calcul de valeurs et de vecteurs propres

Nous abordons dans ce chapitre le problème du calcul de valeurs propres et, éventuellement, de vecteurs propres d'une matrice d'ordre n diagonalisable. C'est un problème beaucoup plus difficile que celui de la résolution d'un système linéaire. En effet, les valeurs propres d'une matrice étant les racines de son polynôme caractéristique [1], on pourrait naïvement penser qu'il suffit de factoriser ce dernier pour les obtenir. On sait cependant (par le théorème d'Abel [2]–Ruffini [3]) qu'il n'est pas toujours possible d'exprimer les racines d'un polynôme de degré supérieur ou égal à cinq à partir des coefficients du polynôme et d'opérations élémentaires (addition, soustraction, multiplication, division et extraction de racines). Par conséquent, il ne peut exister de méthode directe, c'est-à-dire fournissant le résultat en un nombre fini d'opérations, de calcul de valeurs propres d'une matrice et on a recours à des méthodes itératives [4].

Parmi ces méthodes, il convient distinguer celles qui permettent le calcul d'*une* valeur propre (en général celle de plus grand ou de plus petit module, mais pas seulement) de celles qui conduisent à une approximation de l'*ensemble* du spectre d'une matrice. D'autre part, certaines méthodes permettent le calcul de vecteurs propres associés aux valeurs propres obtenues, alors que d'autres non. C'est le cas par exemple de la *méthode de la puissance*, qui fournit une approximation d'un couple particulier de valeur et vecteur propres. Nous abordons enfin les méthodes réservées au cas des matrices réelles symétriques (ou hermitiennes) et discutons dans ce cadre le calcul de la décomposition en valeurs singulières.

1. Réciproquement, on peut vérifier que les racines de tout polynôme $p_n(x) = \sum_{i=0}^{n} a_i\, x^i$ de degré n, avec a_n non nul, sont les valeurs propres de la *matrice compagnon* (*companion matrix* en anglais) de p_n,

$$C(p_n) = \begin{pmatrix} 0 & \dots & \dots & 0 & -\frac{a_0}{a_n} \\ 1 & \ddots & & \vdots & -\frac{a_1}{a_n} \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -\frac{a_{n-2}}{a_n} \\ 0 & \dots & 0 & 1 & -\frac{a_{n-1}}{a_n} \end{pmatrix}.$$

2. Niels Henrik Abel (5 août 1802 - 6 avril 1829) était un mathématicien norvégien. Il est connu pour ses travaux en analyse, notamment sur la semi-convergence des séries numériques, des suites et séries de fonctions, les critères de convergence des intégrales généralisées et sur les intégrales et fonctions elliptiques, et en algèbre, sur la résolution des équations algébriques par radicaux.

3. Paolo Ruffini (22 septembre 1765 - 10 mai 1822) était un médecin et mathématicien italien. Son nom est associé à la démonstration partielle de l'irrésolubilité algébrique des équations de degré strictement supérieur à quatre, à la théorie des groupes et à une règle de division rapide des polynômes.

4. On peut encore indiquer, pour bien comprendre l'intérêt des techniques introduites dans ce chapitre, que la détermination des valeurs propres d'une matrice par la recherche des racines de son polynôme caractéristique par les méthodes présentées dans la section 5.7 peut s'avérer catastrophique en raison de problèmes de stabilité numérique lorsque les calculs sont menés dans une arithmétique en précision finie (voir [Wil59]) et est en général à éviter. S'inspirant de l'exemple du polynôme de Wilkinson (voir la sous-section 1.4.2 du chapitre 1), on peut considérer le calcul du spectre de la matrice diagonale d'ordre 25 dont les coefficients diagonaux (et donc les valeurs propres) sont $1, 2, \dots, 25$. Le calcul des racines de son polynôme caractéristique par la version 3.8.1 de GNU Octave, au moyen de la commande `roots(poly(diag(1:25)))`, donne (en arrondissant les résultats à la cinquième décimale) : 1,00000, 2,00000, 3,00000, 4,00000, 5,00000, 6,00008, 6,99878, 8,01353, 8,91076, 10,19752 − 0,56315i, 10,19752 + 0,56315i, 11,93850 − 1,41358i, 11,93850 + 1,41358i, 13,91694 − 2,15226i, 13,91694 + 2,15226i, 16,14125 − 2,68680i, 16,14125 + 2,68680i, 18,58099 − 2,87523i, 18,58099 + 2,87523i, 21,07846 − 2,54279i, 21,07846 + 2,54279i, 23,30187 − 1,63462i, 23,30187 + 1,63462i, 24,88289 − 0,39812i et 24,88289 + 0,39812i. On constate que seize des approximations obtenues ont une partie imaginaire non négligeable, alors que les valeurs propres recherchées sont toutes réelles !

4.1 Exemples d'application **

ECRIRE UNE INTRODUCTION

4.1.1 Détermination des modes propres de vibration d'une plaque *

MIEUX : REMPLACER le cas traité (corde) par celui d'une plaque (voir [GK12])

On considère une corde homogène sans raideur, de section constante et de longueur ℓ, tendue entre deux extrémités fixes placées le long d'un axe, l'une à l'origine et l'autre au point d'abscisse ℓ. On s'intéresse aux petits mouvements transversaux de cette corde dans le plan vertical, c'est-à-dire que l'on cherche une fonction $u(t,x)$, $0 \leq x \leq \ell$, $t \geq 0$, représentant à l'instant t la déformation verticale de la corde au point d'abscisse x. On montre par des considérations physiques que la fonction u doit satisfaire, en l'absence de force extérieure, l'équation aux dérivées partielles, appelée *équation des ondes* en une dimension d'espace, et les conditions aux limites homogènes, indiquant que la corde est attachée en ses extrémités, suivantes

$$\frac{\partial^2 u}{\partial t^2}(t,x) - c^2 \frac{\partial^2 u}{\partial x^2}(t,x) = 0,\ 0 < x < \ell,\ t > 0, \tag{4.1}$$

$$u(t,0) = u(t,\ell) = 0,\ t \geq 0, \tag{4.2}$$

avec $c = \sqrt{\frac{\tau}{\mu}}$, où τ et μ sont respectivement la tension et la masse linéique de la corde, que l'on complète par des *conditions initiales* donnant la déformation $u(0,x)$ et la vitesse de déformation $\frac{\partial u}{\partial t}(0,x)$, $0 \leq x \leq \ell$ de la corde à l'instant « initial » $t = 0$.

Les *modes propres de vibration* de la corde sont des fonctions non triviales vérifiant (4.1)-(4.2), périodiques en temps de la forme

$$u(t,x) = v(x)\, e^{\mathrm{i}\omega t},$$

où ω désigne la pulsation du mode de vibration considéré. Un calcul simple montre alors que la recherche de ces modes conduit à la détermination de réels λ et de fonctions v non identiquement nulles solutions du problème aux valeurs propres suivant

$$-v''(x) = \lambda\, v(x),\ 0 < x < \ell, \tag{4.3}$$

$$v(0) = v(\ell) = 0, \tag{4.4}$$

les pulsations des modes étant alors données par

$$\omega = c\sqrt{\lambda}.$$

Dans le cas simple d'une corde homogène, on établit facilement que le problème (4.3)-(4.4) a pour seules solutions

$$\lambda_k = \frac{k^2\pi^2}{\ell^2}, v_k = c_k \sin\left(\frac{k\pi}{\ell}\, x\right),\ k \in \mathbb{N}\backslash\{0\},$$

où c_k désigne une constante arbitraire non nulle, mais si la masse linéique varie avec l'abscisse x ou si l'on a affaire aux modes de vibration d'une membrane de forme quelconque, c'est-à-dire un problème posé en dimension deux d'espace, il n'y a plus, en général, de forme explicite pour les solutions du problème aux limites correspondant. Il faut alors calculer numériquement ces solutions, ce que l'on peut faire en les approchant par la méthode des différences finies déjà employée dans la sous-section 2.1.2 du précédent chapitre. Pour le problème (4.3)-(4.4), ceci revient, en posant

$$h = \frac{\ell}{n+1},$$

avec n un entier supérieur ou égal à 1, la discrétisation du problème conduit au système matriciel

$$B_h \boldsymbol{v}_h = \lambda_h \boldsymbol{v}_h, \tag{4.5}$$

avec

$$B_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in M_n(\mathbb{R}),$$

et où $\boldsymbol{v}_h$ désigne le vecteur de $M_{n,1}(\mathbb{R})$ ayant pour composante les valeurs approchées v_i, $1 \leq i \leq n$, de la fonction v aux nœuds du maillage x_i, $1 \leq i \leq n$. En d'autres termes, on approche un couple (λ, v) solution de (4.3)-(4.4) par un couple de valeur et vecteur propres $(\lambda_h, \boldsymbol{v}_h)$ solution de (4.5). Ce dernier problème admet n couples de solutions, la matrice tridiagonale B_h étant réelle symétrique. Évidemment, on est ici en mesure calculer de manière exacte les valeurs et vecteurs propres de la matrice B_h, mais on devra, en général, recourir à des méthodes de calcul approché pour les obtenir.

4.1.2 Évaluation numérique des nœuds et poids des formules de quadrature de Gauss **

Le calcul numérique effectif des nœuds et poids des *formules de quadrature de Gauss* peut se faire avec une complexité de l'ordre de $O(n^2)$ opérations, en résolvant un problème aux valeurs propres faisant intervenir une matrice tridiagonale.

- EXPLIQUER RAPIDEMENT LA PROBLEMATIQUE en renvoyant la section 7.4

- introduire la relation de récurrence à trois termes pour les polynômes orthogonaux (par rapport au produit scalaire pour la mesure de Lebesgue)

- matrice de Jacobi

- théorème de caractérisation

- algorithme de Golub et Welsch [GW69]

REPRENDRE : On introduit pour cela la matrice symétrique, définie positive et d'ordre infini suivante, appelée *matrice de Jacobi*,

$$J_\infty = \begin{pmatrix} \alpha_0 & \sqrt{\beta_1} & & & \\ \sqrt{\beta_1} & \alpha_1 & \sqrt{\beta_2} & & \\ & \sqrt{\beta_2} & \alpha_2 & \sqrt{\beta_3} & \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

dans laquelle les coefficients α_k et β_k, $k \geq 0$, sont ceux donnés par des formules de récurrence à trois termes (A DONNER). On a le résultat suivant pour la matrice J_n, $n \geq 1$, associée au mineur principal d'ordre n de J_∞. (voir [Gau96], theorem 3.1, p. 153)

Théorème 4.1 *Soit λ_k, $k = 1, \ldots, n$, les valeurs propres de la matrices J_n et $\{\boldsymbol{u}_k\}_{k=1,\ldots,n}$ un ensemble de vecteurs propres normalisés associés, c'est-à-dire que*

$$J_n \boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k, \ {\boldsymbol{u}_k}^T \boldsymbol{u}_k = 1, \ 1 \leq k \leq n.$$

Alors les nœuds et poids de la formule de Gauss à n points sont donnés par

$$x_k = \lambda_k \text{ et } \omega_k = \beta_0 (\boldsymbol{u}_k)_1{}^2, \ 1 \leq k \leq n,$$

où $(\boldsymbol{u}_k)_1$ désigne la première composante du vecteur $\boldsymbol{u}_k$ et $\beta_0 = 2$.

4.1.3 PageRank

Lorsqu'un internaute formule une requête sur le site d'un moteur de recherche, ce dernier interroge la base de données dont il dispose pour y répondre. Cette base est construite par une indexation des ressources collectées par des robots qui explorent systématiquement la « Toile » (*World Wide Web* en anglais) en suivant récursivement tous les hyperliens qu'ils trouvent. Le traitement de la requête consiste alors en l'application d'un algorithme identifiant (via l'index) les documents qui correspondent le mieux aux mots apparaissant dans la requête, afin de présenter les résultats de recherche par ordre de pertinence supposée.

De nombreux procédés existent pour améliorer la qualité des réponses fournies par un moteur de recherche. Le plus connu d'entre eux est certainement la technique *PageRank*™ [BP98] employée par Google, qui est une méthode de calcul d'un *indice de notoriété*, associé à chaque page de la « Toile » et servant à affiner le classement préalablement obtenu. Celle-ci est basée sur l'idée simple que plus une page d'un ensemble est la cible d'hyperliens, plus elle est susceptible de posséder un contenu pertinent. Le but de cette sous-section est de donner les grandes

lignes de la modélisation et de la théorie mathématiques sur lesquelles s'appuie PageRank [5] et leur lien avec une des méthodes de calcul numérique de valeurs et vecteurs propres présentées dans ce chapitre, le lecteur intéressé pouvant trouver de très nombreux détails et références bibliographiques sur le sujet dans l'article [LM04].

On peut considérer la « Toile » comme un ensemble de n pages (avec n un entier naturel aujourd'hui extrêmement grand) reliées entre elles par des hyperliens, représentable par un graphe orienté dont les nœuds symbolisent les pages et les arcs les hyperliens. La figure 4.1 présente un minuscule échantillon d'un tel graphe.

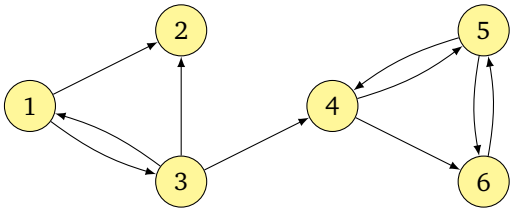


FIGURE 4.1 – Graphe orienté représentant un échantillon de « toile » constitué de six pages.

À ce graphe orienté, on associe une matrice d'adjacence, définie comme la matrice C d'ordre n dont le coefficient c_{ij}, $1 \leq i, j \leq n$, est égal à 1 s'il existe un hyperlien sur la i^{e} page pointant vers la j^{e} page et vaut 0 sinon (on ignore les hyperliens « internes », c'est-à-dire ceux pointant d'une page vers elle-même, de sorte que l'on a $c_{ii} = 0$, $i = 1, \ldots, n$). Pour l'exemple de graphe de la figure 4.1, la matrice d'adjacence est

$$C = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Pour toute page qui en contient, on suppose que chacun des hyperliens possède la même [6] chance d'être suivi. En pondérant les coefficients de la matrice d'adjacence du graphe orienté pour tenir compte de ce fait, on obtient une matrice $\tilde{Q}$ d'ordre n qui, pour l'exemple introduit plus haut est égale à

$$\tilde{Q} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Cette modification de la matrice d'adjacence fait apparaître des similitudes avec la matrice de transition d'une *chaîne* [7] *de Markov* [8] à temps discret et n états, qui est une *matrice stochastique* [9]. Cependant, la structure de la matrice construite à partir des seuls liens réellement existant laisse entrevoir un obstacle à une telle identification. En effet, lorsqu'une page ne possède aucun hyperlien (en anglais, on nomme *dangling node* le nœud du graphe correspondant), la ligne qui lui est associée dans la matrice $\tilde{Q}$ est nulle et cette dernière ne peut dans ce cas être

5. On notera que, en plus d'être formé des mots anglais *page*, qui signifie page, et *rank*, qui signifie rang, ce nom est aussi lié à celui de l'un des inventeurs de cette méthode, Larry Page.

6. On peut en fait utiliser n'importe quelle autre distribution de probabilité que la distribution uniforme, obtenue par exemple en se basant sur des statistiques disponibles pour les pages concernées.

7. Une chaîne de Markov est un processus stochastique vérifiant une propriété caractérisant une « absence de mémoire ». De manière simplifiée, la meilleure prévision que l'on puisse faire, connaissant le passé et le présent, du futur d'un tel processus est identique à la meilleure prévision qu'on puisse faire de ce futur connaissant uniquement le présent.

8. Andrei Andreevich Markov (Андре́й Андре́евич Ма́рков en russe, 2 juin 1856 - 20 juillet 1922) était un mathématicien russe. Il est connu pour ses travaux sur les processus stochastiques en temps discret.

9. Une matrice stochastique est une matrice carrée dont tous les éléments sont des réels positifs compris entre 0 et 1 et dont la somme des éléments d'une même ligne (ou d'une même colonne) est égale à 1. Le spectre d'une matrice stochastique est contenu dans le disque unité.

stochastique. Un remède [10] est de remplacer toutes les lignes en question par $\frac{1}{n}\boldsymbol{e}^\top$, avec $\boldsymbol{e}$ la matrice colonne de $M_{n,1}(\mathbb{R})$ dont les coefficients sont tous égaux à 1, conduisant à la matrice (stochastique) Q. Pour l'exemple de la figure 4.1, on obtient ainsi

$$Q = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

On définit alors l'indice de notoriété d'une page comme la fraction de temps passé sur cette page lorsque le temps tend vers l'infini, correspondant encore à la probabilité fournie par la *distribution stationnaire* (ou *invariante*) de la chaîne de Markov considérée, donnée par un vecteur propre *à gauche*, dont les composantes sont positives et de somme égale à un, associé à la valeur propre (dominante) égale à 1 de la matrice de transition de cette chaîne. Si le *théorème de Perron–Frobenius* [Per07, Fro12] assure qu'un tel vecteur existe toujours pour une matrice stochastique, il faut pour garantir son unicité (ce qui revient à demander que la valeur propre 1 soit d'ordre de multiplicité algébrique simple) que la chaîne de Markov en question soit *irréductible* [11]. Pour que ce soit le cas, on suppose qu'un internaute peut passer de la page sur laquelle il se trouve à

- l'une des pages vers lesquelles mènent les hyperliens qu'elle contient avec une probabilité de valeur $\alpha > 0$, que l'on appelle le *facteur d'amortissement* (*damping factor* en anglais),
- toute page de la « Toile » avec une probabilité uniforme, valant $\frac{1-\alpha}{n}$.

Cette dernière hypothèse conduit à considérer [12] la matrice stochastique suivante

$$A = \alpha Q + \frac{1-\alpha}{n}\boldsymbol{e}\boldsymbol{e}^\top,$$

pour laquelle il existe un unique vecteur $\boldsymbol{\pi}$ de $M_{n,1}(\mathbb{R})$ tel que

$$\boldsymbol{\pi}^\top A = \boldsymbol{\pi}^\top,\ \forall i \in \{1,\dots,n\},\ \pi_i \geq 0,\ \text{et}\ \sum_{i=1}^{n} \pi_i = 1.$$

L'indice de notoriété de la i^{e} page est ainsi donné par la valeur de la composante π_i. Le vecteur $\boldsymbol{\pi}$ étant également un vecteur propre (à droite) associé à la valeur propre dominante de la matrice $A^\top$, Google a recours à la méthode de la puissance (voir la section 4.4), qui ne nécessite que d'effectuer des produits entre la matrice $A^\top$ et des vecteurs donnés. Cependant, la taille de la « Toile » est si gigantesque qu'il est impossible de stocker la matrice A^T ou d'effectuer de manière conventionnelle son produit avec un vecteur. Pour rendre la méthode applicable, il faut observer que la matrice Q est très creuse [13] et tirer parti de cette structure particulière en ne stockant que ses éléments non nuls, tout en adaptant les algorithmes de multiplication matricielle. Même en réalisant ces optimisations, la puissance de calcul requise reste considérable et nécessite des serveurs informatiques adaptés à cette tâche, que Google effectuerait apparemment chaque mois.

Parlons pour finir de la valeur du facteur d'amortissement α. On peut montrer (voir [LM04]) que la valeur propre sous-dominante de la matrice A est égale à α si la matrice Q n'est pas la matrice de transition d'une chaîne de Markov irréductible, strictement inférieure à α si c'est le cas [14]. La vitesse de convergence de la méthode de la puissance se trouve par conséquent directement affectée par le choix de cette valeur. Il y a alors un compromis délicat à trouver entre une convergence rapide de la méthode (c'est-à-dire α choisi proche de 0) et un modèle rendant assez fidèlement compte de la structure de la « Toile » et du comportement des internautes (c'est-à-dire α choisi proche de 1). La valeur utilisée par les fondateurs de Google est $\alpha = 0,85$.

10. Là encore, on peut utiliser n'importe quelle distribution de probabilité autre que la distribution uniforme. Une des premières modifications suggérées du modèle de base a été l'utilisation d'un vecteur « personnalisé » $\boldsymbol{v}$, tel que $v_i \geq 0$, $\forall i \in \{1,\dots,n\}$, et $\sum_{i=1}^{n} v_i = 1$, différenciant les internautes en fonction de leurs habitudes ou de leurs goûts, pour éventuellement proposer des classements adaptés à l'utilisateur qui formule la requête.
11. Une chaîne de Markov est dite irréductible si tout état est accessible à partir de n'importe quel autre état.
12. Dans le cas, évoqué plus haut, d'une personnalisation de la recherche par l'utilisation d'un vecteur $\boldsymbol{v}$, on aura $A = \alpha Q + (1-\alpha)\boldsymbol{e}\boldsymbol{v}^T$.
13. Typiquement, on a seulement de trois à dix termes non nuls sur chacune des lignes de cette matrice.
14. Ce résultat reste vrai en cas d'utilisation d'un vecteur de personnalisation $\boldsymbol{v}$.

4.2 Localisation des valeurs propres

Certaines méthodes permettant d'approcher une valeur propre spécifique d'une matrice, il peut être utile d'avoir une idée de la localisation du spectre dans le plan complexe. Dans ce domaine, une première estimation est donnée par le théorème A.157, dont on déduit que, pour toute matrice carrée A et pour toute norme matricielle $\|\cdot\|$, on a

$$\forall \lambda \in \sigma(A),\ |\lambda| \leq \|A\|.$$

Cette inégalité, bien que souvent trop grossière, montre que toutes les valeurs propres de A sont contenues dans un disque centré en l'origine et de rayon $\|A\|$. Une autre estimation de localisation *a priori* des valeurs propres, plus précise mais néanmoins très simple, est fournie par le théorème 4.3.

Définition 4.2 (« disques de Gershgorin [15] ») *Soit n un entier naturel supérieur ou égal à 2 et A une matrice d'ordre n. Les **disques de Gershgorin** D_i, $i = 1, \ldots, n$, de la matrice A sont les régions du plan complexe définies par*

$$D_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq R_i\},\ \text{avec } R_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|a_{ij}\right|. \tag{4.6}$$

Théorème 4.3 (« théorème des disques de Gershgorin » [Ger31]) *Soit n un entier naturel supérieur ou égal à 2 et A une matrice d'ordre n. On a*

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} D_i,$$

où les D_i sont les disques de Gershgorin de A.

DÉMONSTRATION. Supposons que le nombre complexe λ soit une valeur propre de A. Il existe alors un vecteur non nul $\boldsymbol{v}$ tel que $A\boldsymbol{v} = \lambda \boldsymbol{v}$, c'est-à-dire

$$\forall i \in \{1, \ldots, n\},\ \sum_{j=1}^{n} a_{ij} v_j = \lambda\, v_i.$$

Soit v_k la composante de $\boldsymbol{v}$ ayant le plus grand module (ou l'une des composantes de plus grand module s'il y en a plusieurs). On a d'une part $v_k \neq 0$, puisque $\boldsymbol{v}$ est non nul par hypothèse, et d'autre part

$$|\lambda - a_{kk}|\,|v_k| = |\lambda v_k - a_{kk} v_k| = \left|\sum_{j=1}^{n} a_{kj} v_j - a_{kk} v_k\right| = \left|\sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} v_j\right| \leq |v_k|\, R_k,$$

ce qui prouve, après division par $|v_k|$, que la valeur propre λ est contenue dans le disque de Gershgorin D_k de A. □

Ce théorème assure que toute valeur propre de la matrice A se trouve dans la réunion des disques de Gershgorin de A (voir la figure 4.2). La transposée $A^{\top}$ de A possédant le même spectre que A, on obtient de manière immédiate une première amélioration du résultat.

Proposition 4.4 *Soit n un entier naturel supérieur ou égal à 2 et A une matrice d'ordre n. On a*

$$\sigma(A) \subseteq \left(\bigcup_{i=1}^{n} D_i\right) \cap \left(\bigcup_{j=1}^{n} D'_j\right),$$

où les ensembles D'_j, $j = 1, \ldots, n$, sont tels que

$$D'_j = \left\{z \in \mathbb{C} \mid \left|z - a_{jj}\right| \leq C_j\right\},\ \text{avec } C_j = \sum_{\substack{i=1 \\ i \neq j}}^{n} \left|a_{ij}\right|,$$

et les ensembles D_i sont définis par (4.6)

15. Semyon Aranovich Gershgorin (Семён Аранович Гершгорин en russe, 24 août 1901 - 30 mai 1933) était un mathématicien russe, dont les travaux concernèrent l'algèbre, la théorie des fonctions d'une variable complexe et les méthodes numériques pour la résolution d'équations différentielles.

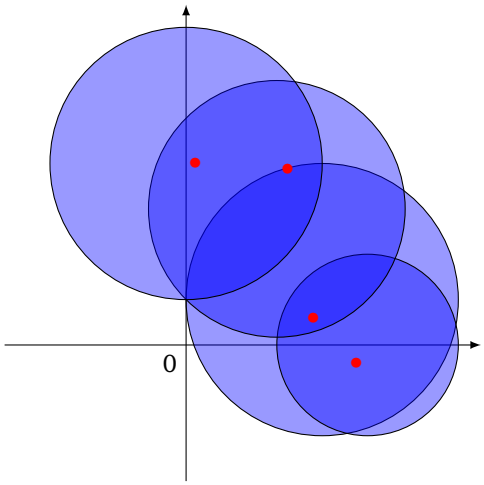


FIGURE 4.2 – Représentation dans le plan complexe des valeurs propres (en rouge) et des disques de Gershgorin (en bleu) de la matrice complexe $A = \begin{pmatrix} 3+\mathrm{i} & -\frac{3}{2} & 0 & \frac{3\mathrm{i}}{2} \\ \frac{1}{2} & 4 & 3+\mathrm{i} & \frac{\mathrm{i}}{2} \\ \sqrt{2} & \sqrt{2}\mathrm{i} & 2+3\mathrm{i} & 0 \\ \mathrm{i} & 1 & \mathrm{i} & 4\mathrm{i} \end{pmatrix}$.

La version suivante du théorème permet d'être encore plus précis sur la localisation des valeurs propres quand la réunion des disques de Gershgorin d'une matrice possède des parties connexes.

Théorème 4.5 (« second théorème de Gershgorin ») *Soit n un entier naturel supérieur ou égal à 2 et A une matrice d'ordre n. On suppose qu'il existe un entier p compris entre 1 et $n-1$ tel que l'on puisse diviser la réunion des disques de Gershgorin de A en deux sous-ensembles disjoints de p et $n-p$ disques. Alors, le premier sous-ensemble contient exactement p valeurs propres de A, chacune étant comptée avec sa multiplicité algébrique, les valeurs propres restantes étant dans le second sous-ensemble.*

DÉMONSTRATION. La preuve est basée sur un argument d'*homotopie*, une notion de topologie algébrique formalisant la notion de déformation continue d'un objet à un autre. On commence par noter $D^{(p)}$ l'union des p disques de l'énoncé, $D^{(q)}$ celle des $q = n-p$ disques restants. Pour toute valeur du paramètre réel ε choisie dans l'intervalle $[0,1]$, on définit ensuite la matrice $B(\varepsilon) = (b_{ij}(\varepsilon))_{1\leq i,j\leq n}$ de $M_n(\mathbb{C})$, telle que

$$b_{ij}(\varepsilon) = \begin{cases} a_{ii} & \text{si } i = j, \\ \varepsilon\, a_{ij} & \text{si } i \neq j. \end{cases}$$

On a alors $B(1) = A$ et $B(0)$ est une matrice diagonale dont les éléments diagonaux coïncident avec ceux de A. Chacune des valeurs propres de $B(0)$ est donc le centre d'un des disques de Gershgorin de A et l'on sait qu'exactement p de ces valeurs propres se trouvent dans l'union de disques $D^{(p)}$. Les valeurs propres de $B(\varepsilon)$ étant les racines de son polynôme caractéristique, dont les coefficients sont des fonctions continues de ε, elles sont des fonctions continues de ε. Par conséquent, lorsque le paramètre ε parcourt l'intervalle $[0,1]$, les valeurs propres de la matrice $B(\varepsilon)$ se déplacent le long de chemins continus du plan complexe et les rayons de ses disques de Gershgorin varient de 0 à ceux des disques de A. Puisque p valeurs propres sont contenues dans l'union $D^{(p)}$ lorsque ε est nul et que les disques de cette union sont disjoints de ceux de $D^{(q)}$, ces p valeurs propres doivent encore se trouver dans $D^{(p)}$ lorsque ε vaut 1. □

Ce dernier résultat se généralise à un nombre de sous-ensembles disjoints de disques de Gershgorin supérieur à deux.

4.3 Conditionnement d'un problème aux valeurs propres

Comme dans le cas de la résolution d'un système linéaire, le calcul numérique de valeurs et de vecteurs propres est affecté par des erreurs d'arrondis. Si la notion de *conditionnement d'un problème aux valeurs propres* fait bien intervenir celle de conditionnement d'une matrice (voir la section A.5.4), il ne s'agit ici pas du conditionnement de la matrice dont on cherche les valeurs propres, mais d'une matrice de passage l'amenant sous forme diagonale, comme le montre le résultat suivant.

Théorème 4.6 (« théorème de Bauer–Fike » [BF60]) *Soit n un entier naturel non nul, A une matrice d'ordre n diagonalisable, P une matrice d'ordre n inversible dont les colonnes sont des vecteurs propres de A, telle que $A = P^{-1}DP$, avec D une matrice d'ordre n diagonale ayant pour coefficients diagonaux les valeurs propres λ_i, $i = 1, \ldots, n$, de A, et δA une matrice de $M_n(\mathbb{C})$. Si μ est une valeur propre de $A + \delta A$, alors*

$$\min_{\lambda \in \sigma(A)} |\lambda - \mu| \leq \mathrm{cond}_p(P)\, \|\delta A\|_p,$$

avec $\|\cdot\|_p$ une norme matricielle subordonnée à une norme p quelconque.

DÉMONSTRATION. Si μ est une valeur propre de A alors l'inégalité du théorème est trivialement vérifiée. On suppose à présent que μ n'est pas une valeur propre de A. Par définition d'une valeur propre, on a $\det(A + \delta A - \mu I_n) = 0$, d'où

$$\begin{aligned} 0 = \det(P^{-1}) \det(D + P^{-1}\delta A P - \mu I_n) \det(P) &= \det(D + P^{-1}\delta A P - \mu I_n) \\ &= \det(D - \mu I_n) \det(I_n + (D - \mu I_n)^{-1} P^{-1} \delta A P), \end{aligned}$$

ce qui implique que -1 est une valeur propre de la matrice $(D - \mu I_n)^{-1} P^{-1} \delta A P$. On a donc, en vertu du théorème A.157, on a

$$1 \leq \|(D - \mu I_n)^{-1} P^{-1} \delta A P\|_p \leq \|(D - \mu I_n)^{-1}\|_p \|\delta A\|_p \|P\|_p \|P^{-1}\|_p,$$

soit encore

$$\frac{1}{\|(D - \mu I_n)^{-1}\|_p} \leq \mathrm{cond}_p(P) \|\delta A\|_p,$$

dont découle l'inégalité, puisque, la matrice $D - \mu I_n$ étant diagonale,

$$\|(D - \mu I_n)^{-1}\|_p = \sup_{\substack{\boldsymbol{v} \in \mathbb{C}^n \\ \boldsymbol{v} \neq \boldsymbol{0}}} \frac{\|(D - \mu I_n)^{-1} \boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \max_{\lambda \in \sigma(A)} \frac{1}{|\lambda - \mu|} = \frac{1}{\min\limits_{\lambda \in \sigma(A)} |\lambda - \mu|}.$$

□

Il résulte de ce théorème que les matrices normales sont très bien conditionnées pour le problème aux valeurs propres puisqu'elles sont diagonalisables au moyen de matrices unitaires, dont le conditionnement relativement à la norme $\|\cdot\|_2$ vaut 1 (voir le théorème A.168).

4.4 Méthode de la puissance

La méthode de la puissance (*power (iteration) method* en anglais) est certainement la méthode la plus simple fournissant une approximation de la valeur propre de plus grand module d'une matrice et d'un vecteur propre associé. Après l'avoir présentée et avoir analysé sa convergence, nous verrons comment, par des modifications adéquates, elle peut être utilisée pour calculer quelques autres couples de valeur et vecteur propres de la même matrice. Dans toute la suite, on considère une matrice A de $M_n(\mathbb{C})$ diagonalisable et on note λ_j, $j = 1, \ldots, n$, ses valeurs propres (comptées avec leurs multiplicités algébriques respectives) et $\boldsymbol{v}_j$, $j = 1, \ldots, n$, des vecteurs propres associés. On suppose de plus que les valeurs propres de A sont ordonnées de la manière suivante

$$|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_n|. \tag{4.7}$$

4.4.1 Approximation de la valeur propre de plus grand module

Faisons l'hypothèse que la valeur propre λ_n est d'ordre de multiplicité algébrique égal à 1 et que la dernière des inégalités de (4.7) est une inégalité stricte. On dit alors que λ_n est la valeur propre *dominante* de A. Pour l'approcher, on peut considérer une méthode itérative, appelée méthode de la puissance. Celle-ci consiste, à partir d'un vecteur initial unitaire $\boldsymbol{q}^{(0)}$ arbitraire, en le calcul des termes des suites définies par

$$\forall k \in \mathbb{N}^*,\ \boldsymbol{z}^{(k)} = A\boldsymbol{q}^{(k-1)},\ \boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|_2} \text{ et } \nu^{(k)} = (\boldsymbol{q}^{(k)})^* A \boldsymbol{q}^{(k)}. \tag{4.8}$$

Analysons les propriétés de la suite $(\boldsymbol{q}^{(k)})_{k \in \mathbb{N}}$. Par une simple récurrence sur l'indice k, on vérifie que

$$\forall k \in \mathbb{N},\ \boldsymbol{q}^{(k)} = \frac{A^k \boldsymbol{q}^{(0)}}{\|A^k \boldsymbol{q}^{(0)}\|_2}, \tag{4.9}$$

et l'on voit alors plus clairement le rôle joué par les puissances de la matrice A, qui donnent son nom à la méthode. En effet, l'ensemble $\{\boldsymbol{v}_j\}_{j=1,\dots,n}$ des vecteurs propres de A formant une base de l'espace, le vecteur $\boldsymbol{q}^{(0)}$ peut s'écrire comme la combinaison linéaire suivante

$$\boldsymbol{q}^{(0)} = \sum_{j=1}^{n} \alpha_j \boldsymbol{v}_j,$$

et l'on a alors, en supposant le coefficient α_n non nul,

$$\forall k \in \mathbb{N},\ A^k \boldsymbol{q}^{(0)} = \sum_{j=1}^{n} \alpha_j (A^k \boldsymbol{v}_j) = \sum_{j=1}^{n} \alpha_j {\lambda_j}^k \boldsymbol{v}_j = \alpha_n {\lambda_n}^k \left(\boldsymbol{v}_n + \sum_{j=1}^{n-1} \frac{\alpha_j}{\alpha_n} \left(\frac{\lambda_j}{\lambda_n} \right)^k \boldsymbol{v}_j \right). \tag{4.10}$$

Comme $\left|\frac{\lambda_j}{\lambda_n}\right| < 1$ pour tout entier j dans $\{1,\dots,n-1\}$, la composante du vecteur $\boldsymbol{q}^{(k)}$ le long de $\boldsymbol{v}_n$ augmente en module avec l'entier k relativement aux composantes le long des autres directions $\boldsymbol{v}_j$, $j = 1,\dots,n-1$. En supposant que les vecteurs de la base $\{\boldsymbol{v}_j\}_{j=1,\dots,n}$ sont de norme euclidienne unitaire, il vient alors

$$\left\| \sum_{j=1}^{n-1} \frac{\alpha_j}{\alpha_n} \left(\frac{\lambda_j}{\lambda_n} \right)^k \boldsymbol{v}_j \right\|_2 \leq \sum_{j=1}^{n-1} \left| \frac{\alpha_j}{\alpha_n} \right| \left| \frac{\lambda_j}{\lambda_n} \right|^k \leq \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^k \sum_{j=1}^{n-1} \left| \frac{\alpha_j}{\alpha_n} \right| = C \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^k,$$

et l'on déduit de (4.9), de (4.10) et de cette dernière inégalité que

$$\forall k \in \mathbb{N},\ \boldsymbol{q}^{(k)} = \frac{\alpha_n {\lambda_n}^k (\boldsymbol{v}_n + \boldsymbol{w}^{(k)})}{\| \alpha_n {\lambda_n}^k (\boldsymbol{v}_n + \boldsymbol{w}^{(k)}) \|_2},$$

où la suite de vecteurs $(\boldsymbol{w}^{(k)})_{k\in\mathbb{N}^*}$ a pour limite le vecteur nul. Le vecteur $\boldsymbol{q}^{(k)}$ devient donc peu à peu colinéaire avec le vecteur propre $\boldsymbol{v}_n$ associé à la valeur propre dominante λ_n quand k tend vers l'infini et ce d'autant plus rapidement que le rapport $\left|\frac{\lambda_{n-1}}{\lambda_n}\right|$ est petit, ce qui correspond à des valeurs propres dominante et sous-dominante bien séparées. La suite des *quotients de Rayleigh* [16]

$$\forall k \in \mathbb{N},\ \nu^{(k)} = (\boldsymbol{q}^{(k)})^* A \boldsymbol{q}^{(k)},$$

converge alors vers la valeur propre λ_n, et on a démontré le résultat suivant.

Théorème 4.7 *Soit n un entier naturel supérieur ou égal à 2 et A une matrice diagonalisable d'ordre n, dont les valeurs propres satisfont*

$$|\lambda_1| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|.$$

On suppose que le vecteur initial $\boldsymbol{q}^{(0)}$ de la méthode de la puissance (4.8) *n'est pas strictement contenu dans le sous-espace vectoriel engendré par les vecteurs propres associés aux valeurs propres autres que la valeur dominante λ_n. Alors, la méthode converge* [17] *et sa vitesse de convergence est d'autant plus rapide que le module du rapport entre λ_{n-1} et λ_n est petit.*

Si la matrice A est réelle symétrique, l'énoncé du théorème précédent se simplifie légèrement puisque A est alors diagonalisable dans une base orthonormale en vertu du théorème A.131 et que la condition sur le vecteur d'initialisation revient à demander qu'il ne soit pas orthogonal au vecteur propre $\boldsymbol{v}_n$. On peut de plus montrer que la convergence de la méthode est plus rapide que dans le cas général.

16. John William Strutt, troisième baron Rayleigh, (12 novembre 1842 - 30 juin 1919) était un physicien anglais. Il est à l'origine de nombreuses découvertes scientifiques. Il expliqua pour la première fois de manière correcte la couleur du ciel en la reliant à la diffusion de la lumière par les molécules d'air, théorisa l'existence des ondes de surface ou encore découvrit, en collaboration avec William Ramsay, l'élément chimique argon, ce qui lui valut de recevoir le prix Nobel de physique en 1904.

17. Pour une valeur propre dominante λ_n *réelle*, la convergence a lieu au sens où l'on a

$$\lim_{k\to+\infty} \nu^{(k)} = \lambda_n \text{ et } \lim_{k\to+\infty} \boldsymbol{q}^{(k)} = \boldsymbol{v}_n \text{ si } \lambda_n > 0 \text{ ou } \lim_{k\to+\infty} (-1)^k \boldsymbol{q}^{(k)} = \boldsymbol{v}_n \text{ si } \lambda_n < 0,$$

le vecteur $\boldsymbol{v}_n$ étant un vecteur propre unitaire associé à λ_n.

Théorème 4.8 *Soit n un entier naturel supérieur ou égal à 2 et A une matrice réelle symétrique d'ordre n, dont les valeurs propres satisfont*

$$|\lambda_1| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|.$$

On suppose que le vecteur initial $\boldsymbol{q}^{(0)}$ de la méthode de la puissance (4.8) *n'est pas orthogonal au sous-espace propre associé à la valeur propre dominante λ_n. Alors, la méthode converge avec un taux quadratique par rapport au quotient $|\lambda_{n-1}/\lambda_n|$, c'est-à-dire que*

$$\forall k \in \mathbb{N},\ \left|\nu^{(k)} - \lambda_n\right| \leq \max_{i\in\{1,\ldots,n-1\}} |\lambda_i - \lambda_n| \tan(\theta^{(0)})^2 \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{2k},$$

avec $\cos(\theta^{(0)}) = \left|\boldsymbol{v}_n{}^\top \boldsymbol{q}^{(0)}\right| \neq 0$.

DÉMONSTRATION. Par définition de la méthode, on a, en utilisant la décomposition du vecteur $\boldsymbol{q}^{(0)}$ dans la base orthonormée $\{\boldsymbol{v}_j\}_{j=1,\ldots,n}$ et le fait que A est symétrique,

$$\forall k \in \mathbb{N},\ \nu^{(k)} = (\boldsymbol{q}^{(k)})^\top A \boldsymbol{q}^{(k)} = \frac{(\boldsymbol{q}^{(0)})^\top A^{2k+1} \boldsymbol{q}^{(0)}}{(\boldsymbol{q}^{(0)})^\top A^{2k} \boldsymbol{q}^{(0)}} = \frac{\sum_{j=1}^n \alpha_j{}^2 \lambda_j{}^{2k+1}}{\sum_{j=1}^n \alpha_j{}^2 \lambda_j{}^{2k}},$$

et donc

$$\forall k \in \mathbb{N},\ \left|\nu^{(k)} - \lambda_n\right| = \left|\frac{\sum_{j=1}^{n-1} \alpha_j{}^2 \lambda_j{}^{2k}(\lambda_j - \lambda_n)}{\sum_{j=1}^n \alpha_j{}^2 \lambda_j{}^{2k}}\right| \leq \max_{i\in\{1,\ldots,n-1\}} |\lambda_i - \lambda_n| \frac{\sum_{j=1}^{n-1} \alpha_j{}^2 \lambda_j{}^{2k}}{\sum_{j=1}^n \alpha_j{}^2 \lambda_j{}^{2k}}.$$

Par hypothèse sur $\boldsymbol{q}^{(0)}$, il vient alors

$$\frac{\sum_{j=1}^{n-1} \alpha_j{}^2 \lambda_j{}^{2k}}{\sum_{j=1}^n \alpha_j{}^2 \lambda_j{}^{2k}} \leq \frac{\sum_{j=1}^{n-1} \alpha_j{}^2 \lambda_j{}^{2k}}{\alpha_n{}^2 \lambda_n{}^{2k}} \leq \frac{1}{\alpha_n{}^2}\left(\sum_{j=1}^{n-1} \alpha_j{}^2\right)\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k} = \frac{1-\alpha_n{}^2}{\alpha_n{}^2}\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k} = \tan(\theta^{(0)})^2\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k},$$

d'où l'inégalité annoncée. □

Bien qu'elle soit difficile à vérifier quand on ne dispose d'aucune information *a priori* sur le sous-espace propre associé à λ_n, on peut remarquer que l'hypothèse faite sur le vecteur initial $\boldsymbol{q}^{(0)}$ n'est pas très contraignante. Tout d'abord, si ce vecteur est tiré aléatoirement selon une loi uniforme, alors l'hypothèse est presque sûrement satisfaite. D'autre part, si celui-ci est en pratique bel et bien strictement contenu dans $\mathrm{Vect}(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}\})$, il est aussi fortement probable que, du fait des erreurs d'arrondi, l'un des vecteurs $\boldsymbol{q}^{(k)}$, avec k un entier naturel non nul, aura une « petite » composante dans la direction de $\boldsymbol{v}_n$, ce qui entraînera alors la convergence de la méthode. Habituellement, on observe ce phénomène après quelques itérations. Par ailleurs, on voit, au moyen de l'expression (4.10), que la suite de vecteurs $(A^k \boldsymbol{q}^{(0)})_{k\in\mathbb{N}}$ n'est, en général, pas convergente. C'est la raison pour laquelle on choisit de travailler avec la suite de vecteurs unitaires $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$.

Si la valeur propre λ_n n'est pas de multiplicité simple tout en étant néanmoins la seule valeur propre de plus grand module de A, on a encore convergence de la suite $(\nu^{(k)})_{k\in\mathbb{N}}$ vers λ_n, alors que la suite $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$ converge vers un élément du sous-espace propre associé. En revanche, s'il existe plusieurs valeurs propres de plus grand module, la méthode de la puissance ne converge généralement pas. Dans le cas de deux valeurs propres dominantes complexes conjuguées, *i.e.*, $\lambda_{n-1} = \overline{\lambda_n}$, la suite $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$ présente notamment un comportement oscillatoire non amorti. Indiquons que supposer l'existence d'une unique valeur propre dominante n'est en principe pas une hypothèse restrictive, puisque l'on peut montrer qu'une matrice possède des valeurs propres de modules distincts de manière générique [18].

Enfin, si la matrice A n'est pas diagonalisable, il reste possible de prouver la convergence de la méthode lorsqu'il existe une unique valeur propre dominante de multiplicité simple, en s'appuyant sur la réduction de Jordan.

Indiquons que la mise en œuvre de la méthode de la puissance ne nécessite que de calculer des produits scalaires entre différents vecteurs, ainsi que des produits entre la matrice A et les vecteurs de la suites $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$. Il n'est en particulier pas obligatoire de stocker la matrice A sous la forme d'un tableau, ce qui peut être particulièrement intéressant lorsque celle-ci est creuse et de grande taille (voir la sous-section 4.1.3 pour un exemple). Dans le cas d'une matrice d'ordre n quelconque, le coût d'une itération de la méthode sera de n^2 multiplications et $n(n-1)$ additions pour effectuer le produit matrice-vecteur, $2n$ multiplications, $n-1$ additions, une division et une extraction de racine carrée pour la normalisation du vecteur courant, n multiplications et $n-1$ additions pour le calcul du quotient de Rayleigh.

18. En topologie, une propriété est dite *générique* si elle est vraie sur un ensemble ouvert dense ou, plus généralement, sur un ensemble *résiduel* (c'est-à-dire un ensemble contenant une intersection dénombrable d'ouverts denses).

4.4.2 Déflation

Pour approcher d'autres valeurs propres que celle de plus grand module, on peut utiliser une technique, nommée *déflation*, consistant en la construction, à partir d'une matrice donnée, d'une matrice possédant le même spectre, à l'exception d'une valeur propre choisie qui se trouve remplacée par 0. Il devient alors possible, en appliquant la méthode de la puissance à la matrice résultant de la déflation par la valeur propre dominante λ_n déjà connue, d'obtenir une suite convergeant vers la valeur propre ayant le *deuxième* plus grand module, puis, éventuellement, de réitérer le procédé pour obtenir les valeurs propres suivantes.

Le procédé de déflation le plus simple, dit *de Hotelling* [19] [Hot33], demande, dans le cas d'une matrice A d'ordre n diagonalisable, de connaître une valeur propre λ_j, l'entier j appartenant à $\{1,\dots,n\}$, un vecteur $\boldsymbol{v}_j$ associé, issu d'une base $\{\boldsymbol{v}_i\}_{i=1,\dots,n}$ formée de vecteurs propres, ainsi que le vecteur $\boldsymbol{u}_j$ correspondant de la *base duale* de $\{\boldsymbol{v}_i\}_{i=1,\dots,n}$, c'est-à-dire le vecteur vérifiant [20]

$$\forall i \in \{1,\dots,n\},\ (\boldsymbol{u}_j)^* \boldsymbol{v}_i = \delta_{ij}.$$

En considérant la perturbation de rang un suivante de A,

$$A_j = A - \lambda_j \boldsymbol{v}_j (\boldsymbol{u}_j)^*,$$

il vient

$$\forall i \in \{1,\dots,n\},\ A_j \boldsymbol{v}_i = A\boldsymbol{v}_i - \lambda_j \boldsymbol{v}_j (\boldsymbol{u}_j)^* \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i - \lambda_j \boldsymbol{v}_j \delta_{ij},$$

ce qui correspond effectivement à la modification annoncée du spectre, les vecteurs propres associés restant inchangés. Lorsque la matrice A est réelle symétrique (ou hermitienne) et que ses valeurs propres sont deux à deux distinctes, on notera que l'on a seulement besoin que de connaître la valeur propre λ_j et un vecteur propre $\boldsymbol{v}_j$ associé, puisque l'on peut dans ce cas définir la matrice A_j par

$$A_j = A - \lambda_j \frac{\boldsymbol{v}_j (\boldsymbol{v}_j)^*}{(\boldsymbol{v}_j)^* \boldsymbol{v}_j}$$

en vertu de la propriété d'orthogonalité des vecteurs propres.

Plus générale, la méthode de déflation due à Wielandt [21] [Wie44b], repose sur la perturbation

$$A_j = A - \boldsymbol{v}_j (\boldsymbol{w}_j)^*,$$

où $\boldsymbol{w}_j$ est un vecteur tel que $(\boldsymbol{w}_j)^* \boldsymbol{v}_j = \lambda_j$. Cette transformation a le même effet sur le spectre et, bien que le vecteur $\boldsymbol{v}_j$ reste associé à la valeur propre rendue nulle, elle modifie en général [22] les vecteurs propres correspondant aux autres valeurs propres. On notera que le choix $\boldsymbol{w}_j = \lambda_j \boldsymbol{u}_j$ conduit à la déflation de Hotelling.

Une troisième méthode, introduite par Collar [23] et Duncan [DC34, FDC38], entraîne une réduction d'ordre de la matrice dont on cherche les valeurs propres. Soit $(\boldsymbol{v}_j)_r$ une composante non nulle du vecteur propre $\boldsymbol{v}_j$. On pose alors

$$A_j = A - \frac{1}{(\boldsymbol{v}_j)_r} \boldsymbol{v}_j \boldsymbol{e_r}^* A,$$

19. Harold Hotelling (29 septembre 1895 - 26 décembre 1973) était un statisticien et économiste américain. En statistique, il est connu pour l'introduction de méthodes d'analyse en composantes principales et son utilisation de la loi de Student pour la validation d'hypothèses et la détermination d'intervalles de confiance.

20. Le spectre de la matrice adjointe de A étant constitué des conjugués des valeurs propres de A, on voit que le vecteur $\boldsymbol{u}_j$ est un vecteur propre de A^*, associé à la valeur propre $\overline{\lambda_j}$ et normalisé de manière convenable. On a en effet

$$\forall i \in \{1,\dots,n\},\ \lambda_j (\boldsymbol{u}_j)^* \boldsymbol{v}_i = (\overline{\lambda_j}\, \boldsymbol{u}_j)^* \boldsymbol{v}_i (A^* \boldsymbol{u}_j)^* \boldsymbol{v}_i = (\boldsymbol{u}_j)^* A \boldsymbol{v}_i = (\boldsymbol{u}_j)^* (A \boldsymbol{v}_i) = \lambda_i (\boldsymbol{u}_j)^* \boldsymbol{v}_i.$$

21. Helmut Wielandt (19 décembre 1910 - 14 février 2001) était un mathématicien allemand. Il est connu pour ses travaux en théorie des groupes, notamment sur les groupes finis et les groupes de permutations, et en théorie des matrices.

22. On a en effet, en supposant tacitement que la valeur propre λ_i, l'entier i appartenant à $\{1,\dots,n\}\backslash\{j\}$, est non nulle,

$$A_j(\boldsymbol{v}_i - \alpha_i \boldsymbol{v}_j) = A\boldsymbol{v}_i - \alpha_i A\boldsymbol{v}_j - \boldsymbol{v}_j (\boldsymbol{w}_j)^* \boldsymbol{v}_i + \alpha_i \boldsymbol{v}_j (\boldsymbol{w}_j)^* \boldsymbol{v}_j = \lambda_i \boldsymbol{v}_i - \boldsymbol{v}_j (\boldsymbol{w}_j)^* \boldsymbol{v}_i = \lambda_i \left(\boldsymbol{v}_i - \frac{(\boldsymbol{w}_j)^* \boldsymbol{v}_i}{\lambda_i} \boldsymbol{v}_j \right).$$

La matrice A_j admet donc λ_i pour valeur propre si $\alpha_i = \frac{(\boldsymbol{w}_j)^* \boldsymbol{v}_i}{\lambda_i}$.

23. Arthur Roderick Collar (22 février 1908 - 12 février 1986) était un ingénieur anglais. On lui doit des contributions significatives dans les domaines de l'aéroélasticité et de l'application de la théorie des matrices en dynamique mécanique.

où $\boldsymbol{e_r}$ est le r^{e} vecteur de la base canonique de $M_{n,1}(\mathbb{C})$. Par construction, la r^{e} ligne de la matrice A_j est nulle et l'on a

$$A_j \boldsymbol{v}_j = A\boldsymbol{v}_j - \frac{1}{(\boldsymbol{v}_j)_r} \boldsymbol{v}_j \boldsymbol{e_r}^* A\boldsymbol{v}_j = \lambda_j \left(1 - \frac{1}{(\boldsymbol{v}_j)_r} \boldsymbol{e_r}^* \boldsymbol{v}_j\right) \boldsymbol{v}_j = \boldsymbol{0}.$$

Le vecteur $\boldsymbol{v}_j$ est donc un vecteur propre de A_j associé à 0. D'autre part, pour toute valeur propre λ_i de A distincte de λ_i et tout vecteur propre associé $\boldsymbol{v}_i$, il vient, en vertu du précédent calcul,

$$A_j(\boldsymbol{v}_i - \boldsymbol{v}_j) = A_j \boldsymbol{v}_i = A\boldsymbol{v}_i - \frac{1}{(\boldsymbol{v}_j)_r} \boldsymbol{v}_j \boldsymbol{e_r}^* A\boldsymbol{v}_i = \lambda_i \left(\boldsymbol{v}_i - \frac{1}{(\boldsymbol{v}_j)_r} (\boldsymbol{e_r}^* \boldsymbol{v}_i) \boldsymbol{v}_j\right).$$

Il en découle que λ_i est une valeur propre de A_j associée au vecteur propre $\boldsymbol{v}_i - \boldsymbol{v}_j$ si $(\boldsymbol{v}_i)_r = (\boldsymbol{v}_j)_r \neq 0$, la r^{e} composante de ce vecteur étant alors nulle.

Cette dernière observation, alliée au fait que la r^{e} ligne de A_j est nulle, montre que l'on peut considérer pour le calcul des valeurs propres distinctes de λ_j et de vecteurs propres associés une matrice d'ordre $n-1$ extraite de A_j, obtenue par supression des r^{es} ligne et colonne.

Utilisées en conjonction avec la méthode de la puissance, ces techniques permettent en théorie d'approcher l'*ensemble* du spectre d'une matrice donnée. Cependant, les valeurs et vecteurs obtenus successivement n'étant en pratique que des approximations des véritables valeurs et vecteurs propres, l'accumulation des erreurs et la mauvaise stabilité numérique de ces deux types de déflation les rendent difficilement utilisables pour le calcul de plus de deux ou trois valeurs propres. On pourra consulter le neuvième chapitre du livre [Wil65] pour une présentation d'approches plus stables de ce procédé, reposant sur des similitudes ou des transformations unitaires.

4.4.3 Méthode de la puissance inverse

On peut facilement adapter la méthode de la puissance pour le calcul de la valeur propre de plus *petit* module, que l'on a noté λ_1, d'une matrice A *inversible* : il suffit de l'appliquer à l'inverse de A, dont la plus *grande* valeur propre en module est ${\lambda_1}^{-1}$. On parle dans ce cas de *méthode de la puissance inverse* (*inverse power method* en anglais).

De manière plus générale, cette variante permet d'approcher la valeur propre de la matrice A la *plus proche* d'un nombre μ donné n'appartenant pas à son spectre. Considérons en effet la matrice $(A-\mu I_n)^{-1}$ dont les valeurs propres sont $(\lambda_i - \mu)^{-1}$, $i = 1, \ldots, n$, et supposons qu'il existe un entier m tel que

$$\forall i \in \{1, \ldots, n\} \setminus \{m\},\ |\lambda_m - \mu| < |\lambda_i - \mu|,$$

ce qui revient à supposer que la valeur propre λ_m qui est la plus proche de μ a une multiplicité algébrique égale à 1 (en particulier, si $\mu = 0$, λ_m sera la valeur propre de A ayant le plus petit module). L'application de la méthode de la puissance inverse pour le calcul de λ_m se résume alors à la construction, étant donné un vecteur initial unitaire $\boldsymbol{q}^{(0)}$ arbitraire, des suites définies par les relations de récurrence suivantes

$$\forall k \in \mathbb{N}^*,\ (A-\mu I_n)\boldsymbol{z}^{(k)} = \boldsymbol{q}^{(k-1)},\ \boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|_2} \text{ et } \nu^{(k)} = (\boldsymbol{q}^{(k)})^* A \boldsymbol{q}^{(k)}. \tag{4.11}$$

Les vecteurs propres de la matrice $A-\mu I_n$ étant ceux de la matrice A, le quotient de Rayleigh ci-dessus fait simplement intervenir A et non $A-\mu I_n$. Le nombre μ peut être vu comme un paramètre permettant de « décaler » (on parle en anglais de *shift*) le spectre de la matrice A de manière à pouvoir approcher toute valeur propre de A dont on possède une estimation *a priori*. De ce point de vue, la méthode de la puissance inverse se prête donc particulièrement bien au raffinement d'une approximation grossière d'une valeur propre obtenue par les techniques de la section 4.2. Par rapport à la méthode de la puissance (4.8), il faut cependant résoudre, à chaque itération de la méthode, un système linéaire (ayant pour matrice $A-\mu I_n$) pour obtenir les vecteurs de la suite $(\boldsymbol{z}^{(k)})_{k\in\mathbb{N}}$. En pratique, on réalise une fois pour toutes la factorisation LU (voir la section 2.5) de cette matrice au début du calcul de manière à n'effectuer par la suite que la résolution de deux systèmes linéaires triangulaires, pour un coût de l'ordre de n^2 opérations, à chaque étape.

S'il est souhaitable que la valeur du paramètre μ soit aussi voisine que possible de la valeur propre λ_m pour que la convergence soit rapide, il faut néanmoins qu'il n'en soit pas proche au point de rendre la matrice $A-\mu I_n$ *numériquement singulière* (cette dernière notion, liée à la présence d'erreurs d'arrondi, étant essentiellement empirique).

4.4.4 Méthode de Lanczos **

La *méthode de Lanczos* [Lan50] est une adaptation de la méthode de la puissance au calcul simultané de plusieurs valeurs et vecteurs propres d'une matrice carrée (ou encore à la décomposition en valeurs singulières d'une matrice rectangulaire). Elle est particulièrement employée dans le cas de très grandes matrices creuses.

A COMPLETER

AJOUTER que la méthode peut aussi servir pour la résolution de systèmes linéaires [Lan52].

4.5 Méthodes pour le calcul de valeurs propres d'une matrice symétrique

Nous avons déjà observé que les matrices réelles symétriques (ou hermitiennes) présentent plusieurs avantages pratiques pour la mise en œuvre de méthodes numériques. Leur structure particulière permet ainsi de réduire le nombre d'opérations arithmétiques requises et l'occupation en mémoire liée à leur stockage. La méthode de Cholesky (voir la sous-section 2.6.2) est un exemple parmi d'autres illustrant ce fait. Certaines de leurs propriétés leur confèrent également une place particulière au sein de l'ensemble des matrices carrées. On sait en effet (voir le théorème A.131) que de telles matrices sont toujours diagonalisables, que leurs valeurs propres sont réelles et que l'on peut former une base orthonormale avec leurs vecteurs propres.

Les méthodes considérées dans cette section tirent parti de façon cruciale d'une (ou de plusieurs) de ces particularités pour permettre le calcul numérique effectif des valeurs et/ou des vecteurs propres de ces matrices.

4.5.1 Méthode de Jacobi pour une matrice réelle

La *méthode de Jacobi* consiste en la construction d'une suite de matrices symétriques convergeant vers la *décomposition de Schur* (voir le théorème A.115), diagonale et orthogonalement semblable, d'une matrice réelle symétrique. Pour ce faire, l'idée est se « rapprocher » en un nombre infini d'étapes, d'une forme diagonale de la matrice en éliminant successivement des couples de coefficients hors diagonaux en position symétrique par utilisation des transformations induites par les *matrices de Givens*.

Matrices de rotation de Givens

Les matrices de Givens sont des matrices orthogonales qui permettent, tout comme les matrices de Householder présentées dans la section 2.6.3, d'annuler certains coefficients d'un vecteur ou d'une matrice. Pour un couple d'indices p et q vérifiant $1 \leq p < q \leq n$, et un nombre réel θ donnés, on définit la matrice de Givens par

$$G(p,q,\theta) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & & & & & & & & & \vdots \\ \vdots & & \ddots & & & & & & & & \vdots \\ \vdots & & & \cos(\theta) & & & & \sin(\theta) & & & \vdots \\ \vdots & & & & 1 & & & & & & \vdots \\ \vdots & & & & & \ddots & & & & & \vdots \\ \vdots & & & & & & 1 & & & & \vdots \\ \vdots & & & -\sin(\theta) & & & & \cos(\theta) & & & \vdots \\ \vdots & & & & & & & & 1 & & \vdots \\ \vdots & & & & & & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} \tag{4.12}$$

$$= I_n - (1-\cos(\theta))(E_{pp} + E_{qq}) + \sin(\theta)(E_{pq} - E_{qp}).$$

Cette matrice représente la rotation d'angle θ (dans le sens trigonométrique) dans le plan engendré par les p^{e} et q^{e} vecteurs de la base canonique de $\mathbb{R}^n$, ce qui constitue une manière de voir qu'elle est orthogonale. D'autres propriétés des matrices de Givens sont résumées dans le résultat suivant.

Proposition 4.9 *Soit p et q deux entiers vérifiant $1 \leq p < q \leq n$ et θ un nombre réel auxquels on associe la matrice définie par* (4.12)*. On a les deux propriétés suivantes.*

1. *Si A est une matrice symétrique, alors la matrice*

$$B = G(p,q,\theta)^\top A G(p,q,\theta), \tag{4.13}$$

également symétrique, vérifie

$$\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2.$$

2. *Si $a_{pq} \neq 0$, alors il existe une unique valeur du nombre θ dans l'ensemble $\left]-\frac{\pi}{4}, 0\right[\cup \left]0, \frac{\pi}{4}\right]$ telle que*

$$b_{pq} = 0,$$

donnée par la seule solution, dans le même ensemble, de l'équation

$$\operatorname{cotan}(2\theta) = \frac{a_{qq} - a_{pp}}{2\, a_{pq}}.$$

On a alors

$$\sum_{i=1}^{n} b_{ii}^2 = \sum_{i=1}^{n} a_{ii}^2 + 2\, a_{pq}^2.$$

DÉMONSTRATION.

1. On remarque que $\sum_{i=1}^n \sum_{j=1}^n b_{ij}^2 = \|B\|_F^2 = \operatorname{tr}(BB^\top)$, où $\|\cdot\|_F$ désigne la norme de Frobenius (définie par (A.8)). On vérifie alors que
$$\operatorname{tr}(G(p,q,\theta)^\top A G(p,q,\theta)(G(p,q,\theta)^\top A G(p,q,\theta))^\top) = \operatorname{tr}(G(p,q,\theta)^\top A G(p,q,\theta)) = \operatorname{tr}(A),$$
d'où l'égalité recherchée.
2. Il suffit de remarquer que la transformation portant sur les éléments d'indices (p,p), (p,q), (q,p) et (q,q) s'écrit sous la forme
$$\begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix},$$
et le même raisonnement qu'en 1. montre que
$$b_{pp}^2 + b_{qq}^2 + 2\, b_{pq}^2 = a_{pp}^2 + a_{qq}^2 + 2\, a_{pq}^2,$$
pour toute valeur de θ. On a par ailleurs
$$b_{pq} = b_{qp} = a_{pq}(\cos(\theta)^2 - \sin(\theta)^2) + (a_{pp} - a_{qq})\cos(\theta)\sin(\theta) = a_{pq}\cos(2\theta) + \frac{a_{pp} - a_{qq}}{2}\sin(2\theta).$$
Le coefficient b_{pq} est donc nul si et seulement si l'angle θ vérifie la relation de l'énoncé, ce qui est toujours possible puisque la fonction $\operatorname{cotan}(2\cdot)$ est surjective sur $\mathbb{R}$. On en déduit que
$$b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2\, a_{pq}^2.$$
Comme tous les autres coefficients diagonaux de la matrice B sont identiques à ceux de la matrice A, on a établi l'égalité désirée.

□

On remarque que seules les p^{es} et q^{es} lignes et colonnes de la matrice A sont modifiées par la transformation fournissant la matrice $B = G(p,q,\theta)^\top A G(p,q,\theta)$. Plus précisément, on a

$$\left\{\begin{array}{rcl} b_{ij} &=& a_{ij} \text{ si } i \neq p,q \text{ et } j \neq p,q, \\ b_{pj} &=& \cos(\theta)\, a_{pj} - \sin(\theta)\, a_{qj} \text{ si } j \neq p,q, \\ b_{qj} &=& \sin(\theta)\, a_{pj} + \cos(\theta)\, a_{qj} \text{ si } j \neq p,q, \\ b_{pp} &=& \cos(\theta)^2\, a_{pp} - 2\cos(\theta)\sin(\theta)\, a_{pq} + \sin(\theta)^2\, a_{qq}, \\ b_{qq} &=& \sin(\theta)^2\, a_{pp} + 2\cos(\theta)\sin(\theta)\, a_{pq} + \cos(\theta)^2\, a_{qq}, \\ b_{pq} &=& \left(\cos(\theta)^2 - \sin(\theta)^2\right) a_{pq} + \cos(\theta)\sin(\theta)\,(a_{pp} - a_{qq}), \\ b_{qp} &=& b_{pq}. \end{array}\right. \tag{4.14}$$

En posant alors $c = \cos(\theta)$ et $s = \sin(\theta)$, on déduit des égalités ci-dessus que les coefficients de la matrice B telle que $b_{pq} = 0$ peuvent être obtenus, de façon stable, à partir de ceux de A par des identités algébriques ne nécessitant pas la détermination de l'angle θ. En effet, si $a_{pq} = 0$, on prend $c = 1$ et $s = 0$, sinon on pose $t = s/c$ et l'on voit que t doit être la racine de plus petit module du trinôme

$$t^2 + 2\tau\, t - 1 = 0, \text{ avec } \tau = \frac{a_{qq} - a_{pp}}{2\, a_{pq}},$$

c'est-à-dire $t = -\tau + \sqrt{\tau^2 + 1}$ si $\tau \geq 0$ et $t = -\tau - \sqrt{\tau^2 + 1}$ sinon, soit encore

$$t = \frac{\text{sign}(\tau)}{|\tau| + \sqrt{\tau^2 + 1}}, \text{ si } \tau \neq 0, \text{ avec } \text{sign}(\tau) = \begin{cases} 1 & \text{si } \tau > 0 \\ -1 & \text{sinon} \end{cases}, \ t = 1 \text{ sinon.}$$

On a alors

$$c = \frac{1}{\sqrt{1 + t^2}} \text{ et } s = c\, t = \frac{t}{\sqrt{1 + t^2}}. \tag{4.15}$$

Méthode de Jacobi

Décrivons à présent la méthode de Jacobi, qui consiste en la construction, après avoir posé $A^{(0)} = A$, d'une suite de matrices symétriques $(A^{(k)})_{k\in\mathbb{N}}$ par la relation de récurrence

$$\forall k \in \mathbb{N},\ A^{(k+1)} = (G^{(k)})^\top A^{(k)} G^{(k)}, \tag{4.16}$$

où $G^{(k)} = G(p^{(k)}, q^{(k)}, \theta^{(k)})$ est une matrice de Givens de la forme (4.12), dont les entiers $p^{(k)}$ et $q^{(k)}$, $1 \leq p^{(k)} < q^{(k)} \leq n$, dépendent de la matrice $A^{(k)}$ selon une stratégie choisie et le réel $\theta^{(k)} \in \left]-\frac{\pi}{4}, 0\right[\cup \left]0, \frac{\pi}{4}\right]$ est déterminé de manière à avoir

$$a^{(k+1)}_{p^{(k)}q^{(k)}} = 0,$$

ce choix particulier portant le nom de *rotation de Jacobi*. Il faut ici remarquer que les coefficients hors-diagonaux annulés à une étape donnée peuvent par la suite être remplacés par des éléments non nuls, puisque l'on arriverait autrement à une matrice diagonale en un nombre *fini* d'itérations, ce qui est impossible. Cependant, en procédant ainsi, on diminue, de façon systématique et à chaque étape, la quantité

$$\text{off}(A^{(k)}) = \sqrt{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^{(k)2}},$$

puisque, en vertu de la proposition 4.9, on a, pour tout $k \in \mathbb{N}$,

$$\text{off}(A^{(k+1)})^2 = \|A^{(k+1)}\|_F^2 - \sum_{i=1}^{n} a_{ij}^{(k+1)} = \|A^{(k)}\|_F^2 - \sum_{i=1}^{n} a_{ij}^{(k)} - 2\, a_{pq}^{(k)2} = \text{off}(A^{(k)})^2 - 2\, a_{pq}^{(k)2} \leq \text{off}(A^{(k)})^2. \tag{4.17}$$

C'est ce dernier fait qui rend la convergence de la méthode possible. En effet, la norme de Frobenius de chacune des matrices de la suite $(A^{(k)})_{k\in\mathbb{N}}$ est la même, alors que, à chaque étape, la somme des carrés des éléments hors-diagonaux est diminuée de la somme des carrés des deux éléments qui viennent d'être annulés. On peut donc espérer que cette suite va converger vers une matrice diagonale, qui sera égale, à des permutations près, à la matrice

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix},$$

où les réels λ_i, $1 \leq i \leq n$, sont les valeurs propres de la matrice A, ordonnées de manière arbitraire.

Avant d'énoncer un résultat de convergence, il faut décider d'une manière de réaliser un choix effectif des indices $p^{(k)}$ et $q^{(k)}$ à chaque étape. Si l'on vise à maximiser la réduction de la quantité $\mathrm{off}(A^{(k)})$, $k \geq 0$, il convient de choisir le couple $(p^{(k)}, q^{(k)})$ comme l'un des couples (p,q) pour lesquels

$$\left|a_{pq}^{(k)}\right| = \max_{1\leq i<j\leq n} \left|a_{ij}^{(k)}\right|.$$

Cette stratégie est à la base de la méthode de Jacobi sous sa forme classique. D'autres façons de procéder seront présentées dans la section suivante.

Nous allons à présent prouver que cette méthode est convergente. Afin d'éviter les situations triviales pour lesquelles on peut avoir convergence en un nombre fini d'étapes de la méthode, nous supposerons implicitement dans ce qui suit que $\max_{1\leq i<j\leq n} \left|a_{ij}^{(k)}\right| \neq 0$, $\forall k \geq 0$.

Théorème 4.10 (convergence vers les valeurs propres pour la méthode de Jacobi) *La suite de matrices $(A^{(k)})_{k\in\mathbb{N}}$ construite par la méthode de Jacobi est convergente et*

$$\lim_{k\to+\infty} A^{(k)} = \begin{pmatrix} \lambda_{\sigma(1)} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_{\sigma(n)} \end{pmatrix},$$

pour une permutation convenable σ de $\mathfrak{S}_n$, où $\mathfrak{S}_n$ désigne le groupe des permutations de l'ensemble $\{1,\dots,n\}$.

Pour démontrer ce théorème, on a besoin du lemme technique suivant.

Lemme 4.11 *Soit X un espace vectoriel normé de dimension finie et $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ une suite bornée dans X, admettant un nombre fini de valeurs d'adhérence et telle que $\lim_{k\to+\infty} \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| = 0$. Alors, la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est convergente.*

DÉMONSTRATION. On note $\boldsymbol{a}_i$, $1 \leq i \leq I$, les valeurs d'adhérence de la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$. Pour tout $\varepsilon > 0$, il existe un entier $k(\varepsilon)$ tel que

$$\forall k \in \mathbb{N},\ k \geq k(\varepsilon) \Rightarrow \boldsymbol{x}^{(k)} \in \bigcup_{i=1}^{I} B(\boldsymbol{a}_i, \varepsilon),$$

où $B(a_i, \varepsilon)$ est la boule fermée de centre $\boldsymbol{a}_i$ et de rayon ε. En effet, si cela n'était pas le cas, il existerait une suite extraite $(\boldsymbol{x}^{(\varphi(k))})_{k\in\mathbb{N}}$ qui convergerait vers un point $\boldsymbol{x}'$ tel que $\boldsymbol{x}' \notin \bigcup_{i=1}^{I} B(\boldsymbol{a}_i, \varepsilon)$. Le point $\boldsymbol{x}'$ serait alors une valeur d'adhérence de la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$, distincte des points $\boldsymbol{a}_i$, $1 \leq i \leq I$, ce qui vient contredire l'hypothèse.
On fait alors le choix particulier

$$\varepsilon_0 = \frac{1}{3} \min_{1\leq i<j\leq I} \|\boldsymbol{a}_i - \boldsymbol{a}_j\| > 0,$$

ce qui conduit à l'existence d'un entier $k(\varepsilon_0)$ tel que

$$\forall k \in \mathbb{N},\ k \geq k(\varepsilon_0) \Rightarrow \boldsymbol{x}^{(k)} \in \bigcup_{i=1}^{I} B(\boldsymbol{a}_i, \varepsilon_0) \text{ et } \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| \leq \varepsilon_0.$$

Ainsi, on a que $\boldsymbol{x}^{(k)} \in B(\boldsymbol{a}_i, \varepsilon_0)$ implique que $\boldsymbol{x}^{(k+1)} \in B(\boldsymbol{a}_i, \varepsilon_0)$ pour tout entier k supérieur ou égal à $k(\varepsilon_0)$ et le résultat est démontré. □

DÉMONSTRATION DU THÉORÈME 4.10. On a déjà établi en (4.17) que

$$\forall k \in \mathbb{N},\ \mathrm{off}(A^{(k+1)})^2 = \mathrm{off}(A^{(k)})^2 - 2\,a_{pq}^{(k)2}.$$

D'autre part, en tenant compte de la stratégie adoptée par la méthode pour le choix du couple $(p^{(k)}, q^{(k)})$ à une itération k donnée, à la majoration

$$\forall k \in \mathbb{N},\ \mathrm{off}(A^{(k)})^2 \leq n(n-1)\,a_{pq}^{(k)2},$$

puisqu'il y a $n(n-1)$ éléments hors-diagonaux. En combinant ces relations, on obtient

$$\forall k \in \mathbb{N},\ \mathrm{off}(A^{(k+1)})^2 \leq \left(1 - \frac{2}{n(n-1)}\right) \mathrm{off}(A^{(k)})^2, \tag{4.18}$$

ce qui montre que

$$\lim_{k\to+\infty} \mathrm{off}(A^{(k)}) = 0. \tag{4.19}$$

Désignons à présent par $D^{(k)}$, $k \in \mathbb{N}$, la matrice diagonale ayant pour coefficients les coefficients diagonaux de la matrice $A^{(k)}$ et montrons que la suite $(D^{(k)})_{k\in\mathbb{N}}$ n'a qu'un nombre fini de valeurs d'adhérence, qui seront nécessairement de la forme

$$\begin{pmatrix} \lambda_{\sigma(1)} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_{\sigma(n)} \end{pmatrix}, \tag{4.20}$$

avec σ une permutation de $\mathfrak{S}_n$. Soit $(D^{(\varphi(k))})_{k\in\mathbb{N}}$ une sous-suite extraite de $(D^{(k)})_{k\in\mathbb{N}}$ convergeant vers une matrice D. On alors, en vertu de (4.19),

$$\lim_{k\to+\infty} A^{(\varphi(k))} = D,$$

et donc

$$\forall \lambda \in \mathbb{C},\ \det(D - \lambda I_n) = \lim_{k\to+\infty} \det(A^{(\varphi(k))} - \lambda I_n).$$

Les matrices A et $A^{(\varphi(k))}$, $k \geq 0$, étant orthogonalement semblables, on a également

$$\forall k \in \mathbb{N},\ \det(A^{(\varphi(k))} - \lambda I_n) = \det(A - \lambda I_n),\ \forall \lambda \in \mathbb{C},$$

et on en déduit que les matrices A et D ont les mêmes polynômes caractéristiques et donc les mêmes valeurs propres, comptées avec leurs multiplicités. Comme D est une matrice diagonale, il existe bien une permutation σ telle que D est de la forme (4.20).

On vérifie ensuite que

$$\forall k \in \mathbb{N},\ a_{ii}^{(k+1)} - a_{ii}^{(k)} = \begin{cases} 0 & \text{si } i \neq p^{(k)}, q^{(k)}, \\ -\tan(\theta^{(k)})\, a^{(k)}_{p^{(k)}q^{(k)}} & \text{si } i = p^{(k)}, \\ \tan(\theta^{(k)})\, a^{(k)}_{p^{(k)}q^{(k)}} & \text{si } i = q. \end{cases}$$

Or, comme $\left|\theta^{(k)}\right| \leq \frac{\pi}{4}$ et $\left|a^{(k)}_{p^{(k)}q^{(k)}}\right| \leq \mathrm{off}(A^{(k)})$, on en conclut, en utilisant une nouvelle fois (4.19), que

$$\lim_{k\to+\infty} (D^{(k+1)} - D^{(k)}) = 0.$$

Enfin, la suite $(D^{(k)})_{k\in\mathbb{N}}$ est bornée puisque

$$\forall k \in \mathbb{N},\ \|D^{(k)}\|_F \leq \|A^{(k)}\|_F = \|A\|_F.$$

La suite $(D^{(k)})_{k\in\mathbb{N}}$, satisfaisant à toutes les hypothèses du lemme 4.11, converge donc vers une de ses valeurs d'adhérence, qui est de la forme (4.20). Le résultat est alors démontré, puisqu'on a

$$\lim_{k\to+\infty} A^{(k)} = \lim_{k\to+\infty} D^{(k)},$$

par (4.19). □

Donnons maintenant un résultat sur la convergence des vecteurs propres.

Théorème 4.12 (convergence vers les vecteurs propres pour la méthode de Jacobi) *On suppose que toutes les valeurs propres de la matrice A sont distinctes. Alors, la suite $(O^{(k)})_{k\in\mathbb{N}}$, avec $O^{(k)} = G^{(0)}G^{(1)}\dots G^{(k)}$, $k \geq 0$, de matrices orthogonales construites par la méthode de Jacobi converge vers une matrice orthogonale dont les colonnes constituent un ensemble de vecteurs propres orthonormaux de A.*

DÉMONSTRATION. Soit $(O^{(\varphi(k))})_{k\in\mathbb{N}}$ une sous-suite de $(O^{(k)})_{k\in\mathbb{N}}$ convergeant vers une matrice orthogonale O'. D'après le théorème précédent, il existe une permutation σ de $\mathfrak{S}_n$ telle que

$$\begin{pmatrix} \lambda_{\sigma(1)} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_{\sigma(n)} \end{pmatrix} = \lim_{k\to+\infty} (O^{(\varphi(k))})^\top A O^{(\varphi(k))} = (O')^\top A O',$$

et les valeurs d'adhérence de la suite $(O^{(k)})_{k\in\mathbb{N}}$ sont donc de la forme $\begin{pmatrix} \boldsymbol{v}_{\sigma(1)} & \dots & \boldsymbol{v}_{\sigma(n)} \end{pmatrix}$, où les vecteurs $\boldsymbol{v}_i$, $1 \le i \le n$, sont les colonnes de la matrice orthogonale O intervenant dans la relation

$$O^\top A O = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix},$$

et sont en nombre fini car les valeurs propres de la matrice A sont simples.

Par construction, le réel $\theta^{(k)}$ vérifie $\left|\theta^{(k)}\right| \le \frac{\pi}{4}$ et

$$\tan(2\theta^{(k)}) = \frac{2\, a^{(k)}_{p^{(k)}q^{(k)}}}{a^{(k)}_{q^{(k)}q^{(k)}} - a^{(k)}_{p^{(k)}p^{(k)}}}.$$

En utilisant le théorème précédent et le fait que les valeurs propres de A sont toutes distinctes, on obtient, pour k assez grand, que

$$\left|a^{(k)}_{q^{(k)}q^{(k)}} - a^{(k)}_{p^{(k)}p^{(k)}}\right| \ge \frac{1}{2} \min_{1\le i<j\le n} \left|\lambda_i - \lambda_j\right| > 0.$$

Comme on sait que $\lim\limits_{k\to+\infty} a^{(k)}_{p^{(k)}q^{(k)}} = 0$, on a établi que $\lim\limits_{k\to+\infty} \theta^{(k)} = 0$ et par conséquent

$$\lim_{k\to+\infty} G^{(k)} = \lim_{k\to+\infty} G(p^{(k)}, q^{(k)}, \theta^{(k)}) = I_n.$$

On a alors

$$\lim_{k\to+\infty} (O^{(k+1)} - O^{(k)}) = \lim_{k\to+\infty} O^{(k)}(G^{(k)} - I_n) = 0,$$

la suite $(O^{(k)})_{k\in\mathbb{N}}$ étant bornée ($\|O^{(k)}\|_F = \sqrt{n}$, $k \ge 0$). On termine la démonstration en appliquant le lemme 4.11. □

Pour une preuve de la convergence de la méthode dans le cas où des valeurs propres possèdent une multiplicité plus grande que un, on consultera [vKem66a].

L'inégalité (4.18) montre que la convergence de la méthode de Jacobi est *linéaire* (voir la section 5.2), mais en pratique, le taux de convergence asymptotique de la méthode est bien meilleur. On peut en fait montrer (voir par exemple [Hen58]) que, pour k assez grand, il existe une constante $C > 0$ telle que

$$\operatorname{off}(A^{(k+n(n-1)/2)}) \le C \operatorname{off}(A^{(k)})^2.$$

On a coutume de donner le nom de *balayage* (*sweep* en anglais) à $\frac{1}{2}n(n-1)$ annulations successives effectuées par des rotations de Jacobi. La dernière inégalité traduit alors le fait que la convergence de la méthode, observée après chaque balayage et un nombre suffisant d'itérations, est *(asymptotiquement) quadratique*.

Abordons à présent la question du coût de la méthode. À chaque itération, il faut tout d'abord déterminer un élément hors-diagonal de plus grande valeur absolue, ce qui requiert de comparer entre eux $\frac{1}{2}n(n-1)$ nombres réels. Une fois le couple d'entier $(p^{(k)}, q^{(k)})$ obtenu, on procède à l'évaluation des quantités $\cos(\theta^{(k)})$ et $\sin(\theta^{(k)})$ au moyen des formules (4.15), ce qui nécessite au plus quatre additions et soustractions, quatre multiplications, trois divisions et deux extractions de racines carrées. Les valeurs de ces coefficients étant connues, il est ensuite inutile d'assembler une matrice de Givens pour l'application de la relation de récurrence (4.16). Les formules (4.14) montrent en effet que, pour arriver à $A^{(k+1)}$ en modifiant quatre lignes et colonnes de la matrice $A^{(k)}$, on a besoin d'environ $4n$ multiplications et $2n$ additions et soustractions, soit $O(n)$ opérations (il en va de même pour le calcul de la matrice $O^{(k+1)}$, qui implique la modification de deux lignes de la matrice $O^{(k)}$).

Il ressort de ce compte d'opérations que la recherche d'un élément hors-diagonal à éliminer s'avère *a priori* être la partie la plus coûteuse à chaque étape, car elle nécessite de l'ordre de $\frac{n^2}{2}$ comparaisons. Une suggestion d'implémentation [24], due à Corbató [25] [Cor63], permet néanmoins d'effectuer cette tâche avec un coût dépendant

24. Celle-ci consiste en l'introduction à chaque étape d'un tableau d'entiers, noté $j^{(k)}$, de taille $n-1$, dont le i^{e} élément $j^{(k)}(i)$ contient le numéro de la colonne dans laquelle se trouve l'élément hors-diagonal possédant la plus grande valeur absolue dans la i^{e} ligne de la matrice $A^{(k)}$. Connaissant ce tableau, il est facile de trouver l'élément à annuler, en réalisant seulement $n-2$ comparaisons. S'il coûte $O(n^2)$ comparaisons pour construire $j^{(0)}$ à la première itération, la mise à jour de ce tableau ne requiert ensuite plus que $O(n)$ comparaisons (on doit en effet parcourir les deux lignes modifiées, d'indices respectifs $p^{(k)}$ et $q^{(k)}$, ainsi que, pour tenir compte des colonnes modifiées, toute ligne dont l'indice i est tel que $j^{(k)}(i) \in \{p^{(k)}, q^{(k)}\}$, cas dont on peut raisonnablement espérer que l'occurrence est rare).

25. Fernando José Corbató (né le 1er juillet 1926) est un informaticien américain, connu pour ses développements dans le domaine des systèmes d'exploitation à temps partagé.

linéairement de l'entier n, moyennant le stockage d'un tableau auxiliaire à une dimension et un effort de programmation supplémentaire (le nombre d'instructions faisant plus que doubler). En procédant de cette manière, chaque balayage demande au total $O(n^3)$ opérations élémentaires, la convergence de la méthode (à une tolérance fixée près) étant typiquement observée après quelques balayages.

Méthode de Jacobi cyclique

On peut se passer de la recherche systématique d'un élément hors-diagonal ayant la plus grande valeur propre en procédant directement à l'annulation de tous les coefficients d'une ligne (resp. d'une colonne), puis en passant à l'annulation de ceux de la suivante, etc. en réalisant toujours dans le même ordre les balayages successifs. On parle alors de *méthode de Jacobi cyclique par lignes* (resp. *par colonnes*). Si, au cours d'un balayage, l'un des coefficients s'avère être déjà nul, on passe simplement au suivant (ce qui équivaut à faire le choix $\theta^{(k)} = 0$). La convergence de la méthode de Jacobi cyclique reste quadratique au sens donné plus haut (voir [Hen58, Sch61, Wil62, vKem66b]).

Une variante de cette méthode, proposée dans l'article [PT57] et nécessitant moins d'opérations arithmétiques, consiste à omettre d'annuler tout coefficient dont la valeur absolue est inférieure à un certain seuil (*threshold* en anglais), fixé à chaque balayage et qui va en diminuant ; il semble en effet superflu d'annuler des éléments déjà « petits » en valeur absolue alors que d'autres sont d'un ordre de grandeur bien plus élevé.

4.5.2 Méthode de Givens–Householder **

Les valeurs propres d'une matrice réelle symétrique étant réelles, on peut en théorie les déterminer en cherchant les racines du polynôme caractéristique qui lui est associé sur la droite réelle. Cependant, comme vu en introduction de chapitre, cette approche est extrêmement sensible aux erreurs d'arrondi dès lors que cette détermination est faite à partir de la connaissance des coefficients du polynôme, le problème sous-jacent étant mal conditionné (voir la sous-section 1.4.2).

Pour pallier ce problème, une approche consiste à évaluer le polynôme caractéristique en divers points pour localiser les valeurs propres, sans jamais pour cela utiliser ses coefficients. Ceci est rendu possible par une propriété particulière des valeurs propres d'une matrice symétrique, qui est l'objet du résultat suivant (dont une preuve élémentaire est proposée dans [Hwa04]).

Théorème 4.13 (« théorème d'entrelacement de Cauchy ») *Soit A une matrice symétrique d'ordre n. Les valeurs propres $\{\lambda_i(A_k)\}_{i=1,\dots,k}$ des sous-matrices*

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix},\ 1 \le k \le n,$$

associées aux mineurs principaux de A vérifient

$$\lambda_1(A_{k+1}) \le \lambda_1(A_k) \le \lambda_2(A_{k+1}) \le \dots \le \lambda_k(A_k) \le \lambda_{k+1}(A_{k+1}),\ 1 \le k \le n-1. \tag{4.21}$$

*On dit encore qu'elle sont **entrelacées** (**interlaced** en anglais).*

La *méthode de Givens–Householder* [Giv54], que nous allons maintenant décrire, exploite cette spécificité. Elle comprend deux étapes distinctes. Tout d'abord, la matrice du problème est ramenée [26] sous forme tridiagonale au moyen de transformations orthogonales. Il est alors aisé d'évaluer, au moyen d'une formule de récurrence, les polynômes caractéristiques des sous-matrices extraites associés aux mineurs principaux de la matrice obtenue, qui forment ce que l'on appelle une *suite de Sturm* [27] [Stu29]. Les propriétés de cette suite permettant de déterminer le nombre de racines du polynôme caractéristique de A (et donc de valeurs propres) distinctes contenues dans un

26. Contrairement au cas de la forme diagonale que l'on cherche à atteindre dans la méthode de Jacobi, une matrice symétrique peut être transformée en une matrice tridiagonale orthogonalement semblable en un nombre *fini* d'opérations arithmétiques.

27. Jacques Charles François Sturm (29 septembre 1803 - 15 décembre 1855) était un mathématicien français d'origine allemande. Il est connu pour le théorème portant son nom, permettant de déterminer le nombre de racines réelles d'un polynôme contenues dans un intervalle donné, et la théorie de Sturm–Liouville, qui concerne les équations différentielles linéaires scalaires du second ordre d'un type particulier. Il s'intéressa également à la compressibilité des liquides et réalisa, avec Jean-Daniel Colladon, la première mesure expérimentale directe de la vitesse du son dans l'eau.

intervalle donné, on est en mesure d'encadrer, avec une précision théoriquement arbitraire, une valeur propre choisie en utilisant la *méthode de dichotomie* (voir la sous-section 5.3.1).

La matrice du problème étant symétrique, sa réduction sous forme tridiagonale se résume à l'application d'une méthode de réduction sous de Hessenberg supérieure (une matrice de Hessenberg symétrique étant nécessairement tridiagonale). Pour cette dernière, on peut envisager d'utiliser soit des rotations de Givens, introduites dans la sous-section 4.5.1) et déjà employées dans la méthode de Jacobi, soit des réflexions de Householder (introduites la sous-section 2.6.3), comme suggéré dans [HB59]. Les deux approches possèdent la même stabilité numérique et l'on choisit généralement de procéder avec des transformations de Householder, dont le coût est environ deux fois moindre de celui des transformations de Givens.

On se trouve alors ramené à considèrer une matrice réelle tridiagonale symétrique d'ordre n

$$A = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & 0 & \beta_{n-1} & \alpha_n \end{pmatrix}, \tag{4.22}$$

sous l'hypothèse [28] d'irréductibilité (au sens de la définition A.112)

$$\forall i \in \{1, \dots, n-1\},\ \beta_i \neq 0. \tag{4.23}$$

Soit la famille de polynômes $\left\{p^{(k)}\right\}_{k=0,\dots,n}$ définie par

$$\forall \lambda \in \mathbb{R},\ p^{(0)}(\lambda) = 1,\ p^{(1)}(\lambda) = \alpha_1 - \lambda, \tag{4.24}$$

ainsi que la relation de récurrence.

$$\forall \lambda \in \mathbb{R},\ \forall k \in \{2, \dots, n\},\ p^{(k)}(\lambda) = (\alpha_k - \lambda)\, p^{(k-1)}(\lambda) - {\beta_{k-1}}^2\, p^{(k-2)}(\lambda). \tag{4.25}$$

On a le résultat suivant.

Lemme 4.14 *Soit A une matrice d'ordre n de la forme* (4.22) *vérifiant* (4.23). *Pour tout entier naturel k appartenant à $\{1, \dots, n\}$ et tout réel λ, le mineur principal d'ordre k de la matrice $A - \lambda I_n$ vaut $p^{(k)}(\lambda)$, le polynôme $p^{(k)}$ étant défini par les relations* (4.24) *et* (4.25).

DÉMONSTRATION. Comme précédemment, pour tout entier k de $\{1, \dots, n\}$, on note A_k la sous-matrice principale d'ordre k extraite de la matrice A. Pour $k = 1$, on a $A_1 = \alpha_1$ et le résultat est évident. Pour $k = 2$, on obtient

$$\forall \lambda \in \mathbb{R},\ \det(A_2 - \lambda I_2) = \begin{vmatrix} \alpha_1 - \lambda & \beta_1 \\ \beta_1 & \alpha_2 - \lambda \end{vmatrix} = (\alpha_1 - \lambda)(\alpha_2 - \lambda) - {\beta_1}^2 = (\alpha_2 - \lambda)\, p^{(1)}(\lambda) - {\beta_1}^2\, p^{(0)}(\lambda) = p^{(2)}(\lambda),$$

par définition de $p^{(2)}$, puisque $p^{(0)}(\lambda) = 1$ et $p^{(1)} = \alpha_1 - \lambda$.

Choisissons à présent un entier k de $\{3, \dots, n\}$ et supposons que le résultat est vérifié pour tout entier naturel compris entre 1 et k. En développant le mineur principal d'ordre k de $A - \lambda I_n$ par rapport à sa dernière ligne, il vient

$$\begin{aligned} \det(A_k - \lambda I_k) &= (-1)^{2k-1} \beta_{k-1} \begin{vmatrix} \alpha_1 - \lambda & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \beta_{k-3} & 0 \\ \vdots & \ddots & \ddots & \alpha_{k-2} - \lambda & 0 \\ 0 & \dots & 0 & \beta_{k-2} & \beta_{k-1} \end{vmatrix} + (-1)^{2k} (\alpha_k - \lambda) \det(A_{k-1} - \lambda I_{k-1}) \\ &= (-1)^{4k-3} {\beta_{k-1}}^2 \det(A_{k-2} - \lambda I_{k-2}) + (\alpha_k - \lambda) \det(A_{k-1} - \lambda I_{k-1}) \\ &= -{\beta_{k-1}}^2 p^{(k-2)}(\lambda) + (\alpha_k - \lambda)\, p^{(k-1)}(\lambda) \\ &= p^{(k)}(\lambda), \end{aligned}$$

28. Dans le cas contraire, la résolution du problème aux valeurs propres se ramène celle de deux (ou davantage de) problèmes aux valeurs propres de taille moindre.

en utilisant l'hypothèse de récurrence et la formule (4.25) définissant $p^{(k)}$. □

Il découle de ce lemme que l'évaluation du polynôme caractéristique associé à la matrice A en un point λ se résume à celle de $p^{(n)}(\lambda)$, ce qui nécessite, par l'utilisation des formules (4.24) et (4.25), $3(n-1)$ multiplications et $2n-1$ soustractions.

Sous l'hypothèse d'irréductibilité de la matrice A, il est par ailleurs possible d'affiner le résultat du théorème 4.13 en montrant que les valeurs propres des sous-matrices A_k associées aux mineurs principaux de A sont en réalité *strictement* entrelacées. Ceci est l'objet de la prochaine proposition.

Proposition 4.15 *Pour entier k dans $\{1,\dots,n-1\}$, le polynôme $p^{(k)}$ possède k racines distinctes qui séparent strictement celles du polynôme $p^{(k+1)}$.*

DÉMONSTRATION. Il découle de (4.24) que le polynôme $p^{(1)}$ admet pour unique racine α_1. Par ailleurs, on a d'une part

$$p^{(2)}(\alpha_1) = -{\beta_1}^2 p^{(0)}(\alpha_1) = -{\beta_1}^2 < 0$$

d'après l'hypothèse d'irréductibilité (4.23), et d'autre part

$$\lim_{|\lambda|\to+\infty} p^{(2)}(\lambda) = +\infty,$$

le monôme de plus haut degré de $p^{(2)}(\lambda)$ étant λ^2. On déduit du théorème des valeurs intermédiaires que le polynôme $p^{(2)}$ possède deux racines, l'une strictement inférieure à α_1, l'autre strictement supérieure.

Supposons à présent que l'assertion est vraie pour un entier k de $\{2,\dots,n-1\}$ et notons $\lambda_1^{(k)} < \lambda_2^{(k)} < \cdots < \lambda_k^{(k)}$ les racines du polynôme $p^{(k)}$. Pour tout entier i de $\{1,\dots,k\}$, on a, en vertu de (4.25),

$$p^{(k+1)}(\lambda_i^{(k)}) = (\alpha_{k+1} - \lambda_i^{(k)})\, p^{(k)}(\lambda_i^{(k)}) - {\beta_k}^2\, p^{(k-1)}(\lambda_i^{(k)}) = -{\beta_k}^2\, p^{(k-1)}(\lambda_i^{(k)}).$$

Il découle alors de (4.23) que l'on a soit $p^{(k+1)}(\lambda_i^{(k)})p^{(k-1)}(\lambda_i^{(k)}) < 0$, soit $p^{(k+1)}(\lambda_i^{(k)}) = p^{(k-1)}(\lambda_i^{(k)}) = 0$. Dans ce dernier cas, un raisonnement par récurrence simple montre qu'alors

$$p^{(k+1)}(\lambda_i^{(k)}) = p^{(k)}(\lambda_i^{(k)}) = \cdots = p^{(1)}(\lambda_i^{(k)}) = p^{(0)}(\lambda_i^{(k)}) = 0,$$

ce qui est absurde puisque $p^{(0)} \equiv 1$. Le polynôme $p^{(k-1)}$ possédant une racine entre $\lambda_i^{(k)}$ et $\lambda_i^{(k)}$, $i \in \{2,\dots,k\}$, il change de signe sur l'intervalle correspondant et l'on en déduit par conséquent qu'il en va de même pour le polynôme $p^{(k+1)}$, qui admet donc $k-1$ racines distinctes et séparées par les racines de $p^{(k)}$. En observant que

$$\forall j \in \{1,\dots,n\},\ \lim_{\lambda\to-\infty} p^{(j)}(\lambda) = +\infty \text{ et } \lim_{\lambda\to+\infty} p^{(j)}(\lambda) = \begin{cases} +\infty & \text{si } j \text{ est pair} \\ -\infty & \text{si } j \text{ est impair} \end{cases},$$

il vient que $p^{(k-1)}(\lambda_1^{(k)}) > 0$, d'où $p^{(k+1)}(\lambda_1^{(k)}) < 0$, et que $p^{(k-1)}(\lambda_k^{(k)}) > 0$ (resp. $p^{(k-1)}(\lambda_k^{(k)}) < 0$) si k est impair (resp. si k est pair), d'où $p^{(k+1)}(\lambda_k^{(k)}) < 0$ (resp. $p^{(k+1)}(\lambda_k^{(k)}) > 0$). Le polynôme $p^{(k+1)}$ admet donc nécessairement une racine située à gauche de $\lambda_1^{(k)}$ et une autre située à droite de $\lambda_k^{(k)}$. □

Cette dernière propriété permet alors de compter le nombre de valeurs propres d'une matrices contenues dans un intervalle donné, en vertu du résultat suivant.

Proposition 4.16 *A REPRENDRE le nombre de changements de signe*[29] *entre des termes consécutifs de la suite $S_\lambda = \left\{p^{(0)}(\lambda), p^{(1)}(\lambda), \dots, p^{(n)}(\lambda)\right\}$ donne le nombre de valeurs propres de A strictement plus petites que λ.*

DÉMONSTRATION. A ECRIRE par récurrence □

A VOIR on dit la famille $p^{(k)}$ forme (ou a la propriété d') une *suite de Sturm* On dispose alors d'un critère permettant d'appliquer la méthode de dichotomie l'approximation des valeurs propres de A.

EXPLICATIONS

initialisation : plus particulièrement adaptée à l'approximation de valeurs propres contenues dans un intervalle déterminé *a priori* (par l'un des résultats de la section 4.2 par exemple).

On applique ensuite la relation eqref à une succession de valeurs du réel x en comptant les changements de signe pour localiser les valeurs propres dans des intervalles arbitrairement petits (coût $O(n)$ à chaque itération)

Détermination d'un vecteur propre une fois déterminée une valeur numérique précise d'une valeur propre : formule explicite en fonction des polynômes $p^{(k)}$: évaluation souvent catastrophique
itération inverse

29. En raison de la propriété ..., on adopte la convention que le signe de $p^{(i)}(\lambda)$ est l'opposé de celui de $p^{(i-1)}(\lambda)$ si $p^{(i)}(\lambda) = 0$.

4.6 Méthodes pour la calcul de la décomposition en valeurs singulières ***

on peut se ramener à un problème aux valeurs propres pour A^*A/AA^* pour obtenir numériquement la décomposition (explications), mais cette approche pose gééralement des problèmes de stabilité numérique.

À la base des techniques utilisées aujourd'hui : la *méthode de Golub–Kahan* [GK65], matrices $m \times n$ avec $m \geq n$: dans une première étape, la matrice est ramenée à une forme bidiagonale, ce qui peut être fait en utilisant des transformations de Householder avec un coût de $4mn^2 - 4n^3/3$ flops, les valeurs singulières sont alors obtenues en appliquant la méthode Givens à une matrice tridiagonale particulière, les vecteurs colonnes formant les matrices U et V sont ensuite obtenues via un procédé de déflation

amélioration possible du compte d'opérations de la bidiagonalisation par le procédé de Lawson–Hanson–Chan : d'abord réduire la matrice A sous forme triangulaire via une factorisation QR, puis d'utiliser les transformations de Householder pour la ramener sous forme bidiagonale, le coût combiné est de $2mn^2 + 2n^3$ flops avantageux si $m > 5n/3$

La seconde étape de détermination peut se faire par une variante de l'algorithme QR pour le calcul de valeurs propres (de l'ordre de $O(m)$ itérations, chacune coûtant $O(m)$ flops), donnant lieu à la *méthode de Golub–Reinsch* [GR70])

méthode pour une haute précision [DK90]

Application en compression d'image (voir Figure 4.3)

FIGURE 4.3 – Exemple de compression d'image utilisant la décomposition en valeurs singulières. La première image numérique tout à gauche de la première ligne est l'originale (ici une version en niveau de gris de la fameuse *Lenna* [30] (ou *Lena*)), constituée de 512 par 512 pixels (abbréviation de la locution anglaise *picture elements*) et représentée en machine par une matrice carrée d'ordre 512 contenant des entiers codant la luminosité de chaque pixel (les valeurs de ces entiers sont ici comprises entre 0, pour le noir, et 255, pour le blanc). Les images suivantes correspondent à la meilleure approximation (au sens de la norme de Frobenius ou de la norme spectrale) de la matrice de l'image d'origine par une matrice de rang respectivement égal à (en allant de gauche à droite) 1, 2, 4 et 8 pour la première ligne et 16, 32, 64, 128 et 256 pour la seconde ; elles ont été obtenues grâce à une décomposition en valeurs singulières.

30. Cette image numérique sert classiquement de test pour les algorithmes de traitement d'image. Son histoire est relatée dans l'article [Hut01].

4.7 Notes sur le chapitre

Il semble que la première utilisation de la méthode de la puissance, pour le calcul de la plus grande valeur propre d'une matrice réelle symétrique définie positive, soit due à Müntz [31] [Mün13], mais cette méthode est généralement attribuée à von Mises [vMP29]. La méthode de la puissance inverse fut introduite par Wielandt en 1944 pour la détermination numérique des vitesses d'écoulement causant l'entrée en résonance d'une aile d'avion [Wie44a].

Lorsque la matrice A n'est pas diagonalisable, il est possible d'obtenir des résultats de convergence pour la méthode de la puissance dans certains cas particulier (voir [PP73]).

AJOUTER un paragraphe sur la *méthode d'Arnoldi* [32] [Arn51]

La méthode de Jacobi, décrite pour la première fois dans [Jac46] où elle est utilisée pour diagonaliser une matrice réelle symétrique d'ordre sept, est certainement la plus ancienne méthode de calcul de l'ensemble des valeurs et vecteurs propres d'une matrice. Elle possède la plupart des caractéristiques des méthodes plus sophistiquées introduites par la suite, en particulier l'utilisation de transformations orthogonales. Quelque peu négligée durant une centaine d'années, elle fut redécouverte vers 1949 (voir [GMvN59]) et connut alors un regain d'intérêt, l'apparition d'ordinateurs permettant sa mise en œuvre sur des matrices de taille plus importante [Gre53]. Elle fût ensuite l'objet de recherches actives (notamment en termes de l'étude de sa vitesse de convergence, de sa généralisation à d'autres types de matrices que les matrices réelles symétriques, etc.) des années 1950 jusqu'au début des années 1990. Bien que convergeant moins rapidement que d'autres méthodes, elle reste d'intérêt en raison du fait qu'elle peut être implantée efficacement sur un calculateur parallèle. Le lecteur intéressé par ces différents aspects pourra consulter, en plus de celles déjà citées, les références fournies dans [GV13, Section 8.5].

La *méthode QR* [Fra61, Fra62, Kub62], basée sur la factorisation du même nom (introduite dans la sous-section 2.6.3), est une des méthodes de référence pour le calcul de toutes les valeurs propres d'une matrice. Elle reprend le principe de transformer la matrice dont on cherche les valeurs propres en une matrice orthogonalement (dans le cas réel, unitairement sinon) semblable, de façon à construire une suite convergeant, moyennant certaines hypothèses, vers la décomposition de Schur de cette matrice. Étant donné une matrice A carrée quelconque, l'algorithme de la méthode est le suivant : à partir de l'initialisation $A^{(0)} = A$, on réalise une factorisation QR

$$\forall k \in \mathbb{N}^*,\ A^{(k-1)} = Q^{(k-1)}R^{(k-1)}, \tag{4.26}$$

et l'on pose

$$\forall k \in \mathbb{N}^*,\ A^{(k)} = R^{(k-1)}Q^{(k-1)},$$

formant ainsi une suite de matrices $(A^{(k)})_{k\in\mathbb{N}}$, orthogonalement ou unitairement selon les cas, semblables à A et ayant le cas échéant pour limite une matrice triangulaire supérieure. La factorisation QR ayant lieu à chaque étape possède (pour une matrice d'ordre n) une complexité en $O(n^3)$ et rend la méthode *a priori* prohibitive, mais ce coût peut être abaissé en effectuant une réduction préliminaire de la matrice $A^{(0)}$ sous la forme d'une matrice de Hessenberg supérieure (voir la définition 2.8). On montre alors que l'étape (4.26) de l'algorithme ne requiert plus que de l'ordre de n^2 opérations. Un des avantages notables de cette méthode est sa stabilité numérique, héritée de la similitude orthogonale (ou unitaire) des matrices de la suite construite : le conditionnement du problème aux valeurs propres pour toute matrice $A^{(k)}$ de la suite est au moins aussi bon que celui pour A. De nombreuses améliorations et variations de cette méthode existent, notamment la *méthode QR avec translations*, et nous renvoyons le lecteur intéressé à la littérature spécialisée.

À l'origine de la méthode QR, on trouve la *méthode LR* [Rut55, Rut58], qui utilise des transformations *non* orthogonales. Dans cette méthode, le fonctionnement de l'algorithme est identique, mais on effectue à chaque étape une factorisation LU [33] (voir une nouvelle fois le chapitre 2, section 2.5) de la matrice courante, *i.e.*

$$\forall k \in \mathbb{N}^*,\ A^{(k-1)} = L^{(k-1)}U^{(k-1)},$$

31. (Chaim) Herman Müntz (28 août 1884 - 17 avril 1956) était un mathématicien allemand, connu pour ses travaux en théorie de l'approximation et en géométrie projective.

32. Walter Edwin Arnoldi (14 décembre 1917 - 5 octobre 1995) était un ingénieur américain. Il s'intéressa à l'aérodynamique, l'acoustique et la modélisation des vibrations des hélices d'aéronefs.

33. En allemand, *matrice triangulaire inférieure* se traduit par *Linksdreiecksmatrix* et *matrice triangulaire supérieure* par *Rechtsdreiecksmatrix*,ce qui explique le nom donné à la méthode.

pour poser ensuite

$$A^{(k)} = U^{(k-1)}L^{(k-1)}.$$

Comme pour la méthode QR, on peut montrer que, sous certaines conditions, la suite de matrices $(A^{(k)})_{k\in\mathbb{N}}$ converge vers la décomposition de Schur de la matrice A. Cette méthode est aujourd'hui rarement employée, à cause de difficultés liées à la factorisation LU, qui peuvent se résoudre simplement en faisant appel à la factorisation $PA = LU$, et surtout de problèmes d'instabilité numérique. L'article [GP11] revient sur la genèse de cette méthode.

L'article [GvdV00] propose un état de l'art des techniques modernes de calcul de valeurs propres, doublé d'une mise en perspective historique et d'une bibliographie exhaustive.

Références

[Arn51] W. E. ARNOLDI. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9(1):17–29, 1951. DOI: 10.1090/qam/42792.

[BF60] F. L. BAUER and C. T. FIKE. Norms and exclusion theorems. *Numer. Math.*, 2(1):137–141, 1960. DOI: 10.1007/BF01386217.

[BP98] S. BRIN and L. PAGE. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN Systems*, 30(1-7):107–117, 1998. DOI: 10.1016/S0169-7552(98)00110-X.

[Cor63] F. J. CORBATÓ. On the coding of Jacobi's method for computing eigenvalues and eigenvectors of real symmetric matrices. *J. ACM*, 10(2):123–125, 1963. DOI: 10.1145/321160.321161.

[DC34] W. J. DUNCAN and A. R. COLLAR. A method for the solution of oscillation problems by matrices. *Lond. Edinb. Dublin Philos. Mag. J. Sci. (7)*, 17(115):865–909, 1934. DOI: 10.1080/14786443409462445.

[DK90] J. DEMMEL and W. KAHAN. Accurate singular values of bidiagonal matrices. *SIAM J. Sci. Statist. Comput.*, 11(5):873–912, 1990. DOI: 10.1137/0911052.

[FDC38] R. A. FRAZER, W. J. DUNCAN, and A. R. COLLAR. *Elementary matrices and some applications to dynamics and differential equations*. Cambridge University Press, 1938. DOI: 10.1017/CBO9780511629211.

[Fra61] J. G. F. FRANCIS. The QR transformation: a unitary analogue to the LR transformation – Part 1. *Comput. J.*, 4(3):265–271, 1961. DOI: 10.1093/comjnl/4.3.265.

[Fra62] J. G. F. FRANCIS. The QR transformation – Part 2. *Comput. J.*, 4(4):332–345, 1962. DOI: 10.1093/comjnl/4.4.332.

[Fro12] G. FROBENIUS. Über Matrizen aus nicht negativen Elementen. *Sitzungsber. Königl. Preuss. Akad. Wiss. Berlin*:456–477, 1912.

[Gau96] W. GAUTSCHI. Orthogonal polynomials: applications and computation. *Acta Numerica*, 5:45–119, 1996. DOI: 10.1017/S0962492900002622.

[Ger31] S. GERSCHGORIN. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk SSSR Ser. Mat.*, 1(6):749–754, 1931.

[Giv54] W. GIVENS. Numerical computation of the characteristic values of a real symmetric matrix. Technical report ORNL 1574, Oak Ridge National Laboratory, 1954. DOI: 10.2172/4412175.

[GK12] M. J. GANDER and F. KWOK. Chladni figures and the Tacoma bridge: motivating PDE eigenvalue problems via vibrating plates. *SIAM Rev.*, 54(3):573–596, 2012. DOI: 10.1137/10081931X.

[GK65] G. GOLUB and W. KAHAN. Calculating the singular values and pseudo-inverse of a matrix. *J. SIAM Ser. B Numer. Anal.*, 2(2):205–224, 1965. DOI: 10.1137/0702016.

[GMvN59] H. H. GOLDSTINE, F. J. MURRAY, and J. von NEUMANN. The Jacobi method for real symmetric matrices. *J. ACM*, 6(1):59–96, 1959. DOI: 10.1145/320954.320960.

[GP11] M. H. GUTKNECHT and B. N. PARLETT. From qd to LR, or, how were the qd and LR algorithms discovered? *IMA J. Numer. Anal.*, 31(3):741–754, 2011. DOI: 10.1093/imanum/drq003.

[GR70] G. H. GOLUB and C. REINSCH. Singular value decomposition and least squares solutions. *Numer. Math.*, 14(5):403–420, 1970. DOI: 10.1007/BF02163027.

[Gre53] R. T. GREGORY. Computing eigenvalues and eigenvectors of a symmetric matrix on the ILLIAC. *Math. Tables Aids Comput.*, 7(44):215–220, 1953. DOI: 10.1090/S0025-5718-1953-0057643-6.

[GV13] G. H. GOLUB and C. F. VAN LOAN. *Matrix computations*. Johns Hopkins University Press, fourth edition, 2013.

[GvdV00] G. H. GOLUB and H. A. van der VORST. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.*, 123(1-2):35–65, 2000. DOI: 10.1016/S0377-0427(00)00413-1.

[GW69] G. H. GOLUB and J. H. WELSCH. Calculation of Gauss quadrature rules. *Math. Comput.*, 23(106):221–230, 1969. DOI: 10.1090/S0025-5718-69-99647-1.

[HB59] A. S. HOUSEHOLDER and F. L. BAUER. On certain methods for expanding the characteristic polynomial. *Numer. Math.*, 1(1):29–37, 1959. DOI: 10.1007/BF01386370.

[Hen58] P. HENRICI. On the speed of convergence of cyclic and quasicyclic Jacobi methods for computing the eigenvalues of hermitian matrices. *SIAM J. Appl. Math.*, 6(2):144–162, 1958. DOI: 10.1137/0106008.

[Hot33] H. HOTELLING. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(3):417–441, 498–520, 1933. DOI: 10.1037/h0071325.

[Hut01] J. HUTCHINSON. Culture, communication, and an information age madonna. *IEEE Prof. Commun. Soc. Newsletter*, 45(3):1, 5–7, 2001.

[Hwa04] S.-G. HWANG. Cauchy's interlace theorem for eigenvalues of Hermitian matrices. *Amer. Math. Monthly*, 111(2):157–159, 2004. DOI: 10.2307/4145217.

[Jac46] C. G. J. JACOBI. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *J. Reine Angew. Math.*, 1846(30):51–94, 1846. DOI: 10.1515/crll.1846.30.51.

[Kub62] V. N. KUBLANOVSKAYA. On some algorithms for the solution of the complete eigenvalue problem. *U.S.S.R. Comput. Math. and Math. Phys.*, 1(3):637–657, 1962. DOI: 10.1016/0041-5553(63)90168-X.

[Lan50] C. LANCZOS. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards*, 45(4):255–282, 1950. DOI: 10.6028/jres.045.026.

[Lan52] C. LANCZOS. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952. DOI: 10.6028/jres.049.006.

[LM04] A. N. LANGVILLE and C. D. MEYER. Deeper inside PageRank. *Internet Math.*, 1(3):335–380, 2004. DOI: 10.1080/15427951.2004.10129091.

[Mün13] C. MÜNTZ. Solution directe de l'équation séculaire et de quelques problèmes analogues transcendants. *C. R. Acad. Sci. Paris*, 156 :43-46, 1913.

[Per07] O. PERRON. Zur Theorie der Matrices. *Math. Ann.*, 64(2):248–263, 1907. DOI: 10.1007/BF01449896.

[PP73] B. N. PARLETT and W. G. POOLE, JR. A geometric theory for the QR, LU and power iterations. *SIAM J. Numer. Anal.*, 10(2):389–412, 1973. DOI: 10.1137/0710035.

[PT57] D. A. POPE and C. TOMPKINS. Maximizing functions of rotations - experiments concerning speed of diagonalization of symmetric matrices using Jacobi's method. *J. ACM*, 4(4):459–466, 1957. DOI: 10.1145/320893.320901.

[Rut55] H. RUTISHAUSER. Une méthode pour la détermination des valeurs propres d'une matrice. *C. R. Acad. Sci. Paris*, 240(1) :34-36, 1955.

[Rut58] H. RUTISHAUSER. Solution of eigenvalue problems with the LR transformation. *Nat. Bur. Standards Appl. Math. Ser.*, 49:47–81, 1958.

[Sch61] A. SCHÖNHAGE. Zur Konvergenz des Jacobi-Verfahrens. *Numer. Math.*, 3(1):374–380, 1961. DOI: 10.1007/BF01386036.

[Stu29] C. STURM. Analyse d'un mémoire sur la résolution des équations numériques. *Bull. Férussac*, 11 :419-422, 1829.

[vKem66a] H. P. M. van KEMPEN. On the convergence of the classical Jacobi method for real symmetric matrices with non-distinct eigenvalues. *Numer. Math.*, 9(1):11–18, 1966. DOI: 10.1007/BF02165224.

[vKem66b] H. P. M. van KEMPEN. On the quadratic convergence of the special cyclic Jacobi method. *Numer. Math.*, 9(1):19–22, 1966. DOI: 10.1007/BF02165225.

[vMP29] R. von MISES und H. POLLACZEK-GEIRINGER. Praktische Verfahren der Gleichungsauflösung. *Z. Angew. Math. Mech.*, 9(1):58–77, 1929. DOI: 10.1002/zamm.19290090105.

[Wie44a] H. WIELANDT. Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme. Teil V: Bestimmung höherer Eigenwerte durch gebrochene Iteration. Technischer Bericht B 44/J/37, Aerodynamische Versuchsanstalt Göttingen, 1944.

[Wie44b] H. WIELANDT. Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben. *Math. Z.*, 50(1):93–143, 1944. DOI: 10.1007/BF01312438.

[Wil59] J. H. WILKINSON. The evaluation of the zeros of ill-conditioned polynomials. Part I. *Numer. Math.*, 1(1):150–166, 1959. DOI: 10.1007/BF01386381.

[Wil62] J. H. WILKINSON. Note on the quadratic convergence of the cyclic Jacobi process. *Numer. Math.*, 4(1):296–300, 1962. DOI: 10.1007/BF01386321.

[Wil65] J. H. WILKINSON. *The algebraic eigenvalue problem*, *Numerical mathematics and scientific computation*. Oxford University Press, 1965.

Deuxième partie

Traitement numérique des fonctions

INTRODUIRE CETTE PARTIE

Chapitre 5

Résolution numérique des équations non linéaires

Nous nous intéressons dans ce chapitre à l'approximation de zéros (ou de racines, dans le cas d'un polynôme [1]) d'une fonction réelle d'une variable réelle, c'est-à-dire, étant donné un intervalle $I \subseteq \mathbb{R}$ et une application f définie sur I et à valeurs réelles, la résolution approchée du problème : *trouver un réel ξ tel que*

$$f(\xi) = 0.$$

Ce problème intervient notamment dans l'étude générale d'une fonction d'une variable réelle, qu'elle soit motivée ou non par des applications, pour laquelle des solutions exactes de ce type d'équation ne sont pas explicitement connues [2].

Toutes les méthodes que nous allons présenter sont itératives et consistent donc en la construction d'une suite de réels $(x^{(k)})_{k\in\mathbb{N}}$ qui, on espère, sera telle que

$$\lim_{k\to+\infty} x^{(k)} = \xi.$$

En effet, à la différence du cas des systèmes linéaires, la convergence de ces méthodes itératives dépend en général du choix de l'approximation initiale $x^{(0)}$. On verra ainsi qu'on ne sait souvent qu'établir des résultats de *convergence locale*, valables lorsque $x^{(0)}$ appartient à un certain voisinage du zéro ξ.

Après avoir caractérisé la convergence de suites engendrées par les méthodes itératives présentées dans ce chapitre, en introduisant notamment la notion d'ordre de convergence, nous introduisons plusieurs méthodes parmi les plus connues et les plus utilisées : tout d'abord les méthodes de dichotomie et de la fausse position qui sont toutes deux des méthodes dites *d'encadrement* (*bracketing methods* en anglais), puis les méthodes de la corde, de Newton–Raphson [3] ou encore de Steffensen [4], qui font partie des *méthodes de point fixe* (*fixed-point methods* en anglais), et enfin la méthode de la sécante. Dans chaque cas, un ou plusieurs résultats de convergence *ad hoc* sont énoncés. Des méthodes adaptées au cas particulier des équations algébriques (c'est-à-dire polynomiales) sont abordées brièvement en fin de chapitre.

5.1 Exemples d'applications *

INTRO autres exemples : utilisation dans les méthodes de tir (voir la section 8.8)

Nous nous contentons ici de donner trois exemples, deux étant issus de la physique et le troisième de l'économie.

1. On commettra parfois dans la suite un abus de langage en appelant « *polynôme* » toute fonction polynomiale, c'est-à-dire toute application associée à un polynôme à coefficients dans un anneau commutatif (le corps $\mathbb{R}$ dans notre cas). Dans ce cas particulier, tout zéro de la fonction est une *racine* du polynôme qui lui est sous-jacent.

2. Même dans le cas d'une équation algébrique, on rappelle qu'il n'existe pas de méthode de résolution générale à partir du degré cinq.

3. Joseph Raphson (v. 1648 - v. 1715) était un mathématicien anglais. Son travail le plus notable est son ouvrage *Analysis aequationum universalis*, publié en 1690 et contenant une méthode pour l'approximation d'un zéro d'une fonction d'une variable réelle à valeurs réelles.

4. Johan Frederik Steffensen (28 février 1873 - 20 décembre 1961) était un mathématicien, statisticien et actuaire danois, dont les travaux furent principalement consacrés au calcul par différences finies et à l'interpolation.

5.1.1 Équation de Kepler

L'*équation de Kepler*[5], développée dans les deux premières décennies du dix-septième siècle, est une équation non linéaire dont la résolution permet celle d'un problème de mécanique céleste. Elle occupe une place importante à la fois dans l'histoire de la physique et dans celle des mathématiques

En analysant les relevés de position de la planète Mars autour du Soleil effectués par Tycho Brahe[6], l'astronome Johannes Kepler détermina que la trajectoire d'un satellite, c'est-à-dire un corps en orbite autour d'un autre corps plus massif sous l'effet de l'attraction gravitationnelle, avait la forme d'une ellipse dont l'un des foyers est occupé par le corps massif. En supposant que l'instant de passage au périapside (le point de l'orbite où la distance par rapport au foyer de cette orbite est minimale) est le temps t_0, la question se pose de connaître la position du satellite au temps $t > t_0$. Pour cela, on peut chercher à résoudre le système d'équations différentielles caractérisant le mouvement des deux corps (voir la sous-section 8.2.1), mais dans ce cadre simple, il est possible de procéder de manière plus directe.

En effet, la trajectoire ayant lieu dans un plan, on peut repérer la position du satellite par l'*anomalie vraie* ν, qui est l'angle, mesuré au foyer de l'orbite, entre la position courante du satellite sur son orbite et la direction du périapside (une valeur de ν égale 2π correspondant à un retour au périapside après une orbite complète). En notant T la *période orbitale*, e l'*excentricité orbitale*, $0 \leq e < 1$, et en introduisant l'*anomalie moyenne* M, définie par

$$M = \frac{2\pi(t - t_0)}{T},$$

qui exprime en radians la fraction de période écoulée depuis le dernier passage au périapside, Kepler obtint par la *loi des aires*[7] la formule

$$\tan\left(\frac{\nu}{2}\right) = \sqrt{\frac{1+e}{1-e}}\tan\left(\frac{E}{2}\right),$$

liant l'anomalie vraie à l'*anomalie excentrique* E, $0 \leq E < 2\pi$, solution de l'équation

$$M = E - e\,\sin(E). \tag{5.1}$$

Cette dernière équation a été longuement et largement étudiée (voir [Col93]). Pour la résoudre numériquement, c'est-à-dire pour trouver une valeur numérique de E dans $[0, 2\,\pi[$ à partir de la donnée de valeurs numériques pour e et pour M, il suffit de trouver le zéro de la fonction

$$f(E) = M - E + e\,\sin(E)$$

par l'une des méthodes présentées dans ce chapitre (voir par exemple la figure 5.5).

5.1.2 Équation d'état de van der Waals pour un gaz

Introduite en 1873 [vdWaa73], l'*équation d'état de van der Waals*[8] est l'une des premières tentatives de prise en compte dans une loi de comportement des interactions qui existent entre les molécules composant un gaz réel. Alors que l'équation d'état d'un gaz parfait (c'est-à-dire l'équation liant les variables d'état que sont la quantité de matière n, la température T, la pression p et le volume V du gaz) est obtenue en supposant que les molécules constituant le gaz n'interagissent les unes avec les autres que par des chocs élastiques et sont de taille négligeable par rapport à la distance intermoléculaire moyenne, et ne fournit donc une approximation correcte qu'à faible

5. Johannes Kepler (27 décembre 1571 - 15 novembre 1630) était un astronome allemand. Il est célèbre pour avoir étudié l'hypothèse héliocentrique de Copernic et découvert que les planètes ne tournaient pas en cercle parfait autour du Soleil, mais en suivant des trajectoires elliptiques.

6. Tycho Brahe (Tyge Ottesen Brahe, 14 décembre 1546 - 24 octobre 1601) était un astronome danois. Ses observations très précises des positions de la planète Mars jouèrent un rôle décisif dans l'établissement par Johannes Kepler des trois lois régissant le mouvement des planètes.

7. La loi des aires (encore appelée deuxième loi de Kepler) énonce que le segment de droite reliant les centres de gravité du corps massif (le Soleil par exemple) et d'un satellite de ce dernier (une planète par exemple) balaie des aires égales pendant des durées égales.

8. Johannes Diderik van der Waals (23 novembre 1837 - 8 mars 1923) était un physicien néerlandais. Ses travaux sur la continuité des états fluides, pour lesquels il est lauréat du prix Nobel de physique de 1910, lui ont permis de découvrir les forces de cohésion à courte distance et de décrire le comportement des gaz à diverses températures au travers de l'équation d'état portant aujourd'hui son nom.

pression, le modèle proposé par van der Waals tient compte d'un volume non nul pour les molécules et d'une force d'attraction entre celles-ci. L'équation d'état de van der Waals s'écrit alors

$$\left(p + a\left(\frac{n}{V}\right)^2\right)(V - nb) = nRT,$$

où les constantes positives a et b dépendent de la nature du gaz considéré et représentent respectivement une pression de cohésion et un covolume et où R est la *constante universelle des gaz parfaits*.

Pour déterminer le volume d'un gaz connaissant sa température et sa pression, on est amené à résoudre une équation non linéaire d'inconnue V et de fonction [9]

$$f(V) = \left(p + a\left(\frac{n}{V}\right)^2\right)(V - nb) - nRT.$$

5.1.3 Calcul du rendement moyen d'un fonds de placement

Admettons que l'on souhaite calculer le taux de rendement annuel moyen R d'un fonds de placement, en supposant que l'on a investi chaque année une somme fixe de V euros dans le fonds et que l'on se retrouve après n années avec un capital d'un montant de M euros. La relation liant M, n, R et V est

$$M = V\sum_{k=1}^{n}(1+R)^k = V\frac{1+R}{R}\left((1+R)^n - 1\right),$$

et on doit alors trouver R tel que $f(R) = M - V\frac{1+R}{R}\left((1+R)^n - 1\right) = 0$.

POUR LA SECTION SUIVANTE : Comme on l'a dit (?), les méthodes de résolution que nous allons présenter sont de type itératif, c'est-à-dire qu'elle fournissent *a priori* la solution après un nombre infini d'étapes.

5.2 Ordre de convergence d'une méthode itérative

Afin de pouvoir évaluer à quelle « vitesse » la suite construite par une méthode itérative converge vers sa limite (ce sera souvent l'un des critères discriminants pour le choix d'une méthode), il nous faut introduire quelques notions.

Définition 5.1 (ordre de convergence d'une suite) *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels convergeant vers une limite ξ. On dit que cette suite est* ***convergente d'ordre*** *$r \geq 1$, s'il existe deux constantes strictement positives C_1 et C_2, avec $C_1 \leq C_2$, et un entier naturel k_0 tels que*

$$\forall k \in \mathbb{N},\ k \geq k_0 \Rightarrow C_1 \leq \frac{\left|x^{(k+1)} - \xi\right|}{|x^{(k)} - \xi|^r} \leq C_2. \tag{5.2}$$

Par extension, une méthode itérative produisant une suite convergente vérifiant les relations (5.2) sera également dite *d'ordre r*. On notera que, dans plusieurs ouvrages, on trouve l'ordre d'une suite simplement défini par le fait qu'il existe une constante $C \geq 0$ et un entier k_0 tels que, pour tout entier k tel que $k \geq k_0 \geq 0$, $\left|x^{(k+1)} - \xi\right| \leq C\left|x^{(k)} - \xi\right|^r$. Il faut cependant observer [10] que cette définition n'assure pas l'unicité de r, l'ordre de convergence pouvant éventuellement être plus grand que r. On préférera donc dire dans ce cas que la suite est d'ordre *r au moins*. On remarquera aussi que, si r est égal à 1, on a nécessairement $C_2 < 1$ dans (5.2), faute de quoi la suite ne pourrait converger.

La définition 5.1 est très générale et n'exige pas que la suite $\left(\frac{\left|x^{(k+1)} - \xi\right|}{|x^{(k)} - \xi|^r}\right)_{k\in\mathbb{N}}$ admette une limite. Lorsque c'est le cas, on a coutume de se servir de la définition suivante.

9. On notera que l'on peut également se ramener à la résolution d'une équation algébrique du troisième degré, dont au moins l'une des racines est réelle.

10. On pourra considérer l'exemple de la suite positive définie par, $\forall k \in \mathbb{N}$, $x^{(k)} = \alpha^{\beta^k}$ avec $0 < \alpha < 1$ et $\beta > 1$. Cette suite est d'ordre β d'après la définition 5.1, alors que, $\forall k \in \mathbb{N}$, $x^{(k+1)} = x^{(k)^\beta} \leq x^{(k)^\gamma}$ pour $1 < \gamma < \beta$.

Définition 5.2 *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels convergeant vers une limite ξ. On dit que cette suite est* ***convergente d'ordre*** *r, avec $r > 1$, vers ξ s'il existe un réel $\mu > 0$, appelé* ***constante asymptotique d'erreur****, tel que*

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|^r} = \mu. \tag{5.3}$$

Dans le cas particulier où $r = 1$, on dit que la suite ***converge linéairement*** *si*

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|} = \mu, \text{ avec } \mu \in]0,1[,$$

et ***super-linéairement*** *(resp.* ***sous-linéairement****) si l'égalité ci-dessus est vérifiée avec $\mu = 0$ (resp. $\mu = 1$).*

Ajoutons que la convergence d'ordre deux est dite *quadratique*, celle d'ordre trois *cubique*. On parle parfois de convergence *q-linéaire*, *q-quadratique*, etc. (*q-linear convergence*, *q-quadratic convergence*, etc. en anglais), la lettre q du préfixe signifiant « quotient ».

Si la dernière caractérisation est particulièrement adaptée à l'étude pratique de la plupart des méthodes itératives que nous allons présenter dans ce chapitre, elle a comme inconvénient de ne pouvoir fournir l'ordre d'une suite dont la « vitesse de convergence » est variable, ce qui se traduit par le fait que la limite (5.3) n'existe pas. On a alors recours à une définition « étendue ».

Définition 5.3 *On dit qu'une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels* ***converge avec un ordre au moins égal à*** *r, avec $r \geq 1$, vers une limite ξ s'il existe une suite positive $(\varepsilon^{(k)})_{k\in\mathbb{N}}$ tendant vers 0 vérifiant*

$$\forall k \in \mathbb{N},\ \left|x^{(k)} - \xi\right| \leq \varepsilon^{(k)}, \tag{5.4}$$

et un réel $\nu > 0$ ($0 < \nu < 1$ si $r = 1$) tel que

$$\lim_{k\to+\infty} \frac{\varepsilon^{(k+1)}}{\varepsilon^{(k)^r}} = \nu.$$

On remarquera l'ajout du qualificatif « au moins » dans la définition 5.3, qui provient du fait que l'on a dû procéder à une majoration par une suite convergeant vers zéro avec un ordre r au sens de la définition 5.2. Bien évidemment, on retrouve la définition 5.2 si l'on a égalité dans (5.4), mais ceci est souvent impossible à établir en pratique. En anglais, on qualifie parfois cette notion de convergence en utilisant le préfixe *r* (*r-linear convergence*, *r-quadratic convergence*, etc.) représentant le mot *"root"*.

Indiquons que les notions d'ordre et de constante asymptotique d'erreur ne sont pas purement théoriques, mais en relation avec le nombre de chiffres exacts obtenus dans l'approximation de ξ. Posons en effet $\delta^{(k)} = -\log_{10}(\left|x^{(k)} - \xi\right|)$; $\delta^{(k)}$ est alors le nombre de chiffres significatifs décimaux exacts de $x^{(k)}$. Pour k suffisamment grand, on a

$$\delta^{(k+1)} \approx r\,\delta^{(k)} - \log_{10}(\mu).$$

On voit donc que si r est égal à un, on ajoute environ $-\log_{10}(\mu)$ chiffres significatifs à chaque itération. Par exemple, si $\mu = 0,999$ alors $-\log_{10}(\mu) \approx 4,34\,10^{-4}$ et il faudra près de 2500 itérations pour gagner une seule décimale. Par contre, si r est strictement plus grand que un, on multiplie environ par r le nombre de chiffres significatifs à chaque itération. Ceci montre clairement l'intérêt des méthodes d'ordre plus grand que un.

5.3 Méthodes d'encadrement

Cette première classe de méthodes repose sur la propriété fondamentale suivante, relative à l'existence d'un zéro pour une application d'une variable réelle à valeurs réelles.

Théorème 5.4 (existence d'un zéro d'une fonction à valeurs réelles continue) *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une application continue de $[a,b]$ dans $\mathbb{R}$ vérifiant $f(a)f(b) < 0$. Alors il existe un réel ξ appartenant à l'intervalle $]a,b[$ tel que $f(\xi) = 0$.*

DÉMONSTRATION. Si $f(a) < 0$, on a $0 \in]f(a), f(b)[$, sinon $f(a) > 0$ et alors $0 \in]f(b), f(a)[$. Dans ces deux cas, le résultat est une conséquence du théorème des valeurs intermédiaires (voir le théorème B.89). □

5.3.1 Méthode de dichotomie

La *méthode de dichotomie*, ou *méthode de la bissection* (*bisection method* en anglais), suppose que la fonction f est continue sur un intervalle $[a,b]$, vérifie $f(a)f(b)<0$ et admet donc (au moins) un zéro ξ dans $]a,b[$.

Son principe est le suivant. On pose $a^{(0)}=a$, $b^{(0)}=b$, on note $x^{(0)}=\frac{1}{2}(a^{(0)}+b^{(0)})$ le milieu de l'intervalle de départ et on évalue la fonction f en ce point. Si $f(x^{(0)})=0$, le point $x^{(0)}$ est le zéro de f et le problème est résolu. Sinon, si $f(a^{(0)})f(x^{(0)})<0$, alors le zéro ξ est contenu dans l'intervalle $]a^{(0)},x^{(0)}[$, alors qu'il appartient à $]x^{(0)},b^{(0)}[$ si $f(x^{(0)})f(b^{(0)})<0$. On réitère ensuite ce processus sur l'intervalle $[a^{(1)},b^{(1)}]$, avec $a^{(1)}=a^{(0)}$ et $b^{(1)}=x^{(0)}$ dans le premier cas, ou $a^{(1)}=x^{(0)}$ et $b^{(1)}=b^{(0)}$ dans le second, et ainsi de suite...

De cette manière, on construit de manière récurrente trois suites $(a^{(k)})_{k\in\mathbb{N}}$, $(b^{(k)})_{k\in\mathbb{N}}$ et $(x^{(k)})_{k\in\mathbb{N}}$ telles que $a^{(0)}=a$, $b^{(0)}=b$ et vérifiant, pour tout entier naturel k,

- $x^{(k)}=\dfrac{a^{(k)}+b^{(k)}}{2}$,
- $a^{(k+1)}=a^{(k)}$ et $b^{(k+1)}=x^{(k)}$ si $f(a^{(k)})f(x^{(k)})<0$,
- $a^{(k+1)}=x^{(k)}$ et $b^{(k+1)}=b^{(k)}$ si $f(x^{(k)})f(b^{(k)})<0$.

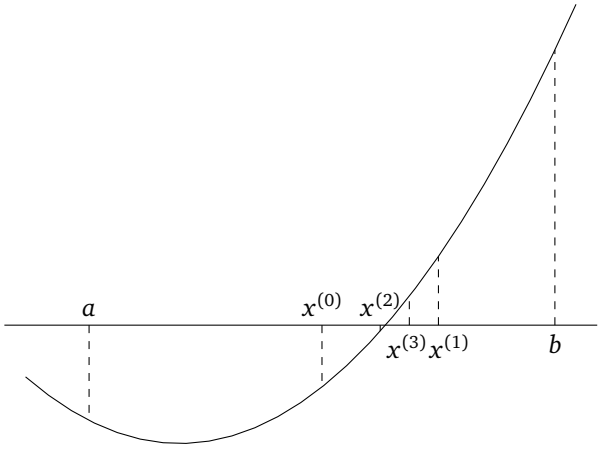


FIGURE 5.1 – Construction des premiers itérés de la méthode de dichotomie.

La figure 5.1 illustre la construction des approximations du zéro produites par cette méthode.

Exemple d'application de la méthode de dichotomie. On utilise la méthode de dichotomie pour approcher la racine du polynôme $f(x)=x^3+2x^2-3x-1$ contenue dans l'intervalle $[1,2]$ (on a en effet $f(1)=-1$ et $f(2)=9$), avec une précision égale à 10^{-4}. Le tableau 5.1 donne les valeurs respectives des bornes $a^{(k)}$ et $b^{(k)}$ de l'intervalle d'encadrement, de l'approximation $x^{(k)}$ de la racine et de $f(x^{(k)})$ en fonction du numéro k de l'itération.

Concernant la convergence de la méthode de dichotomie, on a le résultat suivant, dont la preuve est laissée en exercice.

Proposition 5.5 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une fonction réelle continue sur $[a,b]$, vérifiant $f(a)f(b)<0$ et possédant un unique zéro ξ dans $]a,b[$. Alors, la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode de dichotomie converge vers ξ et on a l'estimation*

$$\forall k\in\mathbb{N},\ \left|x^{(k)}-\xi\right|\leq\frac{b-a}{2^{k+1}}. \tag{5.5}$$

Il ressort de cette proposition que la méthode de dichotomie converge de manière certaine : c'est une méthode *globalement convergente*. L'estimation d'erreur (5.5) fournit par ailleurs directement un critère d'arrêt pour la méthode, puisque, à précision ε donnée, cette dernière permet d'approcher ξ en un nombre prévisible d'itérations. On voit en effet que, pour avoir $|x^{(k)}-\xi|\leq\varepsilon$, il faut que

$$\frac{b-a}{2^{k+1}}\leq\varepsilon\Leftrightarrow k\geq\frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)}-1. \tag{5.6}$$

k	$a^{(k)}$	$b^{(k)}$	$x^{(k)}$	$f(x^{(k)})$
0	1	2	1,5	2,375
1	1	1,5	1,25	0,328125
2	1	1,25	1,125	−0,419922
3	1,125	1,25	1,1875	−0,067627
4	1,1875	1,25	1,21875	0,124725
5	1,1875	1,21875	1,203125	0,02718
6	1,1875	1,203125	1,195312	−0,020564
7	1,195312	1,203125	1,199219	0,003222
8	1,195312	1,199219	1,197266	−0,008692
9	1,197266	1,199219	1,198242	−0,00274
10	1,198242	1,199219	1,19873	0,000239
11	1,198242	1,19873	1,198486	−0,001251
12	1,198486	1,19873	1,198608	−0,000506
13	1,198608	1,19873	1,198669	−0,000133

TABLE 5.1 – Tableau récapitulatif du déroulement de la méthode de dichotomie pour l'approximation (avec une précision égale à 10^{-4}) de la racine du polynôme x^3+2x^2-3x-1 contenue dans l'intervalle $[1,2]$.

Ainsi, pour améliorer la précision de l'approximation du zéro d'un ordre de grandeur, c'est-à-dire trouver $k>j$ tel que $\left|x^{(k)}-\xi\right| = \frac{1}{10}\left|x^{(j)}-\xi\right|$, on doit effectuer $k-j=\frac{\ln(10)}{\ln(2)}\simeq 3,32$ itérations.

On peut par ailleurs remarquer que, sous la seule hypothèse que $f(a)f(b)<0$, le théorème 5.4 garantit l'existence d'*au moins* un zéro de la fonction f dans l'intervalle $]a,b[$ et il se peut donc que ce dernier en contienne plusieurs. Dans ce cas, la méthode converge vers l'un d'entre eux.

Comme on le constate sur la figure 5.2, la méthode de dichotomie ne garantit pas une réduction monotone de l'erreur absolue d'une itération à l'autre. Ce n'est donc pas une méthode d'ordre un au sens de la définition 5.1, mais sa convergence est néanmoins linéaire au sens de la définition 5.3.

On gardera donc à l'esprit que la méthode de dichotomie est une méthode robuste. Si sa convergence est lente, on peut l'utiliser pour obtenir une approximation grossière (mais raisonnable) du zéro recherché servant d'initialisation à une méthode d'ordre plus élevé dont la convergence n'est que *locale*, comme la méthode de Newton–Raphson (voir la section 5.4.4). On peut voir cette approche comme une stratégie de « globalisation » de méthodes localement convergentes.

5.3.2 Méthode de la fausse position

La *méthode de la fausse position* (*false-position method* en anglais), encore appelée *méthode regula falsi*, est une méthode d'encadrement combinant les possibilités de la méthode de dichotomie avec celles de la méthode de la sécante, qui sera introduite dans la section 5.5. L'idée est d'utiliser l'information fournie par les valeurs de la fonction f aux extrémités de l'intervalle d'encadrement pour améliorer la vitesse de convergence de la méthode de dichotomie (cette dernière ne tenant compte que du signe de la fonction).

Comme précédemment, cette méthode suppose connus deux points a et b vérifiant $f(a)f(b)<0$ et servant d'initialisation à la suite d'intervalles $[a^{(k)},b^{(k)}]$, $k\geq 0$, contenant un zéro de la fonction f. Le procédé de construction des intervalles emboîtés est alors le même pour la méthode de dichotomie, à l'exception du choix de $x^{(k)}$, qui est à présent donné par l'abscisse du point d'intersection de la droite passant par les points $(a^{(k)},f(a^{(k)}))$ et $(b^{(k)},f(b^{(k)}))$ avec l'axe des abscisses, c'est-à-dire

$$x^{(k)}=a^{(k)}-\frac{a^{(k)}-b^{(k)}}{f(a^{(k)})-f(b^{(k)})}f(a^{(k)})=b^{(k)}-\frac{b^{(k)}-a^{(k)}}{f(b^{(k)})-f(a^{(k)})}f(b^{(k)})=\frac{f(a^{(k)})b^{(k)}-f(b^{(k)})a^{(k)}}{f(a^{(k)})-f(b^{(k)})}. \qquad (5.7)$$

11. Adrien-Marie Legendre (18 septembre 1752 - 9 janvier 1833) était un mathématicien français. On lui doit d'importantes contributions en théorie des nombres, en statistiques, en algèbre et en analyse, ainsi qu'en mécanique. Il est aussi célèbre pour être l'auteur des *Éléments de géométrie*, un traité publié pour la première fois en 1794 reprenant et modernisant les *Éléments* d'Euclide.

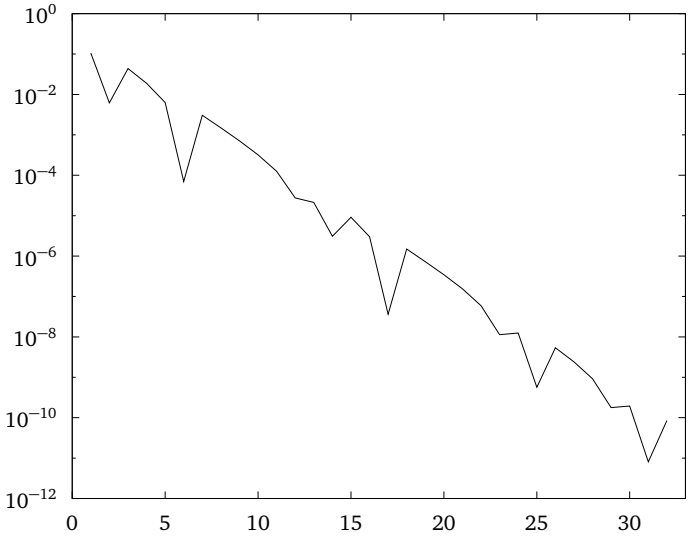


FIGURE 5.2 – Historique de la convergence, c'est-à-dire le tracé de l'erreur absolue $\left|x^{(k)} - \xi\right|$ en fonction k, de la méthode de dichotomie pour l'approximation de la racine $\xi = 0,9061798459...$ du polynôme de Legendre [11] de degré 5, $P_5(x) = \frac{1}{8}\,x\,(63x^4 - 70\,x^2 + 15)$, dont les racines se situent dans l'intervalle $]-1,1[$. On a choisi les bornes $a = 0,6$ et $b = 1$ pour l'intervalle d'encadrement initial et une précision de 10^{-10} pour le test d'arrêt, qui est atteinte après 31 itérations (à comparer à la valeur $30,89735...$ fournie par l'estimation (5.6)). On observe que l'erreur a un comportement oscillant, mais diminue en moyenne de manière linéaire.

La détermination de l'approximation du zéro à chaque étape repose donc sur un procédé d'interpolation linéaire de la fonction f entre les bornes de l'intervalle d'encadrement. Par conséquent, le zéro est obtenu après une seule itération si f est une fonction affine, contre *a priori* une infinité pour la méthode de dichotomie. On observera que la dernière des trois formules équivalentes fournissant l'approximation est la plus indiquée en arithmétique en précision finie, car elle présente moins de risque d'annulation catastrophique que les deux autres.

On a représenté sur la figure 5.3 la construction des premières approximations $x^{(k)}$ ainsi trouvées. Cette méthode apparaît comme plus « flexible » que la méthode de dichotomie, le point $x^{(k)}$ construit étant plus proche de l'extrémité de l'intervalle $[a^{(k)}, b^{(k)}]$ en laquelle la valeur de la fonction $|f|$ est la plus petite.

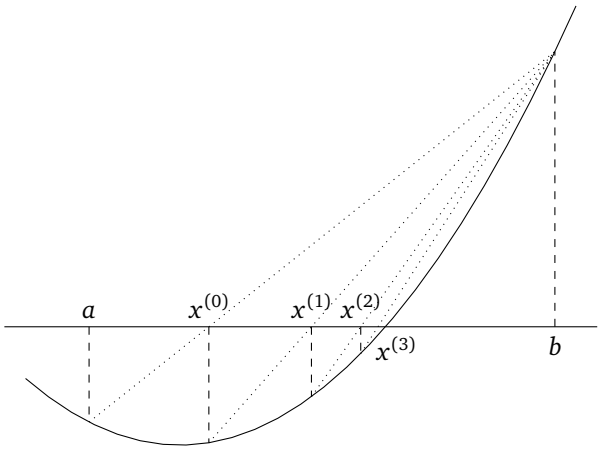


FIGURE 5.3 – Construction des premiers itérés de la méthode de la fausse position.

Indiquons que si la mesure de l'intervalle d'encadrement $[a^{(k)}, b^{(k)}]$ ainsi obtenu décroît bien lorsque k tend vers l'infini, elle ne tend pas nécessairement vers zéro [12], comme c'est le cas pour la méthode de dichotomie. En

12. Pour cette raison, le critère d'arrêt des itérations de la méthode doit être basé soit sur la longueur à l'étape k du plus petit des intervalles

effet, pour une fonction convexe ou concave dans un voisinage du zéro recherché, il apparaît que la méthode conduit inévitablement, à partir d'un certain rang, à l'une des configurations présentées sur la figure 5.4, pour chacune desquelles l'une des bornes de l'intervalle d'encadrement n'est plus jamais modifiée tandis que l'autre converge de manière monotone vers le zéro. On a alors affaire à une *méthode de point fixe* (voir la section 5.4, en comparant en particulier les relations de récurrence (5.8) et (5.10)).

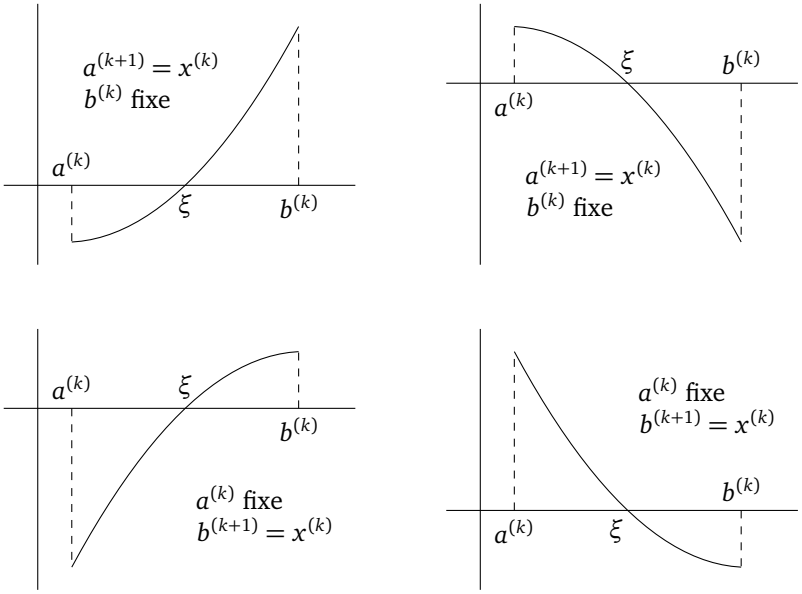


FIGURE 5.4 – Différentes configurations atteintes par la méthode de la fausse position à partir d'un certain rang pour une fonction f supposée convexe ou concave dans un voisinage du zéro ξ.

L'analyse de la méthode de la fausse position est bien moins triviale que celle de la méthode de dichotomie. On peut cependant établir le résultat de convergence *linéaire* suivant moyennant quelques hypothèses sur la fonction f.

Théorème 5.6 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une fonction réelle continue sur $[a,b]$, vérifiant $f(a)f(b)<0$ et possédant un unique zéro ξ dans $]a,b[$. Supposons de plus que f est continûment dérivable sur $]a,b[$ et convexe ou concave dans un voisinage de ξ. Alors, la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode de la fausse position converge au moins linéairement vers ξ.*

DÉMONSTRATION. Compte tenu des hypothèses, l'une des configurations illustrées à la figure 5.4 est obligatoirement atteinte par la méthode de la fausse position à partir d'un certain rang et l'on peut se ramener sans perte de généralité au cas où l'une des bornes de l'intervalle de départ reste fixe tout au long du processus itératif.
On peut ainsi considérer le cas d'une fonction f convexe sur l'intervalle $[a,b]$ et telle que $f(a)<0$ et $f(b)>0$ (ce qui correspond à la première configuration décrite sur la figure 5.4). Dans ces conditions, on montre, en utilisant (5.7) et la convexité de f, que, $\forall k\in\mathbb{N}$, $f(x^{(k)})\leq 0$. Par définition de la méthode, on a, $\forall k\in\mathbb{N}$, $a^{(k+1)}=x^{(k)}$ et $b^{(k+1)}=b$ si $f(x^{(k)})<0$, le point $x^{(k+1)}$ étant alors donné par la formule

$$\forall k\in\mathbb{N},\ x^{(k+1)}=x^{(k)}-\frac{b-x^{(k)}}{f(b)-f(x^{(k)})}f(x^{(k)}), \tag{5.8}$$

ou bien $x^{(k+1)}=x^{(k)}=\xi$ si $f(x^{(k)})=0$, ce dernier cas mettant fin aux itérations.

Supposons à présent que, $\forall k\in\mathbb{N}$, $f(x^{(k)})\neq 0$. Il découle de la relation (5.8) que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est croissante et elle est par ailleurs majorée par b ; elle converge donc vers une limite ℓ, qui vérifie

$$(b-\ell)f(\ell)=0.$$

Puisque, $\forall k\in\mathbb{N}$, $x^{(k)}<\xi$ on a $\ell\leq\xi<b$ et, par voie de conséquence, $f(\ell)=0$, d'où $\ell=\xi$, par unicité du zéro ξ.

$[a^{(k)},x^{(k)}]$ et $[x^{(k)},b^{(k)}]$, $k\geq 0$, soit sur la valeur du résidu $f(x^{(k)})$ (voir la section 5.6 pour plus de détails).

Il reste à prouver que la convergence de la méthode est au moins linéaire. En se servant une nouvelle fois de (5.8) et en faisant tendre k vers l'infini, on trouve que

$$\lim_{k\to+\infty} \frac{x^{(k+1)}-\xi}{x^{(k)}-\xi} = 1 - \frac{b-\xi}{f(b)-f(\xi)} f'(\xi).$$

La fonction f étant supposée convexe sur $[a,b]$, on a, $\forall x \in [a,b]$, $f(x) \geq f(\xi) + (x-\xi)f'(\xi)$; en choisissant $x = a$ et $x = b$ dans cette dernière inégalité, on obtient respectivement que $f'(\xi) > 0$ et $f'(\xi) \leq \frac{f(b)-f(\xi)}{b-\xi}$, d'où la conclusion.

La même technique de démonstration s'adapte pour traiter les trois cas possibles restants, ce qui achève la preuve. □

Exemple d'application de la méthode de la fausse position. Reprenons l'exemple d'application de la section précédente, dans lequel on utilisait la méthode de dichotomie pour approcher la racine du polynôme $f(x) = x^3 + 2x^2 - 3x - 1$. Le tableau 5.2 présente donne les valeurs respectives des bornes $a^{(k)}$ et $b^{(k)}$ de l'intervalle d'encadrement, de l'approximation $x^{(k)}$ de la racine et de $f(x^{(k)})$ en fonction du numéro k de l'itération obtenue avec la méthode de la fausse position (avec une tolérance égale à 10^{-4} pour le test d'arrêt). On observe que la borne de droite de l'intervalle d'encadrement initial est conservée tout au long du calcul.

k	$a^{(k)}$	$b^{(k)}$	$x^{(k)}$	$f(x^{(k)})$
0	1	2	1,1	−0,549
1	1,1	2	1,151744	−0,274401
2	1,151744	2	1,176841	−0,130742
3	1,176841	2	1,188628	−0,060876
4	1,188628	2	1,194079	−0,028041
5	1,194079	2	1,196582	−0,012852
6	1,196582	2	1,197728	−0,005877
7	1,197728	2	1,198251	−0,002685
8	1,198251	2	1,19849	−0,001226
9	1,19849	2	1,1986	−0,00056
10	1,1986	2	1,198649	−0,000255

TABLE 5.2 – Tableau récapitulatif du déroulement de la méthode de la fausse position pour l'approximation (avec une précision égale à 10^{-4}) de la racine du polynôme $x^3 + 2x^2 - 3x - 1$ contenue dans l'intervalle $[1,2]$.

Dans de nombreuses situations, comme pour la résolution approchée de l'équation de Kepler dans le cas d'une orbite elliptique présentée sur la figure 5.5, la méthode de la fausse position converge plus rapidement que la méthode de dichotomie. Ceci n'est cependant pas une règle générale et l'on peut construire des exemples pour lesquels il en va tout autrement (voir la figure 5.6).

5.3.3 Variantes

La convergence linéaire souvent observée de la méthode de la fausse position est due au fait que l'une des bornes de l'intervalle d'encadrement n'est plus jamais modifiée après un certain nombre d'itérations. Plusieurs modifications ont été proposées pour éliminer ce problème de « rétention » et obtenir une convergence superlinéaire.

La première de ces variantes semble avoir été introduite au début des années 1950 au centre de calcul de l'université de l'Illinois, et porte pour cette raison le nom de *méthode Illinois* [DJ71]. Pour la présenter, on adopte le point suivant, qui diffère quelque peu de celui précédemment utilisé. On suppose disposer initialement d'un encadrement d'un zéro de la fonction f avec les bornes $x^{(0)}$ et $x^{(1)}$ et on pose $y^{(0)} = f(x^{(0)})$ et $y^{(1)} = f(x^{(1)})$. À la k^{e} étape de la méthode, k étant un entier naturel non nul, on pose

$$x^{(k+1)} = \frac{x^{(k-1)}y^{(k)} - x^{(k)}y^{(k-1)}}{y^{(k)} - y^{(k-1)}} \text{ et } y^{(k+1)} = f(x^{(k+1)}),$$

le réel $x^{(k+1)}$ correspondant à l'abscisse du point d'intersection de la droite passant par les points $(x^{(k-1)}, y^{(k-1)})$ et $(x^{(k)}, y^{(k)})$ avec l'axe des abscisses, conformément à la construction employée dans la méthode de la fausse

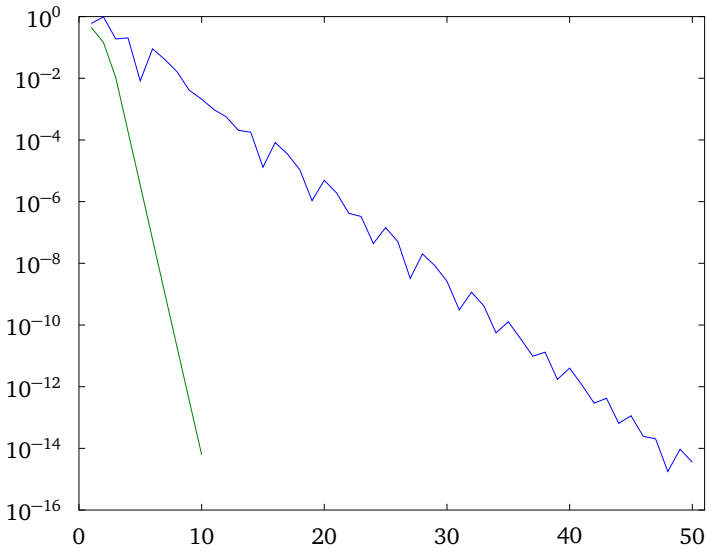


FIGURE 5.5 – Tracés, en fonction du nombre d'itérations, des erreurs absolues de la méthode de dichotomie (en bleu) et de la méthode de la fausse position (en vert) utilisées pour résoudre de manière approchée l'équation (5.1), avec $e = 0,8$ et $M = \frac{4\pi}{3}$, de solution $E = 3,7388733587...$, à partir de l'intervalle d'encadrement initial $[0, 2\pi]$.

position (on pourra comparer la première des égalités ci-dessus avec la dernière de (5.7)). Si $y^{(k+1)}y^{(k)} < 0$, les bornes d'encadrement pour l'étape suivante seront $x^{(k)}$ et $x^{(k+1)}$. En revanche, si $y^{(k+1)}y^{(k)} > 0$, on modifie la valeur de $x^{(k)}$, ainsi que celle de la quantité $y^{(k)}$ associée, de la façon suivante :

$$x^{(k)} = x^{(k-1)} \text{ et } y^{(k)} = \frac{y^{(k-1)}}{2}, \tag{5.9}$$

avant de passer à la prochaine étape. On parle dans ce cas d'étape *modifiée* (par rapport à la méthode de la fausse position). Cette nouvelle méthode restant une méthode d'encadrement, sa convergence est garantie. On peut alors étudier le comportement asymptotique de l'erreur de la méthode en faisant l'hypothèse que la fonction f est deux fois continûment dérivable et que ses dérivées f' et f'' ne s'annulent pas, de sorte que son zéro, noté ξ, est simple. Après une étape non modifiée, on a

$$x^{(k+1)} = x^{(k-1)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k-1)}),$$

et, en utilisant de manière répétée le théorème de Rolle pour une fonction auxiliaire [13] et le théorème des accroissements finis appliqué à f', on montre qu'il existe des réels $\theta^{(k)}$ et $\eta^{(k)}$, strictement compris entre $x^{(k-1)}$ et $x^{(k)}$, tels que

$$x^{(k+1)} - \xi = \frac{1}{2}\frac{f''(\theta^{(k)})}{f'(\eta^{(k)})}(x^{(k)} - \xi)(x^{(k-1)} - \xi).$$

Ceci permet de déduire qu'on a, asymptotiquement et après une étape non modifiée, l'équivalent

$$x^{(k+1)} - \xi \approx \frac{1}{2}\frac{f''(\xi)}{f'(\xi)}(x^{(k)} - \xi)(x^{(k-1)} - \xi).$$

Pour caractériser l'évolution de l'erreur au cours d'une étape modifiée, on fait l'hypothèse (que l'on justifiera *a posteriori*) que l'approximation $x^{(k+2)}$ issue d'une telle étape est précédée par une approximation $x^{(k+1)}$ issue

13. La méthode remplaçant sur l'intervalle d'encadrement courant la fonction f par son polynôme d'interpolation de Lagrange aux nœuds $x^{(k-1)}$ et $x^{(k)}$ pour déterminer $x^{(k+1)}$, on se trouve ici dans un cas particulier d'application du théorème 6.16 pour l'erreur d'interpolation polynomiale.

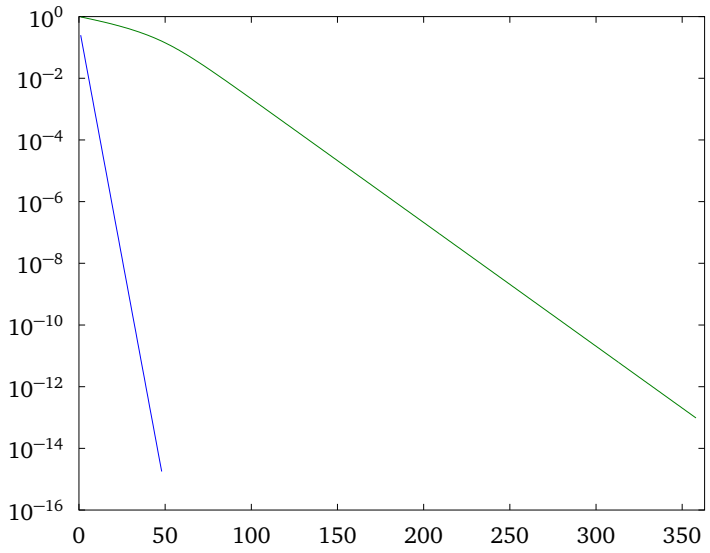


FIGURE 5.6 – Tracés, en fonction du nombre d'itérations, des erreurs absolues de la méthode de dichotomie (en bleu) et de la méthode de la fausse position (en vert) utilisées pour la résolution approchée de l'équation $x^{10}-1=0$ à partir de l'intervalle d'encadrement initial $\left[0,\frac{3}{2}\right]$. Malgré l'accélération observée de la convergence de la méthode de la fausse position durant les premières itérations, la vitesse de cette dernière reste largement inférieure à celle de la méthode de dichotomie.

d'une étape non modifiée. On a dans ce cas

$$x^{(k+2)}-\xi=\left(\frac{f(x^{(k+1)})-f(\xi)}{x^{(k+1)}-\xi}\,\frac{x^{(k-1)}-\xi}{f(x^{(k+1)})-\frac{1}{2}f(x^{(k-1)})}-\frac{f(x^{(k-1)})}{2f(x^{(k+1)})-f(x^{(k-1)})}\right)(x^{(k+1)}-\xi).$$

La convergence de la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ et des développements de Taylor–Lagrange montrent alors que

$$\lim_{k\to+\infty}\frac{f(x^{(k+1)})}{x^{(k-1)}-\xi}=0 \text{ et } \lim_{k\to+\infty}\frac{f(x^{(k-1)})}{2f(x^{(k+1)})-f(x^{(k-1)})}=-1,$$

ce qui conduit à l'équivalent asymptotique

$$x^{(k+2)}-\xi\approx-(x^{(k+1)}-\xi).$$

L'utilisation de ces deux estimations permet de dresser un tableau de signe de l'erreur à chaque étape et du type de l'étape correspondante, qui montre que les étapes modifiées (désignées par la lettre M) et non modifiées (désignées par la lettre U) se succèdent asymptotiquement comme

$k-1$	k	$k+1$	$k+2$	$k+3$	$k+4$	$k+5$	$k+6$	$k+7$	$k+8$	$k+9$	$k+10$	…
−	+	−	−	+	−	−	+	−	−	+	−	…
	U	U	M	U	U	M	U	U	M	U	U	…

si $\frac{f''(\xi)}{f'(\xi)}>0$, ou

$k-1$	k	$k+1$	$k+2$	$k+3$	$k+4$	$k+5$	$k+6$	$k+7$	$k+8$	$k+9$	…
−	+	+	−	+	+	−	+	+	−	+	…
	U	M	U	U	M	U	U	M	U	U	…

si $\frac{f''(\xi)}{f'(\xi)}<0$, de sorte que le motif MUU se trouve indéfiniment répété. Par ailleurs, en combinant les deux équivalents sur ce motif, on arrive à l'estimation

$$x^{(k+2)}-\xi\approx\left(\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right)^{2}(x^{(k-1)}-\xi)^{3},$$

qui fournit une mesure de l'ordre de convergence effectif de la méthode via le calcul de l'*indice d'efficacité computationelle de Traub* (voir [Tra64, Appendix C]), défini par $p^{\frac{1}{d}}$, où p et d sont respectivement l'ordre de la méthode et le nombre d'évaluations de la fonction f sur un motif. On trouve ainsi que la méthode est d'ordre égal à trois sur le motif, ce dernier comportant trois étapes nécessitant chacune une évaluation. L'indice d'efficacité computationnelle de la méthode vaut par conséquent $3^{\frac{1}{3}} \approx 1,44225$, indiquant une convergence superlinéaire.

Des variations de l'étape modifiée ont par la suite été proposées de manière à améliorer l'indice d'efficacité computationnelle. Parmi celles-ci, on peut mentionner la *méthode Pegasus* [DJ72], dans laquelle (5.9) est remplacée par

$$x^{(k)} = x^{(k-1)} \text{ et } y^{(k)} = \frac{y^{(k-1)}y^{(k)}}{y^{(k)} + y^{(k+1)}},$$

dont le motif asymptotique est $MMUU$ et dont l'indice d'efficacité vaut $(7,275)^{\frac{1}{4}} \approx 1,64232$, ou la *méthode d'Anderson–Björck* [AB73], dans laquelle (5.9) est remplacée par

$$x^{(k)} = x^{(k-1)} \text{ et } y^{(k)} = \frac{y^{(k)} - y^{(k+1)}}{y^{(k)}} y^{(k-1)} \text{ si } \frac{y^{(k)} - y^{(k+1)}}{y^{(k)}} > 0,\ y^{(k)} = \frac{y^{(k-1)}}{2} \text{ sinon.}$$

Dans cette dernière, en désignant par la lettre M l'étape modifiée correspondant à celle de la méthode Illinois et par la lettre P celle propre à la méthode d'Anderson–Björck, il s'avère que le motif asymptotique prend selon les cas la forme PUU ou $PPUU$, donnant respectivement lieu à un indice d'efficacité valant $5^{\frac{1}{3}} \approx 1,70997$ ou $8^{\frac{1}{4}} \approx 1,68179$.

La *méthode de Ridders* [Rid79] est un autre type variante de la méthode de la fausse position, visant à réduire à chaque étape l'intervalle d'encadrement en modifiant possiblement les deux bornes. Comme précédemment, on suppose que, à la k^{e} étape de la méthode, le zéro que l'on veut approcher est strictement encadré par les réels $x^{(k-1)}$ et $x^{(k)}$. On introduit alors le point milieu $x_m^{(k)} = \frac{1}{2}\left(x^{(k-1)} + x^{(k)}\right)$ et la fonction $h(x) = f(x)\exp\left(\alpha\left(x - x^{(k-1)}\right)\right)$, dans laquelle le réel α, qui dépend de l'étape, est choisi de manière à ce que

$$h(x_m^{(k)}) = \frac{1}{2}\left(h(x^{(k-1)}) + h(x^{(k)})\right),$$

ce qui signifie encore que la quantité $\exp\left(\alpha\left(x_m^{(k)} - x^{(k-1)}\right)\right)$ est solution de l'équation quadratique

$$\frac{1}{2}f(x^{(k)})\,x^2 - f(x_m^{(k)})\,x + \frac{1}{2}f(x^{(k-1)}) = 0.$$

Le discriminant associé à cette équation vaut $f(x_m^{(k)})^2 - f(x^{(k-1)})f(x^{(k)})$ et est strictement positif en vertu des hypothèses faites sur f et le zéro encadré. Il existe par conséquent deux solutions, la stricte positivité de la fonction exponentielle imposant alors que

$$\exp\left(\alpha\left(x_m^{(k)} - x^{(k-1)}\right)\right) = \frac{f(x_m) - \operatorname{sign}(f(x^{(k-1)}))\sqrt{f(x_m^{(k)})^2 - f(x^{(k-1)})f(x^{(k)})}}{f(x^{(k)})},$$

où $\operatorname{sign}(f(x^{(k-1)}))$ désigne le signe de $f(x^{(k-1)})$. Une nouvelle approximation $x^{(k+1)}$ du zéro est ensuite obtenue en appliquant une étape de la méthode de la fausse position à la fonction h sur l'intervalle courant borné par $x^{(k-1)}$ et $x^{(k)}$, la multiplication de f par une fonction exponentielle laissant inchangé le signe de la fonction aux bornes de l'intervalle. On a alors, en adaptant (5.7) et en se servant de la définition de h et de l'expression ci-dessus,

$$x^{(k+1)} = \frac{h(x^{(k-1)})x_m^{(k)} - h(x_m)x^{(k-1)}}{h(x^{(k-1)}) - h(x_m^{(k)})} = x_m^{(k)} + \operatorname{sign}(f(x^{(k-1)}))\frac{f(x_m^{(k)})(x_m^{(k)} - x^{(k-1)})}{\sqrt{f(x_m^{(k)})^2 - f(x^{(k-1)})f(x^{(k)})}}.$$

Dans le cas favorable où le zéro se trouve entre $x_m^{(k)}$ et $x^{(k+1)}$, ces deux points deviennent les prochaines bornes de l'intervalle d'encadrement. Sinon, le point $x^{(k+1)}$ est pris comme borne tandis que l'une des précédentes bornes, $x^{(k-1)}$ ou $x^{(k)}$, est conservée de sorte que le zéro soit encadré à l'étape suivante. On notera que la méthode nécessite une évaluation supplémentaire de la fonction f à chaque itération par rapport à la méthode de la fausse position.

5.4 Méthodes de point fixe

Les méthodes d'approximation de zéros introduites dans la suite de ce chapitre sont plus générales au sens où elles se passent de l'hypothèse de changement de signe de f en ξ et ne consistent pas en la construction d'une suite d'intervalles contenant le zéro de la fonction ; bien qu'étant aussi des méthodes itératives, ce ne sont pas des méthodes d'encadrement. Rien ne garantit d'ailleurs qu'une suite $(x^{(k)})_{k\in\mathbb{N}}$ produite par l'un des algorithmes présentés prendra ses valeurs dans un intervalle fixé *a priori*.

Au sein de cette catégorie de méthodes itératives, les méthodes de point fixe sont basées sur le fait que tout problème de recherche de zéros d'une fonction peut se ramener à un problème de recherche de points fixes d'une autre fonction. Après avoir rappelé le principe de ces méthodes et étudié leurs propriétés, nous nous penchons sur les cas particuliers des méthodes de la corde et de Newton–Raphson. Cette dernière méthode illustre de manière exemplaire le fait, déjà observé avec la méthode de la fausse position, que la prise en compte dans la méthode de l'information fournie par les valeurs de f et, lorsque cette fonction est différentiable, celles de sa dérivée peut conduire à une vitesse de convergence améliorée [14].

5.4.1 Principe

La famille de méthodes que nous allons maintenant étudier utilise le fait que le problème $f(x) = 0$ peut toujours ramener au problème équivalent $g(x) - x = 0$, pour lequel on a le résultat suivant.

Théorème 5.7 (« théorème du point fixe de Brouwer [15] ») *Soit $[a, b]$ un intervalle non vide de $\mathbb{R}$ et g une application continue de $[a, b]$ dans lui-même. Alors, il existe un point ξ de $[a, b]$, appelé* ***point fixe de la fonction*** *g, vérifiant $g(\xi) = \xi$.*

DÉMONSTRATION. Posons $f(x) = g(x) - x$. On a alors $f(a) = g(a) - a \geq 0$ et $f(b) = g(b) - b \leq 0$, puisque $g(x)$ appartient à $[a, b]$ pour tout x appartenant à $[a, b]$. Par conséquent, la fonction f, continue sur $[a, b]$, est telle que $f(a)f(b) \leq 0$. Le théorème 5.4 assure alors l'existence d'un point ξ dans $[a, b]$ tel que $0 = f(\xi) = g(\xi) - \xi$. □

Bien entendu, toute équation de la forme $f(x) = 0$ peut s'écrire sous la forme $x = g(x)$ en posant $g(x) = x + f(x)$, mais ceci ne garantit en rien que la fonction auxiliaire g ainsi définie satisfait les hypothèses du théorème 5.7. Il existe cependant de nombreuses façons de construire g à partir de f, comme le montre l'exemple ci-après, et il suffit donc de trouver une transformation adaptée.

Exemple. Considérons la fonction $f(x) = e^x - 2x - 1$ sur l'intervalle $[1, 2]$. Nous avons $f(1) < 0$ et $f(2) > 0$, f possède donc bien un zéro sur l'intervalle $[1, 2]$. Soit $g(x) = \frac{1}{2}(e^x - 1)$. L'équation $x = g(x)$ est bien équivalente à $f(x) = 0$, mais g, bien que continue, n'est pas à valeurs de $[1, 2]$ dans lui-même. Posons à présent $g(x) = \ln(2x + 1)$. Cette dernière fonction est continue et croissante sur l'intervalle $[1, 2]$, à valeurs dans lui-même et satisfait donc aux hypothèses du théorème 5.7.

Nous venons de montrer que, sous certaines conditions, approcher les zéros d'une fonction f revient à approcher les points fixes d'une fonction g, sans que l'on sache pour autant traiter ce nouveau problème. Une méthode courante pour la détermination de point fixe se résume à la construction d'une suite $(x^{(k)})_{k\in\mathbb{N}}$ par le procédé itératif suivant : étant donnée une valeur initiale $x^{(0)}$ (appartenant à $[a, b]$), on pose

$$\forall k \in \mathbb{N},\ x^{(k+1)} = g(x^{(k)}). \tag{5.10}$$

On dit que la relation (5.10) est une *itération de point fixe* (*fixed-point iteration* en anglais). La méthode d'approximation résultante est appelée *méthode de point fixe* ou bien encore *méthode des approximations successives*. Si la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par (5.10) converge, cela ne peut être que vers un point fixe de g. En effet, en posant $\lim_{k\to+\infty} x^{(k)} = \xi$, nous avons

$$\xi = \lim_{k\to+\infty} x^{(k+1)} = \lim_{k\to+\infty} g(x^{(k)}) = g\left(\lim_{k\to+\infty} x^{(k)}\right) = g(\xi),$$

la deuxième égalité provenant de la définition (5.10) de la suite récurrente et la troisième étant une conséquence de la continuité de g.

14. Ceci est également vérifié pour la méthode de la sécante (voir la section 5.5).

15. Luitzen Egbertus Jan Brouwer (27 février 1881 - 2 décembre 1966) était un mathématicien et philosophe néerlandais. Ses apports concernèrent principalement la topologie et la logique formelle.

5.4.2 Quelques résultats de convergence

Le choix de la fonction g pour mettre en œuvre cette méthode n'étant pas unique, celui-ci est alors motivé par les exigences du théorème 5.9, qui donne des conditions *suffisantes* sur g pour avoir convergence de la méthode de point fixe définie par (5.10). Avant de l'énoncer, rappelons tout d'abord la notion d'application *contractante*.

Définition 5.8 (application contractante) *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application de $[a,b]$ dans $\mathbb{R}$. On dit que g est une application* ***contractante*** *si et seulement si il existe une constante K telle que $0<K<1$ vérifiant*

$$\forall x\in[a,b],\ \forall y\in[a,b],\ |g(x)-g(y)|\leq K\,|x-y|\,. \tag{5.11}$$

On notera que la constante de Lipschitz de l'application g n'est autre que la plus petite constante K vérifiant la condition (5.11).

Le résultat suivant est une application dans le cas réel du *théorème du point fixe de Banach* [16] [Ban22], dont l'énoncé général vaut pour toute application contractante définie sur un espace vectoriel normé complet (pour la distance issue de sa norme), encore appelé un *espace de Banach*.

Théorème 5.9 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application contractante sur $[a,b]$, telle que l'image de $[a,b]$ par g est incluse dans $[a,b]$. Alors, la fonction g possède un unique point fixe ξ dans $[a,b]$. De plus, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par la relation* (5.10) *converge, pour toute initialisation $x^{(0)}$ dans $[a,b]$, vers ce point fixe et l'on a les estimations suivantes*

$$\forall k\in\mathbb{N},\ \left|x^{(k)}-\xi\right|\leq K^k\left|x^{(0)}-\xi\right|, \tag{5.12}$$

$$\forall k\in\mathbb{N}^*,\ \left|x^{(k)}-\xi\right|\leq\frac{K}{1-K}\left|x^{(k)}-x^{(k-1)}\right|. \tag{5.13}$$

DÉMONSTRATION. On commence par montrer que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est une suite de Cauchy. En effet, on a

$$\forall k\in\mathbb{N}^*,\ \left|x^{(k+1)}-x^{(k)}\right|=\left|g(x^{(k)})-g(x^{(k-1)})\right|\leq K\left|x^{(k)}-x^{(k-1)}\right|,$$

par hypothèse, et on obtient par récurrence que

$$\forall k\in\mathbb{N},\ \left|x^{(k+1)}-x^{(k)}\right|\leq K^k\left|x^{(1)}-x^{(0)}\right|.$$

On en déduit, par une utilisation répétée de l'inégalité triangulaire, que

$$\begin{aligned}\forall k\in\mathbb{N},\ \forall p\in\mathbb{N},\ p>2,\ \left|x^{(k+p)}-x^{(k)}\right|&\leq\left|x^{(k+p)}-x^{(k+p-1)}\right|+\left|x^{(k+p-1)}-x^{(k+p-2)}\right|+\cdots+\left|x^{(k+1)}-x^{(k)}\right|\\&\leq\left(K^{p-1}+K^{p-2}+\cdots+1\right)\left|x^{(k+1)}-x^{(k)}\right|\\&\leq\frac{1-K^p}{1-K}K^k\left|x^{(1)}-x^{(0)}\right|,\end{aligned}$$

le dernier membre tendant vers zéro lorsque l'entier k tend vers l'infini. La suite réelle $(x^{(k)})_{k\in\mathbb{N}}$ converge donc vers une limite ξ dans $[a,b]$. L'application g étant continue [17], on déduit alors par un passage à la limite dans (5.10) que $\xi=g(\xi)$. Supposons à présent que g possède deux points fixes ξ et ζ dans l'intervalle $[a,b]$. On a alors

$$0\leq|\xi-\zeta|=|g(\xi)-g(\zeta)|\leq K\,|\xi-\zeta|\,,$$

d'où $\xi=\zeta$ puisque $K<1$.
La première estimation se prouve alors par un raisonnement par récurrence sur l'entier k, en écrivant que

$$\forall k\in\mathbb{N}^*,\ \left|x^{(k)}-\xi\right|=\left|g(x^{(k-1)})-g(\xi)\right|\leq K\left|x^{(k-1)}-\xi\right|,$$

et la seconde est obtenue en utilisant que

$$\forall k\in\mathbb{N}^*,\ \forall p\in\mathbb{N}^*,\ \left|x^{(k+p)}-x^{(k)}\right|\leq\frac{1-K^p}{1-K}\left|x^{(k+1)}-x^{(k)}\right|\leq\frac{1-K^p}{1-K}K\left|x^{(k)}-x^{(k-1)}\right|,$$

16. Stefan Banach (30 mars 1892 - 31 août 1945) était un mathématicien polonais. Il est l'un des fondateurs de l'analyse fonctionnelle moderne et introduisit notamment des espaces vectoriels normés complets, aujourd'hui appelés *espaces de Banach*, lors de son étude des espaces vectoriels topologiques. Plusieurs importants théorèmes et un célèbre paradoxe sont associés à son nom.

17. C'est par hypothèse une application K-lipschitzienne.

et en faisant tendre l'entier p vers l'infini. □

Sous les hypothèses du théorème 5.9, la convergence des itérations de point fixe est assurée quel que soit le choix de la valeur initiale $x^{(0)}$ dans l'intervalle $[a,b]$: c'est donc un nouvel exemple de convergence *globale*. Par ailleurs, un des intérêts de ce résultat est de donner une estimation de la vitesse de convergence de la suite vers sa limite, la première inégalité montrant en effet que la convergence est *géométrique*. La seconde inégalité s'avère particulièrement utile d'un point de vue pratique, car elle fournit à chaque étape un majorant de la distance à la limite (sans pour autant connaître cette dernière) en fonction d'une quantité connue. Il devient alors possible de majorer le nombre d'itérations que l'on doit effectuer pour approcher le point fixe ξ avec une précision donnée.

Corollaire 5.10 *Considérons la méthode de point fixe définie par la relation* (5.10)*, la fonction g vérifiant les hypothèses du théorème 5.9. Étant données une précision $\varepsilon > 0$ et une initialisation $x^{(0)}$ dans l'intervalle $[a,b]$, soit $k_0(\varepsilon)$ le plus petit entier tel que*

$$\forall k \in \mathbb{N},\ k \geq k_0(\varepsilon) \Rightarrow |x^{(k)} - \xi| \leq \varepsilon.$$

On a alors la majoration

$$k_0(\varepsilon) \leq \left\lfloor \frac{\ln(\varepsilon) + \ln(1-K) - \ln\left(\left|x^{(1)} - x^{(0)}\right|\right)}{\ln(K)} \right\rfloor + 1,$$

où, pour tout réel x, $\lfloor x \rfloor$ désigne la partie entière par défaut de x.

DÉMONSTRATION. En utilisant l'inégalité triangulaire et l'inégalité (5.13) pour $k = 1$, on trouve que

$$\left|x^{(0)} - \xi\right| \leq \left|x^{(0)} - x^{(1)}\right| + \left|x^{(1)} - \xi\right| \leq \left|x^{(0)} - x^{(1)}\right| + K\left|x^{(0)} - \xi\right|,$$

d'où

$$\left|x^{(0)} - \xi\right| \leq \frac{K}{1-K}\left|x^{(0)} - x^{(1)}\right|.$$

En substituant cette expression dans (5.13), on obtient que

$$\left|x^{(k)} - \xi\right| \leq \frac{K^k}{1-K}\left|x^{(0)} - x^{(1)}\right|,$$

et on aura en particulier $|x^{(k)} - \xi| \leq \varepsilon$ si k est tel que

$$\frac{K^k}{1-K}\left|x^{(0)} - x^{(1)}\right| \leq \varepsilon.$$

En prenant le logarithme népérien de chacun des membres de cette dernière inégalité, on arrive à

$$k \geq \frac{\ln(\varepsilon) + \ln(1-K) - \ln\left(\left|x^{(1)} - x^{(0)}\right|\right)}{\ln(K)},$$

dont on déduit le résultat. □

Dans la pratique, vérifier que l'application g est K-lipschitzienne n'est pas toujours aisé. Lorsque g est une fonction de classe $\mathscr{C}^1$ sur l'intervalle $[a,b]$, il est cependant possible d'utiliser la caractérisation suivante.

Proposition 5.11 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une fonction de classe $\mathscr{C}^1$, définie de $[a,b]$ dans lui-même, vérifiant*

$$\forall x \in [a,b],\ \left|g'(x)\right| \leq K < 1.$$

Alors, l'application g est contractante sur $[a,b]$.

DÉMONSTRATION. D'après le théorème des accroissements finis (voir le théorème B.113), pour tous x et y contenus dans l'intervalle $[a,b]$ et distincts, on sait qu'il existe un réel c strictement compris entre x et y tel que

$$|g(x) - g(y)| = \left|g'(c)\right| |x - y|,$$

d'où le résultat. □

La dernière proposition permet alors d'affiner le résultat de convergence globale précédent dans ce cas particulier.

Théorème 5.12 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application satisfaisant les hypothèses de la proposition 5.11. Alors, la fonction g possède un unique point fixe ξ dans $[a,b]$ et la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (5.10) *converge, pour toute initialisation $x^{(0)}$ dans $[a,b]$, vers ce point fixe. De plus, on a*

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi}=g'(\xi), \tag{5.14}$$

la convergence est donc au moins linéaire.

DÉMONSTRATION. La proposition 5.11 établissant que g est une application contractante sur $[a,b]$, les conclusions du théorème 5.9 sont valides et il ne reste qu'à prouver l'égalité (5.14). En vertu du théorème des accroissements finis (voir le théorème B.113), il existe, pour tout entier naturel k, un réel $\eta^{(k)}$ strictement compris entre $x^{(k)}$ et ξ tel que

$$x^{(k+1)}-\xi=g(x^{(k)})-g(\xi)=g'(\eta^{(k)})(x^{(k)}-\xi).$$

La suite $(x^{(k)})_{k\in\mathbb{N}}$ convergeant vers ξ, cette égalité implique que

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi}=\lim_{k\to+\infty}g'(\eta^{(k)})=g'(\xi).$$

□

On notera que ce théorème assure une convergence *au moins linéaire* de la méthode de point fixe. Par comparaison entre (5.3) et (5.14), il apparaît que la quantité $|g'(\xi)|$ est la constante asymptotique d'erreur de la méthode.

En pratique, il est souvent difficile de déterminer *a priori* un intervalle $[a,b]$ sur lequel les hypothèses de la proposition 5.11 sont satisfaites. Il est néanmoins possible de se contenter d'hypothèses plus faibles, au prix d'un résultat moindre de convergence *locale*.

Théorème 5.13 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, g une fonction continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. On suppose de plus que g admet une dérivée continue dans un voisinage de ξ, avec $|g'(\xi)|<1$. Alors, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (5.10) *converge vers ξ, pour toute initialisation $x^{(0)}$ choisie suffisamment proche de ξ.*

DÉMONSTRATION. Par hypothèse sur la fonction g, il existe un réel h strictement positif tel que la dérivée g' est continue sur l'intervalle $[\xi-h,\xi+h]$. Puisque $|g'(\xi)|<1$, on peut alors trouver un intervalle $I_\delta=[\xi-\delta,\xi+\delta]$, avec $0<\delta\le h$, tel que $|g'(x)|\le L$, avec $L<1$, pour tout réel x appartenant à I_δ. Pour cela, il suffit de poser $L=\frac{1}{2}(1+|g'(\xi)|)$ et d'utiliser la continuité de g' pour choisir $\delta\le h$ de manière à ce que

$$\forall x\in I_\delta,\ \left|g'(x)-g'(\xi)\right|\le\frac{1}{2}\left(1-\left|g'(\xi)\right|\right).$$

On en déduit alors que

$$\forall x\in I_\delta,\ \left|g'(x)\right|\le\left|g'(x)-g'(\xi)\right|+\left|g'(\xi)\right|\le\frac{1}{2}\left(1-\left|g'(\xi)\right|\right)+\left|g'(\xi)\right|=L.$$

Supposons à présent que, pour un entier naturel k donné, le terme $x^{(k)}$ de la suite définie par la relation de récurrence (5.10) appartienne à I_δ. On a alors, en vertu du théorème des accroissements finis (voir le théorème B.113),

$$x^{(k+1)}-\xi=g(x^{(k)})-\xi=g(x^{(k)})-g(\xi)=g'(\eta^{(k)})(x^{(k)}-\xi),$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ, d'où

$$\left|x^{(k+1)}-\xi\right|\le L\left|x^{(k)}-\xi\right|,$$

et $x^{(k+1)}$ appartient donc lui aussi à I_δ. On montre alors, en raisonnant par récurrence, que, si $x^{(0)}$ appartient à I_δ, alors tout terme de la suite appartient également à I_δ et

$$\forall k\in\mathbb{N},\ \left|x^{(k)}-\xi\right|\le L^k\left|x^{(0)}-\xi\right|,$$

ce qui implique que la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers ξ. □

On peut observer que, si $|g'(\xi)|>1$ et si $x^{(k)}$ est suffisamment proche de ξ pour avoir $\left|g'(x^{(k)})\right|>1$, on obtient $\left|x^{(k+1)}-\xi\right|>\left|x^{(k)}-\xi\right|$ et la convergence ne peut alors avoir lieu (sauf si $x^{(k)}=\xi$). Dans le cas où $|g'(\xi)|=1$, il peut y avoir convergence ou divergence selon les cas considérés. Cette remarque et le précédent théorème conduisent à l'introduction des notions suivantes.

Définitions 5.14 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, une fonction g continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. On dit que ξ est un point fixe* ***attractif*** *si la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par l'itération de point fixe* (5.10) *converge pour toute initialisation $x^{(0)}$ suffisamment proche de ξ. Réciproquement, si cette suite ne converge pour aucune initialisation $x^{(0)}$ dans un voisinage de ξ, exceptée $x^{(0)}=\xi$, le point fixe est dit* ***répulsif****.*

Exemple. Soit la fonction définie par $g(x)=\frac{1}{2}(x^2+c)$, avec c un réel fixé. Les points fixes de g sont les racines de l'équation du second degré $x^2-2x+c=0$, c'est-à-dire $1\pm\sqrt{1-c}$. Si $c>1$, la fonction n'a donc pas de points fixes réels. Si $c=1$, elle a un unique point fixe dans $\mathbb{R}$ et deux si $c<1$. Supposons que l'on soit dans ce dernier cas et posons $\xi_1=1-\sqrt{1-c}$, $\xi_2=1+\sqrt{1-c}$, de manière à ce que $\xi_1<1<\xi_2$. Puisque $g'(x)=x$, on voit que le point fixe ξ_2 est répulsif, mais que le point ξ_1 est attractif si $-3<c<1$ (on peut d'ailleurs facilement montrer qu'une méthode de point fixe approchera ξ_1 pour toute initialisation $x^{(0)}$ comprise entre $-\xi_2$ et ξ_2).

Au regard de ces définitions, certains point fixes peuvent n'être ni attractif, ni répulsif. Le théorème 5.13 montre que si la fonction g' est continue dans un voisinage de ξ, alors la condition $|g'(\xi)|<1$ suffit pour assurer que ξ est un point fixe attractif. Si $|g'(\xi)|>1$, la convergence n'a en général pas lieu, sauf si $x^{(0)}=\xi$. En effet, en reprenant la preuve du précédent théorème, on montre qu'il existe un voisinage de ξ dans lequel $|g'(x)|\geq L>1$, ce qui implique que, si le point $x^{(k)}$, $k\geq 0$, appartient à ce voisinage, alors il existe un entier $k_0\geq k+1$ telle que $x^{(k_0)}$ se trouve en dehors de celui-ci, ce qui rend la convergence impossible. Enfin, dans le cas où $|g'(\xi)|=1$, on ne peut en général tirer de conclusion : selon le problème considéré, on peut avoir soit convergence, soit divergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$.

Exemple. Soit la fonction définie par $g(x)=x-x^3$ admettant 0 pour point fixe. Bien que $g'(0)=1$, si $x^{(0)}$ appartient à $[-1,1]$ alors la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par (5.10) converge vers 0 (pour $x^{(0)}=\pm 1$, on a même $x^{(k)}=0$ pour $k\geq 1$). Par contre, pour la fonction $g(x)=x+x^3$, qui vérifie $g(0)=0$ et $g'(0)=1$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ diverge pour toute initialisation $x^{(0)}$ différente de 0.

Terminons cette section par un résultat sur l'ordre de convergence des méthodes de point fixe.

Proposition 5.15 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, g une fonction continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. Si g est de classe $\mathscr{C}^{p+1}$, avec p un entier supérieur ou égal à 1, dans un voisinage de ξ et si $g^{(i)}(\xi)=0$ pour $i=1,\dots,p$ et $g^{(p+1)}(\xi)\neq 0$, alors toute suite convergente $(x^{(k)})_{k\in\mathbb{N}}$ définie par la méthode de point fixe* (5.10) *converge avec un ordre égal à $p+1$ et l'on a*

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{(x^{(k)}-\xi)^{p+1}}=\frac{g^{(p+1)}(\xi)}{(p+1)!}.$$

DÉMONSTRATION. Par la formule de Taylor–Lagrange à l'ordre p appliquée à la fonction g au voisinage du point ξ, on obtient

$$\forall k\in\mathbb{N},\ x^{(k+1)}-\xi=\sum_{i=0}^{p}\frac{g^{(i)}(\xi)}{i!}(x^{(k)}-\xi)^i+\frac{g^{(p+1)}(\eta^{(k)})}{(p+1)!}(x^{(k)}-\xi)^{p+1}-g(\xi),$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ. Il vient alors

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{(x^{(k)}-\xi)^{p+1}}=\lim_{k\to+\infty}\frac{g^{(p+1)}(\eta^{(k)})}{(p+1)!}=\frac{g^{(p+1)}(\xi)}{(p+1)!},$$

par convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ et continuité de la fonction $g^{(p+1)}$. □

5.4.3 Méthode de relaxation

Nous avons vu dans la section 5.4.1 que l'on pouvait obtenir de diverses manières une fonction auxiliaire g dont les points fixes sont les zéros de la fonction f. Beaucoup de méthodes de point fixe courantes font cependant le choix de la forme

$$g(x)=x-\beta(x)f(x), \tag{5.15}$$

avec β une fonction satisfaisant $0<|\beta(x)|<+\infty$ sur le domaine de définition de f (ou plus généralement sur un intervalle contenant un zéro de f). Sous cette hypothèse, on vérifie facilement que tout zéro de f est point fixe de g, et vice versa.

Le choix le plus simple pour la fonction β est alors celui d'une fonction constante, ce qui conduit à la *méthode de relaxation* [18]. Cette dernière consiste en la construction d'une suite $(x^{(k)})_{k\in\mathbb{N}}$ satisfaisant la relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \beta\, f(x^{(k)}), \tag{5.16}$$

avec β un réel fixé, la valeur de $x^{(0)}$ étant donnée.

En supposant que ξ est un zéro *simple* de la fonction f, c'est-à-dire que $f(\xi) = 0$ et $f'(\xi) \neq 0$, et que f est continûment différentiable dans un voisinage de ξ, on voit qu'on peut facilement assurer la convergence locale de cette méthode si β est tel que $0 < \beta\, f'(\xi) < 2$. Ceci est rigoureusement établi dans le théorème suivant.

Théorème 5.16 (convergence locale de la méthode de relaxation) *Soit f une fonction réelle de classe $\mathscr{C}^1$ dans un voisinage d'un zéro simple ξ. Alors, il existe un ensemble de réels β tel que la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (5.16) *converge au moins linéairement vers ξ, pour toute initialisation $x^{(0)}$ choisie suffisamment proche de ξ.*

DÉMONSTRATION. Supposons que $f'(\xi) > 0$, la preuve étant identique, à des changements de signe près, si $f'(\xi) < 0$. La fonction f' étant continue dans un voisinage de ξ, on peut trouver un réel $\delta > 0$ tel que, pour tout réel x appartenant à l'intervalle $I_\delta = [\xi - \delta, \xi + \delta]$, $f'(x) \geq \frac{1}{2} f'(\xi)$. Posons alors $M = \max_{x\in I_\delta} f'(x)$. On a alors

$$\forall \beta \in \mathbb{R}_+,\ \forall x \in I_\delta,\ 1 - \beta M \leq 1 - \beta f'(x) \leq 1 - \frac{\beta}{2} f'(\xi).$$

Il suffit ensuite de choisir β de façon à ce que $\beta M - 1 = 1 - \frac{\beta}{2} f'(\xi)$, c'est-à-dire

$$\beta = \frac{4}{2M + f'(\xi)},$$

car, en posant $g(x) = x - \beta f(x)$, on obtient ainsi que

$$\forall x \in I_\delta,\ 0 \leq g'(x) \leq \frac{2M - f'(\xi)}{2M + f'(\xi)} < 1,$$

et la convergence se déduit alors du théorème 5.12. □

D'un point de vue géométrique, le point $x^{(k+1)}$ dans la relation de récurrence (5.16) est, à chaque itération, le point d'intersection entre la droite de pente $1/\beta$ passant par le point $(x^{(k)}, f(x^{(k)}))$ et l'axe des abscisses. Cette technique d'approximation est pour cette raison aussi appelée *méthode de la corde*, le nouvel itéré de la suite étant déterminé par la corde de pente constante joignant un point de la courbe de la fonction f à l'axe des abscisses (voir la figure 5.7). Connaissant un intervalle d'encadrement $[a, b]$ de ξ, on a coutume de définir la méthode en effectuant un choix particulier pour le réel β conduisant à la relation

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{b-a}{f(b) - f(a)}\, f(x^{(k)}), \tag{5.17}$$

avec $x^{(0)}$ donné dans $[a, b]$. Sous les hypothèses du théorème 5.16, la méthode converge si l'intervalle $[a, b]$ est tel que

$$b - a < 2\, \frac{f(b) - f(a)}{f'(\xi)}.$$

On remarque que la méthode de la corde converge en une itération si f est affine.

5.4.4 Méthode de Newton–Raphson

En supposant que la fonction f est de classe $\mathscr{C}^1$ et que le zéro ξ est simple, la *méthode de Newton–Raphson* fait le choix

$$\beta(x) = \frac{1}{f'(x)}$$

dans l'identité (5.15). La relation de récurrence définissant cette méthode est alors

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \tag{5.18}$$

18. On comprendra mieux l'origine de ce nom en essayant de faire le lien entre les relations de récurrence (5.16) et (3.16).

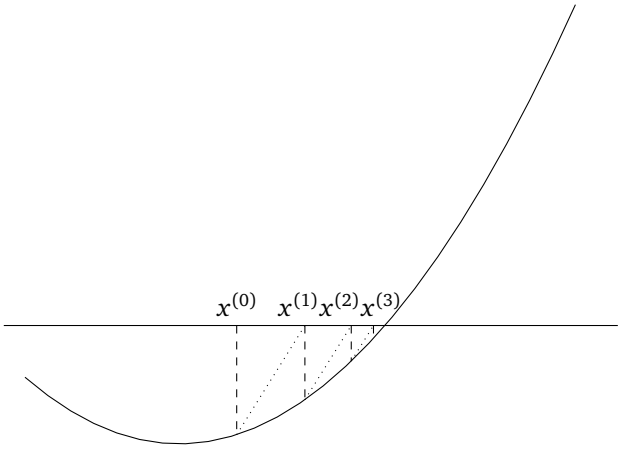


FIGURE 5.7 – Construction des premiers itérés de la méthode de la corde.

l'initialisation $x^{(0)}$ étant donnée.

Dans cette méthode, toute nouvelle approximation du zéro est construite au moyen d'une *linéarisation* de l'équation $f(x) = 0$ autour de l'approximation précédente. En effet, si l'on remplace $f(x)$ au voisinage du point $x^{(k)}$ par l'approximation affine obtenue en tronquant au premier ordre le développement de Taylor de f en $x^{(k)}$ et qu'on résout l'équation linéaire résultante

$$f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)}) = 0,$$

en notant sa solution $x^{(k+1)}$, on retrouve l'égalité (5.18). Il en résulte que, géométriquement parlant, le point $x^{(k+1)}$ est l'abscisse du point d'intersection entre la tangente à la courbe de f au point $(x^{(k)}, f(x^{(k)}))$ et l'axe des abscisses (voir la figure 5.8).

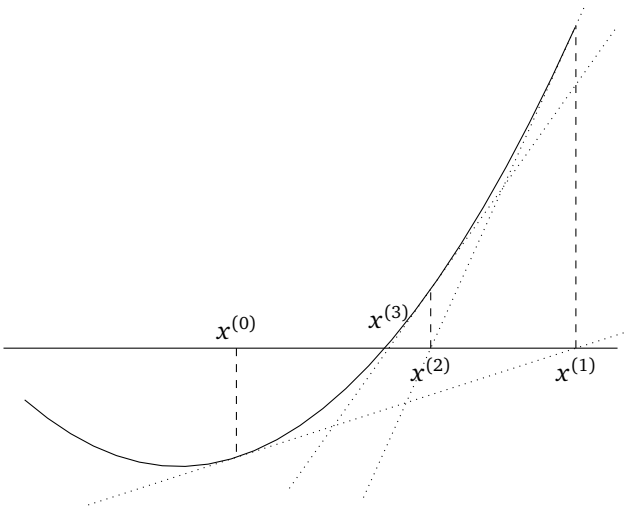


FIGURE 5.8 – Construction des premiers itérés de la méthode de Newton–Raphson.

Par rapport à toutes les méthodes introduites jusqu'à présent, on pourra remarquer que la méthode de Newton nécessite à chaque itération l'évaluation des deux fonctions f et f' au point courant $x^{(k)}$. Cet effort est compensé par une vitesse de convergence accrue, puisque cette méthode est d'ordre deux si le zéro recherché est simple.

Théorème 5.17 (convergence locale de la méthode de Newton–Raphson) *Soit f une fonction réelle de classe $\mathscr{C}^2$ dans un voisinage d'un zéro simple ξ. Alors, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (5.18) *converge au moins quadratiquement vers ξ, pour toute initialisation $x^{(0)}$ choisie suffisamment proche de ce zéro.*

DÉMONSTRATION. Nous allons tout d'abord prouver la convergence locale de la méthode et ensuite obtenir son ordre de convergence. À cette fin, introduisons, pour un réel δ strictement positif, l'ensemble $I_\delta = \{x \in \mathbb{R} \,|\, |x-\xi| \leq \delta\}$ et supposons que f soit de classe $\mathscr{C}^2$ dans ce voisinage de ξ. Définissons alors, pour δ suffisamment petit, la quantité

$$M(\delta) = \max_{\substack{s\in I_\delta \\ t\in I_\delta}} \left| \frac{f''(s)}{2f'(t)} \right|,$$

et supposons que δ soit tel que [19]

$$2\delta M(\delta) < 1. \tag{5.19}$$

Montrons à présent que le réel ξ est l'unique zéro de f contenu dans I_δ. En appliquant la formule de Taylor–Lagrange (voir le théorème B.116) à l'ordre un à la fonction f au voisinage du point ξ, on trouve que

$$f(x) = f(\xi) + (x-\xi)f'(\xi) + \frac{1}{2}(x-\xi)^2 f''(\eta),$$

avec η compris entre x et ξ.Par conséquent, si x appartient à I_δ, on a également que η appartient à I_δ et l'on obtient

$$f(x) = (x-\xi)f'(\xi)\left(1 + (x-\xi)\frac{f''(\eta)}{2f'(\xi)}\right).$$

Si x appartient à I_δ et $x \neq \xi$, les trois facteurs dans le membre de droite sont tous non nuls (le dernier parce que $\left|(x-\xi)\frac{f''(\eta)}{2f'(\xi)}\right| \leq \delta M(\delta) < \frac{1}{2}$) et la fonction f ne s'annule donc qu'en ξ sur l'intervalle I_δ. Prouvons d'autre part que la fonction f' ne s'annule pas sur I_δ. On a en effet

$$f'(x) = f'(\xi) + (x-\xi)f''(\mu),$$

avec μ compris entre x et ξ. Comme précédemment, si x appartient à I_δ, alors μ appartient à I_δ, d'où

$$\forall x \in I_\delta,\ \left|f'(x)\right| \geq \left|f'(\xi)\right| - \left|(x-\xi)f''(\mu)\right| \geq \left|f'(\xi)\right|(1 - 2\delta M(\delta)) > 0,$$

ce qui assure que la méthode de Newton donnée par (5.18) est bien définie quel que soit $x^{(k)}$ dans l'intervalle I_δ.

Montrons à présent que, pour tout choix de $x^{(0)}$ dans I_δ, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode de Newton est contenue dans l'intervalle I_δ et converge vers ξ. Tout d'abord, si, pour un entier naturel k, $x^{(k)}$ appartient à I_δ, il découle de (5.18) et de la formule de Taylor–Lagrange que

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)^2 \frac{f''(\eta^{(k)})}{2f'(x^{(k)})}, \tag{5.20}$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ, d'où

$$\left|x^{(k+1)} - \xi\right| < \frac{1}{2}\left|x^{(k)} - \xi\right| \leq \frac{\delta}{2}.$$

On en conclut que tous les termes de la suite $(x^{(k)})_{k\in\mathbb{N}}$ sont contenus dans I_δ en raisonnant par récurrence sur l'indice k. On obtient également l'estimation suivante

$$\forall k \in \mathbb{N},\ \left|x^{(k)} - \xi\right| \leq \frac{1}{2^k}\left|x^{(0)} - \xi\right| \leq \frac{\delta}{2^k},$$

ce qui implique la convergence de la méthode.

Pour établir le fait que la suite converge quadratiquement, on se sert de (5.20) pour trouver

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|^2} = \lim_{k\to+\infty} \left|\frac{f''(\eta^{(k)})}{2f'(x^{(k)})}\right| = \left|\frac{f''(\xi)}{2f'(\xi)}\right|,$$

en vertu de la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ et de la continuité de f' et f'' sur l'intervalle I_δ. □

On notera que ce théorème ne garantit la convergence de la méthode de Newton–Raphson que si l'initialisation $x^{(0)}$ est *« suffisamment proche »* du zéro recherché. L'exemple suivant montre que la méthode peut en effet diverger lorsque ce n'est pas le cas.

19. Notons que $\lim_{\delta\to 0} M(\delta) = \left|\frac{f''(\xi)}{2f'(\xi)}\right| < +\infty$, puisqu'on a fait l'hypothèse que ξ est un zéro simple de f. On peut donc bien satisfaire la condition (5.19) pour δ assez petit.

Exemple de divergence de la méthode de Newton–Raphson. Considérons la fonction $f(x) = \arctan(x)$ définie sur $\mathbb{R}$ et ayant pour zéro $\xi = 0$. La relation de récurrence définissant la méthode de Newton–Raphson pour la résolution de $f(x) = 0$ est dans ce cas

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \left(1 + (x^{(k)})^2\right) \arctan(x^{(k)}).$$

Il est possible de montrer que, si la valeur de l'initialisation $x^{(0)}$ de la méthode est telle que

$$\arctan\left(\left|x^{(0)}\right|\right) > \frac{2\left|x^{(0)}\right|}{1 + (x^{(0)})^2}, \tag{5.21}$$

alors la suite $\left(\left|x^{(k)}\right|\right)_{k\in\mathbb{N}}$ est divergente (voir la figure 5.9).

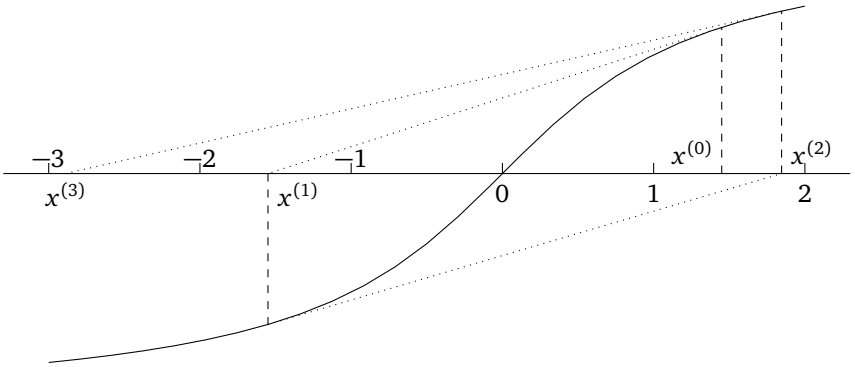


FIGURE 5.9 – Premiers itérés de la méthode de Newton–Raphson pour la résolution de l'équation $\arctan(x) = 0$ avec une initialisation $x^{(0)}$ vérifiant la condition (5.21).

Il est également important d'ajouter que, bien que la méthode de Newton–Raphson converge quadratiquement vers un zéro simple, la notion d'ordre de convergence est *asymptotique* (voir la section 5.2). De fait, on constate parfois que la méthode converge tout d'abord linéairement pour ensuite atteindre une convergence quadratique, une fois arrivé suffisamment près du zéro recherché. La figure 5.10 illustre ce phénomène.

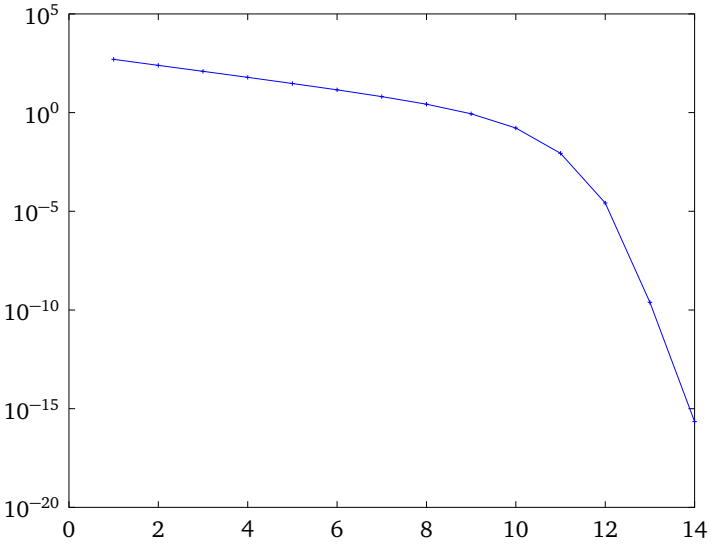


FIGURE 5.10 – Tracé de l'erreur absolue en fonction du nombre d'itérations de la méthode de Newton–Raphson utilisée pour la détermination de la racine positive de l'équation $x^2 - 2 = 0$. On a choisi $x^{(0)} = 1000$, ce qui constitue évidemment une très mauvaise estimation initiale, mais permet de mettre en évidence une période transitoire durant laquelle la convergence de la méthode est seulement *linéaire*.

On peut aussi montrer un résultat de convergence *globale* pour cette méthode, à condition que la fonction f soit strictement croissante (ou décroissante) et strictement convexe (ou concave) sur un intervalle contenant le zéro recherché.

Théorème 5.18 (convergence globale de la méthode de Newton–Raphson) *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une fonction de classe $\mathscr{C}^2$ de $[a,b]$ dans $\mathbb{R}$, changeant de signe sur $[a,b]$ et telle que $f'(x)\neq 0$ et $f''(x)\neq 0$ pour tout x appartenant à $[a,b]$. Alors, pour toute initialisation $x^{(0)}$ dans $[a,b]$ vérifiant $f(x^{(0)})f''(x^{(0)})\geq 0$ la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (5.18) *converge vers l'unique zéro ξ de f dans $[a,b]$.*

DÉMONSTRATION. Tout d'abord, les hypothèses de changement de signe de la fonction continue f et de signe constant de sa dérivée f', également continue, sur $[a,b]$ impliquent qu'il existe un unique zéro ξ appartenant à $[a,b]$. Par conséquent, si $f(x^{(0)})f''(x^{(0)})=0$, on a directement $x^{(0)}=\xi$ et la méthode est (trivialement) convergente. On suppose donc que $f(x^{(0)})f''(x^{(0)})>0$. Puisque f'' garde un signe constant sur l'intervalle $[a,b]$, on doit distinguer deux cas.
Soit, $\forall x\in[a,b]$, $f''(x)>0$ et alors $f(x^{(0)})>0$. Si de plus, $\forall x\in[a,b]$, $f'(x)>0$, on a alors, $\forall x\in[a,\xi[$, $f(x)<0$ et, $\forall x\in]\xi,b]$, $f(x)>0$, d'où $x^{(0)}\in]\xi,b]$. On vérifie alors que

$$\forall x\in]\xi,b],\ g'(x)=\frac{f(x)f''(x)}{(f'(x))^2}>0,$$

la fonction g définissant la méthode est donc strictement croissante sur $]\xi,b]$. On en déduit d'une part que

$$\xi=g(\xi)\leq g(x^{(0)})=x^{(1)},$$

et d'autre part que

$$x^{(1)}=g(x^{(0)})=x^{(0)}-\frac{f(x^{(0)})}{f'(x^{(0)})}<x^{(0)},$$

d'où $\xi\leq x^{(1)}<x^{(0)}$. En raisonnant par récurrence, on obtient que la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par (5.18) est strictement décroissante et minorée par ξ. Elle est donc convergente et a pour limite l'unique point fixe de la fonction g, ξ. Si, $\forall x\in[a,b]$, $f'(x)<0$, un raisonnement identique conduit au fait que la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par (5.18) est strictement croissante et majorée par ξ. De nouveau, cette suite est convergente et a pour limite ξ.
Enfin, si, $\forall x\in[a,b]$, $f''(x)<0$ et $f(x^{(0)})<0$, il suffit de reprendre la preuve ci-dessus en remplaçant f par $-f$ pour établir la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$. □

Exemple de convergence globale de la méthode de Newton–Raphson. On cherche à obtenir une approximation de la racine carrée $\sqrt{a}$ d'un réel strictement positif a en utilisant la méthode de Newton–Raphson pour résoudre l'équation $f(x)=0$, avec $f(x)=x^2-a=0$. Ceci se traduit par la relation de récurrence suivante pour les approximations successives de $\sqrt{a}$

$$\forall k\in\mathbb{N},\ x^{(k+1)}=\frac{1}{2}\left(x^{(k)}+\frac{a}{x^{(k)}}\right),$$

parfois appelée *méthode de Héron* (voir [Hea21]). La fonction

$$g(x)=\frac{1}{2}\left(x+\frac{a}{x}\right)$$

étant strictement convexe sur $]0,+\infty[$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode de Héron est bien définie et converge, en décroissant strictement (sauf lors de la première itération si $0<x^{(0)}<\sqrt{a}$, ou bien si $x^{(0)}=\sqrt{a}$, auquel cas la suite est constante), vers la racine carrée positive de a pour tout choix d'initialisation $x^{(0)}$ strictement positive.

Dans les deux précédents théorèmes, on a supposé que ξ était un zéro simple de la fonction f, c'est-à-dire tel que $f(\xi)=0$ et $f'(\xi)\neq 0$. Si ce zéro est de multiplicité m strictement plus grande que 1, la convergence de la méthode de Newton n'est plus du second ordre. En effet, en supposant que f est de classe $\mathscr{C}^{m+1}$ dans un voisinage de ξ, on déduit de la formule de Taylor–Young (voir le théorème B.117) que, dans ce voisinage, $f(x)=(x-\xi)^m h(x)$, où h est une fonction continue telle que $h(\xi)=\frac{f^{(m)}(\xi)}{m!}\neq 0$. On peut montrer que h est dérivable en tout point de l'intervalle sur lequel elle est définie excepté au point ξ et que $\lim\limits_{x\to\xi}(x-\xi)h'(x)=0$. En cas de convergence de la méthode, on obtient alors facilement que

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi}=g'(\xi)=1-\frac{1}{m}\neq 0,$$

et la convergence n'est donc que linéaire. Cependant, si l'entier m est connu *a priori*, on peut définir la *méthode de Newton–Raphson modifiée*,

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - m\,\frac{f(x^{(k)})}{f'(x^{(k)})}, \tag{5.22}$$

qui convergera quadratiquement le cas échéant.

En conclusion, la méthode de Newton–Raphson est la méthode de choix en termes de vitesse de convergence, puisque, dans les cas favorables, les approximations successives du zéro recherché convergent de manière quadratique, ce qui se traduit, comme on l'a indiqué dans la section 5.2, *grosso modo* par un doublement du nombre de décimales exactes de l'approximation à chaque itération de l'algorithme. Elle nécessite pour cela que, à défaut d'en connaître une expression effective, la dérivée de la fonction f puisse être évaluée en tout point donné. Si cela n'est pas le cas, on utilisera la méthode de la sécante (voir la prochaine section), dont la vitesse de convergence est moindre mais ne requiert pas que la dérivée de f soit connue.

La plus grande difficulté dans l'utilisation de la méthode de Newton–Raphson réside dans le caractère local de sa convergence. Si l'initialisation $x^{(0)}$ est trop éloignée du zéro, la méthode peut très bien diverger. Pour cette raison, il est courant dans les applications de l'associer à une méthode d'encadrement, comme la méthode de dichotomie, cette dernière permettant d'approcher, bien que lentement, le zéro recherché de manière à fournir une « bonne » initialisation pour la méthode de Newton–Raphson.

5.4.5 Méthode de Steffensen

La *méthode de Steffensen* est une autre méthode de point fixe dont la convergence peut être quadratique. Contrairement à la méthode de Newton–Raphson, elle ne nécessite pas de connaître la dérivée de la fonction dont on cherche le zéro, mais elle utilise en revanche deux évaluations de cette fonction à chaque itération. La relation de récurrence la définissant s'écrit

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{\left(f(x^{(k)})\right)^2}{f\left(x^{(k)} + f(x^{(k)})\right) - f(x^{(k)})}, \tag{5.23}$$

l'initialisation $x^{(0)}$ étant donnée.

Pour comprendre l'origine de la formule ci-dessus, il faut tout d'abord parler d'*accélération de convergence de suites numériques*, et plus particulièrement du *procédé Δ^2 d'Aitken* [Ait27]. Celui-ci vise à construire à partir d'une suite numérique $(x^{(k)})_{k\in\mathbb{N}}$ convergente une suite $(y^{(k)})_{k\in\mathbb{N}}$, possédant la même limite et dont la convergence est plus rapide, donnée par

$$\forall k \in \mathbb{N},\ y^{(k)} = x^{(k)} - \frac{(x^{(k+1)} - x^{(k)})^2}{x^{(k+2)} - 2\,x^{(k+1)} + x^{(k)}}. \tag{5.24}$$

On peut justifier la transformation (5.24) en considérant une suite dont la convergence est celle d'une suite géométrique (voir la définition B.46), c'est-à-dire une suite $(x^{(k)})_{k\in\mathbb{N}}$, de limite ξ, pour laquelle il existe un réel μ, avec $|\mu| < 1$, tel que, pour $x^{(k)} \neq \xi$,

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = \mu\,(x^{(k)} - \xi). \tag{5.25}$$

Dans ce cas particulier, on détermine en effet facilement la limite ξ et la constante μ dès que sont connues les valeurs de trois termes consécutifs de la suite $(x^{(k)})_{k\in\mathbb{N}}$, les relations

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = \mu\,(x^{(k)} - \xi) \text{ et } x^{(k+2)} - \xi = \mu\,(x^{(k+1)} - \xi),$$

donnant, par soustraction,

$$\mu = \frac{x^{(k+2)} - x^{(k+1)}}{x^{(k+1)} - x^{(k)}}$$

et, par substitution de cette dernière identité dans la première des relations,

$$\xi = \frac{x^{(k+2)}x^{(k)} - (x^{(k+1)})^2}{x^{(k+2)} - 2\,x^{(k+1)} + x^{(k)}}.$$

En introduisant l'opérateur aux différences Δ tel que $\Delta x^{(i)} = x^{(i+1)} - x^{(i)}$ et $\Delta^2 x^{(i)} = \Delta x^{(i+1)} - \Delta x^{(i)} = x^{(i+2)} - 2\,x^{(i+1)} + x^{(i)}$, $i \in \mathbb{N}$, l'égalité ci-dessus devient

$$\xi = x^{(k)} - \frac{(\Delta x^{(k)})^2}{\Delta^2 x^{(k)}}, \tag{5.26}$$

le procédé tirant son nom de cette écriture. On comprend alors que la définition (5.24) découle de l'application à toute suite $(x^{(k)})_{k\in\mathbb{N}}$ d'une extrapolation basée sur la formule (5.26), avec l'attente que la suite résultante converge plus rapidement, même si l'hypothèse (5.25) n'est pas valable. Le résultat suivant montre que c'est effectivement le cas lorsque la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ n'est qu'*asymptotiquement* géométrique.

Théorème 5.19 *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ pour laquelle il existe une constante μ, $|\mu| < 1$, et une suite $(\delta^{(k)})_{k\in\mathbb{N}}$ telles que, pour $x^{(k)} \neq \xi$,*

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = (\mu + \delta^{(k)})(x^{(k)} - \xi),\ \textit{avec}\ \lim_{k\to+\infty} \delta^{(k)} = 0.$$

Alors, pour k assez grand, les termes de la suite $(y^{(k)})_{k\in\mathbb{N}}$ obtenue par l'application du procédé Δ^2 d'Aitken (5.24) *à $(x^{(k)})_{k\in\mathbb{N}}$ sont bien définis et l'on a*

$$\lim_{k\to+\infty} \frac{y^{(k)} - \xi}{x^{(k)} - \xi} = 0.$$

DÉMONSTRATION. On a, pour tout entier naturel k,

$$\begin{aligned} x^{(k+2)} - 2\,x^{(k+1)} + x^{(k)} &= (x^{(k+2)} - \xi) - 2\,(x^{(k+1)} - \xi) + (x^{(k)} - \xi) \\ &= ((\mu + \delta^{(k+1)})(\mu + \delta^{(k)}) - 2\,(\mu + \delta^{(k)}) + 1)(x^{(k)} - \xi) \\ &= (x^{(k)} - \xi)((\mu-1)^2 + \delta^{(k+1)}\delta^{(k)} + \mu\,(\delta^{(k+1)} + \delta^{(k)}) - 2\,\delta^{(k)}). \end{aligned}$$

On en déduit que, pour k suffisamment grand, $x^{(k+2)} - 2\,x^{(k+1)} + x^{(k)} \neq 0$, puisque $x^{(k)} - \xi \neq 0$, $\mu \neq 1$ et $\lim_{k\to+\infty} \delta^{(k)} = 0$, ce qui garantit que les termes de la suite $(y^{(k)})_{k\in\mathbb{N}}$ sont bien définis à partir d'un certain rang. Étant donné que

$$x^{(k+1)} - x^{(k)} = (x^{(k+1)} - \xi) - (x^{(k)} - \xi) = (\mu - 1 + \delta^{(k)})(x^{(k)} - \xi),$$

il vient

$$y^{(k)} - \xi = (x^{(k)} - \xi) - (x^{(k)} - \xi)\,\frac{(\mu - 1 + \delta^{(k)})^2}{(\mu-1)^2 + \delta^{(k+1)}\delta^{(k)} + \mu\,(\delta^{(k+1)} + \delta^{(k)}) - 2\,\delta^{(k)}},$$

d'où

$$\lim_{k\to+\infty} \frac{y^{(k)} - \xi}{x^{(k)} - \xi} = \lim_{k\to+\infty} \left(1 - \frac{(\mu - 1 + \delta^{(k)})^2}{(\mu-1)^2 + \delta^{(k+1)}\delta^{(k)} + \mu\,(\delta^{(k+1)} + \delta^{(k)}) - 2\,\delta^{(k)}}\right) = 0.$$

□

L'utilisation du procédé Δ^2 avec une suite $(x^{(k)})_{k\in\mathbb{N}}$ issue d'une itération de point fixe (5.10) pour la résolution de l'équation $f(x) = 0$ conduit à l'obtention de la suite $(y^{(k)})_{k\in\mathbb{N}}$ définie par

$$\forall k \in \mathbb{N},\ y^{(k)} = x^{(k)} - \frac{(g(x^{(k)}) - x^{(k)})^2}{g(g(x^{(k)})) - 2\,g(x^{(k)}) - x^{(k)}}.$$

La suggestion, faite par Steffensen dans [Ste33], est d'essayer de tirer un avantage « immédiat » de l'accélération de convergence en introduisant périodiquement (toutes les deux itérations) dans la méthode de point fixe (5.10) les valeurs calculées des termes de la suite $(y^{(k)})_{k\in\mathbb{N}}$. Il en résulte une nouvelle méthode itérative, dont la relation de récurrence s'écrit

$$\forall k \in \mathbb{N},\ x^{(k+1)} = \psi(x^{(k)}), \tag{5.27}$$

avec

$$\psi(x) = x - \frac{(g(x) - x)^2}{g(g(x)) - 2\,g(x) + x} = \frac{x\,g(g(x)) - (g(x))^2}{g(g(x)) - 2\,g(x) + x},$$

la fonction ψ possèdant généralement [20] les mêmes points fixes que la fonction g. Si cette dernière est $p+1$ fois dérivable dans un voisinage d'un point fixe ξ définie une méthode convergeant à l'ordre p, avec $p \in \mathbb{N}^*$, c'est-à-dire que l'on a $g'(\xi) = \cdots = g^{(p-1)} = 0$ si $p > 1$, $g^{(p)}(\xi) \neq 0$ et le développement de Taylor–Lagrange suivant

$$g(x) = \xi + \frac{g^{(p)}(\xi)}{p!}(x-\xi)^p + \frac{g^{(p+1)}(\eta)}{(p+1)!}(x-\xi)^{p+1},$$

avec η un entier compris entre ξ et x. Il vient alors que

$$g(g(x)) = \begin{cases} \xi + O((x-\xi)^{p^2}) & \text{si } p > 1 \\ \xi + (g'(\xi))^2(x-\xi) + O((x-\xi)^2) & \text{si } p = 1 \end{cases}$$

et

$$(g(x))^2 = \ldots$$

d'où

$$\psi(x) = \ldots$$

La fonction ψ définit alors une méthode convergeant localement au moins quadratiquement si $p = 1$, à condition que $g'(\xi) \neq 1$, et à l'ordre $2p-1$ si $p > 1$. Il est intéressant de noter que la méthode est encore quadratiquement convergente même si $|g'(\xi)| > 1$ et que la méthode de point fixe initiale diverge par conséquent.

Pour la recherche de zéros d'une fonction f, le nom de Steffensen est habituellement associé à la méthode obtenue en faisant le choix particulier $g(x) = x + f(x)$, ce qui conduit à la relation de récurrence (5.23).

5.4.6 Classe des méthodes de Householder **

COMPLETER Les *méthodes de Householder* [Hou70, section 4.4], initialement introduites par Schröder [21] [Sch70], forment une classe de méthodes de recherche de zéros d'une fonction d'une variable réelle continûment dérivable $p+1$ fois, où p est l'ordre de la méthode considérée.

Idées concernant leur dérivation à partir d'une fonction supposée analytique dans un voisinage du zéro recherché

relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} + p\,\frac{\left(\frac{1}{f}\right)^{(p-1)}(x^{(k)})}{\left(\frac{1}{f}\right)^{(p)}(x^{(k)})},$$

Leur convergence est d'ordre $p+1$ pour une fonction suffisamment dérivable au voisinage d'un zéro simple.

Pour $p = 1$, on obtient la méthode de Newton–Raphson définie par (5.18)

Pour $p = 2$, on retrouve la *méthode de Halley* [22] [Hal94]

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{2f(x^{(k)})f'(x^{(k)})}{2(f'(x^{(k)}))^2 - f(x^{(k)})f''(x^{(k)})},$$

20. Par définition de ψ, on a $(x-\psi(x))(g(g(x))-2g(x)+x) = (g(x)-x)^2$. Tout point fixe de ψ est donc un point fixe de g. Supposons à présent que ξ soit un point fixe de g, en lequel la fonction est différentiable et telle que $g'(\xi) \neq 0$. L'application de la règle de L'Hôpital (voir le théorème B.111) donne alors

$$\psi(\xi) = \frac{g(g(\xi)) + \xi\, g'(\xi)g'(g(\xi)) - 2\,g'(\xi)g(\xi)}{g'(\xi)g'(g(\xi)) - 2\,g'(\xi) + 1} = \frac{\xi + \xi\,(g'(\xi))^2 - 2\xi\,g'(\xi)}{(g'(\xi))^2 - 2\,g'(\xi) + 1} = \xi.$$

21. Ernst Schröder (25 novembre 1841 - 16 juin 1902) était un mathématicien allemand. Il est un personnage majeur de l'histoire de la logique mathématique, ayant réalisé et publié un important travail de synthèse et de systématisation des divers systèmes de logique formelle de son époque.

22. Edmond Halley (8 novembre 1656 - 14 janvier 1742) était un astronome, géophysicien, mathématicien, météorologue et ingénieur britannique. Il conduisit une des premières missions d'exploration océanographique et détermina la période de la comète portant aujourd'hui son nom.

5.5 Méthode de la sécante et variantes

La *méthode de la sécante* (*secant method* en anglais) peut être considérée comme une variante de la méthode de la corde, dans laquelle la pente de la corde est mise à jour à chaque itération, ou bien une modification de la méthode de la fausse position permettant de se passer de l'hypothèse sur le signe de la fonction f aux extrémités de l'intervalle d'encadrement initial (il n'y a d'ailleurs plus besoin de connaître un tel intervalle). On peut aussi la voir comme une *quasi*-méthode de Newton–Raphson, dans laquelle on aurait remplacé la valeur $f'(x^{(k)})$ par une approximation obtenue par une différence finie. C'est l'une des méthodes que l'on peut employer lorsque la dérivée de f est compliquée, voire impossible [23], à calculer ou encore coûteuse à évaluer.

Plus précisément, à partir de la donnée de deux valeurs initiales $x^{(-1)}$ et $x^{(0)}$, telles que $x^{(-1)} \neq x^{(0)}$, la méthode de la sécante consiste en l'utilisation de la relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}), \tag{5.28}$$

pour obtenir les approximations successives du zéro recherché. Elle tire son nom de l'interprétation géométrique de (5.28) : pour tout entier positif k, le point $x^{(k+1)}$ est le point d'intersection de l'axe des abscisses avec la droite passant par les points $(x^{(k-1)}, f(x^{(k-1)}))$ et $(x^{(k)}, f(x^{(k)}))$ de la courbe représentative de la fonction f (voir la figure 5.11).

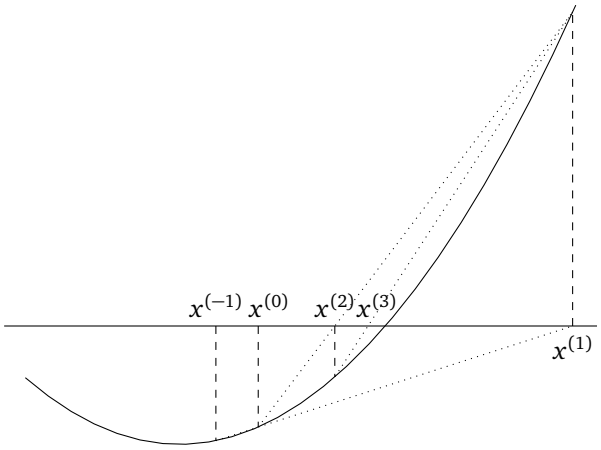


FIGURE 5.11 – Construction des premiers itérés de la méthode de la sécante.

Bien que l'on doive disposer de deux estimations de ξ avant de pouvoir utiliser la relation de récurrence (5.28), cette méthode ne requiert à chaque étape qu'une seule évaluation de fonction, ce qui est un avantage par rapport à la méthode de Newton–Raphson, dont la relation de récurrence (5.18) demande de connaître les valeurs de $f(x^{(k)})$ et de $f'(x^{(k)})$. En revanche, à la différence de la méthode de la fausse position, rien n'assure que, pour tout entier naturel k, au moins un zéro de f se trouve entre les points $x^{(k-1)}$ et $x^{(k)}$. Enfin, comparée à la méthode de la corde, elle nécessite le calcul de « mise à jour » du quotient apparaissant dans (5.28). Le bénéfice tiré de cet effort supplémentaire est une vitesse de convergence *superlinéaire*, mais cette convergence n'est que *locale*, comme le montre le résultat suivant [24].

Théorème 5.20 (convergence locale de la méthode de la sécante) *Soit f une fonction de classe $\mathscr{C}^2$ dans un voisinage d'un zéro simple ξ. Alors, si les données $x^{(-1)}$ et $x^{(0)}$, avec $x^{(-1)} \neq x^{(0)}$, choisies dans ce voisinage, sont suffisamment proches de ξ, la suite définie par* (5.28) *converge vers ξ avec un ordre au moins égal à $\frac{1}{2}(1+\sqrt{5}) = 1,6180339887\ldots$*

DÉMONSTRATION. La démonstration de ce résultat suit essentiellement les mêmes étapes que celle du théorème 5.17. Comme on l'a fait dans cette preuve, on introduit, pour $\delta > 0$, l'ensemble $I_\delta = \{x \in \mathbb{R} \,|\, |x - \xi| \leq \delta\}$ et la constante $M(\delta)$, on suppose δ

23. C'est le cas si la fonction f n'est connue qu'*implicitement*, par exemple lorsque que c'est la solution d'une équation différentielle et que x est un paramètre de la donnée initiale du problème associé.

24. Notons qu'on ne peut utiliser les techniques introduites pour les méthodes de point fixe pour établir un résultat de convergence, la relation de récurrence (5.28) définissant la méthode ne pouvant s'écrire sous la forme (5.10).

suffisamment petit pour avoir

$$\delta M(\delta) < 1.$$

et l'on montre alors que ξ est l'unique zéro de f contenu dans I_δ.

Pour prouver la convergence de la méthode de la sécante, quelles que soient les initialisations $x^{(-1)}$ et $x^{(0)}$, avec $x^{(-1)} \neq x^{(0)}$, dans I_δ, il faut faut montrer d'une part que la méthode est bien définie sur l'intervalle I_δ, c'est-à-dire que deux itérés successifs $x^{(k)}$ et $x^{(k-1)}$, $k \geq 0$ sont distincts (sauf si $f(x^{(k)}) = 0$ pour k donné, auquel cas la méthode aura convergé en un nombre fini d'itérations), et d'autre part que la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode est contenue dans I_δ et converge vers ξ.

Pour cela, on raisonne par récurrence sur l'indice k et l'on suppose que $x^{(k)}$ et $x^{(k-1)}$ appartiennent à I_δ, avec $x^{(k)} \neq x^{(k-1)}$, pour $k \in \mathbb{N}^*$. On se sert alors l'équation (5.28) définissant la méthode pour obtenir une relation faisant intervenir les trois erreurs consécutives $(x^{(i)} - \xi)$, $i = k-1, k, k+1$. En soustrayant ξ dans chaque membre de (5.28) et en utilisant que $f(\xi) = 0$, il vient

$$x^{(k+1)} - \xi = x^{(k)} - \xi - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}) = (x^{(k)} - \xi)\frac{[x^{(k-1)}, x^{(k)}]f - [x^{(k)}, \xi]f}{[x^{(k-1)}, x^{(k)}]f},$$

où l'on a noté, en employant la notation des *différences divisées* (dont on anticipe l'introduction dans le chapitre 6),

$$[x, y]f = \frac{f(x) - f(y)}{x - y}.$$

Par la relation de récurrence (6.16) pour les différences divisées, la dernière égalité s'écrit encore

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)(x^{(k-1)} - \xi)\frac{[x^{(k-1)}, x^{(k)}, \xi]f}{[x^{(k-1)}, x^{(k)}]f}.$$

Par application du théorème des accroissements finis (voir le théorème B.113), il existe $\zeta^{(k)}$, compris entre $x^{(k-1)}$ et $x^{(k)}$, et $\eta^{(k)}$, contenu dans le plus petit intervalle auquel appartiennent $x^{(k-1)}$, $x^{(k)}$ et ξ, tels que

$$[x^{(k-1)}, x^{(k)}]f = f'(\zeta^{(k)}) \text{ et } [x^{(k-1)}, x^{(k)}, \xi]f = \frac{1}{2} f''(\eta^{(k)}).$$

On en déduit que

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)(x^{(k-1)} - \xi)\frac{f''(\eta^{(k)})}{2f'(\zeta^{(k)})}, \tag{5.29}$$

d'où

$$|x^{(k+1)} - \xi| \leq \delta^2 \left|\frac{f''(\eta^{(k)})}{2f'(\zeta^{(k)})}\right| \leq \delta(\delta M(\delta)) < \delta,$$

et $x^{(k+1)}$ appartient donc à I_δ. Par ailleurs, il est clair d'après la relation (5.28) que $x^{(k+1)}$ est différent de $x^{(k)}$, excepté si $f(x^{(k)})$ est nulle.

En revenant à (5.29), il vient alors que

$$\forall k \in \mathbb{N},\ |x^{(k+1)} - \xi| \leq \delta M(\delta)|x^{(k)} - \xi| \leq (\delta M(\delta))^{k+1}|x^{(0)} - \xi|,$$

ce qui permet de prouver que la méthode converge.

On vérifie enfin que l'ordre de convergence de la méthode est au moins égal à $r = \frac{1}{2}(1 + \sqrt{5})$. On remarque tout d'abord que r satisfait

$$r^2 = r + 1.$$

On déduit ensuite de (5.29) que

$$\forall k \in \mathbb{N},\ |x^{(k+1)} - \xi| \leq M(\delta)|x^{(k)} - \xi|\,|x^{(k-1)} - \xi|.$$

En posant, $\forall k \in \mathbb{N}$, $E^{(k)} = M(\delta)|x^{(k)} - \xi|$, on obtient, après multiplication de l'inégalité ci-dessus par $M(\delta)$, la relation

$$\forall k \in \mathbb{N},\ E^{(k+1)} \leq E^{(k)} E^{(k-1)}.$$

Soit $E = \max(E^{(-1)}, {E^{(0)}}^{1/r})$. On va établir par récurrence que

$$\forall k \in \mathbb{N},\ E^{(k)} \leq E^{r^{k+1}}.$$

Cette inégalité est en effet trivialement vérifiée pour $k = 0$. En la supposant vraie jusqu'au rang k, $k \in \mathbb{N}^*$, elle est également vraie au rang $k-1$ et l'on a

$$E^{(k+1)} \leq E^{r^{k+1}} E^{r^k} = E^{r^k(r+1)} = E^{r^k r^2} = E^{r^{k+2}}.$$

Le résultat est donc valable pour tout entier positif k. En revenant à la définition de $E^{(k)}$, on obtient que

$$\forall k \in \mathbb{N},\ |x^{(k)} - \xi| \leq \varepsilon^{(k)},\ \text{avec } \varepsilon^{(k)} = \frac{1}{M(\delta)} E^{r^{k+1}},$$

avec $E < 1$ par hypothèses sur δ, $x^{(-1)}$ et $x^{(0)}$. Il reste à remarquer que

$$\forall k \in \mathbb{N},\ \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^r} = M(\delta)^{r-1},$$

et à utiliser la définition 5.3 pour conclure. □

5.5.1 Méthode de Muller

Une première généralisation de l'approche menant à la méthode de la sécante est fournie par la *méthode de Muller* [Mul56]. Dans cette dernière, disposant à chaque étape de trois approximations précédentes, $x^{(k-2)}$, $x^{(k-1)}$ et $x^{(k)}$, d'un zéro d'une fonction f, on construit le polynôme de degré deux $\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f$, dont le graphe passe par les points $(x^{(k-2)}, f(x^{(k-2)}))$, $(x^{(k-1)}, f(x^{(k-1)}))$ et $(x^{(k)}, f(x^{(k)}))$, et l'on choisit comme nouvelle approximation du zéro la racine de $\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f$ la plus proche de $x^{(k)}$. On notera que cette racine peut être complexe, même si les valeurs d'initialisation $x^{(-2)}$, $x^{(-1)}$ et $x^{(0)}$ sont toutes réelles ; c'est là un aspect important de cette méthode.

Pour obtenir une expression pour le point $x^{(k+1)}$, on fait appel à la forme de Newton du polynôme $\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f$ (voir la sous-section 6.2.2), qui s'écrit

$$\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f(x) = [x^{(k)}]f + [x^{(k-1)}, x^{(k)}]f\,(x - x^{(k)}) + [x^{(k-2)}, x^{(k-1)}, x^{(k)}]f\,(x - x^{(k)})(x - x^{(k-1)}),$$

les différences divisées $[x^{(k)}]f$, $[x^{(k-1)}, x^{(k)}]f$ et $[x^{(k-2)}, x^{(k-1)}, x^{(k)}]f$ étant définies au moyen de la relation de récurrence (6.16). Il suffit alors de poser

$$w^{(k)} = [x^{(k-1)}, x^{(k)}]f + (x^{(k)} - x^{(k-1)})\,[x^{(k-2)}, x^{(k-1)}, x^{(k)}]f$$

pour avoir

$$\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f(x) = [x^{(k)}]f + w^{(k)}\,(x - x^{(k)}) + [x^{(k-2)}, x^{(k-1)}, x^{(k)}]f\,(x - x^{(k)})^2,$$

et être ramené à trouver la valeur $x - x^{(k)}$ de plus petit module telle que $\Pi_{x^{(k-2)},x^{(k-1)},x^{(k)}} f(x) = 0$. On trouve [25]

$$x^{(k+1)} = x^{(k)} - \frac{2 f(x^{(k)})}{w \pm \sqrt{w^2 - 4[x^{(k)}]f\,[x^{(k-2)}, x^{(k-1)}, x^{(k)}]f}},$$

le signe au dénominateur de la fraction étant choisi de manière à maximiser le module de ce dernier.

En supposant la fonction f trois fois continûment différentiable dans un voisinage d'un zéro simple ξ, il vient, par utilisation de la formule de Taylor–Lagrange,

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = (x^{(k)} - \xi)(x^{(k-1)} - \xi)(x^{(k-2)} - \xi)\left(-\frac{f^{(3)}(\eta^{(k)})}{6 f'(\zeta^{(k)})}\right),$$

où $\eta^{(k)}$ et $\zeta^{(k)}$ sont des points situés entre $x^{(k)}$ et ξ. Par une technique de preuve en tout point similaire à celle utilisée pour démontrer la convergence superlinéaire de la méthode de la sécante (voir le théorème 5.20), on peut alors obtenir que la méthode de Muller converge avec un ordre au moins égal à $r \approx 1,8393$, où r est la solution réelle positive de l'équation algébrique $x^3 - x^2 - x - 1 = 0$.

5.5.2 Méthode de Brent **

La *méthode de Brent* [26] [Bre71] combine la méthode de dichotomie et la méthode de la sécante avec une technique d'*interpolation quadratique inverse*, en s'inspirant d'algorithmes introduits précédemment par Dekker [27]

25. On remarquera qu'on a employé la formule réciproque de la formule « standard » pour exprimer la solution recherchée de l'équation quadratique, car elle permet une évaluation numériquement stable de la racine en question en évitant tout problème d'annulation (voir la section 1.4.1).

26. Richard Peirce Brent (né le 20 avril 1946) est un mathématicien et informaticien australien. Ses travaux de recherche concernent notamment la théorie des nombres (et plus particulièrement la factorisation), la complexité et l'analyse des algorithmes, les générateurs de nombres aléatoires et l'architecture des ordinateurs.

27. Theodorus Jozef Dekker (né le 1
ier mars 1927) est un mathématicien hollandais. Il est l'inventeur d'un algorithme d'exclusion mutuelle, permettant à deux processus concurrents de partager une même ressource sans conflit et utilisant un segment de mémoire partagée comme mécanisme de communication.

[Dek69]. Le but de cette hybridation est de profiter des avantages respectifs de ces méthodes tout en s'affranchissant de leurs inconvénients, afin d'obtenir une méthode d'approximation, robuste, efficace et dont la convergence est rapide.

On peut résumer [28] la méthode de la manière suivante.

INTRODUIRE l'interpolation quadratique inverse

f fonction continue [29] à valeurs réelles sur un intervalle $[a, b]$ telle que $f(a)f(b) \leq 0$, th. ref assure l'existence d'au moins un zéro dans l'intervalle

la méthode fonctionne avec trois suites de points $(a^{(k)})$, $(b^{(k)})$ et $(c^{(k)})$ telles que, pour tout k,

$b^{(k)}$ est l'approximation courante du zéro ξ

$a^{(k)} = b^{(k-1)}$, le point $a^{(k)}$ pouvant coïncider avec $c^{(k)}$

$c^{(k)}$ est une ancienne valeur de a ou b telle que ξ est compris entre $b^{(k)}$ et $c^{(k)}$ ($f(b^{(k)})f(c^{(k)}) \leq 0$), avec $\left|f(b^{(k)})\right| \leq \left|f(c^{(k)})\right|$ (initialement $c^{(0)} = a$)

ALGO :

si $f(b^{(k)}) = 0$, on a terminé

si $f(b^{(k)}) \neq 0$, on pose $m = \frac{1}{2}(c^{(k)} - b^{(k)})$ (VOIR pour poser m comme point milieu $m = \frac{1}{2}(b^{(k)} + c^{(k)})$)

- si $|m| \leq \delta$ (voir comment on fixe cette tolérance), on a terminé (n renvoie $0.5(b + c)$

- sinon

- si $c^{(k)} = a^{(k)}$, on utilise la méthode de la sécante pour obtenir un point s - si $a \neq c$, alors les valeurs $f(a^{(k)})$, $f(b^{(k)})$ et $f(c^{(k)})$ sont distinctes (? on a $f(b) \neq f(c)$) et on utilise la procédé d'interpolation quadratique inverse avec les points a, b et c (approximation du zéro obtenue en évaluant en 0 polynôme d'interpolation de Lagrange de degré deux de la fonction réciproque f^{-1} associé aux points $f(a)$, $f(b)$ et $f(c)$) pour obtenir le point s

$$s = \frac{a^{(k)}f(b^{(k)})f(c^{(k)})}{(f(a^{(k)})-f(b^{(k)}))(f(a^{(k)})-f(c^{(k)}))} + \frac{b^{(k)}f(a^{(k)})f(c^{(k)})}{(f(b^{(k)})-f(a^{(k)}))(f(b^{(k)})-f(c^{(k)}))} + \frac{c^{(k)}f(a^{(k)})f(b^{(k)})}{(f(c^{(k)})-f(b^{(k)}))(f(c^{(k)})-f(a^{(k)}))}$$

si s est compris entre $b^{(k)}$ et $b^{(k)} + m$, $b'' = s$, sinon $b'' = b^{(k)} + m$ (dicho)

ensuite $b' = b''$ si $\left|b^{(k)} - b''\right| \geq \delta$, $b' = b^{(k)} + \text{sign}(m)\delta$ (pas en δ) sinon.

on prend $b^{(k+1)} = b'$ (Dekker), mais il peut arriver qu'on ne prenne que des pas en δ, ce qui ralentit fortement la convergence

Pour éviter cela, la modification de Brent est de satisfaire les deux conditions suivante :
$\delta < \left|b^{(k)} - b^{(k-1)}\right|$ si dicho à l'étape précédente ($\delta < \left|b^{(k-1)} - b^{(k-2)}\right|$ sinon) ET $\left|s - b^{(k)}\right| < \frac{1}{2}\left|b^{(k)} - b^{(k-1)}\right|$ si dicho à l'étape précédente ($\left|s - b^{(k)}\right| < \frac{1}{2}\left|b^{(k-1)} - b^{(k-2)}\right|$ sinon) pour accepter le choix $b'' = s$. On utilise $b'' = b^{(k)} + m$ sinon.

+ δ varie à chaque itération

5.6 Critères d'arrêt

Mis à part dans les cas de la méthode de dichotomie et de la méthode de Brent, nous n'avons (volontairement) pas abordé la question du *critère d'arrêt* à utiliser en pratique. En effet, s'il y a convergence, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par une méthode itérative tend vers le zéro ξ quand k tend vers l'infini, et il faut donc, comme nous l'avons fait pour les méthodes de résolution de systèmes linéaires dans le chapitre 3, introduire un critère permettant d'interrompre le processus itératif lorsque l'approximation courante de ξ est jugée « satisfaisante ». Pour cela, on a principalement le choix entre deux types de critères « qualitatifs » (imposer un nombre maximum d'itérations constituant une troisième possibilité strictement « quantitative ») : l'un basé sur l'incrément et l'autre sur le résidu.

Quel que soit le critère retenu, notons $\varepsilon > 0$ la tolérance fixée pour le calcul approché de ξ. Dans le cas d'un *contrôle de l'incrément*, les itérations s'achèveront dès que

$$\left|x^{(k+1)} - x^{(k)}\right| < \varepsilon, \tag{5.30}$$

28. L'algorithme proposé dans l'article est passablement plus compliqué que la présentation qui en est faite ici, l'auteur ayant porté un grand soin dans la prise en compte des problèmes causés par les erreurs d'arrondi et les débordements en arithmétique en précision finie.

29. VOIR remarque dans l'article de Brent sur fonction discontinue

alors qu'on mettra fin au calcul dès que

$$\left|f(x^{(k)})\right| < \varepsilon, \tag{5.31}$$

si l'on choisit de *contrôler le résidu*.

Selon les configurations, chacun de ces critères peut s'avérer plus ou moins bien adapté. Pour s'en convaincre, considérons la suite $(x^{(k)})_{k\in\mathbb{N}}$ produite par la méthode de point fixe (5.10), en supposant la fonction g continûment différentiable dans un voisinage de ξ. Par un développement au premier ordre, on obtient

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = g(x^{(k)}) - g(\xi) = g'(\eta^{(k)})(x^{(k)} - \xi),$$

avec $\eta^{(k)}$ un réel compris entre $x^{(k)}$ et ξ. On a alors

$$\forall k \in \mathbb{N},\ x^{(k+1)} - x^{(k)} = x^{(k+1)} - \xi - (x^{(k)} - \xi) = (g'(\eta^{(k)}) - 1)(x^{(k)} - \xi),$$

dont on déduit, en cas de convergence, le comportement asymptotique suivant

$$x^{(k)} - \xi \simeq \frac{1}{g'(\xi) - 1}(x^{(k+1)} - x^{(k)}).$$

Par conséquent, le critère d'arrêt (5.30), basé sur l'incrément, sera indiqué si $-1 < g'(\xi) \leq 0$ (il est d'ailleurs optimal pour une méthode de point fixe dont la convergence est au moins quadratique, c'est-à dire pour laquelle $g'(\xi) = 0$), mais très peu satisfaisant si $g'(\xi)$ est proche de 1.

Considérons maintenant le cas d'un critère basé sur le résidu, en supposant la fonction f continûment différentiable dans un voisinage d'un zéro simple ξ. En cas de convergence de la méthode et pour $k \geq 0$ assez grand, il vient, par la formule de Taylor–Young (voir le théorème B.117),

$$f(x^{(k)}) = f'(\xi)(\xi - x^{(k)}) + (\xi - x^{(k)})\epsilon(\xi - x^{(k)}),$$

avec $\epsilon(x)$ une fonction définie dans un voisinage de l'origine et tendant vers 0 quand x tend vers 0, dont on déduit l'estimation

$$\left|x^{(k)} - \xi\right| \lesssim \frac{\left|f(x^{(k)})\right|}{|f'(\xi)|}.$$

Le critère (5.31) fournira donc un test d'arrêt adéquat lorsque $|f'(\xi)| \simeq 1$, mais s'avérera trop restrictif si $|f'(\xi)| \gg 1$ ou en revanche trop optimiste si $|f'(\xi)| \ll 1$.

5.7 Méthodes pour les équations algébriques

Dans cette dernière section, nous considérons la résolution numérique d'équations algébriques, c'est-à-dire le cas pour lequel l'application f est un élément p_n de l'ensemble $\mathbb{P}_n$ des fonctions polynomiales de degré $n \geq 0$ associées aux polynômes de $\mathbb{R}_n[X]$, *i.e.*

$$p_n(x) = \sum_{i=0}^{n} a_i\, x^i, \tag{5.32}$$

les coefficients a_i, $i = 0, \ldots, n$, étant des nombres réels donnés. Les solutions de ce type d'équation, réelles ou complexes, sont encore appelées racines.

S'il est trivial de résoudre les équations algébriques du premier degré [30] et que la forme des solutions des équations du second degré [31] est bien connue, il existe aussi des expressions analytiques pour les solutions des équations de degré trois et quatre, publiées par Cardano [32] en 1545 dans son *Artis Magnæ, Sive de Regulis Algebraicis*

30. Ce sont les équations du type $a\,x + b = 0$, avec $a \neq 0$, dont la solution est donnée par $x = -\frac{b}{a}$.

31. Ce sont les équations de la forme $a\,x^2 + b\,x + c = 0$, avec $a \neq 0$, dont les solutions sont données par $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

32. Girolamo Cardano (24 septembre 1501 - 21 septembre 1576) était un mathématicien, médecin et astrologue italien. Ses travaux en algèbre, et plus précisément ses contributions à la résolution des équations algébriques du troisième degré, eurent pour conséquence l'émergence des nombres imaginaires.

Liber Unus (les formules étant respectivement dues à del Ferro [33] et Tartaglia [34] pour le troisième degré et à Ferrari [35] pour le quatrième degré). En revanche, le théorème d'Abel–Ruffini indique qu'il existe des polynômes de degré supérieur ou égal à cinq dont les racines ne s'expriment pas par radicaux. Le recours à une approche numérique se trouve par conséquent complètement motivé.

On peut directement utiliser la plupart des méthodes précédemment présentées pour la recherche de racines réelles de p_n, et certaines, comme les méthodes basées sur les itérations de point fixe, peuvent être facilement adaptées pour les racines complexes. Le besoin de déterminer l'ensemble des racines ou, au contraire, certaines racines d'un polynôme survient cependant de manière courante dans les applications, justifiant l'élaboration de méthodes tirant parti de la forme très particulière de la fonction dont on cherche les zéros. Comme on va le voir, cette dernière a inspiré une grande variété d'approches originales.

Après avoir donné des outils permettant de localiser des racines ou d'estimer leur nombre à l'intérieur d'un intervalle donné, nous commençons par introduire la *méthode de Horner* [36] servant à l'évaluation numérique efficace d'une fonction polynomiale en un point donné. Nous nous intéressons ensuite à quelques méthodes classiques et largement représentées dans la littérature (au moins pour leur intérêt théorique). Celles-ci permettent, selon les cas, la détermination d'une, de plusieurs ou de toutes les racines, de manière simultanée ou non, d'un polynôme à coefficients réels.

5.7.1 Localisation des racines **

A VOIR : borne de Cauchy (1829), suites de Sturm, etc.

5.7.2 Évaluation des fonctions polynomiales et de leurs dérivées

Nous allons à présent décrire la méthode de Horner (*Horner's rule* en anglais) [Hor19], qui permet l'évaluation efficace d'une fonction polynomiale et de ses dérivées en un point donné. Celle-ci repose sur le fait que toute fonction polynomiale $p_n \in \mathbb{P}_n$ peut s'écrire sous la forme

$$p_n(x) = a_0 + x\,(a_1 + x\,(a_2 + \cdots + x\,(a_{n-1} + a_n\,x)\ldots)). \tag{5.33}$$

Si les formes (5.32) et (5.33) sont algébriquement équivalentes, l'évaluation de la première nécessite n additions et $2n-1$ multiplications alors que celle de la seconde ne requiert que n additions et n multiplications [37].

Évaluation d'une fonction polynomiale en un point

L'algorithme pour évaluer la fonction polynomiale p_n en un point z se résume au calcul de n constantes b_i, $i = 0,\ldots,n$, définies de la manière suivante :

$$\begin{aligned} b_n &= a_n, \\ b_i &= a_i + b_{i+1}\,z,\ i = n-1, n-1, \ldots, 0, \end{aligned}$$

avec $b_0 = p_n(z)$.

33. Scipione del Ferro (6 février 1465 - 5 novembre 1526) était un mathématicien italien. Il est célèbre pour avoir été le premier à trouver la méthode de résolution des équations algébriques du troisième degré sans terme quadratique.

34. Niccolò Fontana Tartaglia (vers 1499 - 13 décembre 1557) était un mathématicien italien. Il fût l'un des premiers à utiliser les mathématiques en balistique, pour l'étude des trajectoires de boulets de canon.

35. Lodovico Ferrari (2 février 1522 - 5 octobre 1565) était un mathématicien italien. Élève de Cardano, il est à l'origine de la méthode de résolution des équations algébriques du quatrième degré.

36. William George Horner (1786 - 22 septembre 1837) était un mathématicien britannique. Il est connu pour sa méthode permettant l'approximation des racines d'un polynôme et pour l'invention en 1834 du *zootrope*, un appareil optique donnant l'illusion du mouvement.

37. La méthode de Horner est optimale, au sens où tout autre algorithme pour l'évaluation d'une fonction polynomiale arbitraire en un point donné requerra au moins autant d'opérations, en termes du nombre d'opérations arithmétiques (addition et multiplication) requises (voir [Pan66]). Pour les fonctions polynomiales de degré strictement supérieur à 4, on peut trouver des méthodes qui nécessitent moins de multiplications, mais utilisent des calculs préliminaires de coefficients. Ces dernières sont par conséquent à réserver aux situations dans lesquelles on cherche à évaluer une même fonction polynomiale en plusieurs points et la méthode de Horner reste la méthode la plus généralement employée.

Application de la méthode de Horner pour l'évaluation d'une fonction polynomiale en un point. Évaluons la fonction polynomiale $p(x) = 7x^4+5x^3-2x^2+8$ au point $z = 0,5$ par la méthode de Horner. On a $b_4 = 7$, $b_3 = 5+7\times 0,5 = 8,5$, $b_2 = -2+8,5\times 0,5 = 2,25$, $b_1 = 0+2,25\times 0,5 = 1,125$ et $b_0 = 8+1,125\times 0,5 = 8,5625$, d'où la valeur $p(z) = 8,5625$.

Il est à noter qu'on peut organiser ces calculs successifs de cet algorithme dans un tableau, ayant pour premère ligne les coefficients a_i, $i = n, n-1, \ldots, 0$, de la fonction à évaluer et comme seconde ligne les coefficients b_i, $i = n, n-1, \ldots, 0$. Ainsi, chaque élément de la seconde ligne est obtenu en multipliant l'élément situé à sa gauche par z et en ajoutant au résultat l'élément situé au dessus. Pour l'exemple d'application précédent, on trouve ainsi le tableau suivant [38]

	7	5	−2	0	8
0	7	8,5	2,25	1,125	8,5625

.

Division euclidienne d'un polynôme par un monôme

Remarquons que les opérations employées par la méthode sont celles d'un procédé de *division synthétique*. En effet, si l'on réalise la division euclidienne de $p_n(x)$ par $x-z$, il vient

$$p_n(x) = (x-z)\,q_{n-1}(x;z) + r_0, \tag{5.34}$$

où le quotient $q_{n-1}(\cdot;z) \in \mathbb{P}_{n-1}$ est une fonction dépendant de z par l'intermédiaire de ses coefficients, puisque, par identification,

$$q_{n-1}(x;z) = \sum_{i=1}^{n} b_i\, x^{i-1}, \tag{5.35}$$

et où le reste r_0 est une constante telle que $r_0 = b_0 = p_n(z)$. Ainsi, la méthode de Horner fournit un moyen simple d'effectuer très rapidement la division euclidienne d'un polynôme par un monôme de degré un.

Application de la méthode de Horner pour la division euclidienne d'un polynôme. Effectuons la division euclidienne du polynôme $4x^3-7x^2+3x-5$ par $x-2$. En construisant un tableau comme précédemment, soit

	4	−7	3	−5
0	4	1	5	5

,

on obtient $4x^3-7x^2+3x-5 = (x-2)(4x^2+x+5)+5$.

Évaluation des dérivées successives d'une fonction polynomiale en un point

Appliquons de nouveau la méthode pour effectuer la division du polynôme $q_{n-1}(x;z)$ par $(x-z)$. On trouve

$$q_{n-1}(x;z) = (x-z)\,q_{n-2}(x;z) + r_1,$$

avec $q_{n-2}(\cdot;z) \in \mathbb{P}_{n-2}$ et r_1 une constante, avec

$$q_{n-2}(x;z) = \sum_{i=2}^{n} b_i\, x^{i-2} \text{ et } r_1 = c_1,$$

les coefficients c_i, $i = 1, \ldots, n$, étant définis par

$$\begin{aligned} c_n &= b_n, \\ c_i &= b_i + c_{i+1}\,z,\ i = n-1, n-1, \ldots, 1. \end{aligned}$$

On a par ailleurs

$$p_n(x) = (x-z)^2\,q_{n-2}(x;z) + r_1\,(x-z) + r_0,$$

38. Dans ce tableau, on a ajouté une première colonne contenant 0 à la deuxième ligne afin de pouvoir réaliser la même opération pour obtenir tous les coefficients b_i, $i = 0, \ldots, n$, y compris b_n.

et, en dérivant cette dernière égalité, on trouve que $r_1 = c_1 = p_n{}'(z)$. On en déduit un procédé itératif permettant d'évaluer toutes les dérivées de la fonction polynomiale p_n au point z. On arrive en effet à

$$p_n(x) = r_n(x-z)^n + \cdots + r_1(x-z) + r_0, \tag{5.36}$$

après $n+1$ itérations de la méthode, que l'on peut résumer dans un tableau synthétique comme on l'a déjà fait

$$\begin{array}{c|cccccc} & a_n & a_{n-1} & \dots & a_2 & a_1 & a_0 \\ \hline 0 & b_n & b_{n-1} & \dots & b_2 & b_1 & r_0 \\ 0 & c_n & c_{n-1} & \dots & c_2 & r_1 & \\ \cdot & \cdot & \cdot & \dots & r_2 & & \\ \vdots & \vdots & \vdots & \iddots & & & \\ \cdot & \cdot & r_{n-1} & & & & \\ 0 & r_n & & & & & \end{array} \tag{5.37}$$

dans lequel tous les éléments n'appartenant pas à la première ligne (contenant les seuls coefficients connus initialement) ou à la première colonne sont obtenus en multipliant l'élément situé à gauche par z et en ajoutant le résultat de cette opération à l'élément situé au dessus. Par dérivations successives de (5.36), on montre alors que

$$r_j = \frac{1}{j!}\, p_n^{(j)}(z),\ j = 0, \dots, n,$$

où $p_n^{(j)}$ désigne la j^{e} dérivée de p_n.

On notera que le calcul de l'ensemble des coefficients du tableau (5.37) demande $\frac{1}{2}(n+1)n$ additions et autant de multiplications.

Stabilité numérique de la méthode de Horner

A ECRIRE (voir Wilkinson 1963) La valeur $\hat{q_0}$ calculée au point x est la valeur exacte en x d'un polynôme obtenu en perturbant avec une erreur relative d'au plus γ_{2n} les coefficients de p_n :

$$\hat{q_0} = (1+\theta_1)\,a_0 + (1+\theta_3)\,a_1 x + \cdots + (1+\theta_{2n-1})\,a_{n-1}x^{n-1} + (1+\theta_{2n})\,a_n x^n,$$

avec $|\theta_k| \leq \frac{ku}{1-ku} = \gamma_k$

erreur directe

$$|p_n(x) - \hat{q_0}| \leq \gamma_{2n} \sum_{i=0}^{n} |a_i|\,|x|^i$$

erreur relative

$$\frac{|p_n(x) - \hat{q_0}|}{|p_n(x)|} \leq \gamma_{2n} \sum_{i=0}^{n} \frac{|a_i|\,|x|^i}{|p_n(x)|} = \gamma_{2n}\psi(p_n, x)$$

$\psi(p_n, x)$ peut être arbitrairement grand, mais on a $\psi(p_n, x) = 1$ si $a_i \geq 0\ \forall i$ et $x \geq 0$, ou si $(-1)^i a_i \geq 0\ \forall i$ et $x \leq 0$.

5.7.3 Méthode de Newton–Horner

À la lecture de la sous-section précédente, on voit immédiatement que la méthode de Horner peut judicieusement être exploitée dans une implémentation de la méthode de Newton–Raphson pour l'approximation d'une racine, réelle ou complexe, d'un polynôme p_n de degré n. En effet, en déduisant de (5.34) que

$$p_n'(x) = q_{n-1}(x; z) + (x-z)\,q_{n-1}'(x; z),$$

où $q_{n-1}(\cdot; z) \in \mathbb{P}_{n-1}$ est le polynôme défini par (5.35) et p_n' et $q_{n-1}'(\cdot; z)$ désignent respectivement les dérivées de p_n et $q_{n-1}(\cdot; z)$ par rapport à x, on obtient la forme suivante pour (5.18)

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p_n'(x^{(k)})} = x^{(k)} - \frac{p_n(x^{(k)})}{q_{n-1}(x^{(k)}; x^{(k)})}, \tag{5.38}$$

qui est la relation de récurrence de la *méthode de Newton–Horner*. On remarque qu'on a seulement besoin d'avoir calculé les coefficients b_i, $i = 1, \ldots, n$, c'est-à-dire la première ligne du tableau (5.37), pour être en mesure d'évaluer le quotient dans le dernier membre de droite de (5.38). Une fois ces coefficients obtenus, le coût de chaque itération de la méthode est de $2n$ additions, $2n-1$ multiplications et une division.

Indiquons que si la racine que l'on cherche à approcher à une partie imaginaire non nulle, il est nécessaire de travailler en arithmétique complexe et de choisir une donnée initiale $x^{(0)}$ dont la partie imaginaire est non nulle.

5.7.4 Déflation

La méthode de Horner permettant d'effectuer des divisions euclidiennes de polynômes, il sera possible de calculer des approximations de l'*ensemble* des racines d'un polynôme donné p_n de degré n en opérant comme suit : une fois l'approximation d'une première racine z du polynôme p_n obtenue, on effectue une division de p_n par le monôme $(x-z)$ et on applique de nouveau une méthode de recherche de zéros au polynôme quotient pour obtenir une autre racine, et ainsi de suite... Ce procédé itératif, qui permet d'approcher successivement *toutes* les racines d'un polynôme, porte le nom de *déflation* [39]. Associé à la méthode de Newton–Horner (voir la sous-section précédente), il exploite pleinement l'efficacité de la méthode de Horner, mais on peut plus généralement faire appel à toute méthode de détermination de zéros pour la recherche des racines.

Il est important de mentionner que la déflation se trouve affectée par l'accumulation des erreurs d'arrondi au cours de chaque cycle de recherche d'une nouvelle racine. En effet, toute racine étant déterminée de manière approchée, le polynôme quotient effectivement obtenu après chaque étape a pour racines des perturbations des racines restant à trouver du polynôme initialement considéré. Les approximations des dernières racines obtenues à l'issue du processus de déflation peuvent ainsi présenter des erreurs importantes. Pour améliorer la stabilité numérique du procédé, on peut commencer par approcher la racine de plus petit module, puis continuer avec les suivantes jusqu'à celle de plus grand module. De plus, on peut, à chaque étape, améliorer la qualité de l'approximation d'une racine déjà trouvée en s'en servant comme donnée initiale de la méthode de recherche de zéros utilisée appliquée au polynôme p_n. On parle alors de phase de *raffinement*.

Lorsque l'on utilise la méthode de Newton–Horner néanmoins, il est possible de se passer de la déflation en utilisant une procédure due à Maehly [Mae54], basée sur l'observation que la dérivée du polynôme

$$q_{n-j}(x) = \frac{p_n(x)}{(x-\xi_1)\ldots(x-\xi_j)},$$

où les nombres ξ_i, $1 \leq i \leq j$ avec j un entier positif inférieur ou égal à n, sont des racines du polynôme p_n, est de la forme

$$q'_{n-j}(x) = \frac{p'_n(x)}{(x-\xi_1)\ldots(x-\xi_j)} - \frac{p_n(x)}{(x-\xi_1)\ldots(x-\xi_j)} \sum_{i=1}^{j} \frac{1}{(x-\xi_i)}.$$

Pour j strictement inférieur à n, la relation de récurrence de la méthode de Newton–Raphson pour la recherche d'une racine de q_{n-j} s'écrit alors

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{p_n(x^{(k)})}{p'_n(x^{(k)}) - \sum_{i=1}^{j} \frac{p_n(x^{(k)})}{(x^{(k)}-\xi_j)}}.$$

La méthode obtenue, parfois dite *de Newton–Maehly*, est insensible à d'éventuels problèmes d'approximation des racines $\xi_1, \ldots, \xi_j$ déjà obtenues et sa convergence reste quadratique.

5.7.5 Méthode de Bernoulli *

La *méthode de Bernoulli* [40] [Ber32] exploite le lien existant entre une *équation aux différences linéaire* (voir la sous-section 8.4.1) et son polynôme caractéristique pour obtenir la racine dominante de ce dernier. On peut en

39. Une technique similaire, portant le même nom, est utilisée pour la détermination de l'ensemble du spectre d'une matrice par la méthode de la puissance (voir la sous-section 4.4.2).

40. Daniel Bernoulli (8 février 1700 - 17 mars 1782) était un médecin, physicien et mathématicien suisse. Il est notamment célèbre pour ses applications des mathématiques à la mécanique, et plus particulièrement à l'hydrodynamique et à la théorie cinétique des gaz, et pour ses travaux en probabilités (à la base de la théorie économique de l'aversion du risque) et en statistique.

effet associer au polynôme défini par (5.32) l'équation homogène

$$\forall k \in \mathbb{N},\ a_n u_{k+n} + a_{n-1} u_{k+n-1} + \cdots + a_1 u_{k+1} + a_0 u_k = 0,$$

définissant une suite $(u_k)_{k\in\mathbb{N}}$ à partir de la donnée de n valeurs initiales u_j, $j = 0, \ldots, n-1$. Lorsque les racines de p_n sont simples, la solution de cette équation est donnée par

$$u_k =,\ k \geq n$$

les coefficients ... dépendant des valeurs initiales. En supposant alors qu'il est possible d'ordonner les racines de la manière suivante

$$|\xi_1| \leq \cdots \leq |\xi_{n-1}| < |\xi_n|,$$

c'est-à-dire en faisant l'hypothèse que la racine dominante ξ_n est unique, et que .. $\neq 0$, on peut écrire

...

et l'on a

$$\lim_{k\to+\infty} \frac{u_{k+1}}{u_k} = \xi_n.$$

Le principe de la méthode de Bernoulli est de calculer, à partir de conditions initiales adéquates, les valeurs de la suite $(u_k)_{k\in\mathbb{N}}$ de manière à approcher la racine ξ_n au moyen d'une suite de quotient. Par conditions initiales adéquates, on entend une sélection de valeurs initiales ... garantissant la non nullité du coefficient ... Pour cela, on génère des valeurs à partir du système

$$a_n u_m + a_{n-1} u_{m-1} + \cdots + a_1 u_{n-m+1} + a_0 u_{m-n} = 0,\ m = 1, \ldots, n,$$

garantissant [41] que tous les coefficients sont tous égaux à 1.

si la racine sous-dominante est la seule ayant ce module, on peut accélérer la convergence par le procédé d'Aitken

si les racines non dominantes ne sont pas simples, les propriétés de convergence restent les mêmes, mais si la racine dominante n'est pas simple, convergence ralentie

racines dominantes complexes conjuguées : on peut modifier la méthode pour déterminer le module ρ et l'argument θ

Une extension moderne de la méthode de Bernoulli est due à Rutishauser [42] et permet d'obtenir toutes les racines simultanément. Il s'agit de l'*algorithme différence-quotient* (*quotient-difference (qd) algorithm* en anglais) [Rut54].

Cette méthode déboucha sur la méthode LR pour le calcul des valeurs propres d'une matrice (voir la section 4.7).

5.7.6 Méthode de Gräffe

La *méthode de Gräffe* [43] est une méthode permettant d'approcher simultanément *toutes* les racines d'un polynôme, dont la convergence est globale [44] et quadratique. Elle repose sur la construction d'équations algébriques successives, dont chacune a pour solutions les carrés des solutions de l'équation la précédant, permettant, à partir d'un certain rang, d'obtenir facilement des approximations des racines recherchées, ou au moins de leurs modules, en utilisant les relations de Viète qui existent entre les coefficients d'un polynôme et ses zéros.

Bien que cette méthode puisse aussi bien s'appliquer à des équations algébriques à coefficients réels que complexes, nous n'allons ici considérer que le premier de ces deux cas en supposant que l'on cherche à déterminer les racines, notées ξ_i, $i = 1, \ldots, n$, du polynôme *normalisé* de degré n à coefficients réels

$$p_n(x) = x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

41. preuve ?

42. Heinz Rutishauser (30 janvier 1918 - 10 novembre 1970) était un mathématicien suisse, pionnier de l'analyse numérique et du calcul scientifique. Il est l'inventeur de plusieurs algorithmes remarquables et l'un des contributeurs à l'élaboration du premier compilateur pour le langage de programmation ALGOL 58.

43. Karl Heinrich Gräffe (7 novembre 1799 - 2 décembre 1873) était un mathématicien allemand. Il est connu pour la méthode de résolution numérique des équations algébriques qu'il développa pour répondre à une question de l'académie des sciences de Berlin.

44. Cette méthode ne nécessite en effet aucune approximation préalable des racines recherchées.

tel que le réel a_0 est non nul. Lorsque toutes les racines de ce polynôme sont simples, réelles et de valeurs absolues distinctes, des approximations des racines sont directement fournies par la méthode de Gräffe. Dans toute autre situation (existence de racines multiples et/ou complexes et/ou de modules égaux), l'obtention de ces valeurs approchées repose, comme nous allons le voir, sur diverses règles plus ou moins complexes qui font que cette technique est loin de posséder le caractère générique que l'on souhaiterait.

Le principe de la méthode consiste en l'application d'un procédé récursif de mise au carré des racines par définition d'une suite $(p_n^{(k)})_{k\in\mathbb{N}}$ de polynômes de degré n par la relation de récurrence

$$\forall k \in \mathbb{N},\ p_n^{(k+1)}(x^2) = (-1)^n\, p_n^{(k)}(x)\, p_n^{(k)}(-x), \tag{5.39}$$

et en posant $p_n^{(0)} = p_n$. On vérifie aisément que le polynôme $p_n^{(k+1)}$ ainsi défini a pour racines les carrés des racines de $p_n^{(k)}$; l'équation $p_n^{(k)}(x) = 0$ a par conséquent pour solutions les nombres $(\xi_i)^{2^k}$, $i = 1, \ldots, n$. En pratique, on ne va pas réaliser de produit de polynômes mais simplement calculer les coefficients du polynôme obtenu à chaque étape, ces derniers étant donnés par $a_i^{(0)} = a_i$, $i = 0, \ldots, n-1$, et

$$\forall k \in \mathbb{N},\ a_i^{(k+1)} = (-1)^{n-i}\left((a_i^{(k)})^2 + 2\sum_{j=1}^{\min(i,n-i)} (-1)^j a_{i-j}^{(k)} a_{i+j}^{(k)}\right),\ i = 0, \ldots, n-1.$$

On met fin aux itérations lorsque les coefficients du polynôme à l'étape courante sont égaux, à une tolérance fixée, aux carrés de ceux du polynôme obtenu à l'étape précédente.

Décrivons la mise en œuvre de cette méthode lorsque les racines recherchées sont simples, réelles et de valeurs absolues différentes, c'est-à-dire que l'on a $\xi_i \in \mathbb{R}$, $i = 1, \ldots, n$, et $|\xi_1| < |\xi_2| < \cdots < |\xi_n|$. Après r itérations du procédé, avec r un entier suffisamment grand, on trouve que

$$|\xi_n|^{2r} \approx a_{n-1}^{(r)},\ |\xi_{n-i}|^{2r} \approx \frac{a_{n-i-1}^{(r)}}{a_{n-i}^{(r)}},\ i = 1, \ldots, n-1, \tag{5.40}$$

dont on déduit des approximation des valeurs absolues des racines en réalisant n extractions de racines et $n-1$ divisions. La détermination du signe de chaque racine se fait ensuite par simple substitution de l'une des deux possibilités dans l'équation algébrique originelle $p_n(x) = 0$.

Supposons maintenant que les racines soient encore réelles et de valeurs absolues différentes, mais que l'une d'entre elles soit double, $\xi_{s+1} = \xi_s$, $1 \le s \le n-1$. Dans ce cas, le coefficient $a_{s-1}^{(k+1)}$ du polynôme $p_n^{(k+1)}$ va, à partir d'un certain rang, être approximativement égal à *la moitié* du carré du coefficient $a_{s-1}^{(k)}$ lui correspondant dans le polynôme $p_n^{(k)}$. La racine en question satisfait alors

$$|\xi_s|^{2r} \approx \frac{a_s^{(r)}}{a_{s-2}^{(r)}} \text{ et } |\xi_s|^r \approx \frac{a_{s-1}^{(r)}}{2\, a_{s-2}^{(r)}},$$

après r itérations ce qui permet de déterminer une approximation de sa valeur absolue en utilisant l'une ou l'autre de ces relations.

Si le polynôme p_n possède au moins une paire de racines complexes conjuguées, $\xi_{s+1} = \overline{\xi_s}$, une oscillation du signe du coefficient est observée à chaque itération, de façon à ce que

$$|\xi_s|^{2r} \approx \frac{a_s^{(r)}}{a_{s-2}^{(r)}} \text{ et } 2\,|\xi_s|^r \cos(r \arg(\xi_s)) \approx \frac{a_{s-1}^{(r)}}{a_{s-2}^{(r)}},$$

après r itérations si aucune autre racine ne possède le même module. Une approximation du module est $|\xi_s|$ alors facilement obtenue, tandis que la détermination d'une valeur approchée de $\arg(\xi_s)$ nécessite de tester chaque possibilité dans l'équation algébrique originelle. S'il n'existe qu'une seule paire de racines conjuguées, on peut éviter cette dernière étape en se rappelant la somme des racines est égale au coefficient $-a_{n-1}$,

$$-a_{n-1} = \xi_1 + \cdots + \xi_{s-1} + 2\,\mathrm{Re}(\xi_s) + \xi_{s+2} + \cdots + \xi_n,$$

ce qui fournit une approximation de la partie réelle des racines conjuguées une fois les valeurs approchées des $n-2$ autres racines (réelles) déterminées. Lorsque deux paires de racines conjuguées, $\xi_{s+1}=\overline{\xi_s}$ et $\xi_{t+1}=\overline{\xi_t}$, sont présentes, on a de la même manière

$$2\,(\mathrm{Re}(\xi_s)+\mathrm{Re}(\xi_t))=-(a_{n-1}+\xi_1+\cdots+\xi_{s-1}+\xi_{s+2}+\cdots+\xi_{t-1}+\xi_{t+2}+\cdots+\xi_n),$$

et on peut alors utiliser que la somme des inverses des racines est égale à $-\dfrac{a_1}{a_0}$,

$$2\left(\frac{\mathrm{Re}(\xi_s)}{|\xi_s|}+\frac{\mathrm{Re}(\xi_t)}{|\xi_t|}\right)=-\left(\frac{a_1}{a_0}+\frac{1}{\xi_1}+\cdots+\frac{1}{\xi_{s-1}}+\frac{1}{\xi_{s+2}}+\cdots+\frac{1}{\xi_{t-1}}+\frac{1}{\xi_{t+2}}+\cdots+\frac{1}{\xi_n}\right),$$

pour trouver successivement les parties réelles et imaginaires des approximations des quatre racines complexes, les modules $|\xi_s|$ et $|\xi_t|$ étant fournis par (5.40).

Il reste possible de faire appel à des arguments similaires à ceux que nous venons d'exposer dans d'autres cas de figure, mais leur implémentation dans un programme informatique reste peu évidente. Une procédure plus systématique consiste à appliquer la méthode de Gräffe à la fois à la résolution de $p_n(x)=0$ et à celle de l'équation de $p_n(x+\epsilon)=0$, avec ϵ un réel positif fixé suffisamment petit, et à considérer, pour toute racine complexe ξ, les points d'intersection dans le plan complexe des cercles respectivement centrés en l'origine et en ϵ et de rayons respectifs $|\xi|$ et $|\xi+\epsilon|$. Pour éviter des choix du réel ϵ conduisant à des correspondances incorrectes, une utilisation directe des informations obtenues en faisant tendre ϵ vers 0 a été proposée par Brodetsky et Smeal [BS24] et complétée par Lehmer [45] [Leh63].

Terminons en mentionnant le danger de débordements vers l'infini ou vers zéro en arithmétique en précision finie (voir la sous-section 1.3.2), causés par la croissance (dans le cas de racines de module strictement plus grand que 1) ou l'évanescence (dans le cas de racines de module strictement inférieur à 1) de certains coefficients de la suite de polynômes, que l'on peut éradiquer en recourant à une technique de mise à l'échelle [Gra63].

5.7.7 Méthode de Laguerre

La *méthode de Laguerre* [46] [Lag80] permet d'approcher l'une des racines d'un polynôme, généralement de manière globale et avec une convergence cubique si cette racine est simple. Étant donné un polynôme p_n de degré n, la méthode construit, à partir d'une estimation quelconque $x^{(0)}$ de l'une des racines de p_n, une suite de valeurs approchées $(x^{(k)})_{k\in\mathbb{N}}$ en utilisant la relation de récurrence

$$\forall k\in\mathbb{N},\ x^{(k+1)}=x^{(k)}-\frac{n\,p_n(x^{(k)})}{p_n'(x^{(k)})\pm\sqrt{(n-1)\big((n-1)(p_n'(x^{(k)}))^2-n\,p_n(x^{(k)})p_n''(x^{(k)})\big)}},\tag{5.41}$$

dans laquelle le signe au dénominateur de la fraction est choisi de façon à minimiser [47] la valeur de l'incrément $\left|x^{(k+1)}-x^{(k)}\right|$. En supposant que toutes les racines de p_n sont réelles et que l'approximation $x^{(k)}$ est comprise entre deux racines (ou entre $-\infty$ et la plus petite racine ou encore entre la plus grande racine et $+\infty$), l'idée présidant à la définition (5.41) est de construire une parabole coupant l'axe réel en deux points, dont les abscisses respectives appartiennent à l'intervalle considéré pour $x^{(k)}$, de manière à ce que l'abscisse de l'un de ces points soit la plus proche possible d'un zéro de p_n.

Pour des équations algébriques dont toutes les solutions sont réelles, la convergence est globale au sens où la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge pour toute valeur réelle de l'initialisation $x^{(0)}$. Supposons en effet que les racines ξ_i, $i=1,\ldots,n$ de p_n soient ordonnées de la manière suivante : $\xi_1\le\xi_2\le\cdots\le\xi_n$, et que $x^{(0)}\neq\xi_i$, $i=1,\ldots,n$. Les nombres $(x^{(0)}-\xi_i)^{-1}$, $i=1,\ldots,n$, sont alors [48] contenus dans un intervalle dont les extrémités sont $a(x^{(0)})$ et

45. Derrick Henry Lehmer (23 février 1905 - 22 mai 1991) était un mathématicien américain. Auteur de plusieurs résultats remarquables en théorie des nombres, il s'intéressa aussi aux aspects informatiques de cette discipline, notamment en inventant un test de primalité et en proposant un algorithme pour la factorisation euclidienne.

46. Edmond Nicolas Laguerre (9 avril 1834 - 14 août 1886) était un mathématicien français, qui travailla principalement dans les domaines de la géométrie et de l'analyse. Il reste surtout connu pour l'introduction des polynômes orthogonaux portant aujourd'hui son nom.

47. Lorsque la racine approchée est réelle, le signe choisi est celui de $p_n'(x^{(k)})$.

48. Il suffit de prouver que ce résultat est vrai pour une valeur donnée de l'entier i entre 1 et n, disons $i=1$. Posons $x^{(0}=z$; en écrivant que $p_n(z)=c\prod_{i=1}^n(z-\xi_i)$, $c\neq0$, on constate que l'assertion à démontrer ne dépend pas de la valeur de la constante c et l'on peut donc supposer

$b(x^{(0)})$, avec

$$a(x) = \frac{p_n'(x) - \sqrt{(n-1)\big((p_n'(x))^2 - n\,p_n(x)p_n''(x)\big)}}{n\,p_n(x)}$$

et

$$b(x) = \frac{p_n'(x) + \sqrt{(n-1)\big((p_n'(x))^2 - n\,p_n(x)p_n''(x)\big)}}{n\,p_n(x)}.$$

Si l'on a $\xi_j < x^{(0)} < \xi_{j+1}$ pour un certain entier j compris entre 1 et $n-1$, on fait l'hypothèse que $p_n(x)$ prend des valeurs strictement positives sur l'intervalle $]\xi_j, \xi_{j+1}[$ (le cas contraire se traitant de manière similaire). Il vient alors

$$a(x^{(0)}) \leq (x^{(0)} - \xi_{j+1})^{-1} < 0 < (x^{(0)} - \xi_j)^{-1} \leq b(x^{(0)})$$

d'où

$$\xi_j \leq x^{(0)} - \frac{1}{b(x^{(0)})} < x^{(0)} < x^{(0)} - \frac{1}{a(x^{(0)})} \leq \xi_{j+1}.$$

On voit dans ce cas que les deux points correspondant aux choix possibles pour $x^{(1)}$ appartiennent à l'intervalle $[\xi_j, \xi_{j+1}]$. Si l'un d'entre eux coïncide avec l'une des racines ξ_j ou ξ_{j+1}, une racine de p_n aura été obtenu. Sinon, on réitère le procédé et l'on est formellement conduit à considérer deux suites $(y^{(k)})_{k\in\mathbb{N}}$ et $(z^{(k)})_{k\in\mathbb{N}}$, respectivement définies par

$$y^{(0)} = x^{(0)},\ \forall k \in \mathbb{N},\ y^{(k+1)} = y^{(k)} - \frac{1}{b(y^{(k)})} \text{ et } z^{(0)} = x^{(0)},\ \forall k \in \mathbb{N},\ z^{(k+1)} = z^{(k)} - \frac{1}{a(z^{(k)})}.$$

La suite $(y^{(k)})_{k\in\mathbb{N}}$ est décroissante et minorée par ξ_j. Elle est donc convergente et l'on a

$$\lim_{k\to+\infty} \frac{1}{b(y^{(k)})} = 0,$$

dont on déduit que

$$\lim_{k\to+\infty} p_n(y^{(k)}) = 0$$

et donc que $\lim_{k\to+\infty} y^{(k)} = \xi_j$. De manière analogue, on obtient que la suite $(z^{(k)})_{k\in\mathbb{N}}$ a pour limite ξ_{j+1}.
Il reste à traiter les cas pour lesquels la valeur de $x^{(0)}$ n'est pas comprise entre deux racines. Supposons que $x^{(0)} > \xi_n$, le même raisonnement s'appliquant si $x^{(0)} < \xi_1$. On fait alors l'hypothèse que $p_n(x)$ prend des valeurs strictement positives sur l'intervalle $]\xi_n, +\infty[$ (le cas contraire se traitant là encore de manière similaire). On a alors

$$0 < (x^{(0)} - \xi_n)^{-1} \leq b(x^{(0)})$$

d'où

$$\xi_n \leq x^{(0)} - \frac{1}{b(x^{(0)})} < x^{(0)}.$$

En introduisant la suite $(y^{(k)})_{k\in\mathbb{N}}$ et procédant comme précédemment, on montre que la méthode converge vers la racine ξ_n.

---

que $c = 1$. En introduisant les quantités $\alpha = \sum_{i=1}^n (z-\xi_i)^{-1}$ et $\beta = \sum_{i=1}^n (z-\xi_i)^{-2}$, on a, en vertu de l'inégalité de Cauchy–Schwarz,

$$(\alpha - (z-\xi_1)^{-1})^2 = ((z-\xi_2)^{-1} + \cdots + (z-\xi_n)^{-1})^2 \leq (n-1)((z-\xi_2)^{-2} + \cdots + (z-\xi_n)^{-2}) = (n-1)(\beta - -(z-\xi_1)^{-2}).$$

Il en découle que

$$n(z-\xi_1)^{-2} - 2\,\alpha(z-\xi_1)^{-1} + \alpha^2 - (n-1)\,\beta \leq 0.$$

La fonction quadratique définie par $q(x) = n\,x^2 - 2\,\alpha x + \alpha^2 - (n-1)\,\beta$ est donc négative ou nulle en $x = (z-\xi_1)^{-1}$, tout en étant clairement strictement positive pour $|x|$ tendant vers $+\infty$. La quantité $(z-\xi_1)^{-1}$ est donc contenue entre les deux zéros de q, donnés par $n^{-1}(\alpha \pm \sqrt{(n-1)(n\beta - \alpha^2)})$. On conclut en observant que l'on a

$$\frac{p_n'(z)}{p_n(z)} = \alpha \text{ et } \frac{(p_n'(z))^2 - p_n(z)p_n''(z)}{(p_n(z))^2} = \beta.$$

Au voisinage d'une racine simple, la convergence de la méthode est cubique [49]. Elle est linéaire pour une racine multiple.

Pour un polynôme possédant des racines complexes, il n'existe pas de résultat de convergence globale de la méthode, mais de bonnes propriétés de convergence sont néanmoins observées en pratique [50]. En particulier, si $x^{(0)} = 0$, la méthode converge vers la racine de plus petit module du polynôme. Dans le cas d'un racine complexe simple, la convergence de la méthode reste d'ordre trois, cette propriété étant de nature algébrique.

5.7.8 Méthode de Durand–Kerner **

déjà utilisée par Wierstrass en 1891, redécouverte et analysée par Durand vers 1960 [Dur60] et Kerner en 1966 [Ker66]

approximation de toutes les racines du polynôme, basée sur la méthode de Newton

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = x_i^{(k)} - \frac{p_n(x_i^{(k)})}{\prod_{\substack{j=1 \\ j\neq i}}^{n} (x_i^{(k)} - x_j^{(k)})},\ i = 1, \ldots, n,$$

5.7.9 Méthode de Bairstow

La *méthode de Bairstow* [51] est utilisée pour la détermination approchée de racines d'un polynôme à coefficients réels, pour lequel les racines complexes vont par paires de valeurs conjuguées. Pour tout entier n supérieur ou égal à 2, elle consiste à écrire la fonction polynomiale $p_n(x)$ de degré n sous la forme

$$p_n(x) = (x^2 + u\,x + v)\,q_{n-2}(x) + r\,x + s, \tag{5.42}$$

où q_{n-2} une fonction polynomiale de degré $n-2$, $q_{n-2}(x) = \sum_{i=0}^{n-2} b_i\,x^i$, et à faire en sorte que les coefficients r et s soient nuls par ajustement du choix des coefficients de la fonction quadratique $x^2 + u\,x + v$, qui fournira alors deux racines de p_n. Pour cela, il suffit de voir les réels r et s comme des fonctions implicites de u et de v et de résoudre le système de deux équations non linéaires à deux inconnues

$$r(u, v) = 0 \text{ et } s(u, v) = 0,$$

ce que Bairstow suggère de faire par la méthode de Newton–Raphson généralisée aux systèmes d'équations : étant données des initialisations [52] $u^{(0)}$ et $v^{(0)}$, on cherche u et v comme les limites respectives des suites $(u^{(k)})_{k\in\mathbb{N}}$ et $(v^{(k)})_{k\in\mathbb{N}}$ définies par

$$\forall k \in \mathbb{N},\ \begin{pmatrix} u^{(k+1)} \\ v^{(k+1)} \end{pmatrix} = \begin{pmatrix} u^{(k)} \\ v^{(k)} \end{pmatrix} - J(u^{(k)}, v^{(k)})^{-1} \begin{pmatrix} r(u^{(k)}, v^{(k)}) \\ s(u^{(k)}, v^{(k)}) \end{pmatrix},$$

avec

$$J(u^{(k)}, v^{(k)}) = \begin{pmatrix} \dfrac{\partial r}{\partial u}(u^{(k)}, v^{(k)}) & \dfrac{\partial r}{\partial v}(u^{(k)}, v^{(k)}) \\ \dfrac{\partial s}{\partial u}(u^{(k)}, v^{(k)}) & \dfrac{\partial s}{\partial v}(u^{(k)}, v^{(k)}) \end{pmatrix},$$

49. Pour démontrer ce résultat, on utilise le fait que la relation de récurrence (5.41) est une itération de point fixe, de fonction

$$g(x) = x - \frac{n\,p_n(x)}{p_n'(x) \pm \sqrt{(n-1)\big((n-1)(p_n'(x))^2 - n\,p_n(x)p_n''(x)\big)}}.$$

Pour toute racine réelle simple ξ, on a alors

$$g'(\xi) = 1 - \frac{n\,p_n'(\xi)}{p_n'(\xi) \pm (n-1)\left|p_n'(\xi)\right|} \text{ et } g''(\xi) = \frac{-n\,p_n''(\xi)}{p_n'(\xi) \pm (n-1)\left|p_n'(\xi)\right|}\left(1 - \frac{2\,p_n'(\xi)}{p_n'(\xi) \pm (n-1)\left|p_n'(\xi)\right|}\left(1 \pm \frac{n-2}{2}\,\frac{p_n'(\xi)}{\left|p_n'(\xi)\right|}\right)\right),$$

ces deux quantités étant nulles si le signe choisi est celui de $p_n'(\xi)$. On peut en revanche montrer que $g'''(\xi) \neq 0$ et l'ordre de convergence de la méthode est donc égal à trois, en vertu de la proposition 5.15.

50. La méthode est en effet capable d'approcher des racines complexes, même lorsque l'initialisation est choisie réelle, la quantité apparaissant sous la racine carrée dans la relation (5.41) pouvant dans ce cas être négative.

51. Leonard Bairstow (25 juin 1880 - 8 septembre 1963) était un mécanicien britannique. Il s'intéressa principalement à l'aérodynamique ainsi qu'aux mathématiques appliquées à l'aéronautique.

52. Un choix courant est $u^{(0)} = \frac{a_{n-1}}{a_n}$ et $v^{(0)} = \frac{a_{n-2}}{a_n}$.

soit encore

$$\forall k \in \mathbb{N},\ \left\{ \begin{array}{rcl} u^{(k+1)} & = & u^{(k)} - \dfrac{1}{J^{(k)}} \left(r(u^{(k)}, v^{(k)}) \dfrac{\partial s}{\partial v}(u^{(k)}, v^{(k)}) - s(u^{(k)}, v^{(k)}) \dfrac{\partial r}{\partial v}(u^{(k)}, v^{(k)}) \right) \\ v^{(k+1)} & = & v^{(k)} - \dfrac{1}{J^{(k)}} \left(s(u^{(k)}, v^{(k)}) \dfrac{\partial r}{\partial u}(u^{(k)}, v^{(k)}) - r(u^{(k)}, v^{(k)}) \dfrac{\partial s}{\partial u}(u^{(k)}, v^{(k)}) \right) \end{array} \right. , \tag{5.43}$$

où

$$J^{(k)} = \det\left(J(u^{(k)}, v^{(k)})\right) = \frac{\partial r}{\partial u}(u^{(k)}, v^{(k)}) \frac{\partial s}{\partial v}(u^{(k)}, v^{(k)}) - \frac{\partial r}{\partial v}(u^{(k)}, v^{(k)}) \frac{\partial s}{\partial u}(u^{(k)}, v^{(k)}).$$

Pour utiliser ces formules, il faut alors être en mesure d'évaluer les dérivées de r et de s par rapport à u et à v. En identifiant (5.32) avec (5.42), on trouve que

$$b_{n-2} = a_n,\ b_{n-3} = a_{n-1} - u\, a_n = a_{n-1} - u\, b_{n-2},\ b_i = a_{i+2} - u\, b_{i+1} - v\, b_{i+2},\ i = n-4, \ldots, 0, \tag{5.44}$$

$$r = a_1 - u\, b_0 - v\, b_1,\ s = a_0 - v\, b_0,$$

et, par différentiation, il vient

$$\frac{\partial b_{n-2}}{\partial u} = 0,\ \frac{\partial b_{n-3}}{\partial u} = -b_{n-2},\ \frac{\partial b_i}{\partial u} = -b_{i+1} - u \frac{\partial b_{i+1}}{\partial u} - v \frac{\partial b_{i+2}}{\partial u},\ i = n-4, \ldots, 0, \tag{5.45}$$

$$\frac{\partial b_{n-2}}{\partial v} = 0,\ \frac{\partial b_{n-3}}{\partial v} = 0,\ \frac{\partial b_i}{\partial v} = -b_{i+2} - u \frac{\partial b_{i+1}}{\partial v} - v \frac{\partial b_{i+2}}{\partial v},\ i = n-4, \ldots, 0, \tag{5.46}$$

et

$$\frac{\partial r}{\partial u} = -b_0 - u \frac{\partial b_0}{\partial u} - v \frac{\partial b_1}{\partial u},\ \frac{\partial s}{\partial u} = -v \frac{\partial b_0}{\partial u},\ \frac{\partial r}{\partial v} = -b_1 - u \frac{\partial b_0}{\partial v} - v \frac{\partial b_1}{\partial v},\ \frac{\partial s}{\partial v} = -b_0 - v \frac{\partial b_0}{\partial v}.$$

En introduisant les coefficients c_i, $i = 0, \ldots, n-1$, définis par la relation de récurrence

$$c_{n-1} = c_{n-2} = 0,\ c_i = -b_{i+1} - u\, c_{i+1} - v\, c_{i+2},\ i = n-3, \ldots, 0, \tag{5.47}$$

on trouve en comparant respectivement (5.47) à (5.45) et (5.46) que

$$\frac{\partial b_i}{\partial u} = c_i \text{ et } \frac{\partial b_i}{\partial v} = c_{i+1},\ 0 \leq i \leq n-2,$$

d'où

$$\frac{\partial r}{\partial u} = -b_0 - u\, c_0 - v\, c_1,\ \frac{\partial s}{\partial u} = -v\, c_0,\ \frac{\partial r}{\partial v} = -b_1 - u\, c_1 - v\, c_2,\ \frac{\partial s}{\partial v} = -b_0 - v\, c_1.$$

À chaque nouvelle itération de la méthode de Newton–Raphson, il faut donc calculer les suites $(b_i)_{0 \leq i \leq n-2}$ et $(c_i)_{0 \leq i \leq n-1}$ à partir des valeurs courantes $u^{(k)}$ et $v^{(k)}$ via les relations (5.44) et (5.47) pour réaliser la mise à jour (5.43), ce procédé prenant fin lorsque une fois un critère de convergence satisfait avec une tolérance fixée.

De par sa construction, la méthode de Bairstow hérite du caractère local [53] et quadratique de la convergence de la méthode de Newton–Raphson, sauf si la multiplicité du facteur quadratique est plus grande que un, auquel cas cette convergence n'est plus que linéaire.

5.7.10 Méthode de Jenkins–Traub **

[JT70b] (pour un polynôme réel [JT70a])

53. On peut en effet montrer que la matrice jacobienne $J(u, v)$ intervenant dans la méthode est inversible dans voisinage d'un point (u^*, v^*) tel que que le facteur quadratique $x^2 + u^* x + v^*$ a pour racines deux zéros *simples* de p_n (*i. e.*, $r(u^*, v^*) = s(u^*, v^*) = 0$). En notant ces deux racines ξ_i et ξ_j, $i, j \in \{1, \ldots, n\}$, $\xi_i \neq \xi_j$, et en dérivant l'identité (5.42) par rapport à u et v, on arrive à un système de quatre égalités, qui est équivalent à l'identité matricielle

$$\begin{pmatrix} \xi_i\, q_{n-2}(\xi_i) & q_{n-2}(\xi_i) \\ \xi_j\, q_{n-2}(\xi_j) & q_{n-2}(\xi_j) \end{pmatrix} = \begin{pmatrix} \xi_i & 1 \\ \xi_j & 1 \end{pmatrix} J(u^*, v^*).$$

Le fait que la matrice $J(u^*, v^*)$ soit inversible découle alors du fait que le déterminant de la matrice dans le membre de gauche de l'équation ci-dessus, égal à $(\xi_i - \xi_j)\, q_{n-2}(\xi_i)\, q_{n-2}(\xi_j)$ est non nul par hypothèse sur les zéros ξ_i et ξ_j.

5.7.11 Recherche des valeurs propres d'une matrice compagnon **

recherche des zéros d'un polynôme par recherche des valeurs propres de sa matrice compagnon. Bien qu'il existe des algorithmes stables de recherche de valeurs propres, ce problème est extrêmement mal conditionné (à cause du choix de la base de représentation des polynômes), sauf lorsque les racines sont toutes situées sur ou à proximité du cercle unité. Dans les autres cas, il faut faire appel à une autre base, issue d'une famille de polynômes orthogonaux. La matrice associée est dite *collègue* (*colleague matrix* en anglais) [Spe60, Goo61] pour le choix des polynômes de Chebyshev, *camarade* (*comrade matrix* en anglais) pour les autres [Spe57, Bar75a, Bar75b]). Les problèmes ainsi obtenus sont bien conditionnés.

5.8 Notes sur le chapitre

La méthode de la fausse position apparaît dans un texte indien intitulé *Vaishali Ganit* datant approximativement du troisième siècle avant J.-C.. On la retrouve, utilisée pour la résolution d'équations linéaires uniquement, dans le septième des *« neuf chapitres sur l'art mathématique »*, déjà mentionnés dans la section 2.8.

La méthode de Newton–Raphson est l'une des plus célébrées des mathématiques appliquées et des plus utilisées en pratique. Elle fut décrite pour la première fois par Newton dans *De analysi per aequationes numero terminorum infinitas*, écrit en 1669, sous une forme considérablement différente de celle connue aujourd'hui, car la notion de dérivée (et donc de linéarisation) d'une fonction n'était pas encore définie à l'époque. Dans un exemple d'application, Newton s'en servit pour affiner une estimation grossière de l'une des racines de l'équation algébrique $x^3 - 2x - 5 = 0$. On la retrouve par la suite dans les deuxième et troisième éditions de l'ouvrage *Philosophiae naturalis principia mathematica* du même auteur, utilisée sous une forme géométrique pour la résolution de l'équation de Kepler. En 1690, Raphson publia dans *Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, et expedita, ex nova infinitarum serierum doctrina deducta ac demonstrata* une description simplifiée de la méthode complétée de nombreux exemples impliquant uniquement des polynômes, car il considérait cette technique de résolution comme étant purement algébrique. C'est en fait à Simpson [54] que l'on doit, dans *Essays on several curious and useful subjects, in speculative and mix'd mathematicks* paru en 1740, la première formulation de la méthode en tant que procédé itératif de résolution d'équations non linéaires générales basé sur l'utilisation du calcul de *« fluxions »*, ce dernier terme étant celui utilisé par Newton pour désigner la dérivée d'une fonction. Pour de nombreuses autres informations sur le développement historique de la méthode de Newton–Raphson, on pourra consulter l'article [Ypm95], ou encore la courte note [BJ79] sur l'origine du nom actuel de cette méthode.

Ajoutons que l'application de cette méthode se généralise naturellement à la résolution d'une équation (non linéaire) d'une variable complexe, d'un système [55] d'équations (non linéaires) ou même d'une équation fonctionnelle (non linéaire) dans un espace de Banach (la dérivée étant alors entendue au sens de la dérivée de Fréchet [56]). Elle est ainsi un élément essentiel de la démonstration [57] du *théorème de plongement isométrique de Nash* [58] [Nas56] ou

54. Thomas Simpson (20 août 1710 - 14 mai 1761) était un inventeur et mathématicien anglais, connu principalement pour la méthode d'intégration numérique portant son nom.

55. On peut par exemple l'utiliser pour calculer l'inverse d'une matrice carrée inversible A en ne se servant que de sommes et de produits de matrices, donnant lieu à la *méthode de Newton–Schulz* [Sch33] définie par la relation de récurrence

$$X^{(n+1)} = 2X^{(k)} - X^{(k)}AX^{(k)},\ k \geq 0,$$

la matrice $X^{(0)}$ étant donnée.

56. Maurice René Fréchet (2 septembre 1878 - 4 juin 1973) était un mathématicien français. Très prolifique, il fit plusieurs importantes contributions en topologie, où il introduisit par exemple le concept d'espace métrique, en analyse, en probabilités et en statistique.

57. La méthode permet plus précisément d'obtenir, par un procédé la combinant avec une technique de lissage, un théorème d'inversion locale formulé dans une classe particulière d'espaces de Fréchet. Ce résultat, qui porte le nom de *théorème de Nash–Moser* suite aux articles [Mos66a, Mos66b] et généralise le théorème des fonctions implicites, sert notamment à prouver l'unicité locale de solutions d'équations aux dérivées partielles non linéaires dans des espaces de fonctions régulières.

58. John Forbes Nash, Jr. (13 juin 1928 - 23 mai 2015) était un mathématicien américain. Il s'est principalement intéressé à la théorie des jeux, la géométrie différentielle et aux équations aux dérivées partielles. Il a partagé, avec Reinhard Selten et John Harsanyi, en 1994 le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel pour ses travaux en théorie des jeux.

encore de celle du fameux *théorème de Kolmogorov*[59]*–Arnold*[60]*–Moser*[61] [Kol54, Arn63, Mos62], qui répond à des questions d'existence et de stabilité de solutions presque périodiques de certains systèmes dynamiques (notamment ceux de la mécanique céleste). Elle est aussi utilisée pour la résolution numérique du problème d'optimisation non linéaire sans contraintes

$$\min_{x\in\mathbb{R}^d} f(x),$$

dans lequel f est supposée régulière, dont l'équation d'optimalité s'écrit

$$\boldsymbol{\nabla} f(x) = 0,$$

où $\nabla f(x)$ est le gradient de f au point x. Cette dernière équation est en effet un système de d équations à d inconnues que l'on peut résoudre par la méthode de Newton. Dans ce cas particulier, il est important de noter que la méthode construit une suite convergeant vers un point stationnaire de la fonction f, sans faire de distinction entre les minima ou les maxima. Il faut donc en général procéder à des modifications adéquates de la méthode pour la contraindre à éviter les points stationnaires qui ne sont pas des minima de f, ce qui n'est pas une tâche aisée. En partie pour cette raison, la littérature sur les applications de la méthode de Newton (et de toutes ses variantes) en optimisation est très riche. Nous renvoyons le lecteur intéressé à l'ouvrage [BGL$^+$06] en guise d'introduction.

Enfin, lorsque l'on se sert de la méthode de Newton–Raphson pour la recherche dans le plan complexe des racines d'un polynôme p, celle-ci présente ce que l'on appelle des *bassins de convergence* ou *d'attraction*. Ce sont des régions du plan complexe associées à l'une des solutions de l'équation $p(z) = 0$ de la manière suivante : un point z du plan appartient au bassin de convergence G_ξ associé à la racine ξ si la suite définie par la méthode de Newton avec z comme donnée initiale, c'est-à-dire $z^{(0)} = z$ et

$$\forall k \in \mathbb{N},\ z^{(k+1)} = z^{(k)} - \frac{p(z^{(k)})}{p'(z^{(k)})},$$

converge vers ξ. Les frontières de ces régions sont alors constituées des points pour lesquels la suite $(z^{(k)})_{k\in\mathbb{N}}$ ne converge pas. Fait remarquable, cet ensemble est une *fractale* (c'est plus précisément l'ensemble de Julia[62] associé à la fonction méromorphe[63] $z \mapsto z - \frac{p(z)}{p'(z)}$) et sa représentation donne lieu, selon le polynôme considéré, à des images surprenantes (voir la figure 5.12).

La méthode de Steffensen est particulièrement intéressante pour la résolution de systèmes d'équations non linéaires. Elle se généralise à des équations faisant intervenir des opérateurs non linéaires dans des espaces de Banach [Che64].

L'ordre de convergence obtenu pour la méthode de la sécante n'est autre que le *nombre d'or*, encore appelé la « *divine proportion* » d'après l'ouvrage *De divina proportione* de Pacioli[64], défini comme l'unique rapport entre deux longueurs telles que le rapport de la somme de ces longueurs sur la plus grande soit égal à celui de la plus grande sur la plus petite. Ce nombre irrationnel intervient, par exemple, dans la construction du pentagone régulier et ses propriétés algébriques le lient à la fameuse *suite de Fibonacci*[65].

59. Andrei Nikolaevich Kolmogorov (Андре́й Никола́евич Колмого́ров en russe, 25 avril 1903 - 20 octobre 1987) était un mathématicien russe. Ses apports aux mathématiques sont considérables et touchent divers domaines, au nombre desquels figurent la théorie des probabilités, la topologie, la logique intuitionniste, la théorie algorithmique de l'information, mais aussi la mécanique classique, la théorie des systèmes dynamiques et la théorie de la turbulence.

60. Vladimir Igorevich Arnold (Влади́мир И́горевич Арно́льд en russe, 12 juin 1937 - 3 juin 2010) était un mathématicien russe. On lui doit d'importantes contributions à la théorie des systèmes dynamiques, la topologie, la géometrie algébrique, les théories des catastrophes et des singularités, ainsi qu'à la mécanique classique.

61. Jürgen Kurt Moser (4 juillet 1928 - 17 décembre 1999) était un mathématicien américain d'origine allemande. Ses recherches portèrent sur les équations différentielles, la théorie spectrale, la mécanique céleste et la théorie de la stabilité. Il apporta des contributions fondamentales à l'étude des systèmes dynamiques.

62. Gaston Maurice Julia (3 février 1893 - 19 mars 1978) était un mathématicien français, spécialiste des fonctions d'une variable complexe. Il est principalement connu pour son remarquable *Mémoire sur l'itération des fractions rationnelles*.

63. On rappelle qu'une fonction d'une variable complexe est dite *méromorphe* si elle est holomorphe (c'est-à-dire définie et dérivable) dans tout le plan complexe, sauf éventuellement sur un ensemble de points isolés dont chacun est un pôle pour la fonction. On définit de manière informelle l'*ensemble de Julia* d'une telle fonction comme l'ensemble des nombres complexes pour lesquels une perturbation arbitrairement petite peut modifier de façon drastique la suite des valeurs itérées de la fonction qui en est issue.

64. Luca Bartolomeo de Pacioli (v. 1445 - 19 juin 1517) était un moine et mathématicien italien. Il est considéré, grâce à son recueil *Summa de arithmetica, geometria, proportioni et proportionalita* publié à Venise en 1494, comme le premier codificateur de la comptabilité moderne.

65. Leonardo Pisano, dit Fibonacci, (v. 1175 - v. 1250) était un mathématicien italien. Il reste connu de nos jours pour un problème conduisant aux nombres et à la suite qui portent son nom, mais, en son temps, ce furent surtout les applications de l'arithmétique au calcul commercial (calcul du profit des transactions, conversion entre monnaies de différents pays) qui le rendirent célèbre.

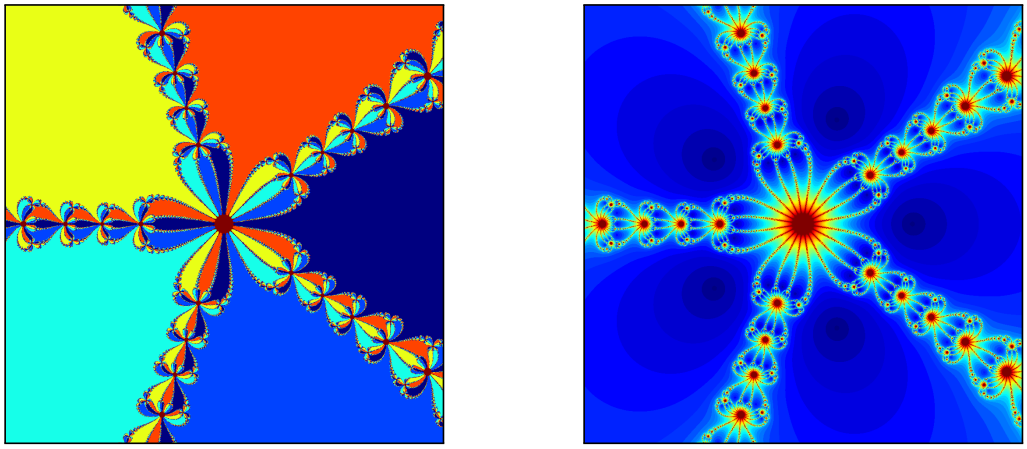


FIGURE 5.12 – Illustration de l'utilisation de la méthode de Newton pour la recherche des racines complexes de l'équation $z^5-1=0$. À gauche, on a représenté les bassins de convergence de la méthode : chaque point $z^{(0)}$ (choisi ici tel que $|\mathrm{Re}(z^{(0)})| \leq 2$ et $|\mathrm{Im}(z^{(0)})| \leq 2$) servant d'initialisation est coloré en fonction de la racine atteinte en cas de convergence (une sixième couleur étant attribuée s'il n'y a pas convergence). À droite, on a coloré ces mêmes points en fonction du nombre d'itérations requis pour atteindre la convergence avec une tolérance égale à 10^{-3} pour le critère d'arrêt. La structure fractale des frontières des bassins de convergence est clairement observée.

Les généralisations possibles de la méthode de la sécante à la résolution d'un système d'équations non linéaires sont à l'origine des classes des *quasi-méthodes de Newton* (*quasi-Newton methods* en anglais), dans lesquelles on substitue à l'inverse de la matrice jacobienne de la fonction que l'on cherche à annuler une approximation facilement mise à jour à chaque itération et à laquelle on peut choisir d'imposer certaines propriétés. Parmi les méthodes ainsi obtenues, on peut citer la *méthode de Broyden* [66] [Bro65] ou la *méthode de Davidon–Fletcher–Powell* [Dav59, FP63], utilisée pour la résolution de problèmes d'optimisation non linéaire.

La possibilité d'élever au carré de façon élémentaire les racines d'une équation algébrique afin d'en accélérer la détermination par la méthode de Newton–Raphson ou de la fausse position a été suggérée par Dandelin [67] en 1826 [Dan26], ce dernier mentionnant simplement que le procédé pouvait être réitéré et utilisé de manière à obtenir les modules des zéros ou les zéros eux-mêmes. Lobachevskii [68] redécouvrit cette méthode en 1834 et Gräffe en proposa en 1837 un algorithme pratique de mise en œuvre [Grä37]. La méthode de Gräffe est pour ces raisons parfois appelée *méthode de Dandelin–Gräffe* ou *méthode de Lobachesvkii* par certains auteurs (voir [Hou59]).

La méthode de Bairstow fut initialement introduite dans l'annexe du livre [Bai20] pour la détermination des racines d'une équation algébrique du huitième degré intervenant dans l'étude de la stabilité d'un avion.

Pour un aperçu historique et une présentation d'algorithmes récents concernant la résolution des équations algébriques, on pourra consulter l'article de Pan [Pan97].

AJOUTER des explications relatives à la référence [Boy02]

Références

[AB73] N. ANDERSON and Å. BJÖRCK. A new high order method of regula falsi type for computing a root of an equation. *BIT*, 13(3):253–264, 1973. DOI: 10.1007/BF01951936.

[Ait27] A. C. AITKEN. On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinburgh*, 46:289–305, 1927. DOI: 10.1017/S0370164600022070.

[Arn63] V. I. ARNOLD. Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian (russian). *Uspehi Mat. Nauk*, 18(5(113)):13–40, 1963.

66. Charles George Broyden (3 février 1933 - 20 mai 2011) était un mathématicien anglais, spécialiste de la résolution numérique de problèmes d'optimisation non linéaire et d'algèbre linéaire.

67. Germinal Pierre Dandelin (12 avril 1794 - 15 février 1847) était un mathématicien belge. Ses travaux portèrent sur la géométrie et plus particulièrement sur les coniques.

68. Nikolai Ivanovich Lobachevskii (Никола́й Ива́нович Лобаче́вский en russe, 1er décembre 1792 - 24 février 1856) était un mathématicien russe, inventeur d'une géométrie hyperbolique non-euclidienne.

[Bai20] L. BAIRSTOW. *Applied aerodynamics*. Longmans, Green and co., 1920.

[Ban22] S. BANACH. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. Math.*, 3 :133-181, 1922.

[Bar75a] S. BARNETT. A companion matrix analogue for orthogonal polynomials. *Linear Algebra and Appl.*, 12(3):197–202, 1975. DOI: 10.1016/0024-3795(75)90041-5.

[Bar75b] S. BARNETT. Some applications of the comrade matrix. *Internat. J. Control*, 21(5):849–855, 1975. DOI: 10.1080/00207177508922039.

[Ber32] D. BERNOULLI. Methodus universalis determinandae curvaturae fili a potentiis quamcunque legem inter se observantibus extensi, una cum solutione problematum quorundam novorum eo pertinentium. *Comm. Acad. Sci. Imper. Petropolitanae*, 3:62–69, 1732.

[BGL+06] J. F. BONNANS, J. C. GILBERT, C. LEMARÉCHAL, and C. A. SAGASTIZÁBAL. *Numerical optimization, theoretical and practical aspects*, *Universitext*. Springer, second edition, 2006. DOI: 10.1007/978-3-540-35447-5.

[BJ79] N. BIĆANIĆ and K. H. JOHNSON. Who was '–Raphson'? *Internat. J. Numer. Methods Engrg.*, 14(1):148–152, 1979. DOI: 10.1002/nme.1620140112.

[Boy02] J. P. BOYD. Computing zeros on a real interval through Chebyshev expansion and polynomial rootfinding. *SIAM J. Numer. Anal.*, 40(5):1666–1682, 2002. DOI: 10.1137/S0036142901398325.

[Bre71] R. P. BRENT. An algorithm with guaranteed convergence for finding a zero of a function. *Comput. J.*, 14(4):422–425, 1971. DOI: 10.1093/comjnl/14.4.422.

[Bro65] C. G. BROYDEN. A class of methods for solving nonlinear simultaneous equations. *Math. Comput.*, 19(92):577–593, 1965. DOI: 10.1090/S0025-5718-1965-0198670-6.

[BS24] S. BRODETSKY and G. SMEAL. On Graeffe's method for complex roots of algebraic equations. *Math. Proc. Cambridge Philos. Soc.*, 22(2):83–87, 1924. DOI: 10.1017/S0305004100002802.

[Che64] K.-W. CHEN. Generalization of Steffensen's method for operator equations in Banach space. *Comment. Math. Univ. Carolinae*, 5(2):47–77, 1964.

[Col93] P. COLWELL. *Solving Kepler's equation over three centuries*. Willmann–Bell, 1993.

[Dan26] G. DANDELIN. Recherches sur la résolution des équations numériques. In *Nouveaux mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*. Tome 3, pages 7-71. P. J. De Mat, imprimeur de l'Académie Royale, Bruxelles, 1826.

[Dav59] W. C. DAVIDON. Variable metric method for minimization. Technical report ANL-5990, A.E.C. Research and Development, 1959. DOI: 10.2172/4252678.

[Dek69] T. J DEKKER. Finding a zero by means of successive linear interpolation. In B. DEJON and P. HENRICI, editors, *Constructive aspects of the fundamental theorem of algebra*, pages 37–48. Wiley-Interscience, 1969.

[DJ71] M. DOWELL and P. JARRATT. A modified regula falsi method for computing the root of an equation. *BIT*, 11(2):168–174, 1971. DOI: 10.1007/BF01934364.

[DJ72] M. DOWELL and P. JARRATT. The "Pegasus" method for computing the root of an equation. *BIT*, 12(4):503–508, 1972. DOI: 10.1007/BF01932959.

[Dur60] E. DURAND. *Solutions numériques des équations algébriques. Tome I équations du type $F(x) = 0$, racines d'un polynôme*. Masson, 1960.

[FP63] R. FLETCHER and M. J. D. POWELL. A rapidly convergent descent method for minimization. *Comput. J.*, 6(2):163–168, 1963. DOI: 10.1093/comjnl/6.2.163.

[Goo61] I. J. GOOD. The colleague matrix, a Chebyshev analogue of the companion matrix. *Quart. J. Math. Oxford Ser. (2)*, 12(1):61–68, 1961. DOI: 10.1093/qmath/12.1.61.

[Grä37] C. H. GRÄFFE. Die Auflösung der höheren numerischen Gleichungen, als Beantwortung einer von der Königl. Akademie der Wissenschaften zu Berlin aufgestellten Preisfrage, 1837.

[Gra63] A. A. GRAU. On the reduction of number range in the use of the Graeffe process. *J. Assoc. Comput. Mach.*, 10(4):538–544, 1963. DOI: 10.1145/321186.321198.

[Hal94] E. HALLEY. Methodus nova accurata et facilis inveniendi radices aequationum quarumcumque generaliter, sine praevia reductione. *Philos. Trans. Roy. Soc. London*, 18:136–148, 1694. DOI: 10.1098/rstl.1694.0029.

[Hea21] T. HEATH. *A history of Greek mathematics. Volume II from Aristarchus to Diophantus*. Clarendon Press, 1921.

[Hor19] W. G. HORNER. A new method of solving numerical equations of all orders, by continuous approximation. *Philos. Trans. Roy. Soc. London*, 109:308–335, 1819. DOI: 10.1098/rstl.1819.0023.

[Hou59] A. S. HOUSEHOLDER. Dandelin, Lobachévskiĭ, or Graeffe? *Amer. Math. Monthly*, 66(6):464–466, 1959. DOI: 10.2307/2310626.

[Hou70] A. S. HOUSEHOLDER. *The numerical treatment of a single nonlinear equation*. McGraw-Hill, 1970.

[JT70a] M. A. JENKINS and J. F. TRAUB. A three-stage algorithm for real polynomials using quadratic iteration. *SIAM J. Numer. Anal.*, 7(4):545–566, 1970. DOI: 10.1137/0707045.

[JT70b] M. A. JENKINS and J. F. TRAUB. A three-stage variable-shift iteration for polynomial zeros and its relation to generalized Rayleigh iteration. *Numer. Math.*, 14(3):252–263, 1970. DOI: 10.1007/BF02163334.

[Ker66] I. O. KERNER. Ein Gesamtschrittverfahren zur Berechnung der Nullstellen von Polynomen. *Numer. Math.*, 8(3):290–294, 1966. DOI: 10.1007/BF02162564.

[Kol54] A. N. KOLMOGOROV. On conservation of conditionally periodic motions for a small change in Hamilton's function (russian). *Dokl. Akad. Nauk SSSR*, 98:527–530, 1954.

[Lag80] E. N. LAGUERRE. Sur une méthode pour obtenir par approximation les racines d'une équation algébrique qui a toutes ses racines réelles. *Nouv. Ann. Math. (2)*, 19 :193-202, 1880.

[Leh63] D. H. LEHMER. The complete root-squaring method. *SIAM J. Appl. Math.*, 11(3):705–717, 1963. DOI: 10.1137/0111053.

[Mae54] H. J. MAEHLY. Zur iterativen Auflösung algebraischer Gleichungen. *Z. Angew. Math. Phys.*, 5(3):260–263, 1954. DOI: 10.1007/BF01600333.

[Mos62] J. MOSER. On invariant curves of area-preserving mappings of an annulus. *Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II*, 1962(1):1–20, 1962.

[Mos66a] J. MOSER. A rapidly convergent iteration method and non-linear differential equations - I. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 20(2):265–315, 1966.

[Mos66b] J. MOSER. A rapidly convergent iteration method and non-linear differential equations - II. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 20(3):499–535, 1966.

[Mul56] D. E. MULLER. A method for solving algebraic equations using an automatic computer. *Math. Tables Aids Comp.*, 10(56):208–215, 1956. DOI: 10.1090/S0025-5718-1956-0083822-0.

[Nas56] J. NASH. The imbedding problem for Riemannian manifolds. *Ann. Math. (2)*, 63(1):20–63, 1956. DOI: 10.2307/1969989.

[Pan66] V. Y. PAN. On means of calculating values of polynomials (russian). *Uspehi Mat. Nauk*, 21(1):103–134, 1966.

[Pan97] V. Y. PAN. Solving a polynomial equation: some history and recent progress. *SIAM Rev.*, 39(2):187–220, 1997. DOI: 10.1137/S0036144595288554.

[Rid79] C. J. F. RIDDERS. A new algorithm for computing a single root of a real continuous function. *IEEE Trans. Circuits Syst.*, 26(11):979–980, 1979. DOI: 10.1109/TCS.1979.1084580.

[Rut54] H. RUTISHAUSER. Der Quotienten-Differenzen-Algorithmus. *Z. Angew. Math. Phys.*, 5(3):233–251, 1954. DOI: 10.1007/BF01600331.

[Sch33] G. SCHULZ. Iterative Berechung der reziproken Matrix. *Z. Angew. Math. Mech.*, 13(1):57–59, 1933. DOI: 10.1002/zamm.19330130111.

[Sch70] E. SCHRÖDER. Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen. *Math. Ann.*, 2(2):317–365, 1870. DOI: 10.1007/BF01444024.

[Spe57] W. SPECHT. Die Lage der Nullstellen eines Polynoms. III. *Math. Nachr.*, 16(5-6):369–389, 1957. DOI: 10.1002/mana.19570160509.

[Spe60] W. SPECHT. Die Lage der Nullstellen eines Polynoms. IV. *Math. Nachr.*, 21(3-5):201–222, 1960. DOI: 10.1002/mana.19600210307.

[Ste33] J. F. STEFFENSEN. Remarks on iteration. *Skand. Aktuar.*, 1933(1):64–72, 1933. DOI: 10.1080/03461238.1933.10419209.

[Tra64] J. F. TRAUB. *Iterative methods for the solution of equations*. Prentice-Hall, 1964.

[vdWaa73] J. D. van der WAALS. Over de continuiteit van den gas- en vloeistoftoestand, 1873.

[Ypm95] T. J. YPMA. Historical development of the Newton–Raphson method. *SIAM Rev.*, 37(4):531–551, 1995. DOI: 10.1137/1037125.

Chapitre 6

Interpolation polynomiale

Historiquement [1], le terme d'*interpolation* regroupe l'ensemble des techniques permettant de construire une courbe d'un type donné passant par un nombre fini de points donnés du plan. D'un point de vue applicatif, les ordonnées de ces points peuvent représenter les valeurs aux abscisses d'une fonction arbitraire, que l'on cherche dans ce cas à remplacer par une fonction plus simple à manipuler lors d'un calcul numérique, ou encore de données expérimentales, pour lesquelles on vise à obtenir empiriquement une loi de distribution lorsque leur nombre est important. Sous sa forme la plus simple, l'interpolation *linéaire*, ce procédé est bien connu des utilisateurs de tables de logarithmes, qui furent massivement employées pour les calculs avant l'arrivée des calculatrices. C'est aussi un ingrédient essentiel de nombreuses et diverses [2] méthodes numériques, ainsi que de techniques d'estimation statistique (comme le *krigeage* [3] en géostatistique par exemple).

Nous nous limiterons dans ces pages à des problèmes d'interpolation *polynomiale*, ce qui signifie que la courbe que l'on cherche à obtenir est le graphe d'une fonction polynomiale (éventuellement par morceaux). Ce choix n'est, de loin, pas le seul possible : l'*interpolation trigonométrique*, basée sur les polynômes trigonométriques, est en effet largement utilisée pour l'interpolation des fonctions périodiques et la mise en œuvre de techniques en lien avec l'analyse de Fourier [4], l'*interpolation rationnelle* se sert de quotients de polynômes, etc. Cependant, les nombreuses propriétés analytiques et algébriques des polynômes, alliées à la facilité que l'on a à les dériver, les intégrer ou les évaluer numériquement en un point, en font une classe de fonctions extrêmement intéressante en pratique. L'interpolation polynomiale est pour cette raison un outil numérique de premier ordre pour l'*approximation polynomiale* des fonctions réelles d'une variable réelle, dont nous rappelons en introduction plusieurs résultats fondamentaux.

Après avoir ainsi en partie motivé la problématique de l'interpolation, nous étudions en détail l'*interpolation de Lagrange* [5], qui constitue certainement la base théorique principale de l'interpolation polynomiale, et son application à l'approximation d'une fonction réelle. Des généralisations de ce procédé sont ensuite explorées et quelques exemples d'*interpolation par morceaux* concluent le chapitre.

1. Le lecteur intéressé par l'histoire de l'interpolation pourra consulter l'article [Mei02] qui en retrace les développements de l'Antiquité jusqu'au début du XXIe siècle.

2. Parmi les méthodes présentées dans ces notes et faisant intervenir l'interpolation, on peut citer les méthodes de recherche de zéro de la sous-section 5.3.2 et de la section 5.5, les formules de quadrature du chapitre 7 ou les méthodes à pas multiples linéaires de la sous-section 8.3.3.

3. A REPRENDRE Le krigeage est une méthode d'interpolation spatiale dans laquelle les valeurs interpolées sont modélisées par un processus gaussien gourverné par unz covariance entre points déterminée par interprétation d'un variogramme expérimental. Sous des hypothèses convenables sur les ..., cette méthode fournit le meilleur estimateur linéaire non-biaisé des valeurs intermédiaires.

4. Joseph Fourier (21 mars 1768 - 16 mai 1830) était un mathématicien et physicien français, connu pour ses travaux sur la décomposition de fonctions périodiques en séries trigonométriques convergentes et leur application au problème de la propagation de la chaleur.

5. Joseph Louis Lagrange (Giuseppe Lodovico Lagrangia en italien, 25 janvier 1736 - 10 avril 1813) était un mathématicien et astronome franco-italien. Fondateur, avec Euler, du calcul des variations, il a également produit d'importantes contributions tant en analyse, en géométrie et en théorie des groupes qu'en mécanique.

6.1 Quelques résultats concernant l'approximation polynomiale

Dans la suite, pour toute fonction g à valeurs réelles définie sur un intervalle $[a,b]$ borné et non vide de $\mathbb{R}$, la *norme de la convergence uniforme de g sur $[a,b]$* sera notée

$$\|g\|_\infty = \max_{x\in[a,b]} |g(x)|.$$

INTRODUIRE ici les normes $\|\cdot\|_\infty$ et $\|\cdot\|_2$ (à poids ?) sur $\mathscr{C}([a,b])$ (intervalle sous-jacent clair d'après le contexte), remarque sur l'absence d'équivalence (dimension infinie), le choix de la norme influera fortement sur le résultat obtenu et dépendra *a priori* de l'application visée...

6.1.1 Polynômes et fonctions polynomiales

COMPLETER notation de $\mathbb{P}_n$ et amalgame notation polynôme/fonction polynomiale associée

6.1.2 Approximation uniforme

Le bien-fondé de l'approximation polynomiale repose sur le résultat suivant, qui montre qu'il est possible d'approcher, sur un intervalle borné et de manière arbitraire relativement à la norme de la convergence uniforme, toute fonction continue par une fonction polynomiale de degré suffisamment élevé.

Théorème 6.1 (« théorème d'approximation de Weierstrass [6] » [Wei85]) *Soit f une fonction d'une variable réelle à valeurs réelles, continue sur un intervalle $[a,b]$ borné et non vide de $\mathbb{R}$. Alors, pour tout réel ε strictement positif, il existe une fonction polynomiale p telle que*

$$\|f-p\|_\infty < \varepsilon.$$

Il existe de nombreuses démonstrations de ce résultat. La preuve originelle de Weierstrass est basée sur l'analyticité d'intégrales singulières, obtenues par convolution d'une extension de la fonction à approcher avec une fonction gaussienne, solutions d'une équation de la chaleur sur la droite réelle. Une démonstration constructive célèbre, suivant une approche probabiliste, est due à Bernstein [7] [Ber13]. Nous reproduisons ici une preuve particulièrement simple proposée par Kuhn [Kuh64].

DÉMONSTRATION. On observe que, par un changement de variable, il suffit de démontrer le résultat pour une fonction f définie et continue sur l'intervalle $[0,1]$. Nous allons tout d'abord montrer qu'il existe une fonction affine par morceaux approchant uniformément f sur $[0,1]$. La fonction f étant uniformément continue sur $[0,1]$ en vertu du théorème de Heine (voir le théorème B.95), il existe, pour tout réel $\varepsilon>0$ donné, un entier naturel n, que l'on choisit strictement plus grand que 1, tel que, pour tous x et y appartenant à $[0,1]$, $|f(x)-f(y)|\leq\frac{\varepsilon}{4}$ si $|x-y|\leq\frac{1}{n}$. Posons alors $x_i=\frac{i}{n}$, $i=0,\dots,n$, ces points définissant une partition de l'intervalle $[0,1]$, et introduisons la fonction g, affine sur chacun des intervalles $[x_i,x_{i+1}]$, $i=0,\dots,n-1$, et telle que $g(x_i)=f(x_i)$, $i=0,\dots,n$. Pour tout x dans $[0,1]$, il existe un entier i compris entre 0 et $n-1$ tel que x appartient à l'intervalle $[x_i,x_{i+1}]$ et l'on a donc $|f(x)-f(x_i)|\leq\frac{\varepsilon}{4}$. D'autre part, une fonction affine sur un intervalle atteignant ses bornes aux extrémités de l'intervalle, la valeur $g(x)$ est comprise entre $g(x_i)=f(x_i)$ et $g(x_{i+1})=f(x_{i+1})$ et par conséquent $|f(x_i)-g(x)|\leq|f(x_i)-f(x_{i+1})|\leq\frac{\varepsilon}{4}$. L'application de l'inégalité triangulaire conduit alors à

$$|f(x)-g(x)|\leq|f(x)-f(x_i)|+|f(x_i)-g(x)|\leq\frac{\varepsilon}{4}+\frac{\varepsilon}{4}=\frac{\varepsilon}{2}.$$

Nous allons maintenant exhiber une fonction polynomiale approchant la fonction g de manière uniforme sur l'intervalle $[0,1]$. Pour cela, nous remarquons que g peut s'écrire explicitement

$$\forall x\in[-1,1],\ g(x)=g_1(x)+\sum_{i=1}^{n-1}(g_{i+1}(x)-g_i(x))h(x-x_i), \tag{6.1}$$

6. Karl Theodor Wilhelm Weierstraß (31 octobre 1815 - 19 février 1897) était un mathématicien allemand, souvent cité comme le « père de l'analyse moderne ». On lui doit l'introduction de plusieurs définitions et de formulations rigoureuses, comme les notions de limite et de continuité, et ses contributions au développement d'outils théoriques en analyse ouvrirent la voie à l'étude du calcul des variations telle que nous la connaissons aujourd'hui.

7. Sergei Natanovich Bernstein (Серге́й Ната́нович Бернште́йн en russe, 5 mars 1880 - 26 octobre 1968) était un mathématicien russe. Ses travaux portèrent sur l'approximation constructive des fonctions et la théorie des probabilités.

où g_i, $1 \leq i \leq n$, est la fonction affine définie par

$$\forall x \in \mathbb{R},\ g_i(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}(x - x_{i-1}),$$

et h est la *fonction échelon de Heaviside* [8], telle que

$$\forall x \in \mathbb{R},\ h(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}.$$

Compte tenu de l'expression (6.1), on voit que le problème se ramène à celui de l'approximation polynomiale de la fonction h. Pour tout entier naturel m, considérons la fonction polynomiale p_m définie par

$$p_m(x) = q_m\left(\frac{1-x}{2}\right),\ \text{avec } q_m(x) = (1-x^m)^{2^m}.$$

Pour m strictement positif, la fonction q_m décroît de manière monotone sur l'intervalle $[0,1]$, prenant la valeur 1 en $x = 0$ et 0 en $x = 1$. Soit un point x de l'intervalle $[0, \frac{1}{2}[$. On a alors

$$1 \geq q_m(x) = (1-x^m)^{2^m} \geq 1 - (2x)^m$$

d'après l'*inégalité de Bernoulli* [9], d'où

$$\forall x \in \left[0, \frac{1}{2}\right[,\ \lim_{m \to +\infty} q_m(x) = 1.$$

Soit x un point de $]\frac{1}{2}, 1[$. On a cette fois

$$\frac{1}{q_m(x)} = \left(\frac{1}{1-x^m}\right)^{2^m} = \left(1 + \frac{x^m}{1-x^m}\right)^{2^m} \geq 1 + \frac{(2x)^m}{1-x^n} > (2x)^m,$$

dont on déduit que

$$\forall x \in \left]\frac{1}{2}, 1\right],\ \lim_{m \to +\infty} q_m(x) = 0.$$

Ainsi, la suite de fonctions polynomiales $(p_m)_{m \in \mathbb{N}}$, bornée sur l'intervalle $[-1,1]$, converge vers la fonction h sur $[-1,-\delta] \cup [\delta, 1]$, pour tout $\delta > 0$, la décroissance de la fonction q_m assurant que cette convergence est uniforme. Faisons à présent le choix d'un réel $\delta > 0$ suffisamment petit pour que les intervalles $[x_i - \delta, x_i + \delta]$, $i = 1, \ldots, n-1$ soient disjoints, puis d'un entier m suffisamment grand pour que l'on ait, pour tout x tel que $0 < \delta \leq |x| \leq 1$, $|h(x) - p_m(x)| \leq \frac{\varepsilon}{4s}$ avec $s = \sum_{i=1}^{n-1} |g_{i+1}(x) - g_i(x)|$, et définissons la fonction polynomiale

$$p(x) = g_1(x) + \sum_{i=1}^{n-1} (g_{i+1}(x) - g_i(x))\, p_m(x - x_i).$$

Pour tout point x dans l'intervalle $[0,1]$, il vient

$$|g(x) - p(x)| \leq \sum_{\substack{i=1 \\ i \neq k}}^{n-1} |g_{i+1}(x) - g_i(x)|\, |h(x - x_i) - p_m(x - x_i)| + |g_{k+1}(x) - g_k(x)|\, |h(x - x_k) - p_m(x - x_k)|$$

$$\leq s\frac{\varepsilon}{4s} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2} \text{ si } x \in [x_k - \delta, x_k + \delta], 1 \leq k \leq n-1,$$

et

$$|g(x) - p(x)| \leq \sum_{i=1}^{n-1} |g_{i+1}(x) - g_i(x)|\, |h(x - x_i) - p_m(x - x_i)| \leq s\frac{\varepsilon}{4s} = \frac{\varepsilon}{4} \text{ sinon.}$$

On conclut alors en utilisant de nouveau l'inégalité triangulaire. □

REMARQUE sur l'approximation par les *polynômes de Bernstein*

$$\forall x \in [0,1],\ B_n f(x) = \sum_{j=0}^{n} f\left(\frac{j}{n}\right)\binom{n}{j} x^j (1-x)^{n-j},$$

8. Oliver Heaviside (18 mai 1850 - 3 février 1925) était un ingénieur, mathématicien et physicien britannique autodidacte. Il est notamment à l'origine des simplifications algébriques conduisant à la forme des équations de Maxwell connue aujourd'hui en électromagnétisme.

9. Jakob (ou Jacques) Bernoulli (27 décembre 1654 - 16 août 1705) était un mathématicien et physicien suisse. Il s'intéressa principalement à l'analyse fonctionnelle, aux calculs différentiel et intégral, dont il se servit pour résoudre de célèbres problèmes de mécanique. Il posa par ailleurs les bases du calcul des probabilités, qu'il appliqua à l'étude des jeux de hasard.

utilisée en pratique et sa lente convergence ($\|f - B_n(f)\|_\infty \leq \frac{1}{8n}\|f''\|_\infty$, $\forall f \in \mathscr{C}^2([a,b])$, optimal)

on peut faire mieux :

Théorème 6.2 (« inégalité de Jackson [10] ») *Une fonction de classe $\mathscr{C}^k$ peut être approchée par une suite de fonctions polynomiales de degré n croissant de manière à ce que l'erreur d'approximation uniforme soit au pire comme $\frac{C}{n^k}$ lorsque $n \to +\infty$, la constante ne dépendant que de k.*

DÉMONSTRATION. A ECRIRE □

NOTE : résultat initialement établi par Jackson dans sa thèse pour les polynômes trigonométriques et algébriques

Corollaire 6.3 *inégalité faisant intervenir le module de continuité (c'est de cette inégalité qu'on a besoin)*

$$\|f - p^*\|_\infty \leq \frac{C(k)}{n^k}\,\omega\left(f^{(k)}, \frac{1}{n}\right)$$

DÉMONSTRATION. A ECRIRE □

Polynôme de meilleure approximation uniforme

Compte tenu des précédents résultats et étant donné une fonction continue sur un intervalle et un entier naturel n, on peut naturellement chercher à déterminer la fonction polynomiale de degré inférieur ou égal à n approchant au mieux cette fonction en norme uniforme sur l'intervalle considéré. Ce problème donne lieu à la définition d'un *polynôme de meilleure interpolation uniforme*.

Définition 6.4 (polynôme de meilleure approximation uniforme) *Soit f une fonction d'une variable réelle à valeurs réelles, continue sur un intervalle $[a,b]$ borné et non vide de $\mathbb{R}$. Pour tout entier naturel n, on appelle* ***polynôme de meilleure approximation uniforme de degré*** *n* ***de*** *f* ***sur*** *$[a,b]$ la fonction polynôme p^* de degré n réalisant*

$$\|f - p^*\|_\infty = \min_{q \in \mathbb{P}_n} \|f - q\|_\infty.$$

FAIRE une remarque sur l'extension de cette définition à d'autres ensembles que des intervalles et des types plus généraux de « polynômes ».

Cette définition se trouve justifiée par le résultat ci-après.

Théorème 6.5 (existence du polynôme de meilleure interpolation uniforme) *Pour toute fonction f de $\mathscr{C}([a,b])$, avec $[a,b]$ un intervalle borné et non vide de $\mathbb{R}$, et tout entier positif n, il existe un polynôme de meilleure approximation uniforme de degré n de f sur $[a,b]$.*

DÉMONSTRATION. A ECRIRE □

Une caractérisation d'un polynôme de meilleure approximation uniforme, esquissée en 1853 [Tch54], puis abordé plus en détail en 1857 [Tch59], par Chebyshev (mais seulement démontrée par Borel [Bor05, Chapitre IV] après l'introduction de la notion de compacité) est donnée par le théorème suivant.

Théorème 6.6 (propriété d'équi-oscillation du polynôme de meilleure approximation uniforme [11]) *RE-PRENDRE Soit f une fonction d'une variable réelle à valeurs réelles, continue sur un intervalle $[a,b]$ borné et non vide de $\mathbb{R}$ et n un entier naturel. La fonction polynomiale $p \in \mathbb{P}_n$ est un polynôme de meilleure approximation uniforme de degré n de f sur $[a,b]$ si et seulement si l'erreur d'approximation absolue atteint son maximum en $n+2$ points distincts $x_0 < x_1 < \cdots < x_{n+1}$:*

$$f(x_i) - p(x_i) = \sigma(-1)^i \|f - p\|_\infty,\ i = 0, \ldots, n+1$$

où σ est le signe de $f(x_0) - p(x_0)$ ($\sigma = \pm 1$).

DÉMONSTRATION. A ECRIRE □

Ce résultat permet de montrer que le polynôme de meilleure approximation uniforme est unique.

10. Dunham Jackson (24 juillet 1888 - 6 novembre 1946) était un mathématicien américain. Ses travaux portèrent sur la théorie de l'approximation, et plus particulièrement les polynômes trigonométriques et orthogonaux.

11. Ce résultat fut d'abord .

Corollaire 6.7 (unicité du polynôme de meilleure approximation uniforme) *Le polynôme de meilleure approximation uniforme est unique.*

DÉMONSTRATION. A ECRIRE □

La propriété d'équi-oscillation est fondamentale pour le calcul pratique du polynôme de meilleure approximation uniforme par l'*algorithme de Remes* [12] [Rem34a, Rem34c, Rem34b], ce dernier utilisant de manière essentielle le résultat suivant, qui compare l'erreur oscillante d'un polynôme à l'erreur de meilleure approximation uniforme.

Théorème 6.8 (« théorème de de la Vallée Poussin [13] » [dlVal10]) *REPRENDRE $f \in \mathscr{C}([a,b])$, $n \geq 0$ et p polynôme de degré inférieur ou égal à n tel que la différence $f-p$ prend alternativement des valeurs positives et négatives en $n+2$ points x_j de $[a,b]$ ($a \leq x_0 < x_1 < \cdots < x_{n+1} \leq b$). Alors*

$$\min_{0\leq j\leq n+1} \left|f(x_j)-p(x_j)\right| \leq \|f-p^*\|_\infty.$$

DÉMONSTRATION. On raisonne par l'absurde. Si le résultat était faux, il existerait une fonction polynomiale q de degré n telle que

$$\|f-q\|_\infty < \min_{0\leq j\leq n+1} \left|f(x_j)-p(x_j)\right|.$$

Il en résulte que la fonction polynomiale $q-p=f-p-(f-q)$ est en alternance strictement positive et strictement négative aux points $x_0,\ldots,x_{n+1}$. En vertu du théorème des valeurs intermédiaires (voir le théorème B.89), cela entraine qu'elle possède au moins $n+1$ zéros et donc que $q=p$, d'où une contradiction. □

Algorithme de Remes

Le théorème 6.8 fournit un moyen effectif de mesurer la qualité d'approximation uniforme d'une fonction f par une polynomiale p. Il suffit en effet de d'évaluer $\|f-p\|_\infty$ d'une part et $\min_{0\leq j\leq n+1} \left|f(x_j)-p(x_j)\right|$ d'autre part. Si ces deux valeurs sont suffisamment proches l'une de l'autre, on peut alors considérer p comme proche de p^* puisque l'on a

$$\frac{\|f-p\|_\infty - \|f-p^*\|_\infty}{\|f-p^*\|_\infty} \leq \frac{\|f-p\|_\infty - \min_{0\leq j\leq n+1} \left|f(x_j)-p(x_j)\right|}{\min_{0\leq j\leq n+1} \left|f(x_j)-p(x_j)\right|}.$$

Ceci est mis à profit pour le calcul numérique du polynôme de meilleure interpolation uniforme : on obtient ce dernier de manière itérative en cherchant à chaque étape les $n+2$ points d'extrema locaux de la différence entre le polynôme courant et la fonction f et en déterminant les coefficients d'un nouveau polynôme de manière à ce que celui-ci minimise le maximum de la valeur absolue de sa différence avec la fonction f en ces mêmes points.

Remes proposa deux stratégies pour construire l'ensemble des points d'extrema à chaque itération. Décrivons celle qui est souvent appelé le *second* algorithme de Remes.

DESCRIPTION

Pour certaines fonctions, on peut montrer que la convergence de cet algorithme est quadratique (voir [Vei60]), mais nous nous contentons ici de prouver un résultat de convergence linéaire.

Théorème 6.9 *Hypothèses ? ? La suite de fonctions polynomiales $(p^{(k)})_{k\in\mathbb{N}}$ construite par le second algorithme de Remes converge uniformément vers la fonction polynomiale de meilleure approximation p^* et est telle qu'il existe une constante strictement positive C et un réel θ strictement compris entre 0 et 1, tous deux indépendants de l'entier k, satisfaisant*

$$\forall k \in \mathbb{N},\ \|p^{(k)}-p^*\|_\infty \leq C\theta^k.$$

DÉMONSTRATION. A ECRIRE □

REMARQUE La non-linéarité du problème de meilleure approximation uniforme rend l'algorithme de Remes coûteux et celui-ci est généralement peu utilisé [14]. Ceci est aussi dû au fait que, comme on le verra plus loin,

12. Evgenii Yakovlevich Remez (Евге́ний Я́ковлевич Ре́мез en russe, 17 février 1896 - 21 août 1975) était un mathématicien russe. Il est connu pour ses apports à la théorie constructive des fonctions, en particulier l'algorithme et l'inégalité portant aujourd'hui son nom.

13. Charles-Jean Étienne Gustave Nicolas de la Vallée Poussin (14 août 1866 - 2 mars 1962) était un mathématicien belge. Il est connu pour avoir démontré, simultanément mais indépendamment de Hadamard, le théorème des nombres premiers à l'aide de méthodes issues de l'analyse complexe.

14. Le traitement numérique du signal fait exception à la règle, l'algorithme ayant été adapté avec succès pour la conception de filtres à réponse impulsionnelle finie, sous la forme d'une variante connue sous le nom d'*algorithme de Parks et McClelland* [PM72].

l'interpolation de Lagrange aux points de Chebychev fournit souvent une approximation quasiment optimale bien plus aisément calculable.

En lien avec ce dernier point, indiquons que l'on peut estimer la « qualité » d'une approximation polynomiale de degré donné d'une fonction continue en comparant l'erreur d'approximation, mesurée en norme de la convergence uniforme, qui lui correspond avec celle commise par le polynôme de meilleure approximation uniforme de même degré. En notant P_n l'opérateur de projection de $\mathscr{C}([a,b])$ dans lui-même qui associe à toute fonction continue sur l'intervalle $[a,b]$ ladite approximation polynomiale p_n de degré n, *i.e.*, $\forall f \in \mathscr{C}([a,b])$, $P_n(f) = p_n$ et $P_n(p_n) = p_n$, on trouve

$$\begin{aligned}\|f - p_n\|_\infty = \|f - P_n(f)\|_\infty &\leq \|f - p_n^*\|_\infty + \|p_n^* - P_n(f)\|_\infty \\ &= \|f - p_n^*\|_\infty + \|P_n(p_n^* - f)\|_\infty \leq (1 + \Lambda_n)\,\|f - p_n^*\|, \quad (6.2)\end{aligned}$$

en faisant simplement appel à l'inégalité triangulaire et en introduisant la *constante de Lebesgue* [15] de l'opérateur P_n,

$$\Lambda_n = \|P_n\|_\infty = \sup_{f \in \mathscr{C}([a,b])} \frac{\|p_n\|_\infty}{\|f\|_\infty} = \sup_{\substack{f \in \mathscr{C}([a,b]) \\ \|f\|_\infty \leq 1}} \|p_n\|_\infty,$$

qui n'est autre que sa norme d'opérateur, évaluée dans la norme de la convergence uniforme.

6.1.3 Meilleure approximation au sens des moindres carrés

définition d'un problème analogue au précédent : approximation en moyenne quadratique

on peut voir que la notion de meilleure approximation dans cette norme est liée à celles d'espace préhilbertien/de produit scalaire/d'orthogonalité

construction du polynôme par décomposition dans la base canonique, matrice de Hilbert, mauvais conditionnement du système linéaire

choix d'une base dans laquelle la matrice du système linéaire est diagonale : introduction des polynômes orthogonaux, définition, théorie (propriétés dont relation de récurrence à trois termes)

Note : the polynomial of best approximation in the 2-norm for a function $f \in \mathscr{C}([a,b])$ is also a near-best approximation in the ∞-norm for f on $[a,b]$

6.2 Interpolation de Lagrange

Soit n un entier naturel. Dans l'ensemble de cette section, on suppose que la famille $\{(x_i, y_i)\}_{i=0,\ldots,n}$, est un ensemble de $n+1$ points du plan euclidien dont les abscisses sont *toutes deux à deux distinctes*.

6.2.1 Définition du problème d'interpolation

Le problème d'interpolation de Lagrange s'énonce en ces termes : *étant donné une famille de $n+1$ couples (x_i, y_i), $i = 0, \ldots, n$, distincts de nombres réels, trouver un polynôme Π_n de degré inférieur ou égal à n dont le graphe de la fonction polynomiale associée passe par les $n+1$ points du plan ainsi définis*. Plus concrètement, ceci signifie que le polynôme Π_n solution de ce problème, appelé *polynôme d'interpolation*, ou *interpolant*, *de Lagrange associé aux points* $\{(x_i, y_i)\}_{i=0,\ldots,n}$, satisfait les contraintes

$$\Pi_n(x_i) = y_i,\ i = 0, \ldots, n. \quad (6.3)$$

On dit encore qu'il *interpole* les quantités y_i aux *nœuds* x_i, $i = 0, \ldots, n$.

Commençons par montrer que ce problème de détermination est bien posé, c'est-à-dire (voir la sous-section 1.4.2) qu'il admet une unique solution.

Théorème 6.10 (existence et unicité du polynôme d'interpolation de Lagrange) *Soit n un entier naturel. Étant donné $n+1$ points distincts $x_0, \ldots, x_n$ et $n+1$ valeurs $y_0, \ldots, y_n$, il existe un unique polynôme Π_n de $\mathbb{P}_n$ satisfaisant* (6.3).

15. Henri-Léon Lebesgue (28 juin 1875 - 26 juillet 1941) était un mathématicien français. Il révolutionna le calcul intégral en introduisant une théorie des fonctions mesurables en 1901 et une théorie générale de l'intégration l'année suivante.

DÉMONSTRATION. Le polynôme Π_n recherché étant de degré n, on peut poser

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \sum_{j=0}^{n} a_j\, x^j, \tag{6.4}$$

et ramener le problème d'interpolation à la détermination des coefficients a_j, $j = 0, \ldots, n$. En utilisant les conditions $\Pi_n(x_i) = y_i$, $i = 0, \ldots, n$, on arrive à un système linéaire à $n+1$ équations et $n+1$ inconnues :

$$a_0 + a_1\, x_i + \cdots + a_n {x_i}^n = y_i,\ i = 0, \ldots, n. \tag{6.5}$$

Ce système possède une unique solution si et seulement si la matrice carrée qui lui est associée est inversible. Or, il se trouve que cette dernière est une matrice de Vandermonde dont le déterminant vaut

$$\begin{vmatrix} 1 & x_0 & \ldots & {x_0}^n \\ 1 & x_1 & \ldots & {x_1}^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & {x_n}^n \end{vmatrix} = \prod_{0 \le i < j \le n} (x_j - x_i) = \prod_{i=0}^{n-1} \left(\prod_{j=i+1}^{n} (x_j - x_i) \right).$$

Les nœuds d'interpolation étant tous distincts, ce déterminant est non nul. □

On notera qu'il est également possible de prouver l'unicité du polynôme d'interpolation en supposant qu'il existe un autre polynôme Ψ_m, de degré m inférieur ou égal à n, tel que $\Psi_m(x_i) = y_i$ pour $i = 0, \ldots, n$. La différence $\Pi_n - \Psi_m$ s'annulant en $n+1$ points distincts, il découle du théorème fondamental de l'algèbre qu'elle est nulle.

Pour obtenir les coefficients du polynôme Π_n dans la base canonique de l'anneau des polynômes, il suffit donc de résoudre le système linéaire (6.5). On peut cependant montrer que les matrices de Vandermonde sont généralement très mal conditionnées, quel que soit le choix de nœuds d'interpolation (voir les articles [Gau75, Bec00]) et la résolution numérique des systèmes associés par une méthode directe est alors sujette à des problèmes de stabilité, en plus de s'avérer coûteuse lorsque le nombre de nœuds est important [16]. Indiquons qu'il existe néanmoins des méthodes efficaces et numériquement stables dédiées à la résolution de systèmes de Vandermonde, comme celle de Björck et Pereyra [BP70] que l'on présente dans la sous-section 6.2.2.

A VOIR : amélioration possible du conditionnement par translation et mise à l'échelle des fonctions de base ($\phi_i(x) = \left(\frac{x-c}{d}\right)^{i-1}$, avec $c = \frac{x_0+x_n}{2}$ et $d = \frac{x_n-x_0}{2}$)

Sous la forme (6.4), le polynôme d'interpolation de Lagrange peut être évalué en tout point distinct d'un nœud d'interpolation par la méthode de Horner avec n additions et n multiplications.
STABILITE, il faut considérer le conditionnement de cette représentation du polynôme d'interpolation de Lagrange relativement à des perturbations de ses coefficients, lien avec le conditionnement de la matrice Vandermonde ?

Notons que la constante de Lebesgue du problème d'interpolation de Lagrange défini plus haut est simplement la norme de l'opérateur linéaire, dit d'interpolation, L_n de $\mathbb{R}^{n+1}$ dans $\mathbb{P}_n$, qui associe à tout jeu de $n+1$ valeurs liées aux nœuds d'interpolation le polynôme d'interpolation de Lagrange de degré n correspondant,

$$\Lambda_n = \|L_n\|_{\infty,\infty} = \max_{\boldsymbol{y} \in \mathbb{R}^{n+1}} \frac{\|\Pi_n\|_\infty}{\|\boldsymbol{y}\|_\infty}.$$

Une expression particulièrement simple pour la constante de Lebesgue sera donnée dans la prochaine sous section et l'on verra dans la sous-section 6.2.3 comment, vue comme une fonction de l'entier n, elle se comporte selon le choix de distribution des nœuds d'interpolation sur l'intervalle $[a, b]$. Cette question s'avère en effet fondamentale lorsque l'on cherche à approcher au mieux une fonction donnée par son polynôme d'interpolation de Lagrange, les valeurs y_i, $i = 0, \ldots, n$, étant dans ce cas les valeurs aux nœuds de la fonction en question, et sera abordée dans la sous-section 6.2.3.

A VOIR : la constante de Lebesque permet entre autres choses d'aborder la question de la sensibilité du problème d'interpolation relativement au choix des nœuds

16. On a vu dans le chapitre 2 que le coût de la résolution d'un système d'ordre n par la méthode d'élimination de Gauss était de l'ordre de $\frac{2}{3}n^3$ opérations arithmétiques.

6.2.2 Différentes représentations du polynôme d'interpolation de Lagrange

REPRENDRE Nous venons de voir comment obtenir le polynôme d'interpolation de Lagrange dans la base canonique de l'anneau des polynômes et les inconvénients de cette approche (...). Une autre possibilité consiste à utiliser une représentation différente de ce polynôme, de manière à ce que sa détermination soit rendue particulièrement aisée. C'est ce que l'on fait en adaptant la base choisie de façon à ce que la matrice du système linéaire associé au problème soit diagonale ou triangulaire.

Forme de Lagrange

Commençons par introduire les *polynômes de Lagrange* et leurs propriétés.

Définition 6.11 *Soit n un entier naturel non nul. On appelle* ***polynômes de Lagrange associés aux nœuds*** $\{x_i\}_{i=0,\dots,n}$ *les $n+1$ polynômes $l_i \in \mathbb{P}_n$, $i = 0, \dots, n$, définis par*

$$\forall i \in \{0, \dots, n\},\ \forall x \in \mathbb{R},\ l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{6.6}$$

Bien que communément employée pour ne pas alourdir les écritures, la notation l_i, $i = 0, \dots, n$, utilisée pour les polyômes de Lagrange ne fait pas explicitement apparaître leur degré, la valeur de l'entier n étant fixée et généralement claire compte tenu du contexte. Il faudra cependant garder cette remarque à l'esprit, puisque l'on peut être amené à faire tendre tendre cette valeur vers l'infini (voir la section 6.2.3). Ajoutons que, si l'on a exigé que l'entier n soit supérieur ou égal à 1 dans la définition, le cas trivial $n = 0$ peut être inclus dans tout ce qui va suivre en posant $l_0 \equiv 1$ si $n = 0$.

Proposition 6.12 *Soit n un entier naturel. Les polynômes de Lagrange $\{l_i\}_{i=0,\dots,n}$ sont de degré n, vérifient $l_i(x_k) = \delta_{ik}$, $i, k = 0, \dots, n$, où δ_{ik} désigne le symbole de Kronecker, et forment une base de $\mathbb{P}_n$.*

DÉMONSTRATION. Le résultat est évident si l'entier n est nul. S'il est non nul, les deux premières propriétés découlent directement de la définition (6.6) des polynômes de Lagrange. On déduit ensuite de la deuxième propriété que, si le polynôme $\sum_{i=0}^{n} \lambda_i l_i$, les coefficients $\lambda_1, \dots, \lambda_n$ étant réels, est identiquement nul, alors on a

$$\forall j \in \{1, \dots, n\},\ 0 = \sum_{i=0}^{n} \lambda_i l_i(x_j) = \lambda_j.$$

La famille $\{l_i\}_{i=0,\dots,n}$ est donc libre et forme une base de $\mathbb{P}_n$. □

À titre d'illustration, on a représenté sur la figure 6.1 les graphes sur l'intervalle $[-1, 1]$ des polynômes de Lagrange associés aux nœuds -1, $-\frac{1}{2}$, 0, $\frac{1}{2}$ et 1.

On déduit de la proposition 6.12 le résultat suivant.

Théorème 6.13 (« formule d'interpolation de Lagrange ») *Soit n un entier naturel. Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ valeurs $y_0, \dots, y_n$, le polynôme d'interpolation Π_n de $\mathbb{P}_n$ tel que $\Pi_n(x_i) = y_i$, $i = 0, \dots, n$, est donné par*

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \sum_{i=0}^{n} y_i l_i(x). \tag{6.7}$$

DÉMONSTRATION. Pour établir (6.7), on utilise que la famille de polynômes $\{l_i\}_{i=0,\dots,n}$ forment une base de $\mathbb{P}_n$. La décomposition de Π_n dans cette base s'écrit $\Pi_n = \sum_{i=0}^{n} \mu_i l_i$, et on a alors

$$\forall j \in \{0, \dots, n\},\ y_j = \Pi_n(x_j) = \sum_{i=0}^{n} \mu_i l_i(x_j) = \mu_j.$$

□

Il découle de cette formule que la constante de Lebesgue Λ_n précédemment introduite s'exprime très simplement en fonction des polynôme de Lagrange. Il vient en effet

$$\|\Pi_n\|_\infty = \max_{x \in [a,b]} \left| \sum_{i=0}^{n} y_i l_i(x) \right| \leq \|y\|_\infty \max_{x \in [a,b]} \sum_{i=0}^{n} |l_i(x)|,$$

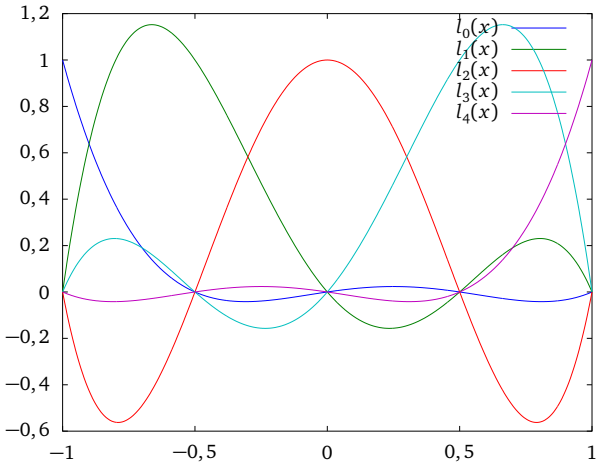


FIGURE 6.1 – Graphes des polynômes de Lagrange l_i, $i = 0, \ldots, 4$, associés à des nœuds équidistribués sur l'intervalle $[-1, 1]$.

et l'on peut alors montrer que (PREUVE ?)

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)| .$$

A VOIR : the Lebesgue constant can be viewed as the relative condition number of the operator mapping each coefficient vector $\boldsymbol{y}$ to the set of the values of the polynomial with coefficients $\boldsymbol{y}$ in the Lagrange form.

$$\|\Pi_n - \tilde{\Pi}_n\|_\infty = \max_{x\in[a,b]} \left| \sum_{i=0}^{n} (y_i - \tilde{y}_i) l_i(x) \right| \leq \max_{0\leq i\leq n} |y_i - \tilde{y}_i| \max_{x\in[a,b]} \left| \sum_{i=0}^{n} l_i(x) \right| = \Lambda_n \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_\infty$$

L'évaluation du polynôme d'interpolation Π_n sous sa forme de Lagrange (6.7) en un point autre que l'un des nœuds d'interpolation demande d'évaluer chacun des polynômes de Lagrange l_i, $i = 0, \ldots, n$, en ce point et nécessite au total n additions, $\frac{1}{2}(n+2)(n+1)$ soustractions, $(2n+1)(n+1)$ multiplications et $n+1$ divisions. Ce calcul peut par ailleurs s'avérer numériquement instable lorsque la valeur de n est élevée (REFERENCE ?). La « mise à jour [17] » de ce même polynôme, c'est-à-dire l'opération consistant à obtenir le polynôme Π_{n+1} associé à $n+2$ couples (x_i, y_i), $i = 0, \ldots, n+1$, à partir de la donnée du polynôme Π_n associé aux paires (x_i, y_i), $i = 0, \ldots, n$, et de celle du couple (x_{n+1}, y_{n+1}), est malaisée, car la base des polynômes de Lagrange servant à écrire Π_{n+1} est différente de celle utilisée pour Π_n. Pour ces raisons, la formule d'interpolation de Lagrange (6.7) est généralement considérée comme un outil théorique, peu utile en pratique, et le recours à la *forme de Newton* du polynôme d'interpolation est souvent recommandé.

Forme de Newton

La forme de Newton du polynôme d'interpolation offre une alternative à la formule (6.7) qui facilite à la fois l'évaluation et la mise à jour du polynôme d'interpolation après que certaines quantités, indépendantes du point auquel on évalue le polynôme, ont été calculées. Afin de l'expliciter, nous allons chercher à écrire le polynôme d'interpolation de Lagrange Π_n, avec n un entier strictement positif, associé aux nœuds d'interpolation $x_0, \ldots, x_n$, comme la somme du polynôme Π_{n-1}, tel que $\Pi_{n-1}(x_i) = y_i$ pour $i = 0, \ldots, n-1$, et d'un polynôme de degré n, qui dépendra des nœuds $x_0, \ldots, x_{n-1}$ et d'un seul autre coefficient que l'on devra déterminer. Posons ainsi

$$\Pi_n(x) = \Pi_{n-1}(x) + q_n(x), \tag{6.8}$$

17. Cette opération est primordiale dans le cadre de l'approximation polynomiale d'une fonction par le biais de l'interpolation (voir la sous-section 6.2.3). En effet, lorsqu'on ne sait *a priori* pas combien de points sont nécessaires pour approcher une fonction donnée par son polynôme d'interpolation de Lagrange avec une précision fixée, il est particulièrement utile de pouvoir introduire, un à un, de nouveaux nœuds d'interpolation jusqu'à satisfaction.

où q_n appartient à $\mathbb{P}_n$. Puisque $q_n(x_i) = \Pi_n(x_i) - \Pi_{n-1}(x_i) = 0$ pour $i = 0, \ldots, n-1$, on a nécessairement

$$q_n(x) = a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}).$$

Notons alors

$$\forall x \in \mathbb{R},\ \omega_n(x) = \prod_{j=0}^{n-1}(x - x_j) \tag{6.9}$$

le *polynôme de Newton de degré n associé aux nœuds* $\{x_i\}_{i=0,\ldots,n-1}$ et déterminons le coefficient a_n. Puisque $\Pi_n(x_n) = y_n$, on déduit de (6.8) que

$$a_n = \frac{y_n - \Pi_{n-1}(x_n)}{\omega_n(x_n)}.$$

Le coefficient a_n donné par la formule ci-dessus est appelée la *n^e différence divisée de Newton* et se note généralement [18]

$$a_n = [x_0, x_1, \ldots, x_n]\boldsymbol{y}.$$

On a par conséquent

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \Pi_{n-1}(x) + [x_0, x_1, \ldots, x_n]\boldsymbol{y}\ \omega_n(x). \tag{6.10}$$

En posant $[x_0]\boldsymbol{y} = y_0$ et $\omega_0 \equiv 1$, on obtient, à partir de (6.10) et en raisonnant par récurrence sur le degré n, que

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \sum_{i=0}^{n}[x_0, \ldots, x_i]\boldsymbol{y}\ \omega_i(x), \tag{6.11}$$

qui est, en vertu de l'unicité du polynôme d'interpolation, le même polynôme que celui défini par la formule (6.7). La forme (6.11) est appelée *formule des différences divisées de Newton du polynôme d'interpolation*. Ce n'est autre que l'écriture de Π_n dans la base [19] de $\mathbb{P}_n$ formée par la famille de polynômes de Newton $\{\omega_i\}_{i=0,\ldots,n}$. On remarquera que, écrite dans la base des polynômes de Newton, la matrice du système linéaire associé au problème d'interpolation de Lagrange est

$$\begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 1 & x_1 - x_0 & 0 & & & \vdots \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) & 0 & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n-1} - x_0 & (x_{n-1} - x_0)(x_{n-1} - x_1) & \ldots & \prod_{j=0}^{n-2}(x_{n-1} - x_j) & 0 \\ 1 & x_n - x_0 & (x_n - x_0)(x_n - x_1) & \ldots & \prod_{j=0}^{n-2}(x_n - x_j) & \prod_{j=0}^{n-1}(x_n - x_j) \end{pmatrix}. \tag{6.12}$$

Les différences divisées $[x_0]\boldsymbol{y}, [x_0, x_1]\boldsymbol{y}, \ldots, [x_0, \ldots, x_n]\boldsymbol{y}$ sont donc solution d'un système triangulaire inférieur et peuvent par conséquent être obtenues au moyen d'une méthode de descente (voir la section 2.3), après calcul des coefficients de la matrice ci-dessus.

La figure 6.2 présente les graphes sur l'intervalle $[-1, 1]$ la famille de polynômes de Newton associés aux nœuds -1, $-\frac{1}{2}$, 0, $\frac{1}{2}$ et 1.

Les différences divisées possèdent plusieurs propriétés algébriques. Tout d'abord, on peut vérifier, à titre d'exercice, que l'on a

$$\forall i \in \{0, \ldots, n\},\ \forall x \in \mathbb{R},\ l_i(x) = \frac{\omega_{n+1}(x)}{\omega'_{n+1}(x_i)(x - x_i)}, \tag{6.13}$$

où ω_{n+1} est le polynôme de Newton de degré $n+1$ associé aux nœuds $x_0, \ldots, x_n$, et la formule (6.7) se réécrit donc

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \omega_{n+1}(x)\sum_{i=0}^{n}\frac{y_i}{\omega'_{n+1}(x_i)(x - x_i)}. \tag{6.14}$$

18. On peut trouver dans la littérature de nombreuses notations pour les différences divisées. Celle choisie dans ces pages, à savoir $[\ldots]\boldsymbol{y}$ ou, plus loin, $[\ldots]f$ lorsque la différence divisée est appliquée aux valeurs prises aux nœuds par une fonction f continue, vise à mettre en avant le fait que $[\ldots]$ désigne un opérateur, dépendant de l'ensemble des nœuds d'interpolation apparaissant en place de $\ldots$.

19. On montre en effet par récurrence que $\{\omega_i\}_{i=0,\ldots,n}$ est une famille de $n+1$ polynômes *échelonnée en degré* (*i.e.*, que le polynôme ω_i, $i = 0, \ldots, n$, est de degré i).

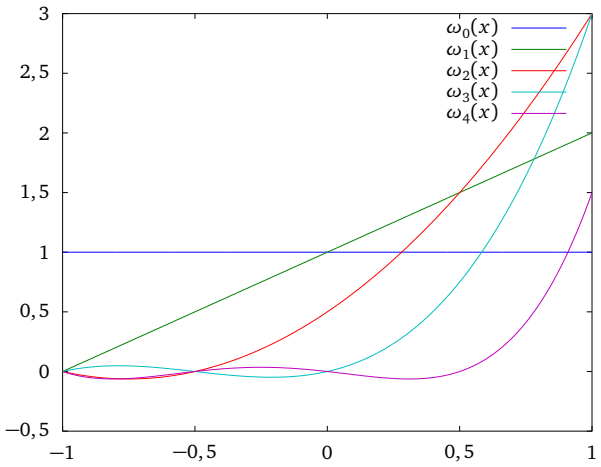


FIGURE 6.2 – Graphes des polynômes de Newton ω_i, $i = 0, \ldots, 4$, associés à des nœuds équidistribués sur l'intervalle $[-1, 1]$.

En utilisant alors l'écriture (6.11) pour identifier le coefficient $[x_0, \ldots, x_n]y$ avec celui qui lui correspond dans l'identité (6.14), on en déduit la forme explicite

$$[x_0, \ldots, x_n]y = \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)} = \sum_{i=0}^{n} \frac{y_i}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)} \tag{6.15}$$

pour la n^{e} différence divisée. Parmi toutes les conséquences de cette dernière expression, il en est une particulièrement importante pour la mise en œuvre de la forme de Newton du polynôme d'interpolation. En effet, par une manipulation algébrique, il vient

$$[x_0, \ldots, x_n]y = \frac{[x_1, \ldots, x_n]y - [x_0, \ldots, x_{n-1}]y}{x_n - x_0},$$

dont on déduit la formule de récurrence

$$\forall i \in \{0, \ldots, n\}, \forall k \in \{0, \ldots, i\},\ [x_{i-k}, \ldots, x_i]y = \frac{[x_{i-k+1}, \ldots, x_i]y - [x_{i-k}, \ldots, x_{i-1}]y}{x_i - x_{i-k}}, \tag{6.16}$$

de laquelle les différences divisées tirent leur nom et qui fournit un procédé pour leur calcul effectif. Ce dernier consiste en la construction du tableau suivant

$$\begin{array}{c|cccccc} x_0 & [x_0]y & & & & \\ x_1 & [x_1]y & [x_0,x_1]y & & & \\ x_2 & [x_2]y & [x_1,x_2]y & [x_0,x_1,x_2]y & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ x_n & [x_n]y & [x_{n-1},x_n]y & [x_{n-2},x_{n-1},x_n]y & \cdots & [x_0,\ldots,x_n]y \end{array} \tag{6.17}$$

au sein duquel les différences divisées sont disposées de manière à ce que leur évaluation se fasse de proche en proche en observant la règle suivante : la valeur d'une différence est obtenue en soustrayant à la différence placée immédiatement à sa gauche celle située au dessus de cette dernière, puis en divisant le résultat par la différence entre les deux points de l'ensemble $\{x_i\}_{i=0,\ldots,n}$ situés respectivement sur la ligne de la différence à calculer et sur la dernière ligne atteinte en remontant diagonalement dans le tableau à partir de cette même différence.

Les différences divisées apparaissant dans la forme de Newton (6.11) du polynôme d'interpolation de Lagrange sont les $n+1$ coefficients diagonaux du tableau (6.17). Leur obtention requiert par conséquent $(n+1)n$ soustractions

et $\frac{1}{2}(n+1)n$ divisions. Si ce coût est du même ordre que celui requis par la résolution d'un système linéaire triangulaire, il s'avère que la construction du tableau des différences divisées est bien moins susceptible de produire des débordements vers l'infini ou vers zéro en arithmétique à virgule flottante que le calcul des éléments de la matrice (6.12).

Exemple de calcul de la forme de Newton du polynôme d'interpolation. Calculons le polynôme d'interpolation de Lagrange prenant les valeurs 4, −1, 4 et 6 aux points respectifs −1, 1, 2 et 3, en tirant parti de (6.11) et de la méthode de calcul des différences divisées basée sur la formule (6.16). Nous avons

$$\begin{array}{c|cccc} -1 & 4 & & & \\ 1 & -1 & (-1-4)/(1-(-1))=-5/2 & & \\ 2 & 4 & (4-(-1))/(2-1)=5 & (5-(-5/2))/(2-(-1))=5/2 & \\ 3 & 6 & (6-4)(3-2)=2 & (2-5)/(3-1)=-3/2 & (-3/2-5/2)/(3-(-1))=-1 \end{array}$$

d'où

$$\Pi_3(x) = 4 - \frac{5}{2}(x+1) + \frac{5}{2}(x+1)(x-1) - (x+1)(x-1)(x-2).$$

Il découle enfin de la représentation (6.15) que les différences divisées sont des *fonctions symétriques de leurs arguments*. On a en effet

$$[x_0, \ldots, x_n]\boldsymbol{y} = [x_{\sigma(0)}, \ldots, x_{\sigma(n)}]\boldsymbol{y}, \tag{6.18}$$

pour toute permutation σ de l'ensemble $\{0, \ldots, n\}$.

Une fois les différences divisées calculées, l'évaluation du polynôme d'interpolation Π_n, sous sa forme de Newton, en un point autre que l'un des nœuds d'interpolation se fait au moyen d'une généralisation de la méthode de Horner introduite dans la sous-section 5.7.2, en remarquant que l'on a

$$\Pi_n(x) = (\ldots([x_0, \ldots, x_n]\boldsymbol{y}\,(x - x_{n-1}) + [x_0, \ldots, x_{n-1}]\boldsymbol{y})(x - x_{n-2}) + \cdots + [x_0, x_1]\boldsymbol{y})(x - x_0) + [x_0]\boldsymbol{y}.$$

Le calcul de la valeur de $\Pi_n(x)$ nécessite alors n additions, n soustractions et n multiplications. Pour mettre à jour le polynôme d'interpolation, il suffit simplement, disposant d'une valeur y_{n+1} associée à un nœud x_{n+1}, de calculer et d'ajouter la ligne supplémentaire $[x_{n+1}]\boldsymbol{y}\ [x_n, x_{n+1}]\boldsymbol{y}\ \cdots\ [x_0, \ldots, x_{n+1}]\boldsymbol{y}$ au tableau des différences divisées existant, ce qui nécessite $2(n+1)$ soustractions et $n+1$ divisions.

Stabilité numérique : deux étapes, évaluation des coefficients de la forme c_i puis évaluation du polynôme en un point connaissant ces coefficients. Le calcul des différences divisées dépend fortement de l'ordonnancement des nœuds d'interpolation

Si l'on cherche à calculer les différences divisées aussi précisément que possible ou à minimiser les résidus $|\Pi_n(x_i) - \mathrm{fl}(\Pi_n(x_i))|$, $i = 0, \ldots, n$,, l'ordonnancement $x_0 < x_1 < \cdots < x_n$ (ou $x_0 > x_1 > \cdots > x_n$) fournit des estimations d'erreur "optimales". En revanche, si l'on cherche à minimiser $|\Pi_n(x) - \mathrm{fl}(\Pi_n(x))|$ pour tout $x \neq x_i$, $i = 0, \ldots, n$, il vaut mieux arranger les nœuds d'interpolation comme des *points de Leja*[20] [Lej57]. Dans ce dernier cas :

$$|x_0| = \max_{0 \leq i \leq n} |x_i|\,, \prod_{k=0}^{j-1} \left|x_j - x_k\right| = \max_{i \leq j} \prod_{k=0}^{j-1} |x_i - x_k|\,,\ j = 1, \ldots, n-1.$$

étant donné $n+1$ points x_i, $i = 0, \ldots, n$, cet ordonnancement peut-être calculé en $O(n^2)$ opérations.

Résultats : - borne sur le conditionnement (qui croît moins vite qu'exponentiellement tant que m (le nombre de points de l'ensemble) est relativement plus petit que n) si les nœuds d'interpolation sont des points de Leja (voir [Rei90])

- ordonnancement optimal des points de Fejér (points de Chebyshev par exemple) en les ordonnant selon une *suite de van der Corput*[21] [vdCor35] (voir [FR89])

20. Franciszek Leja (27 janvier 1885 - 11 octobre 1979) était un mathématicien polonais. Il s'intéressa aux fonctions analytiques et plus particulièrement aux méthodes de points d'extremum et aux diamètres transfinis.

21. Johannes Gualtherus van der Corput (4 septembre 1890 - 16 september 1975) était un mathématicien hollandais. Il travailla dans le domaine de la théorie analytique des nombres.

Une application de la forme de Newton du polynôme d'interpolation de Lagrange : l'algorithme de Björck–Pereyra. La forme de Newton du polynôme d'interpolation de Lagrange est à la base d'une méthode de résolution d'un système linéaire de Vandermonde, à la fois efficace en termes du nombre d'opérations arithmétiques et du stockage requis, puisqu'il nécessite $\frac{3}{2}n(n-1)+1$ soustractions, $\frac{1}{2}n(n-1)+1$ multiplications et $\frac{1}{2}n(n-1)$ divisions pour la résolution d'un système d'ordre n et que la matrice n'a pas besoin d'être stockée, et numériquement stable, la solution numérique obtenue pouvant être très précise malgré un mauvais conditionnement de la matrice (on trouvera une analyse expliquant ce phénomène dans l'article [Hig87]).

L'algorithme en question, introduit par Björck et Pereyra dans [BP70], effectue la résolution du système linéaire[22] $V\boldsymbol{a}=\boldsymbol{y}$, où V est la matrice de Vandermonde d'ordre n liée à des nombres distincts $x_0,\dots,x_{n-1}$ et $\boldsymbol{y}$ est une matrice colonne à n coefficients, en procédant de la façon suivante. Tout d'abord, les n différences divisées $c_0=[x_0]\boldsymbol{y}$, $c_1=[x_0,x_1]\boldsymbol{y},\dots$, $c_{n-1}=[x_0,\dots,x_{n-1}]\boldsymbol{y}$ apparaissant dans la forme de Newton du polynôme d'interpolation de Lagrange Π_{n-1} associé aux couples (x_i,y_i), $i=0,\dots,n-1$, sont calculées, ce qui nécessite de construire le tableau de la forme (6.17) correspondant pour un coût de $n(n-1)$ soustractions et $\frac{1}{2}n(n-1)$ divisions. Une fois les coefficients c_i, $=0,\dots,n-1$, connus, on peut se servir, comme on l'a vu plus haut, la méthode de Horner pour obtenir, par identification, un à un les coefficients de Π_{n-1} dans la base canonique de l'anneau des polynômes, avec un coût de $\frac{1}{2}n(n-1)+1$ soustractions et autant de multiplications. On remarquera que les vecteurs de coefficients $\boldsymbol{c}$ et $\boldsymbol{a}$ peuvent être stockés tour à tour dans le vecteur de valeurs $\boldsymbol{y}$ (les composantes étant remplacées au fur et à mesure grâce à un ordonnancement adéquat des calculs), ce qui rend inutile tout tableau de stockage supplémentaire. Une mise en œuvre de la méthode est proposée ci-après avec l'algorithme 15.

Algorithme 15 : Algorithme de Björck–Pereyra pour la résolution du système linéaire $V\boldsymbol{a}=\boldsymbol{y}$.

Entrée(s) : les tableaux contenant le vecteur $\boldsymbol{x}$ des nœuds associés à la matrice de Vandermonde V et le vecteur $\boldsymbol{y}$.
Sortie(s) : le vecteur $\boldsymbol{a}$ solution du système $V\boldsymbol{a}=\boldsymbol{y}$, contenu dans le tableau dans lequel était stocké le vecteur $\boldsymbol{y}$ en entrée.

```
pour j = 1 à n−1 faire
    pour i = n à j+1 faire
        y(i) = (y(i)−y(i−1))/(x(i)−x(i−j))
    fin
fin
pour j = n−1 à 1 faire
    pour i = j à n−1 faire
        y(i) = y(i)−y(i+1)∗x(j)
    fin
fin
```

Considérons l'application de cet algorithme sur un exemple. Supposons que l'on souhaite résoudre le système linéaire

$$\begin{pmatrix}1&1&1&1\\1&2&3&4\\1&4&9&16\\1&8&27&64\end{pmatrix}\begin{pmatrix}a_0\\a_1\\a_2\\a_3\end{pmatrix}=\begin{pmatrix}10\\26\\58\\112\end{pmatrix}.$$

L'algorithme de Björck–Pereyra calcule tout d'abord le vecteur $\boldsymbol{c}=\begin{pmatrix}10&16&8&1\end{pmatrix}^\top$ des coefficients de la forme de Newton du polynôme d'interpolation de Lagrange Π_3 associé à ce problème,

$$\Pi_3(x)=10+16(x-1)+8(x-1)(x-2)+(x-1)(x-2)(x-3),$$

pour trouver ensuite $\boldsymbol{a}=\begin{pmatrix}4&3&2&1\end{pmatrix}^\top$.

Formes barycentriques

Si la forme de Newton du polynôme d'interpolation s'avère plus commode à manipuler que celle de Lagrange d'un point de vue pratique, il est néanmoins possible de réécrire cette dernière de façon à permettre une évaluation et une mise à jour de l'interpolant Π_n avec un nombre d'opérations proportionnel au degré n. Pour cela, il faut considérer la formule (6.7) et, en se servant la définition (6.6) des polynômes de Lagrange, y mettre en facteur la quantité $\omega_{n+1}(x)=\prod_{j=0}^{n}(x-x_j)$. On trouve alors ce qu'on appelle communément la *première forme de la formule*

22. Une variante de l'algorithme permet de résoudre le système linéaire dual $V^\top\boldsymbol{b}=\boldsymbol{z}$.

d'interpolation barycentrique,

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \omega_{n+1}(x) \sum_{i=0}^{n} \frac{w_i}{x - x_i}\, y_i, \tag{6.19}$$

dans laquelle les *poids barycentriques* w_i, $i = 0, \ldots, n$, sont définis par

$$w_i = \frac{1}{\displaystyle\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)}, \tag{6.20}$$

ou encore, par identification de (6.19) avec (6.14),

$$w_i = \frac{1}{\omega'_{n+1}(x_i)}.$$

Il est important d'observer que les poids barycentriques ne dépendent ni du point x, ni des valeurs y_i, $i = 0, \ldots, n$, représentant les données à interpoler. Leur calcul demande $\frac{1}{2}n(n-1)$ soustractions, $(n+1)(n-1)$ multiplications et $n+1$ divisions. Une fois cette tâche accomplie, l'évaluation du polynôme d'interpolation Π_n sous la forme barycentrique (6.19) en tout point ne coïncidant pas avec un nœud d'interpolation requiert n additions, $n+1$ soustractions, $2n+1$ multiplications et $n+1$ divisions. De même, la prise en compte d'un nœud d'interpolation x_{n+1} et d'une valeur y_{n+1} supplémentaires pour construire le polynôme Π_{n+1} à partir de Π_n se fait en divisant respectivement chacun des poids w_i, $i = 0, \ldots, n$, associés à Π_n par $x_i - x_{n+1}$ et en calculant w_{n+1} via la formule (6.20), pour un total de $n+1$ soustractions, n multiplications et $n+2$ divisions. Le coût de ces opérations est donc de l'ordre de grandeur de celui obtenu avec la forme de Newton du polynôme d'interpolation.

La formule (6.19) peut cependant être rendue encore plus « élégante ». Il suffit en effet de remarquer que, si les valeurs y_i, $i = 0, \ldots, n$, sont celles prises par la fonction constante égale à 1 aux nœuds d'interpolation, il vient

$$1 = \sum_{i=0}^{n} l_i(x) = \omega_{n+1}(x) \sum_{i=1}^{n} \frac{w_i}{x - x_i}.$$

En divisant alors l'égalité (6.19) par cette dernière identité, on arrive à la *seconde (ou vraie) forme de la formule d'interpolation barycentrique* suivante

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \frac{\displaystyle\sum_{i=0}^{n} \frac{w_i}{x - x_i}\, y_i}{\displaystyle\sum_{i=0}^{n} \frac{w_i}{x - x_i}}, \tag{6.21}$$

qui est généralement celle implémentée en pratique, car elle permet des réductions lors du calcul des poids. En effet, les quantités w_i, $i = 0, \ldots, n$, intervenant de manière identique (à un facteur multiplicatif près) au numérateur et au dénominateur du membre de droite de l'égalité (6.21), tout facteur commun à chacun des poids peut, à profit, être simplifié sans que la valeur du polynôme d'interpolation s'en trouve affectée. Nous allons illustrer ce principe sur quelques exemples de distributions de nœuds d'interpolation pour lesquelles on connaît une formule explicite pour les poids barycentriques. Dans le cas de nœuds équidistribués sur un intervalle $[a, b]$ borné de $\mathbb{R}$, on a

$$w_i = \frac{(-1)^{n-i} n^n}{n!(b-a)^n} \binom{n}{i},\ i = 0, \ldots, n,$$

où $\binom{n}{i} = \frac{n!}{i!(n-i)!}$ désigne le *coefficient binomial* donnant le nombre de sous-ensembles distincts à i éléments que l'on peut former à partir d'un ensemble contenant n éléments. Les poids « réduits » correspondants sont alors $\widetilde{w}_i = (-1)^i \binom{n}{i}$, $i = 0, \ldots, n$. Pour les familles de points de Chebyshev sur l'intervalle $[-1, 1]$ respectivement donnés par les racines des polynômes de Chebyshev de première espèce ($x_i = \cos\left(\frac{2i+1}{2(n+1)}\,\pi\right)$, $i = 0, \ldots, n$) et de deuxième espèce ($x_i = \cos\left(\frac{i}{n}\pi\right)$, $i = 0, \ldots, n$), on a comme poids réduits

$$\widetilde{w}_i = (-1)^i \sin\left(\frac{2i+1}{2(n+1)}\,\pi\right),\ i = 0, \ldots, n,$$

et [23] (voir [Sal72])

$$\widetilde{w}_0 = \frac{1}{2},\ \widetilde{w}_i = (-1)^i,\ 1 \leq i \leq n-1 \text{ et } \widetilde{w}_n = \frac{(-1)^n}{2}.$$

Sous la forme (6.21), l'évaluation du polynôme d'interpolation en un point autre que l'un des nœuds d'interpolation ne nécessite plus que $2n$ additions, $n+1$ soustractions, $n+1$ multiplications et $n+2$ divisions.

STABILITE NUMERIQUE : [Hig04] la formule (6.19) est stable au sens inverse, c'est-à-dire que, sous le hypothèses habituelles sur l'arithmétique à virgule flottante, la valeur calculée $\mathrm{fl}(\Pi_n(x_i))$ obtenue en utilisant cette formule est la valeur exacte du polynôme d'interpolation au point x pour un jeu de données légèrement perturbées, les perturbations relatives n'étant pas d'amplitude plus grande que $5nu$, où u est la précision machine. En revanche, la formule (6.21) n'est pas stable au sens inverse mais vérifie une estimation de stabilité au sens direct plus restrictive. mais différence ne sera visible que pour un "mauvais" choix de nœuds d'interpolation et/ou un jeu de données issu d'une fonction particulière.

Mentionnons enfin que, pour toute distribution de $n+1$ nœuds d'interpolation, on a l'inégalité (voir [BM97] pour une preuve)

$$\Lambda_n \geq \frac{1}{2n^2} \frac{\max\limits_{0\leq i\leq n} |w_i|}{\min\limits_{0\leq i\leq n} |w_i|},$$

ce qui permet de calculer très facilement, une fois les poids barycentriques connus, une borne inférieure pour la constante de Lebesgue Λ_n de l'opérateur d'interpolation de Lagrange associé à ces nœuds.

Algorithme de Neville

Si l'on ne cherche pas à construire le polynôme d'interpolation de Lagrange, mais simplement à connaître sa valeur en un point donné et distinct des nœuds d'interpolation, on peut envisager d'employer une méthode itérative basée sur des interpolations linéaires successives entre polynômes. Ce procédé particulier repose sur le résultat suivant.

Lemme 6.14 *Soit n un entier naturel, $x_{i_0}, \ldots, x_{i_n}$ $n+1$ nœuds distincts et $y_{i_0}, \ldots, y_{i_n}$ $n+1$ valeurs. On note $\Pi_{x_{i_0}, x_{i_1}, \ldots, x_{i_n}}$ le polynôme d'interpolation de Lagrange de degré n tel que*

$$\Pi_{x_{i_0}, x_{i_1}, \ldots, x_{i_n}}(x_{i_k}) = y_{i_k},\ k = 0, \ldots, n.$$

Étant donné x_i, x_j, $x_{i_0}, \ldots, x_{i_n}$, $n+3$ nœuds distincts et y_i, y_j, $y_{i_0}, \ldots, y_{i_n}$ $n+3$ valeurs, on a

$$\forall z \in \mathbb{R},\ \Pi_{x_{i_0}, \ldots, x_{i_n}, x_i, x_j}(z) = \frac{(z - x_j)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_i}(z) - (z - x_i)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_j}(z)}{x_i - x_j}. \tag{6.22}$$

DÉMONSTRATION. Soit $q(z)$ le membre de droite de l'égalité (6.22). Les polynômes $\Pi_{x_{i_0}, \ldots, x_{i_n}, x_i}$ et $\Pi_{x_{i_0}, \ldots, x_{i_n}, x_j}$ étant tous deux de degré $n+1$, le polynôme q est de degré inférieur ou égal à $n+2$. On vérifie ensuite que

$$q(x_{i_k}) = \frac{(x_{i_k} - x_j)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_i}(x_{i_k}) - (x_{i_k} - x_i)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_j}(x_{i_k})}{x_i - x_j} = y_{i_k},\ k = 0, \ldots, n,$$

et

$$q(x_i) = \frac{(x_i - x_j)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_i}(x_i)}{x_i - x_j} = y_i,\ q(x_j) = -\frac{(x_j - x_i)\,\Pi_{x_{i_0}, \ldots, x_{i_n}, x_j}(x_j)}{x_i - x_j} = y_j.$$

On en déduit que $q = \Pi_{x_{i_0}, \ldots, x_{i_n}, x_i, x_j}$ par unicité du polynôme d'interpolation. □

Dans la classe de méthodes faisant usage de l'identité (6.22), l'une des plus connues est l'*algorithme de Neville* [24] [Nev34], qui consiste à calculer de proche en proche les valeurs au point z considéré de polynômes d'interpolation

23. Pour ces derniers points, on a en effet $w_0 = \frac{2^{n-2}}{n}$, $w_i = (-1)^i\,\frac{2^{n-1}}{n}$, $1 \leq i \leq n-1$, et $w_n = (-1)^n\,\frac{2^{n-2}}{n}$ (voir [Rie16]).

24. Eric Harold Neville (1er janvier 1889 - 22 août 1961) était un mathématicien britannique. Ses travaux les plus notables concernent la géométrie différentielle et les fonction elliptiques de Jacobi. Il joua, à la demande de son collègue Godfrey Harold Hardy, un rôle prépondérant dans la venue en Angleterre en 1914 du mathématicien indien Srinivasa Ramanujan.

de degré croissant, associés à des sous-ensembles des points $\{(x_i, y_i)\}_{i=0,\dots,n}$. À la manière de ce que l'on a fait pour le calcul des différences divisées, cette construction peut s'organiser dans un tableau synthétique :

$$\begin{array}{c|ccccc} x_0 & \Pi_{x_0}(z)=y_0 & & & & \\ x_1 & \Pi_{x_1}(z)=y_1 & \Pi_{x_0,x_1}(z) & & & \\ x_2 & \Pi_{x_2}(z)=y_2 & \Pi_{x_1,x_2}(z) & \Pi_{x_0,x_1,x_2}(z) & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ x_n & \Pi_{x_n}(z)=y_n & \Pi_{x_{n-1},x_n}(z) & \Pi_{x_{n-2},x_{n-1},x_n}(z) & \cdots & \Pi_{x_0,\dots,x_n}(z) \end{array} . \tag{6.23}$$

Le point z étant fixé, les éléments de la deuxième colonne du tableau sont les valeurs prescrites y_i associées aux nœuds d'interpolation x_i, $i = 0, \dots, n$. À partir de la troisième colonne, tout élément est obtenu à partir de deux éléments situés immédiatement à sa gauche (respectivement sur la même ligne et sur la ligne précédente) par application de la relation (6.22). Par exemple, la valeur $\Pi_{x_0,x_1,x_2}(z)$ est donnée par

$$\Pi_{x_0,x_1,x_2}(z) = \frac{(z-x_2)\Pi_{x_0,x_1}(z) - (z-x_0)\Pi_{x_1,x_2}(z)}{x_0 - x_2}.$$

Pour obtenir la valeur de $\Pi_{x_0,\dots,x_n}(z)$, on doit ainsi effectuer $(n+1)^2$ soustractions, $(n+1)n$ multiplications et $\frac{1}{2}(n+1)n$ divisions.

Il existe plusieurs variantes de l'algorithme de Neville permettant d'améliorer son efficacité ou sa précision (voir par exemple [SB02]). Il n'est lui-même qu'une modification de l'*algorithme d'Aitken* [Ait32], qui utilise des polynômes d'interpolation intermédiaires différents et conduit au tableau suivant

$$\begin{array}{c|ccccc} x_0 & \Pi_{x_0}(z)=y_0 & & & & \\ x_1 & \Pi_{x_1}(z)=y_1 & \Pi_{x_0,x_1}(z) & & & \\ x_2 & \Pi_{x_2}(z)=y_2 & \Pi_{x_0,x_2}(z) & \Pi_{x_0,x_1,x_2}(z) & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ x_n & \Pi_{x_n}(z)=y_n & \Pi_{x_0,x_n}(z) & \Pi_{x_0,x_1,x_n}(z) & \cdots & \Pi_{x_0,\dots,x_n}(z) \end{array} . \tag{6.24}$$

Exemple d'application des algorithmes de Neville et d'Aitken. Construisons les tableaux de valeurs des algorithmes de Neville et d'Aitken pour l'évaluation du polynôme d'interpolation de Lagrange associé aux valeurs de la fonction $f(x) = e^x$ aux nœuds $x_i = 1 + \frac{i}{4}$, $i = 0, \dots, 5$ au point $z = 1,8$. On a (en arrondissant les résultats à la cinquième décimale) respectivement

$$\begin{array}{c|cccccc} 1 & 2,71828 & & & & & \\ 1,25 & 3,49034 & 5,18888 & & & & \\ 1,5 & 4,48169 & 5,67130 & 5,96076 & & & \\ 1,75 & 5,75460 & 6,00919 & 6,04297 & 6,04845 & & \\ 2 & 7,38906 & 6,08149 & 6,05257 & 6,05001 & 6,04970 & \\ 2,25 & 9,48774 & 5,71011 & 6,04435 & 6,04928 & 6,04961 & 6,04964 \end{array}$$

pour l'algorithme de Neville et

$$\begin{array}{c|cccccc} 1 & 2,71828 & & & & & \\ 1,25 & 3,49034 & 5,18888 & & & & \\ 1,5 & 4,48169 & 5,53973 & 5,96076 & & & \\ 1,75 & 5,75460 & 5,95702 & 6,03384 & 6,04845 & & \\ 2 & 7,38906 & 6,45490 & 6,11729 & 6,05468 & 6,04970 & \\ 2,25 & 9,48774 & 7,05073 & 6,21290 & 6,06161 & 6,04977 & 6,04964 \end{array}$$

pour celui d'Aitken. L'approximation de $e^{1,8}$ obtenue est $6,04964$.

6.2.3 Interpolation polynomiale d'une fonction

L'intérêt de remplacer une fonction quelconque par un polynôme l'approchant aussi précisément que voulu sur un intervalle donné est évident d'un point de vue numérique et informatique, puisqu'il est très aisé de stocker et de manipuler, c'est-à-dire additionner, multiplier, dériver ou intégrer, des polynômes dans un calculateur. Pour ce faire, il semble naturel de chercher à utiliser un polynôme d'interpolation de Lagrange associé aux valeurs prises par la fonction en des nœuds choisis.

Polynôme d'interpolation de Lagrange d'une fonction

Cette dernière idée conduit à l'introduction de la définition suivante.

Définition 6.15 *Soit n un entier naturel, $x_0, \ldots, x_n$ $n+1$ nœuds distincts et f une fonction réelle donnée, définie en chacun des nœuds. On appelle* ***polynôme d'interpolation*** *(ou* ***interpolant)*** ***de Lagrange de degré*** *n* ***de la fonction*** *f, et on note $\Pi_n f$, le polynôme d'interpolation de Lagrange de degré n associé aux points $(x_i, f(x_i))_{i=0,\ldots,n}$.*

Exemple de polynôme d'interpolation de Lagrange d'une fonction. Construisons le polynôme d'interpolation de Lagrange de degré deux de la fonction $f(x) = e^x$ sur l'intervalle $[-1, 1]$, avec comme nœuds d'interpolation les points $x_0 = -1$, $x_1 = 0$ et $x_2 = 1$. Nous avons tout d'abord

$$l_0(x) = \frac{1}{2}x(x-1),\ l_1(x) = 1-x^2 \text{ et } l_2(x) = \frac{1}{2}x(x+1),$$

la forme de Lagrange du polynôme d'interpolation est donc la suivante

$$\Pi_2 f(x) = \frac{1}{2}x(x-1)e^{-1} + (1-x^2) + \frac{1}{2}x(x+1)e.$$

Pour la forme de de Newton de ce même polynôme d'interpolation, il vient

$$\omega_0(x) = 1,\ \omega_1(x) = (x+1) \text{ et } \omega_2(x) = (x+1)x,$$

ainsi que, en étendant quelque peu la notation utilisée pour les différences divisées,

$$[x_0]f = e^{-1},\ [x_0, x_1]f = 1-e^{-1} \text{ et } [x_0, x_1, x_2]f = \frac{1}{2}(e-2+e^{-1}) = \cosh(1)-1,$$

d'où

$$\Pi_2 f(x) = e^{-1} + (1-e^{-1})(x+1) + (\cosh(1)-1)(x+1)x.$$

Enfin, les poids barycentriques associés aux nœuds d'interpolation valent

$$w_0 = \frac{1}{2},\ w_1 = -\frac{1}{2} \text{ et } w_2 = \frac{1}{2},$$

et la première forme barycentrique du polynôme d'interpolation est par conséquent

$$\Pi_2 f(x) = (x+1)x(x-1)\left(\frac{1}{2}\frac{e^{-1}}{x+1} - \frac{1}{2}\frac{1}{x} + \frac{1}{2}\frac{e}{x-1}\right).$$

On remarquera que $\Pi_2 f$ s'écrit encore

$$\Pi_2 f(x) = 1 + \sinh(1)\,x + (\cosh(1)-1)\,x^2$$

en utilisant les fonction polynomiales associées aux éléments de la base canonique de l'anneau des polynômes.

Erreur d'interpolation polynomiale

En termes de théorie de l'approximation, on peut voir le polynôme d'interpolation de Lagrange de la fonction f aux nœuds $x_0, \ldots, x_n$ comme le polynôme de degré n minimisant l'*erreur d'approximation* $\|f - p_n\|$, avec p_n un polynôme de $\mathbb{P}_n$, mesurée avec la semi-norme

$$\|f\| = \sum_{i=0}^{n} |f(x_i)|\,.$$

Bien que les valeurs de f et de son polynôme d'interpolation coïncident aux nœuds d'interpolation, elles diffèrent en général en tout autre point et il convient donc d'étudier l'*erreur d'interpolation* $f - \Pi_n f$ sur l'intervalle auquel appartiennent les nœuds d'interpolation. En supposant la fonction f suffisamment régulière, on peut établir le résultat suivant, qui donne une estimation de cette différence.

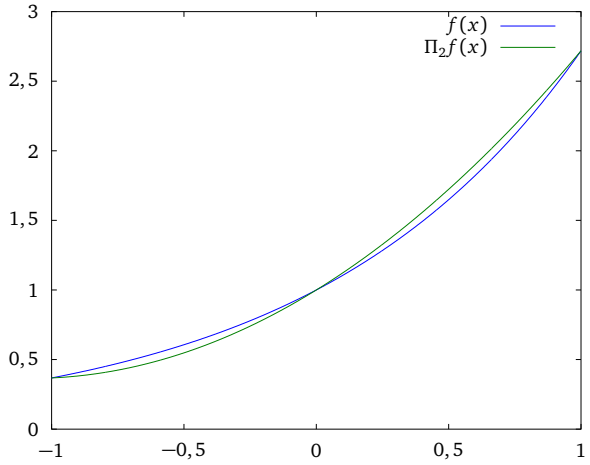


FIGURE 6.3 – Graphes de la fonction $f(x) = e^x$ et de son polynôme d'interpolation de Lagrange de degré deux à nœuds équirépartis sur l'intervalle $[-1, 1]$.

Théorème 6.16 *Soit n un entier naturel, $[a, b]$ un intervalle non vide borné de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^{n+1}$ sur $[a, b]$ et $n+1$ nœuds distincts $x_0, \ldots, x_n$, contenus dans l'intervalle $[a, b]$. Alors, pour tout réel x appartenant à $[a, b]$, il existe un point ξ contenu dans l'intérieur de I_x, le plus petit intervalle contenant $x_0, \ldots, x_n$ et x, tel que l'erreur d'interpolation au point x est donnée par*

$$f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x), \tag{6.25}$$

où ω_{n+1} est le polynôme de Newton de degré $n+1$ associé à la famille $\{x_i\}_{i=0,\ldots,n}$.

DÉMONSTRATION. Si le point x coïncide avec l'un des nœuds d'interpolation, les deux membres de (6.25) sont nuls et l'égalité est trivialement vérifiée. Supposons donc que x est un point distinct des nœuds $x_0, \ldots, x_n$ et introduisons la fonction auxiliaire

$$\forall t \in I_x,\ \varphi(t) = f(t) - \Pi_n f(t) - \frac{f(x) - \Pi_n f(x)}{\omega_{n+1}(x)}\,\omega_{n+1}(t).$$

Celle-ci est de classe $\mathscr{C}^{n+1}$ sur I_x (en vertu des hypothèses sur la fonction f) et s'annule en $n+2$ points (puisque $\varphi(x) = \varphi(x_0) = \cdots = \varphi(x_n) = 0$). D'après le théorème de Rolle (voir théorème B.112 en annexe), la fonction φ' possède au moins $n+1$ zéros distincts appartenant à l'intérieur de l'intervalle I_x et, en raisonnant par récurrence, on en déduit que, pour tout entier naturel j inférieur ou égal à $n+1$, la dérivée $\varphi^{(j)}$ admet au moins $n+2-j$ zéros distincts, tous contenus à l'intérieur de I_x. Par conséquent, il existe un réel ξ, appartenant à l'intérieur de I_x, tel que $\varphi^{(n+1)}(\xi) = 0$, ce qui s'écrit encore

$$f^{(n+1)}(\xi) - \frac{f(x) - \Pi_n f(x)}{\omega_{n+1}(x)}\,(n+1)! = 0$$

et dont on déduit l'identité (6.25). □

En utilisant la continuité de $f^{(n+1)}$ et en considérant les bornes supérieures des valeurs absolues des deux membres de (6.25) sur l'intervalle $[a, b]$, on obtient comme corollaire immédiat

$$\|f - \Pi_n f\|_\infty \leq \frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty \|\omega_{n+1}\|_\infty \leq \frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty (b-a)^{n+1}. \tag{6.26}$$

La première majoration montre en particulier que l'amplitude de l'erreur d'interpolation dépend à la fois de la quantité $\|f^{(n+1)}\|_\infty$, qui peut être importante si la fonction f est très oscillante, et de la quantité $\|\omega_{n+1}\|_\infty$, dont la valeur est liée à la distribution des nœuds d'interpolation dans l'intervalle $[a, b]$. Nous reviendrons sur ce second point.

La forme de Newton du polynôme d'interpolation conduit à une expression de l'erreur d'interpolation polynomiale autre que (6.25). Pour le voir, considérons $\Pi_n f$ le polynôme d'interpolation de f aux nœuds $x_0, \ldots, x_n$ et

x_{n+1} un nœud arbitraire distinct des précédents. Si l'on désigne par $\Pi_{n+1}f$ le polynôme interpolant f aux nœuds $x_0,\ldots,x_{n+1}$, on a, en utilisant (6.11),

$$\Pi_{n+1}f(x)=\Pi_n f(x)+[x_0,\ldots,x_n,x_{n+1}]f\,(x-x_0)\ldots(x-x_n),$$

d'où, en posant $x_{n+1}=x$ et en tenant compte de la définition (6.9) des polynômes de Newton,

$$f(x)-\Pi_n f(x)=[x_0,\ldots,x_n,x]f\;\omega_{n+1}(x). \tag{6.27}$$

Cette nouvelle représentation de l'erreur d'interpolation s'avère être une tautologie, puisque, si elle ne fait intervenir aucune dérivée, elle utilise des valeurs de f dont celle au point x... Néanmoins, en supposant vraies les hypothèses du théorème 6.16 et en comparant (6.27) avec (6.25), il vient

$$[x_0,\ldots,x_n,x]f=\frac{f^{(n+1)}(\zeta)}{(n+1)!}, \tag{6.28}$$

avec $\min(x_0,\ldots,x_n,x)<\zeta<\max(x_0,\ldots,x_n,x)$. Cette dernière identité, due à Cauchy [Cau40], est notamment utile pour évaluer l'ordre de grandeur des différences divisées. Elle se généralise au cas de points non supposés distincts (voir le corollaire 2 de la section 1 du chapitre 6 de [IK94]).

Théorème 6.17 *Soit n un entier naturel, $[a,b]$ un intervalle non vide borné de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^{n+1}$ sur $[a,b]$ et $\{x_i\}_{i=1,\ldots,n}$ un ensemble de $n+1$ points contenus dans $[a,b]$. On a*

$$[x_0,\ldots,x_n]f=\frac{f^{(n)}(\zeta)}{n!},$$

avec $\min(x_0,\ldots,x_n)<\zeta<\max(x_0,\ldots,x_n)$.

Quelques propriétés des différences divisées associées à une fonction

Nous allons dans cette section établir des propriétés de continuité et de dérivabilité pour la fonction de la variable réelle x définie par $[x_0,x_1,\ldots,x_n,x]f$, où les points $x_0,\ldots,x_n$ sont distincts et contenus dans un intervalle $[a,b]$ borné de $\mathbb{R}$ et x appartient à $[a,b]$.

Tout d'abord, on obtient, en explicitant (6.27),

$$f(x)-\sum_{i=1}^{n}f(x_i)\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}=\left(\prod_{j=0}^{n}(x-x_j)\right)[x_0,\ldots,x_n,x]f,$$

d'où

$$[x_0,\ldots,x_n,x]f=\sum_{i=0}^{n}\frac{[x,x_i]f}{\displaystyle\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)}. \tag{6.29}$$

Si f est une fonction continue sur $[a,b]$, alors, pour tout entier i dans $\{0,\ldots,n\}$, l'application $x\mapsto[x,x_i]f$ est continue sur $[a,b]\backslash\{x_i\}$ et prolongeable par continuité en x_i en posant $[x_i,x_i]f=f'(x_i)$ si f est dérivable en ce point, puisque, en vertu de la relation (6.16), $[x,x_i]f$ représente le taux d'accroissement de la fonction f en x_i. On déduit alors de (6.29) que $[x_0,\ldots,x_n,x]f$ est une fonction continue sur $[a,b]$ si f est dérivable sur cet intervalle.

En supposant de plus que f est de classe $\mathscr{C}^1$ sur $[a,b]$ et que la dérivée f'' est définie et continue sur un intervalle (arbitrairement petit) contenant le nœud d'interpolation x_i, le calcul de $\frac{\mathrm{d}}{\mathrm{d}x}([x,x_i]f)$ pour $x\neq x_i$, suivi d'un développement de Taylor au point x_i dans lequel on fait tendre x vers x_i, montre que $\frac{\mathrm{d}}{\mathrm{d}x}([x,x_i]f)$ est une fonction continue sur $[a,b]$ pour tout entier naturel i inférieur ou égal à n. On a alors montré le résultat suivant.

Lemme 6.18 *Soit n un entier naturel, f une fonction de classe $\mathscr{C}^2$ sur un intervalle $[a,b]$ non vide borné de $\mathbb{R}$ et $n+1$ points $x_0,\ldots,x_n$ distincts et contenus dans $[a,b]$. Alors, l'application de $[a,b]$ dans $\mathbb{R}$ définie par*

$$x\mapsto[x_0,\ldots,x_n,x]f$$

est de classe $\mathscr{C}^1$ sur $[a,b]$.

Une conséquence de ce résultat est que l'on peut définir, pour tout x appartenant à $[a,b]\backslash\{x_i\}_{i=0,\dots,n}$, la quantité $[x_0,\dots,x_n,x,x]f$ en posant

$$[x_0,\dots,x_n,x,x]f = \lim_{t\to 0}[x_0,\dots,x_n,x,x+t]f.$$

Par utilisation des propriétés (6.16) et (6.18) des différences divisées, le membre de gauche de cette égalité peut encore s'écrire

$$[x_0,\dots,x_n,x,x+t]f = \frac{[x_0,\dots,x_n,x+t]f-[x_0,\dots,x_n,x]f}{(x+t)-x} = \frac{[x_0,\dots,x_n,x+t]f-[x_0,\dots,x_n,x]f}{t}$$

et l'on a alors l'identité suivante

$$[x_0,\dots,x_n,x,x]f = \frac{\mathrm{d}}{\mathrm{d}x}([x_0,\dots,x_n,x]f). \tag{6.30}$$

L'application $x\mapsto[x_0,\dots,x_n,x,x]f$ est donc une fonction continue sur $[a,b]$, en vertu du lemme 6.18.

Convergence des polynômes d'interpolation et contre-exemple de Runge

Intéressons-nous à la question de la convergence uniforme du polynôme d'interpolation d'une fonction vers cette dernière lorsque le nombre de nœuds d'interpolation tend vers l'infini. Comme ce polynôme dépend de la distribution des nœuds d'interpolation, il est nécessaire de formuler ce problème de manière plus précise. Nous supposerons ici que l'on fait le choix, particulièrement simple, d'une répartition *uniforme* des nœuds (on dit que les nœuds sont *équirépartis* ou encore *équidistribués*) sur un intervalle $[a,b]$ non vide de $\mathbb{R}$, en posant

$$\forall n\in\mathbb{N}^*,\ x_i = a+i\,\frac{(b-a)}{n},\ i=0,\dots,n.$$

Au regard de l'estimation (6.26), il apparaît clairement que la convergence de la suite $(\Pi_n f)_{n\in\mathbb{N}^*}$ des polynômes d'interpolation d'une fonction f de classe $\mathscr{C}^\infty$ sur $[a,b]$ est liée au comportement de $\|f^{(n+1)}\|_\infty$ lorsque n augmente. En effet, si

$$\lim_{n\to+\infty}\frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty\|\omega_{n+1}\|_\infty = 0,$$

il vient immédiatement que

$$\lim_{n\to+\infty}\|f-\Pi_n f\|_\infty = 0,$$

c'est-à-dire qu'on a convergence vers f, uniformément sur $[a,b]$, de la suite des polynômes d'interpolation de f associés à des nœuds équirépartis sur l'intervalle $[a,b]$ quand n tend vers l'infini.

Malheureusement, il existe des fonctions, que l'on qualifiera de « pathologiques », pour lesquelles le produit $\|f^{(n+1)}\|_\infty\|\omega_{n+1}\|_\infty$ tend vers l'infini *plus rapidement* que $(n+1)!$ lorsque n tend vers l'infini. Un exemple célèbre, dû à Runge [25] [Run01], considère la convergence du polynôme d'interpolation de la fonction

$$f(x) = \frac{1}{1+x^2} \tag{6.31}$$

à nœuds équirépartis sur l'intervalle $[-5,5]$. Les valeurs du maximum de la valeur absolue de l'erreur d'interpolation pour cette fonction sont présentées dans la table 6.1 pour quelques valeurs paires du degré d'interpolation n. On observe une croissance exponentielle de l'erreur avec l'entier n.

La figure 6.4 présente les graphes de la fonction f et des polynômes d'interpolation $\Pi_2 f$, $\Pi_4 f$, $\Pi_6 f$, $\Pi_8 f$ et $\Pi_{10} f$ associés à des nœuds équirépartis sur l'intervalle $[-5,5]$ et met visuellement en évidence un phénomène de divergence de l'erreur d'interpolation au voisinage des extrémités de l'intervalle, parfois nommé *phénomène de Runge* (*Runge's phenomenon* en anglais).

25. Carl David Tolmé Runge (30 août 1856 - 3 janvier 1927) était un mathématicien et physicien allemand. Il est connu pour avoir développé une méthode de résolution numérique des équations différentielles ordinaires très utilisée. On lui doit également d'importants travaux expérimentaux sur les spectres des éléments chimiques pour des applications en spectrocopie astronomique.

degré n	$\max_{x\in[-5,5]} \lvert f(x)-\Pi_n f(x)\rvert$
2	0,64623
4	0,43836
6	0,61695
8	1,04518
10	1,91566
12	3,66339
14	7,19488
16	14,39385
18	29,19058
20	59,82231
22	123,62439
24	257,21305

TABLE 6.1 – Valeur (arrondie à la cinquième décimale) de l'erreur d'interpolation de Lagrange à nœuds équirépartis en norme de la convergence uniforme en fonction du degré d'interpolation pour la fonction de Runge $f(x)=\frac{1}{1+x^2}$ sur l'intervalle $[-5,5]$.

Ce comportement de la suite des polynômes d'interpolation n'a rien à voir avec un éventuel « défaut » de régularité de la fonction que l'on interpole [26], qui est de classe $\mathscr{C}^\infty$ sur $\mathbb{R}$. Il est en revanche lié au fait [27] que, vue comme une fonction d'une variable complexe, la fonction f, bien qu'analytique sur l'axe réel, possède deux pôles sur l'axe imaginaire en $z=\pm\mathrm{i}$.

D'autres choix de nœuds d'interpolation permettent néanmoins d'établir un résultat de convergence uniforme du polynôme d'interpolation dans ce cas. C'est, par exemple, le cas des points de Chebyshev (voir la table 6.2 et la figure 6.5), donnés sur tout intervalle $[a,b]$ non vide de $\mathbb{R}$ par

$$\forall n\in\mathbb{N}^*,\ x_i=\frac{a+b}{2}+\frac{b-a}{2}\cos\left(\frac{2i+1}{2(n+1)}\pi\right),\ i=0,\dots,n.$$

Il découle d'une définition des polynômes de Chebyshev de première espèce que l'on a pour ces points [28]

$$\|\omega_{n+1}\|_\infty=2\left(\frac{b-a}{4}\right)^{n+1}.$$

Cette valeur est minimale parmi toutes les distributions de nœuds possibles et bien inférieure à l'estimation

$$\|\omega_{n+1}\|_\infty\sim\left(\frac{b-a}{e}\right)^{n+1}$$

obtenue pour des nœuds équirépartis, e étant la constante de Napier ($e=2,718281828459\dots$). On voit ainsi que l'interpolation de Lagrange aux points de Chebychev est généralement plus précise que celle en des nœuds

26. La situation peut en revanche être pire lorsque la fonction n'est pas régulière. Dans [Ber18], Bernstein montre en effet que la suite des polynômes d'interpolation de Lagrange à nœuds équidistribués de la fonction valeur absolue sur tout intervalle $[-a,a]$, $a>0$, diverge en tout point de cet intervalle différent de $-a$, 0 ou a.

27. Le lecteur intéressé par la compréhension de ce phénomène pourra trouver plus de détails dans l'article [Epp87]. Indiquons simplement que l'on peut démontrer qu'on n'aura la convergence de la valeur $\Pi_n f(z)$ du polynôme d'interpolation de Lagrange à nœuds équirépartis sur un intervalle $[a,b]$ d'une fonction f vers la valeur $f(z)$ lorsque n tend vers l'infini que si la fonction considérée est analytique sur $[a,b]$ et que le nombre complexe z appartient à un contour $C(\rho)$, $\rho>0$, défini par

$$C(\rho)=\{z\in\mathbb{C}\,|\,\sigma(z)=\rho\},\ \text{avec } \sigma(z)=\exp\left(\frac{1}{b-a}\int_a^b\ln(|z-s|)\,\mathrm{d}s\right),$$

à l'intérieur duquel f est analytique. Or, dans le cas de l'exemple de Runge, pour tout nombre réel z tel que $|z|>3,6333843024$, tout contour $C(\rho)$ contenant z doit nécessairement inclure les pôles de la fonction.

28. On a en effet

$$\omega_{n+1}(x)=2^{-n}\left(\frac{b-a}{2}\right)^{n+1}T_{n+1}\left(\frac{2x-(a+b)}{b-a}\right),$$

où T_{n+1} est le polynôme de Chebyshev de première espèce de degré $n+1$, défini sur l'intervalle $[-1,1]$ par $T_{n+1}(x)=\cos((n+1)\arccos(x))$.

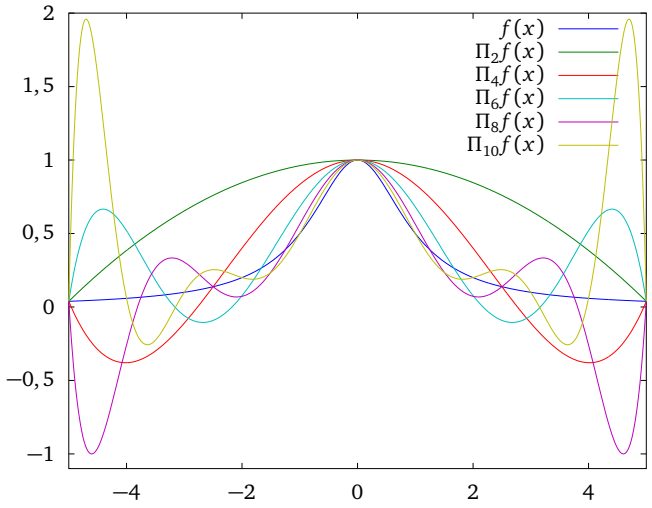


FIGURE 6.4 – Graphes de la fonction de Runge $f(x) = \dfrac{1}{1+x^2}$ et de cinq de ses polynômes d'interpolation à nœuds équirépartis sur l'intervalle $[-5,5]$.

équidistribués. Il est cependant possible d'aller plus loin en établissant un résultat de convergence, moyennant une hypothèse de régularité sur la fonction interpolée. Pour ce faire, il suffit de faire appel à l'inégalité (6.2) et d'obtenir des estimations sur la constante de Lebesgue Λ_n associée à l'opérateur d'interpolation de Lagrange.

Tout d'abord, des résultats d'Erdős [29] [Erd61] et de Brutman [Bru78] montrent que, pour toute distribution de $n+1$ nœuds sur l'intervalle $[-1,1]$, on a

$$\frac{2}{\pi}\ln(n+1) + \frac{2}{\pi}\left(\ln\left(\frac{4}{\pi}\right)+\gamma\right) < \Lambda_n,$$

où γ désigne la *constante d'Euler* [30]*–Mascheroni* [31] ($\gamma = 0,5772156649015\ldots$), ce qui implique la constante de Lebesgue de l'opérateur croît au moins logarithmiquement avec le nombre de points d'interpolation. En revanche, il vient (voir [Riv90])

$$\Lambda_n \leq \frac{2}{\pi}\ln(n+1) + 1 \tag{6.32}$$

pour les racines des polynômes de Chebyshev de première et de deuxième espèces. Ceci explique l'excellent comportement de l'interpolation de Lagrange aux points de Chebyshev dans de nombreux cas pratiques et nous conduit au résultat suivant.

Théorème 6.19 ***(convergence de l'interpolation de Lagrange aux points de Chebyshev)*** *A ECRIRE la suite des polynômes d'interpolation aux points de Chebyshev converge uniformément vers la fonction f à interpoler dès que cette dernière satisfait une condition de Dini* [32]*–Lipschitz, à savoir que $\lim_{\delta\to 0^+}\omega(f,\delta)\ln(\delta) = 0$, avec $\omega(f,\cdot)$ le module de continuité de f*

29. Paul Erdős (Erdős Pál en hongrois, 26 mars 1913 - 20 septembre 1996) était un mathématicien hongrois. Il posa et résolut de nombreux problèmes et conjectures en théorie des nombres, en combinatoire, en théorie des graphes, en probabilité, en théorie des ensembles et en analyse. Son œuvre prolifique (composée d'environ 1500 publications scientifiques) a donné naissance au concept humoristique de *nombre d'Erdős*, représentant une « distance de collaboration » entre Erdős et une personne donnée.

30. Leonhard Paul Euler (15 avril 1707 - 18 septembre 1783) était un mathématicien et physicien suisse. Il est considéré comme l'un des plus grands scientifiques de tous les temps. Il fit de nombreuses découvertes dans des domaines aussi variés que les mathématiques, la mécanique, l'optique et l'astronomie. En mathématiques, il apporta de très importantes contributions, notamment en analyse, en théorie des nombres, en géométrie et en théorie des graphes.

31. Lorenzo Mascheroni (13 mai 1750 - 14 juillet 1800) était un mathématicien italien. Il démontra que tout point du plan constructible à la règle et au compas l'est également au compas seul.

32. Ulisse Dini (14 novembre 1845 - 28 octobre 1918) était un mathématicien et homme politique italien. On lui doit des résultats importants sur les séries de Fourier, sur l'intégration des fonctions d'une variable complexe et en géométrie différentielle.

degré n	$\max_{x\in[-5,5]} \lvert f(x)-\Pi_n f(x)\rvert$
2	0,60060
4	0,20170
6	0,15602
8	0,17083
10	0,10915
12	0,06921
14	0,04660
16	0,03261
18	0,02249
20	0,01533
22	0,01036
24	0,00695

TABLE 6.2 – Valeur (arrondie à la cinquième décimale) de l'erreur d'interpolation de Lagrange utilisant les points de Chebyshev en norme de la convergence uniforme en fonction du degré d'interpolation pour la fonction de Runge $f(x)=\frac{1}{1+x^2}$ sur l'intervalle $[-5,5]$.

DÉMONSTRATION. REPRENDRE, utilise l'inégalité (6.2) combinée avec la majoration (6.32) pour donner

$$\|f-\Pi_n f\|_\infty \leq \left(\frac{2}{\pi}\ln(n+1)+2\right)\|f-p_n^*\|_\infty \tag{6.33}$$

puis le corollaire 6.3 (disant que $\|f-p_n^*\|_\infty \leq \omega(f,\frac{1}{n})$ pour arriver à

$$\|f-\Pi_n f\|_\infty \leq C\left(\frac{2}{\pi}\ln(n+1)+2\right)\omega\left(f,\frac{1}{n}\right)$$

avec $\lim_{n\to+\infty}\omega(f,\frac{1}{n})\ln(n)=0$ par hypothèse, ce qui donne le résultat □

Revenons sur la quasi-optimalité de l'interpolation de Lagrange aux points de Chebyshev à laquelle il a été fait allusion dans la section 6.1. On voit avec l'inégalité (6.33) que l'erreur d'interpolation aux points de Chebychev ne peut être pire que l'erreur de meilleure approximation uniforme que d'un facteur multiplicatif égal à $\frac{2}{\pi}\ln(n+1)+2$. Pour $n=10^5$, cette constante vaut environ $9,32936$ et la non-optimalité de l'approximation se traduit par une perte de précision correspondant à un chiffre (deux si l'on va jusqu'à $n=10^{66}$, nombre amplement suffisant en pratique, le facteur valant alors environ $98,7475$) par rapport à la valeur de l'erreur de meilleure approximation. Cette observation fait que l'erreur d'approximation du polynôme d'interpolation de Lagrange aux points de Chebyshev est qualifiée de *presque optimale* (*near best* en anglais) dans la littérature.

À l'opposé de ce résultat positif, on a l'estimation asymptotique suivante pour des nœuds équidistribués sur l'intervalle $[-1,1]$ (on pourra consulter [TW91] pour un historique de ce résultat)

$$\Lambda_n \underset{n\to+\infty}{\sim} \frac{2^{n+1}}{en\ln(n)}.$$

Dans ce cas, la croissance exponentielle de la constante de Lebesgue conduit à proscrire ce choix de distribution des nœuds dès que leur nombre dépasse quelques dizaines d'unités : l'amplification des erreurs d'arrondi est en effet telle que, même en considérant une fonction pour laquelle la convergence des polynômes d'interpolation a lieu en théorie, une divergence est systématiquement observée, au moins pour des points proches des extrémités de l'intervalle, en arithmétique en précision finie (voir [PTK11]).

6.2.4 Généralisations *

Dans cette sous-section, nous considérons deux extensions, de généralité croissante, du problème d'interpolation de Lagrange.

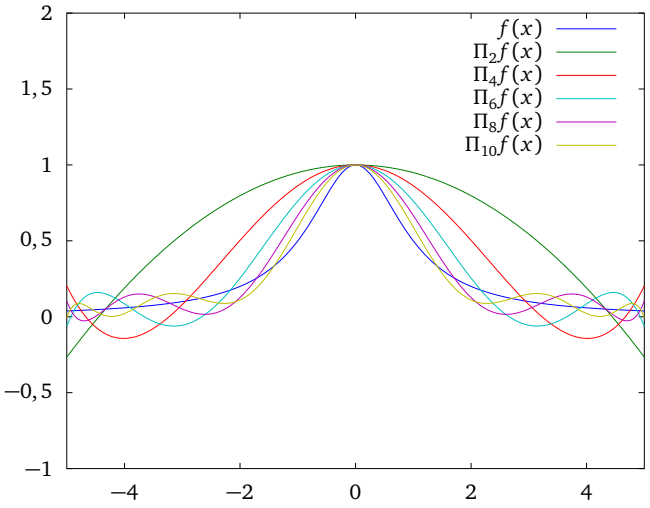


FIGURE 6.5 – Graphes de la fonction de Runge $f(x) = \dfrac{1}{1+x^2}$ et de cinq de ses polynômes d'interpolation aux points de Chebyshev sur l'intervalle $[-5,5]$.

Interpolation de Hermite

Soit k un entier naturel, $\{x_i\}_{i=0,\dots,k}$ un ensemble de $k+1$ nombres réels deux à deux distincts et des valeurs $y_i^{(j)}$, $j=0,\dots,n_i-1$, $i=0,\dots,k$, les entiers n_i étant strictement positifs. Le problème d'*interpolation de Hermite* [33] [Her78] (dans l'article : formule d'interpolation avec contour dans le plan complexe...) associé à ces données est le suivant : *trouver le polynôme* Π_n *de degré inférieur à n, avec n l'entier tel que*

$$n+1 = \sum_{i=0}^{k} n_i, \tag{6.34}$$

vérifiant les conditions

$$\Pi_n{}^{(j)}(x_i) = y_i^{(j)},\ j=0,\dots,n_i-1,\ i=0,\dots,k. \tag{6.35}$$

En d'autres mots, ce nouveau problème d'interpolation vise à imposer non seulement la valeur du polynôme, mais aussi celles de ses n_i-1 premières dérivées successives, en chacun des nœuds d'interpolation. On parle parfois de polynôme *osculateur* (adjectif issu du verbe latin *osculo* signifiant *embrasser*), le graphe du polynôme épousant « le mieux possible » celui de la fonction interpolée aux nœuds d'interpolation.

On peut immédiatement observer que le problème d'interpolation de Lagrange, défini par (6.3), correspond au cas particulier $n_i = 1$, $i=0,\dots,k$, avec $k=n$.

A VOIR : On peut généraliser la notion de matrice de Vandermonde de manière à permettre aux valeurs des points x_i, $i=0,\dots,n-1$, dont dépend la matrice d'être égales entre elles (ce cas de figure se présente par exemple avec le problème d'interpolation de Hermite (voir la sous-section 6.2.4)) et ainsi obtenir une *matrice de Vandermonde confluente* (*confluent Vandermonde matrix* en anglais). Si $x_i = x_{i+1} = \cdots = x_{i+k}$ et que $x_i \neq x_{i-1}$, la $(i+k)^{\text{e}}$ ligne d'une telle matrice est donnée par

$$V_{i+k\,j} = \begin{cases} 0 & \text{si } j \leq k \\ \frac{(j-1)!}{(j-k-1)!} x_i{}^{j-k-1} & \text{si } j > k \end{cases}.$$

On a le résultat suivant.

Théorème 6.20 ***(existence et unicité du polynôme d'interpolation de Hermite)*** *REPRENDRE Le polynôme d'interpolation de Hermite* Π_n *solution du problème défini plus haut existe et est unique.*

33. Charles Hermite (24 décembre 1822 - 14 janvier 1901) était un mathématicien français. Il s'intéressa principalement à la théorie des nombres, aux formes quadratiques, à la théorie des invariants, aux polynômes orthogonaux et aux fonctions elliptiques.

DÉMONSTRATION. REPRENDRE En faisant le choix d'une base pour la représentation du polynôme d'interpolation Π_n, on peut exprimer les conditions (6.35) sous la forme d'un système linéaire de $n+1$ équations pour les composantes de Π_n. Celui-ci possède une unique solution pour tout second membre si et seulement si le système homogène associé a pour seule solution la solution nulle. Supposons que ce ne soit pas le cas. Le polynôme d'interpolation correspondant à une solution non triviale du système homogène est alors tel que

$$\Pi_n{}^{(j)}(x_i) = 0,\ j = 0, \ldots, n_i - 1,\ i = 0, \ldots, k,$$

et chaque nœud d'interpolation x_i est donc un zéro de multiplicité n_i, $i = 0, \ldots, k$, de Π_n. En comptant ces zéros avec leur multiplicité, on arrive à $n+1$ racines pour un polynôme de degré inférieur ou égalà n, ce qui achève la preuve. □

cas remarquable de l'interpolation de Hermite d'une fonction f régulière : $k = 0$ et $n_0 = n+1$. On obtient

$$\Pi_n f(x) = \sum_{j=0}^{n} \frac{f^{(j)}((x_0)}{j!} (x - x_0)^j,$$

qui n'est autre que le polynôme de Taylor de la fonction associé au point x_0.

expression du polynôme d'interpolation de Hermite dans la base canonique de l'anneau de polynôme conduit à un système linéaire ayant pour matrice une *matrice de Vandermonde confluente*

forme de Lagrange : base de polynômes de Lagrange généralisés

propriétés de continuité et de différentiabilité pour les différences divisées

tableau modifié pour le calcul des différences divisées

EXEMPLE : polynôme d'interpolation de Hermite cubique

erreur d'interpolation

Interpolation de Birkhoff *

Dans le précédent problème d'interpolation, le fait d'imposer la valeur d'une dérivée d'ordre donné du polynôme d'interpolation en un nœud implique que les valeurs en ce point des dérivées d'ordre inférieur et du polynôme lui-même sont également imposées. Le problème d'*interpolation de Birkhoff*[34] [Bir06] se passe de cette obligation en autorisant de fixer les valeurs de dérivés non successives en un même nœud, expliquant pourquoi ce procédé d'interpolation est parfois qualifé de *lacunaire*. Pour le définir précisement, introduisons une *matrice d'incidence (non orientée) E*, c'est-à-dire dont les coefficients valent soit 0 soit 1, de taille $k \times m$ et considérons l'entier naturel n tel que

$$n + 1 = \sum_{i=1}^{k} \sum_{j=1}^{m} e_{ij}.$$

On cherche alors un polynôme Π_n de degré inférieur ou égal à n vérifiant

$$\Pi_n{}^{(j)}(x_i) = y_i^{(j)} \text{ si } e_{i+1\,j+1} = 1,\ i = 0, \ldots, k,\ j = 0, \ldots, m-1 \tag{6.36}$$

où les points (x_i, y_i) et les entiers positifs n_i sont donnés. Cette approche diffère de celle de l'interpolation de Hermite dans le sens où il est possible de fixer une valeur d'une dérivée de p en un point sans pour autant fixer les valeurs des dérivées d'ordre inférieur ou du polynôme lui-même en ce même point.

Conséquence : ce problème peut ne pas avoir de solution.

exemple : valeur du polynôme en -1 et 1 et de sa dérivée en 0, polynôme de degré deux

Dans l'article [Sch66], Schoenberg[35] proposa de déterminer l'ensemble des matrices d'incidence, dites *régulières*, pour lesquelles le problème d'interpolation défini par (6.36) admet toujours (c'est-à-dire pour tout choix de nœuds et de valeurs imposées) une solution, mais le cas de nœuds fixes pose déjà d'intéressantes questions théoriques.

6.3 Interpolation par morceaux

Jusqu'à présent, nous n'avons envisagé le problème de l'approximation d'une fonction f sur un intervalle $[a, b]$ par l'interpolation de Lagrange qu'en un sens *global*, c'est-à-dire en cherchant à n'utiliser qu'une seule expression

34. George David Birkhoff (21 mars 1884 - 12 novembre 1944) est un mathématicien américain. Plusieurs de ses travaux eurent une portée considérable, en particulier ceux concernant les systèmes dynamiques et la théorie ergodique.

35. Isaac Jacob Schoenberg (21 avril 1903 - 21 février 1990) était un mathématicien roumain, connu pour ses travaux sur les splines.

analytique de l'interpolant (un seul polynôme) sur $[a,b]$. Pour obtenir une approximation que l'on espère plus précise, on n'a alors d'autre choix que d'augmenter le degré du polynôme d'interpolation. L'exemple de Runge évoqué dans la section précédente montre que la convergence uniforme de la suite $(\Pi_n f)_{n\in\mathbb{N}}$ vers f n'est cependant pas garantie pour toute distribution arbitraire des nœuds d'interpolation.

Une alternative à cette première approche est de construire une partition de l'intervalle $[a,b]$ en sous-intervalles sur chacun desquels une interpolation polynomiale de bas degré est employée. On parle alors d'*interpolation polynomiale par morceaux*. L'idée naturelle suivie est que toute fonction peut être approchée de manière arbitrairement précise par des polynômes de bas degré (un ou même zéro par exemple), de manière à limiter les phénomènes d'oscillations observés avec l'interpolation de haut degré, sur des intervalles *suffisamment petits*.

Dans toute cette section, on désigne par $[a,b]$ un intervalle non vide borné de $\mathbb{R}$ et par f une application de $[a,b]$ dans $\mathbb{R}$. On considère également un entier naturel non nul m et $m+1$ points $x_0,\dots,x_m$, tels que $a = x_0 < x_1 < \cdots < x_m = b$, réalisant une partition $\mathscr{T}_h$ de $[a,b]$ en m sous-intervalles $[x_{j-1},x_j]$ de longueur $h_j = x_j - x_{j-1}$, $1 \le j \le m$, dont on caractérise la « *finesse* » par

$$h = \max_{1\le j\le m} h_j.$$

Après avoir brièvement introduit l'*interpolation de Lagrange par morceaux*, nous allons nous concentrer sur une classe de méthodes d'interpolation par morceaux possédant des propriétés de régularité globale intéressantes : les *splines d'interpolation*.

6.3.1 Interpolation de Lagrange par morceaux

L'interpolation de Lagrange par morceaux d'une fonction f donnée relativement à une partition $\mathscr{T}_h$ d'un intervalle $[a,b]$ consiste en la construction d'un *polynôme d'interpolation par morceaux* coïncidant sur chacun des sous-intervalles $[x_{j-1},x_j]$, $j=1,\dots,m$, de $\mathscr{T}_h$ avec le polynôme d'interpolation de Lagrange de f en des nœuds fixés de ce sous-intervalle. La fonction interpolante ainsi obtenue est, en général, simplement continue sur $[a,b]$.

On notera qu'on peut *a priori* choisir un polynôme d'interpolation de degré différent sur chaque sous-intervalle (il en va de même la répartition des nœuds lui correspondant). Cependant, en pratique, on utilise très souvent la même interpolation, de bas degré, sur tous les sous-intervalles pour des raisons de commodité. Dans ce cas, on note $\Pi_h^n f$ le polynôme d'interpolation par morceaux obtenu en considérant sur chaque sous-intervalle $[x_{j-1},x_j]$, $j=1,\dots,m$, d'une partition $\mathscr{T}_h$ de $[a,b]$ une interpolation de Lagrange de f aux nœuds $x_j^{(0)},\dots,x_j^{(n)}$, par exemple équirépartis, avec n un entier naturel non nul et « petit ». Puisque la restriction de $\Pi_h^n f$ à chaque sous-intervalle $[x_{j-1},x_j]$, $j=1,\dots,m$, est le polynôme d'interpolation de Lagrange de f de degré n associé aux nœuds $x_j^{(0)},\dots,x_j^{(n)}$, on déduit aisément, si f est de classe $\mathscr{C}^{n+1}$ sur $[a,b]$, du théorème 6.16 une majoration de l'erreur d'interpolation $\left|f(x)-\Pi_h^n f(x)\right|$ sur chaque sous-intervalle de $\mathscr{T}_h$, conduisant à une estimation d'erreur globale de la forme

$$\|f-\Pi_h^n f\|_\infty \le C\, h^{n+1}\|f^{(n+1)}\|_\infty,$$

avec C une constante strictement positive dépendant de l'entier n. On observe alors qu'on peut rendre arbitrairement petite l'erreur d'interpolation dès lors que la partition $\mathscr{T}_h$ de $[a,b]$ est suffisamment fine (*i.e.*, le réel h est suffisamment petit).

6.3.2 Interpolation par des fonctions splines

L'interpolation polynomiale de Lagrange par morceaux introduite dans la section précédente fait partie de la classe plus large d'*interpolation par des fonctions splines*. Étant donnée une partition $\mathscr{T}_h$ de l'intervalle $[a,b]$, une *fonction spline* est une fonction qui possède une certaine régularité globale prescrite et dont la restriction à chacun des sous-intervalles de $\mathscr{T}_h$ est un polynôme de degré également prescrit. On suppose habituellement qu'une fonction spline est au minimum continue [36] et qu'elle est continûment dérivable jusqu'à un certain ordre. Une famille de splines importante pour les applications est celle des fonctions splines de degré n, avec $n \ge 1$, dont les dérivées jusqu'à l'ordre $n-1$ sont continues, mais des splines possédant moins de régularité sont également couramment employées [37].

36. Cette condition n'est évidemment pas nécessaire.

37. Par exemple, le langage informatique PostScript, qui sert pour la description des éléments d'une page (textes, images, polices, couleurs, etc.), utilise des splines cubiques qui sont seulement continûment dérivables.

Après quelques remarques générales sur les fonctions splines, nous considérons plus en détail deux exemples d'interpolation d'une fonction par des fonctions splines respectivement linéaires et cubiques, en mettant l'accent sur leur construction effective et leur propriétés. Nous renvoyons le lecteur intéressé par la théorie des splines d'interpolation et leur mise en œuvre pratique à l'ouvrage de de Boor [dBoo01] pour une présentation exhaustive.

Généralités

Commençons par donner une définition générale des fonctions splines.

Définition 6.21 *Soit trois entiers naturels k, m et n, avec $k \leq n$, un intervalle $[a,b]$ non vide de $\mathbb{R}$ et $\mathcal{T}_h$ une partition de $[a,b]$ en m sous-intervalles $[x_j, x_{j+1}]$, $j = 0,\ldots,m-1$. On définit la* ***classe*** $\mathcal{S}_n^k(\mathcal{T}_h)$ ***des fonctions splines de degré*** n ***et de classe*** $\mathcal{C}^k$ ***sur l'intervalle*** $[a,b]$ ***relativement à la partition*** $\mathcal{T}_h$ *par*

$$\mathcal{S}_n^k(\mathcal{T}_h) = \left\{ s \in \mathcal{C}^k([a,b]) \mid s_{|_{[x_{j-1},x_j]}} \in \mathbb{P}_n,\ j = 1,\ldots,m \right\}.$$

Dans la suite, on abrégera toujours la notation $\mathcal{S}_n^k(\mathcal{T}_h)$ en $\mathcal{S}_n^k$, la dépendance par rapport à la partition étant sous-entendue. On note qu'une définition plus générale est possible si l'on n'impose pas la continuité *globale* de la fonction spline sur $[a,b]$ (l'entier k pouvant alors prendre la valeur -1). Dans ce cas, la classe $\mathcal{S}_n^{-1}$ désignera l'ensemble des fonctions polynomiales par morceaux, relativement à $\mathcal{T}_h$, de degré n, qui ne sont pas nécessairement des fonctions continues.

Il ressort aussi de cette définition que tout polynôme de degré n sur $[a,b]$ est une fonction spline [38], mais une fonction spline s est en général, et même quasiment toujours en pratique, constituée de polynômes différents sur chacun des sous-intervalles $[x_{j-1}, x_j]$, $j = 1,\ldots,m$, et la dérivée k^{e} de s peut donc présenter une discontinuité en chacun des *nœuds internes* $x_1,\ldots,x_{m-1}$. Un nœud en lequel se produit une telle discontinuité est dit *actif*.

La restriction d'une fonction spline s de $\mathcal{S}_n^k$ à un sous-intervalle $[x_{j-1}, x_j]$, $1 \leq j \leq m$, étant un polynôme de degré n, on voit que l'on a besoin de déterminer $(n+1)m$ coefficients pour caractériser complètement s. La condition de régularité $s \in \mathcal{C}^k([a,b])$ revient à imposer des raccords en valeurs des dérivées d'ordre 0 jusqu'à k de la fonction s en chaque nœud interne de la partition, ce qui fournit $k(m-1)$ équations pour ces coefficients. Si s est par ailleurs une spline d'interpolation de la fonction f aux nœuds de la partition, elle doit vérifier les contraintes

$$s(x_i) = f(x_i),\ i = 0,\ldots,m,$$

et il reste donc $(m-1)(n-k-1)+n-1$ paramètres à fixer pour obtenir s (dans le cas très courant pour lequel $k = n-1$, on aura donc $n-1$ conditions supplémentaires à imposer). C'est le choix (arbitraire) de ces $(m-1)(n-k-1)+n-1$ dernières conditions qui définit alors le type des fonctions splines d'interpolation utilisées.

Interpolation par une fonction spline linéaire

Nous considérons dans cette section le problème de l'interpolation d'une fonction f aux nœuds d'une partition $\mathcal{T}_h$ d'un intervalle $[a,b]$ par une fonction spline linéaire, c'est-à-dire de degré un, continue, c'est-à-dire l'unique fonction s de $\mathcal{S}_1^0$ vérifiant

$$s(x_i) = f(x_i),\ i = 0,\ldots,m. \tag{6.37}$$

En remarquant que ceci correspond à une interpolation de Lagrange par morceaux de degré un de f relativement à $\mathcal{T}_h$ (voir la section 6.3.1), ce problème est trivialement résolu, car l'on déduit immédiatement de la forme de Newton du polynôme d'interpolation que

$$\forall x \in [x_{j-1}, x_j],\ s(x) = f(x_{j-1}) + [x_{j-1}, x_j]f\,(x - x_{j-1}),\ j = 1,\ldots,m-1.$$

L'étude de l'erreur d'interpolation commise est alors très simple, puisque, si la fonction f est de classe $\mathcal{C}^2$ sur $[a,b]$, il découle du théorème 6.16 qu'il existe $\xi_j \in]x_{j-1}, x_j[$, $j = 1,\ldots,m$, tel que

$$\forall x \in [x_{j-1}, x_j],\ f(x) - s(x) = \frac{f''(\xi_j)}{2}(x - x_{j-1})(x - x_j),$$

38. Ceci correspond au choix $k = n = m$. Dans ce cas, la fonction spline interpolant une fonction régulière f est simplement donnée par le polynôme d'interpolation de Lagrange de f aux nœuds de la partition introduit dans la section 6.2.3.

d'où

$$\forall x \in [x_{j-1}, x_j],\ |f(x)-s(x)| \leq \frac{h_j^2}{8} \max_{t\in[x_{j-1},x_j]} \left|f''(t)\right|,$$

ce qui conduit à

$$\|f-s\|_\infty \leq \frac{h^2}{8}\,\|f''\|_\infty.$$

Cette estimation montre que l'erreur d'interpolation peut être rendue aussi petite que souhaité, de manière uniforme sur l'intervalle $[a,b]$, en prenant h suffisamment petit.

Plutôt que de définir la fonction spline linéaire interpolant f par ses restrictions aux sous-intervalles $[x_{j-1}, j]$, $j = 1,\ldots,m$, de la partition $\mathscr{T}_h$, on peut chercher à l'écrire sous la forme d'une combinaison linéaire de $m+1$ fonctions φ_i, $0 \leq i \leq m$, appropriées, dites *fonctions de base*[39],

$$\forall x \in [a,b],\ s(x) = \sum_{i=0}^{m} f(x_i)\,\varphi_i(x), \tag{6.38}$$

où l'on exige de chaque fonction φ_i, $i = 0,\ldots,m$, qu'elle soit une fonction spline linéaire continue, s'annulant en tout nœud de la partition excepté le i^{e}, en lequel elle vaut 1, soit encore

$$\varphi_i(x_j) = \delta_{ij},\ 1 \leq i,j \leq m, \tag{6.39}$$

où δ_{ij} est le symbole de Kronecker. On a, plus explicitement, les définitions suivantes

$$\varphi_i(x) = \begin{cases} \dfrac{x-x_{i-1}}{h_i} & \text{si } x_{i-1} \leq x \leq x_i \\ \dfrac{x_{i+1}-x}{h_{i+1}} & \text{si } x_i \leq x \leq x_{i+1} \\ 0 & \text{sinon} \end{cases},\ 1 \leq i \leq m-1,$$

et

$$\varphi_0(x) = \begin{cases} \dfrac{x_1-x}{h_1} & \text{si } x_0 \leq x \leq x_1 \\ 0 & \text{sinon} \end{cases},\ \varphi_m(x) = \begin{cases} \dfrac{x-x_{m-1}}{h_m} & \text{si } x_{m-1} \leq x \leq x_m \\ 0 & \text{sinon} \end{cases},$$

qui expliquent pourquoi ces fonctions sont parfois appelées *« fonctions chapeaux »* (*"hat functions"* en anglais) en raison de l'allure de leur graphe (voir la figure 6.6).

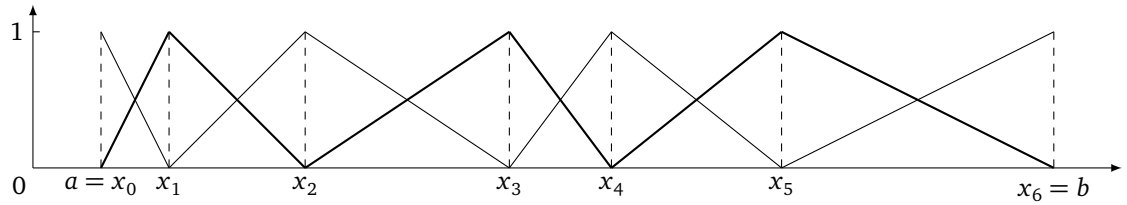


FIGURE 6.6 – Graphes des fonctions de base $\{\varphi_i\}_{i=0,\ldots,6}$ associées aux nœuds $\{x_i\}_{i=0,\ldots,6}$ d'une partition de l'intervalle $[a,b]$.

Interpolation par une fonction spline cubique

Les *fonctions splines d'interpolation cubiques* sont les fonctions splines de plus petit degré permettant une interpolation de classe $\mathscr{C}^2$ d'une fonction f régulière donnée. Nous allons maintenant nous intéresser à l'interpolation aux nœuds de la partition $\mathscr{T}_h$ de l'intervalle $[a,b]$ d'une fonction f de classe $\mathscr{C}^2$ sur $[a,b]$ par une fonction spline s de $\mathscr{S}_3^2$, c'est-à-dire une fonction spline de degré trois et deux fois continûment dérivable satisfaisant (6.37).

39. De telles fonctions forment en effet une base de $\mathscr{S}_1^0$, car on peut facilement montrer, en utilisant la propriété (6.39), qu'elles sont linéairement indépendantes et engendrent cet espace, en vertu de l'égalité (6.38). On peut d'ailleurs voir (6.38) comme un analogue de la formule d'interpolation de Lagrange (6.7).

Nous allons tout d'abord voir comment déterminer une telle fonction. Pour cela, nous introduisons la notation

$$M_i = s''(x_i),\ i = 0, \ldots, m,$$

pour les valeurs de la dérivée seconde de s aux nœuds de la partition $\mathscr{T}_h$, que l'on appelle encore les *moments* de la fonction spline. Puisque la restriction de s à chacun des sous-intervalles de la partition est un polynôme de degré trois, la restriction de sa dérivée seconde est une fonction affine et l'on a

$$\forall x \in [x_{j-1}, x_j],\ s''(x) = M_{j-1}\,\frac{x_j - x}{h_j} + M_j\,\frac{x - x_{j-1}}{h_j},\ j = 1, \ldots, m.$$

En intégrant deux fois cette relation et en utilisant les conditions (6.37) aux nœuds x_{j-1} et x_j, il vient

$$\forall x \in [x_{j-1}, x_j],\ s(x) = M_{j-1}\,\frac{(x_j - x)^3}{6h_j} + M_j\,\frac{(x - x_{j-1})^3}{6h_j} + \left([x_{j-1}, x_j]f - \frac{h_j}{6}\,(M_j - M_{j-1})\right)(x - x_{j-1}) - \frac{h_j^2}{6}\,M_{j-1} + f(x_{j-1}),\ j = 1, \ldots, m.$$

Il faut maintenant imposer la continuité de la dérivée première de s en chacun des nœuds intérieurs x_i, $i = 1, \ldots, m-1$. On obtient alors

$$s'(x_i^-) = \frac{h_i}{6}\,M_{i-1} + \frac{h_i}{3}\,M_i + [x_{i-1}, x_i]f = \frac{h_{i+1}}{6}\,M_i + \frac{h_{i+1}}{3}\,M_{i+1} + [x_i, x_{i+1}]f = s'(x_i^+),\ i = 1, \ldots, m-1,$$

où $s'(x_i^{\pm}) = \lim\limits_{\substack{t \to 0 \\ t > 0}} s'(x_i \pm t)$. Ceci conduit au système d'équations linéaires suivant

$$\mu_i\,M_{i-1} + 2\,M_i + \lambda_i\,M_{i+1} = b_i,\ i = 1, \ldots, m-1, \tag{6.40}$$

où l'on a posé

$$\mu_i = \frac{h_i}{h_i + h_{i+1}},\ \lambda_i = \frac{h_{i+1}}{h_i + h_{i+1}} \text{ et } b_i = \frac{6}{h_i + h_{i+1}}\,([x_i, x_{i+1}]f - [x_{i-1}, x_i]f),\ i = 1, \ldots, m-1. \tag{6.41}$$

On a ainsi obtenu un système de $m-1$ équations à $n+1$ inconnues, auquel deux (c'est-à-dire $3-1$) conditions supplémentaires, généralement appelées *conditions aux extrémités* (*end conditions* en anglais), doivent être ajoutées. Divers choix sont possibles dont voici quelques uns parmi les plus courants (voir également l'illustration à la figure 6.7).

(i) Lorsque les valeurs de la dérivée première de f en a et b sont connues, on peut imposer les conditions aux extrémités

$$s'(a) = f'(a) \text{ et } s'(b) = f'(b),$$

ce qui revient à ajouter les équations

$$\frac{h_1}{3}\,M_0 + \frac{h_1}{6}\,M_1 = [x_0, x_1]f - f'(x_0) \text{ et } \frac{h_m}{6}\,M_{m-1} + \frac{h_m}{3}\,M_m = f'(x_m) - [x_{m-1}, x_m]f.$$

La fonction spline interpolante est alors dite *complète*.

(ii) De la même manière, si l'on connaît les valeurs de la dérivée seconde de f en a et b, on peut choisir d'ajouter les contraintes

$$s''(a) = f''(a) \text{ et } s''(a) = f''(a), \tag{6.42}$$

c'est-à-dire $M_0 = f''(a)$ et $M_m = f''(b)$.

(iii) Si on n'a, en revanche, aucune information sur les dérivées de la fonction f, on peut utiliser les conditions homogènes $s''(a) = s''(b) = 0$, *i.e.*, $M_0 = M_m = 0$. La fonction spline interpolante correspondant à ce choix est dite *naturelle*.

(iv) On peut également chercher à imposer la continuité de s''' aux nœuds internes x_1 et x_{m-1}, ce qui correspond à faire respectivement coïncider les restrictions de la fonction spline sur les deux premiers et les deux derniers sous-intervalles de la partition $\mathscr{T}_h$. Ceci se traduit par les conditions suivantes

$$h_2 M_0 - (h_1 + h_2) M_1 + h_1 M_2 = 0 \text{ et } h_m M_{m-2} - (h_{m-1} + h_m) M_{m-1} + h_{m-1} M_m = 0.$$

Les moments M_1 et M_{m-1} peuvent alors être éliminés du système linéaire obtenu. Les nœuds x_1 et x_{m-1} n'intervenant par conséquent pas dans la construction de la fonction spline interpolante, ce ne sont pas des nœuds actifs et on parle en anglais de *not-a-knot spline*.

(v) Enfin, on peut imposer la condition $s''(a) = s''(b)$, *i.e.*, $M_0 = M_m$. Dans ce cas, la fonction spline d'interpolation obtenue est dite *périodique*.

Les cas de l'interpolation par une fonction spline cubique *not-a-knot* ou périodique mis à part [40], chacun des choix de conditions énumérés ci-dessus conduit [41] à la résolution d'un système linéaire d'ordre $m+1$, dont les inconnues sont les moments de la fonction spline s, de la forme

$$\begin{pmatrix} 2 & \lambda_0 & & & \\ \mu_1 & 2 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{m-1} & 2 & \lambda_{m-1} \\ & & & \mu_m & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_{m-1} \\ M_m \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{m-1} \\ b_m \end{pmatrix}. \tag{6.43}$$

Dans tous les cas cependant, l'existence et l'unicité de la solution du système pour les moments résulte du fait que sa matrice est *à diagonale strictement dominante par lignes* (on a notamment $\mu_i > 0$, $\lambda_i > 0$ et $\mu_i + \lambda_i = 1$ pour $i = 1, \ldots, m-1$), quels que soient la partition $\mathscr{T}_h$ et le choix des conditions aux extrémités. On rappellera pour finir qu'un système linéaire dont la matrice est tridiagonale peut être efficacement résolu par l'algorithme de Thomas présenté dans la section 2.5.4.

Une autre approche pour la construction d'une fonction spline interpolante cubique réside dans l'utilisation d'une base de $\mathscr{S}_3^2$, qui est un espace de dimension $m+3$. Le procédé général [42] le plus courant consiste à utiliser des fonctions polynomiales par morceaux particulières nommées fonctions *B-splines* (ce dernier terme étant une abbréviation de *basis spline* en anglais), qui ont le même degré et la même régularité que la spline que l'on cherche à construire et possèdent de plus les propriétés d'être positives et non nulles sur seulement quelques sous-intervalles contigus de la partition $\mathscr{T}_h$ (on dit qu'elles sont *à support compact*). On trouvera des définitions et de nombreux autres détails sur les fonctions B-splines dans l'ouvrage [dBoo01].

Parmi les propriétés des fonctions splines cubiques interpolantes, on peut mentionner une remarquable propriété de minimisation vérifiée par les fonctions splines *naturelles*.

Théorème 6.22 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^2$ sur $[a,b]$. Si s désigne la fonction spline cubique naturelle interpolant f relativement à une partition $\mathscr{T}_h$ de $[a,b]$, alors, pour toute fonction g de classe $\mathscr{C}^2$ sur $[a,b]$ interpolant la fonction f aux mêmes nœuds que s, on a*

$$\int_a^b \big(g''(x)\big)^2 \, \mathrm{d}x \geq \int_a^b \big(s''(x)\big)^2 \, \mathrm{d}x, \tag{6.44}$$

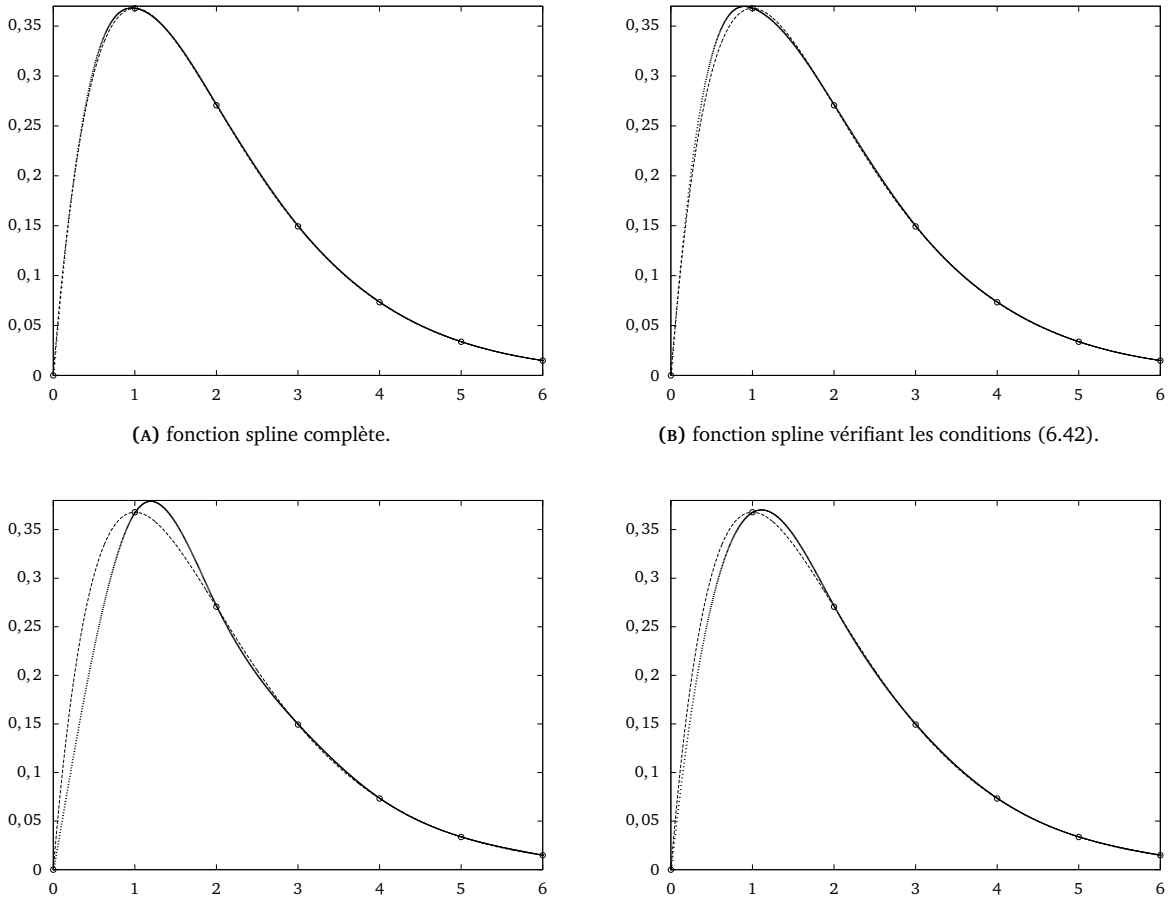
avec égalité si et seulement si $g = s$.

40. Dans le cas d'un choix de conditions périodiques, le système linéaire obtenu est d'ordre m (car $M_0 = M_m$) et n'est pas tridiagonal, puisque sa matrice est

$$\begin{pmatrix} 2 & \lambda_1 & & & \mu_1 \\ \mu_1 & 2 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{m-1} & 2 & \lambda_{m-1} \\ \lambda_m & & & \mu_m & 2 \end{pmatrix}.$$

41. Par exemple, on a $\lambda_0 = 1$, $b_0 = \frac{6}{h_1}\big([x_0, x_1]f - f'(x_0)\big)$, $\mu_m = 1$ et $b_m = \frac{6}{h_m}\big(f'(x_m) - [x_{m-1}, x_m]f\big)$ pour le choix (i), ou bien encore $\lambda_0 = 0$, $b_0 = 0$, $\mu_m = 0$ et $b_m = 0$ pour le choix (iii).

42. La base de $\mathscr{S}_1^0$ exhibée dans la section précedente est par exemple constituée de fonctions B-splines.



(A) fonction spline complète.

(B) fonction spline vérifiant les conditions (6.42).

(C) fonction spline naturelle.

(D) fonction spline *not-a-knot*.

FIGURE 6.7 – Graphes illustrant l'interpolation de la fonction $f(x) = x\,e^{-x}$ sur l'intervalle $[0,6]$ par différents types de fonctions splines cubiques relativement à une partition uniforme de pas égal à 1 (le graphe de la fonction est en trait discontinu, celui de la fonction spline interpolante en pointillé).

DÉMONSTRATION. Pour prouver ce résultat, il suffit d'établir que

$$\int_a^b \big(g''(x)\big)^2\,\mathrm{d}x = \int_a^b \big(g''(x)-s''(x)\big)^2\,\mathrm{d}x + \int_a^b \big(s''(x)\big)^2\,\mathrm{d}x. \tag{6.45}$$

L'inégalité (6.44) découle en effet directement de cette relation et l'on voir égalité dans (6.44) si et seulement si $g''-s''=0$ sur $[a,b]$, ce qui implique, après deux intégrations entre a et x et utilisation des conditions satisfaites par les fonctions g et s au nœud a, que $g=s$ sur $[a,b]$.

On remarque alors que (6.45) est équivalente à

$$\int_a^b s''(x)\big(g''(x)-s''(x)\big)\,\mathrm{d}x = 0.$$

En intégrant par partie, il vient alors

$$\int_a^b s''(x)\big(g''(x)-s''(x)\big)\,\mathrm{d}x = \big[s''(x)\big(g'(x)-s'(x)\big)\big]_{x=a}^b - \int_a^b s'''(x)\big(g'(x)-s'(x)\big)\,\mathrm{d}x$$
$$= -\int_a^b s'''(x)\big(g'(x)-s'(x)\big)\,\mathrm{d}x,$$

puisque $s''(a) = s''(b) = 0$. La fonction s''' étant constante par morceaux relativement à la partition $\mathscr{T}_h$, on a alors

$$\int_a^b s'''(x)\big(g'(x)-s'(x)\big)\,\mathrm{d}x = \sum_{i=0}^{m-1} s'''(x_i^+)\int_{x_i}^{x_{i+1}} \big(g'(x)-s'(x)\big)\,\mathrm{d}x$$
$$= \sum_{i=0}^{m-1} s'''(x_i^+)(g(x_{i+1})-s(x_{i+1})-(g(x_i)-s(x_i))) = 0,$$

avec $s'''(x_i^+) = \lim\limits_{\substack{t\to 0\\ t>0}} s'''(x_i+t)$, $i = 0,\dots,m-1$. □

Cette caractérisation variationnelle est due à Holladay [Hol57]. On dit que la fonction spline cubique naturelle est l'interpolant *« le plus régulier »* d'une fonction f de classe $\mathscr{C}^2$ sur l'intervalle $[a,b]$, au sens où la norme $L^2(]a,b[)$ de sa dérivée seconde,

$$\|s''\|_{L^2(]a,b[)} = \left(\int_a^b \big|s''(x)\big|^2\,\mathrm{d}x\right)^{1/2}, \text{ avec } L^2(]a,b[) = \Big\{f\ :]a,b[\to\mathbb{R}\mid \int_a^b |f(x)|^2\,\mathrm{d}x\Big\},$$

est la plus petite parmi celles de toutes[43] les fonctions g de classe $\mathscr{C}^2$ interpolant f aux nœuds d'une partition donnée de l'intervalle $[a,b]$. Cette propriété est d'ailleurs à l'origine du terme « spline », qui est un mot d'origine anglaise désignant une latte souple en bois, ou *cerce*, servant à tracer des courbes. En effet, en supposant que la forme prise par une cerce que l'on contraint à passer par des points de contrôle donnés $(x_i, f(x_i))$, $i = 0,\dots,n$, est une courbe d'équation $y = g(x)$, avec $a \le x \le b$, on observe que la configuration adoptée minimise, parmi toutes celles satisfaisant les mêmes contraintes, l'énergie de flexion

$$E(g) = \int_a^b \frac{(g''(x))^2}{\big(1+|g'(x)|^2\big)^3}\,\mathrm{d}x.$$

Pour une courbe variant lentement, c'est-à-dire pour $\|g'\|_\infty \ll 1$, on voit que cette dernière propriété est très proche de (6.44).

Ajoutons que l'inégalité (6.44) est encore valable si s est une fonction spline cubique complète et si la fonction g vérifie $g'(a) = f'(a)$ et $g'(b) = f'(b)$, en plus d'interpoler f aux nœuds de $\mathscr{T}_h$.

Nous terminons cette section un résultat d'estimation d'erreur pour les fonctions splines interpolantes *complètes*.

Théorème 6.23 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^4$ sur $[a,b]$, $\mathscr{T}_h$ une partition de $[a,b]$ et $K \ge 1$ la constante définie par*

$$K = \max_{1\le j\le m} \frac{h}{h_j}.$$

Si s est la fonction spline cubique complète interpolant f aux nœuds x_i, $i = 0,\dots,m$, de $\mathscr{T}_h$, alors il existe des constantes $C_k \le 2$, ne dépendant pas de $\mathscr{T}_h$, telles que

$$\|f^{(k)} - s^{(k)}\|_\infty \le C_k K h^{4-k}\|f^{(4)}\|_\infty,\ k = 0,1,2,3.$$

Pour démontrer ce théorème, nous aurons besoin du résultat suivant.

Lemme 6.24 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^4$ sur $[a,b]$ et $\mathscr{T}_h$ une partition de $[a,b]$. En notant $\boldsymbol{M}$ le vecteur des moments M_i, $i = 0,\dots,m$, de la fonction spline cubique complète interpolant f aux nœuds de $\mathscr{T}_h$ et en posant*

$$\boldsymbol{F} = \begin{pmatrix} f''(x_0)\\ \vdots \\ f''(x_m)\end{pmatrix},$$

on a

$$\|\boldsymbol{M}-\boldsymbol{F}\|_\infty \le \frac{3}{4}h^2\|f^{(4)}\|_\infty.$$

43. On remarquera que l'inégalité (6.44) est en particulier vraie si $g = f$.

DÉMONSTRATION. Posons $\boldsymbol{r} = A(\boldsymbol{M}-\boldsymbol{F}) = \boldsymbol{b}-A\boldsymbol{F}$, où A et $\boldsymbol{b}$ désignent respectivement la matrice et le second membre du système linéaire (6.43) associé aux moments de la fonction spline cubique complète interpolant f. On a alors

$$r_0 = \frac{6}{h_1}\left([x_0,x_1]f - f'(x_0)\right) - 2f''(x_0) - f''(x_1).$$

En utilisant la formule de Taylor–Lagrange pour exprimer $f(x_1)$ et $f''(x_1)$ en termes des valeurs de f et de ses dérivées au point x_0, il vient

$$\begin{aligned} r_0 = \frac{6}{h_1}\left(f'(x_0) + \frac{h_1}{2}f''(x_0) + \frac{h_1^2}{6}f'''(x_0) + \frac{h_1^3}{24}f^{(4)}(\eta_1) - f'(x_0)\right) \\ - 2f''(x_0) - \left(f''(x_0) + h_1 f'''(x_0) + \frac{h_1}{2}f^{(4)}(\eta_2)\right) = \frac{h_1^2}{4}f^{(4)}(\eta_1) - \frac{h_1}{2}f^{(4)}(\eta_2), \end{aligned}$$

avec η_1 et η_2 appartenant à $]x_0,x_1[$, d'où

$$|r_0| \le \frac{3}{4}h^2\|f^{(4)}\|_\infty.$$

De manière analogue, on obtient

$$|r_m| \le \frac{3}{4}h^2\|f^{(4)}\|_\infty.$$

Pour les composantes restantes du vecteur $\boldsymbol{r}$, on trouve

$$r_i = \frac{6}{h_i+h_{i+1}}\left([x_i,x_{i+1}]f - [x_{i-1},x_i]f\right) - \frac{h_i}{h_i+h_{i+1}}f''(x_{j-1}) - 2f''(x_i) - \frac{h_{i+1}}{h_i+h_{i+1}}f''(x_{i+1}),\ 1 \le i \le m-1,$$

et, par la formule de Taylor–Lagrange,

$$\begin{aligned} r_i = \frac{1}{h_i+h_{i+1}}\Bigg[6\Bigg(f'(x_i) + \frac{h_{i+1}}{2}f''(x_i) + \frac{h_{i+1}^2}{6}f'''(x_i) + \frac{h_{i+1}^3}{24}f^{(4)}(\eta_1) \\ - f'(x_j) - \frac{h_i}{2}f''(x_i) - \frac{h_i^2}{6}f'''(x_i) - \frac{h_i^3}{24}f^{(4)}(\eta_2)\Bigg) \\ - h_i\left(f''(x_i) - h_i f'''(x_i) + \frac{h_i^2}{2}f^{(4)}(\eta_3)\right) - 2(h_i+h_{i+1})f''(x_i) - h_{i+1}\left(f''(x_i) + h_{i+1}f'''(x_i) + \frac{h_{i+1}}{2}f^{(4)}(\eta_4)\right)\Bigg] \\ = \frac{1}{h_i+h_{i+1}}\left(\frac{h_{i+1}^3}{4}f^{(4)}(\eta_1) - \frac{h_i^3}{4}f^{(4)}(\eta_2) - \frac{h_i^3}{2}f^{(4)}(\eta_3) - \frac{h_{i+1}^3}{2}f^{(4)}(\eta_4)\right), \end{aligned}$$

avec η_1, η_2, η_3 et η_4 appartenant à $]x_{i-1},x_{i+1}[$. Il vient alors

$$|r_i| \le \frac{3}{4}\frac{h_i^3+h_{i+1}^3}{h_i+h_{i+1}}\|f^{(4)}\|_\infty \le \frac{3}{4}h^2\|f^{(4)}\|_\infty.$$

Pour conclure, il suffit alors de remarquer que, pour tous vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ de $\mathbb{R}^{m+1}$, de composantes respectives u_i et v_i, $i = 0,\dots,m$, on a

$$A\boldsymbol{u} = \boldsymbol{v} \ \Rightarrow\ \|\boldsymbol{u}\|_\infty \le \|\boldsymbol{v}\|_\infty.$$

En effet, en notant i_0 l'indice tel que $\left|u_{i_0}\right| = \|\boldsymbol{u}\|_\infty = \max\limits_{0\le i\le m}|u_i|$, on a, en posant $\mu_0 = 0$ et $\lambda_m = 0$,

$$\mu_{i_0}u_{i_0-1} + 2u_{i_0} + \lambda_{i_0}u_{i_0+1} = v_{i_0},$$

d'où, en se servant de (6.41),

$$\|\boldsymbol{v}\|_\infty \ge \left|v_{i_0}\right| \ge 2\left|u_{i_0}\right| - \mu_{i_0}\left|u_{i_0-1}\right| - \lambda_{i_0}\left|u_{i_0+1}\right| \ge (2-\mu_{i_0}-\lambda_{i_0})\left|u_{i_0}\right| \ge \left|u_{i_0}\right| = \|\boldsymbol{u}\|_\infty.$$

□

DÉMONSTRATION DU THÉORÈME 6.23. Prouvons tout d'abord le résultat pour $k = 3$. Pour tout point x dans l'intervalle $[x_{j-1},x_j]$, $j = 1,\dots,m$, on a

$$\begin{aligned} s'''(x) - f'''(x) = \frac{M_j - M_{j-1}}{h_j} - f'''(x) \\ = \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1}-f''(x_{j-1})}{h_j} + \frac{f''(x_j) - f''(x) - (f''(x_{j-1}) - f''(x))}{h_j} - f'''(x). \end{aligned}$$

En vertu de la formule de Taylor–Lagrange et du lemme 6.24, il vient alors

$$\left|f'''(x)-s'''(x)\right| \leq \frac{3}{2}\frac{h^2}{h_j}\|f^{(4)}\|_\infty + \frac{1}{h_j}\left|(x_j-x)f'''(x)+\frac{(x_j-x)^2}{2}f^{(4)}(\eta_1)-(x_{j-1}-x)f'''(x)-\frac{(x_{j-1}-x)^2}{2}f^{(4)}(\eta_2)-h_j f'''(x)\right|,$$

avec η_1 et η_2 appartenant à $]x_{j-1},x_j[$, d'où

$$\left|f'''(x)-s'''(x)\right| \leq 2\frac{h^2}{h_j}\|f^{(4)}\|_\infty.$$

Puisqu'on a, par définition, $\dfrac{h}{h_j} \leq K$, on trouve finalement

$$\left|f'''(x)-s'''(x)\right| \leq 2Kh\|f^{(4)}\|_\infty.$$

Pour $k=2$, on remarque que, pour tout x dans $]a,b[$, il existe un entier $j=j(x)$ tel que

$$\left|x_{j(x)}-x\right| \leq \frac{h}{2}.$$

On a alors

$$f''(x)-s''(x) = f''(x_{j(x)})-s''(x_{j(x)})+\int_{x_{j(x)}}^{x}(f''(t)-s''(t))\,\mathrm{d}t,$$

et, puisque $K\geq 1$,

$$\forall x\in[a,b],\ \left|f''(x)-s''(x)\right| \leq \frac{3}{4}h^2\|f^{(4)}\|_\infty+\frac{h}{2}2Kh\|f^{(4)}\|_\infty \leq \frac{7}{4}h^2\|f^{(4)}\|_\infty.$$

Considérons maintenant le cas $k=1$. En plus des extrémités $\xi_0=a$ et $\xi_{m+1}=b$, il existe, en vertu du théorème de Rolle, m points $\xi_j\in]x_{j-1},x_j[$, $j=1,\ldots,m$, vérifiant

$$f'(\xi_j)=s'(x_j),\ j=0,\ldots,m+1.$$

Pour chaque point x de l'intervalle $[a,b]$, il existe donc un entier $j=j(x)$ tel que

$$\left|\xi_{j(x)}-x\right| \leq h,$$

et l'on a par conséquent

$$f'(x)-s'(x)=\int_{\xi_{j(x)}}^{x}(f''(t)-s''(t))\,\mathrm{d}t.$$

On en déduit finalement que

$$\left|f'(x)-s'(x)\right| \leq h\frac{7}{4}h^2\|f^{(4)}\|_\infty = \frac{7}{4}h^3\|f^{(4)}\|_\infty,\ \forall x\in[a,b].$$

Reste le cas $k=0$. Comme

$$f(x)-s(x)=\int_{x_{j(x)}}^{x}(f'(t)-s'(t))\,\mathrm{d}t,\ \forall x\in[a,b],$$

il vient, en utilisant l'inégalité établie pour $k=1$,

$$|f(x)-s(x)| \leq \frac{h}{2}\frac{7}{4}h^3\|f^{(4)}\|_\infty = \frac{7}{8}h^4\|f^{(4)}\|_\infty,\ \forall x\in[a,b].$$

□

La constante $K\geq 1$ introduite dans ce théorème « mesure » la déviation de la partition $\mathscr{T}_h$ par rapport à la partition uniforme de $[a,b]$ ayant le même nombre de sous-intervalles. Ce résultat montre alors, sous réserve que K soit uniformément bornée lorsque h tend vers zéro, la fonction spline cubique complète et ses trois premières dérivées convergent, uniformément sur $[a,b]$, vers f et ses dérivées[44] lorsque le nombre de points d'interpolation tend vers l'infini. On ajoutera que ces estimations peuvent être améliorées. Sous les mêmes hypothèses, on montre (voir pour cela [HM76]) en effet que

$$\|f^{(k)}-s^{(k)}\|_\infty \leq c_k h^{4-r}\|f^{(4)}\|_\infty,\ k=0,1,2,3,$$

avec $c_0=\dfrac{5}{384}$, $c_1=\dfrac{1}{24}$, $c_2=\dfrac{3}{8}$ et $c_3=\dfrac{1}{2}(K+K^{-1})$, les constantes c_0 et c_1 étant optimales.

44. On remarquera au passage que f''' est alors approchée par une suite de fonctions en général discontinues (car constantes par morceaux).

6.4 Notes sur le chapitre

Pour plus de détails sur la génèse du théorème d'approximation de Weierstrass, on pourra consulter l'article de revue [Pin00], dans lequel sont présentées plusieurs preuves alternatives et généralisations de ce résultat.

A VOIR polynômes de Bernstein et courbes de Bézier, algorithme de De Casteljau

On retrouve la forme de Lagrange du polynôme d'interpolation, ainsi que les polynômes de Lagrange, dans une leçon donnée par Lagrange à l'École normale en 1795 [Lag77], mais sa découverte semble due à Waring[45], qui en fit la publication seize ans auparavant [War79]. Les différences divisées et la forme dite de Newton du polynôme d'interpolation ont pour leur part été introduites dans la preuve du lemme V, traitant de la résolution du problème d'interpolation polynômiale, du troisième livre des *Philosophiæ naturalis principia mathematica*, l'œuvre maîtresse de Newton publiée en 1687.

La première forme de la formule d'interpolation barycentrique apparaît déjà dans la thèse de Jacobi, intitulée *Disquisitiones analyticae de fractionibus simplicibus* et soutenue en 1825. La seconde forme n'arrive en revanche que bien plus tard, dans l'article [Tay45], son utilisation étant alors restreinte à un choix de nœuds équidistribués. La dénomination d'interpolation « barycentrique » semble pour sa part provenir de la note [Dup48]. Pour plus de détails sur ces formules d'interpolation et de nombreuses références bibliographiques associées, on pourra consulter l'excellent article de synthèse [BT04b].

Le fait que la suite des constantes de Lebesgue $(\Lambda_n)_{n\in\mathbb{N}}$ de toute distribution de nœuds a pour limite $+\infty$ est à relier à un théorème dû à Faber[46] [Fab14] (voir également [Ber18]), qui répond par la négative à la question légitime : *« existe-t-il un tableau triangulaire infini de points appartenant à l'intervalle* $[a,b]$*,*

$$\begin{array}{ccccccc}
x_0^{(0)} & & & & & & \\
x_0^{(1)}, & x_1^{(1)} & & & & & \\
x_0^{(2)}, & x_1^{(2)}, & x_2^{(2)} & & & & \\
\vdots & \vdots & \vdots & \ddots & & & \\
x_0^{(n)}, & x_1^{(n)}, & x_2^{(n)}, & \cdots & x_n^{(n)} & & \\
\vdots & \vdots & \vdots & & \vdots & \ddots &
\end{array},$$

tel que, pour n'importe quelle fonction f *continue sur* $[a,b]$*, la suite* $(\Pi_n f)_{n\in\mathbb{N}}$ *des polynômes d'interpolation de Lagrange associés aux nœuds des lignes de ce tableau converge, en norme de la convergence uniforme, vers* f *? »*, en stipulant que, pour chaque tableau de points, il existe une fonction continue pour laquelle la suite de polynômes d'interpolation associés diverge. Des résultats encore plus pessimistes furent obtenus par la suite, notamment par Bernstein [Ber31] (A VERIFIER : Pour tout tableau, il existe une fonction continue et un point c de $[a,b]$ tels que $\Pi_n f(c)$ ne converge pas vers $f(c)$ quand n tend vers $+\infty$) et par Erdős et Vértesi [EV80] (A VERIFIER : Pour tout tableau, il existe une fonction continue telle que $\Pi_n f(c)$ ne converge pas vers $f(c)$ quand n tend vers $+\infty$ pour presque tout c dans $[a,b]$).

L'interpolation de Lagrange aux points de Chebyshev de fonctions via les formules barycentriques est à la base de `Chebfun` [BT04a, DHT14], une collection de programmes *open source* écrits pour MATLAB permettant d'effectuer diverses opérations numériques sur des fonctions continues ou continues par morceaux via la surcharge de commandes normalement réservées à la manipulation de vecteurs et de matrices.

A VOIR : autre généralisation (mean value interpolation) de l'interpolation de Lagrange, principalement à portée théorique car difficile à mettre en œuvre, comme l'*interpolation de Kergin* [Ker80]

L'introduction de la théorie de l'approximation par des splines est due à Schoenberg avec la publication de deux articles fondateurs en 1946 [Sch46a, Sch46b], mais ce n'est réellement qu'à partir des années 1970, avec la découverte, de manière indépendante par Cox [Cox72] et de Boor [dBoo72] en 1972, d'une formule de récurrence numériquement stable pour le calcul de fonctions B-splines, qu'elle ne furent employées pour des applications.

45. Edward Waring (v. 1736 - 15 août 1798) était un mathématicien anglais. Il traita dans son célèbre ouvrage *Meditationes algebraicæ*, publié en 1770, des solutions des équations algébriques et de la théorie des nombres, mais il travailla aussi notamment sur l'approximation des racines, l'interpolation et la géométrie des coniques.

46. Georg Faber (5 avril 1877 - 7 mars 1966) était un mathématicien allemand. Il travailla de manière importante sur la représentation des fonctions complexes par des séries de polynômes à l'intérieur de courbes régulières et prouva, indépendamment de Krahn, une conjecture de Rayleigh concernant la forme minimisant, à volume donné, la plus petite valeur propre de l'opérateur laplacien muni d'une condition aux limites de Dirichlet homogène.

Ajoutons que l'interpolation polynomiale se généralise très simplement au cas multidimensionnel lorsque le domaine d'interpolation est un produit tensoriel d'intervalles. Associée à des nœuds choisis comme étant les racines de polynômes orthogonaux, elle est à l'origine de plusieurs *méthodes spectrales* d'approximation (voir par exemple [Tre00]). L'interpolation polynomiale par morceaux est pour sa part extrêmement flexible et permet, une fois étendue au cas multidimensionnel, de prendre en compte facilement des domaines de forme complexe (typiquement tout polygone lorsqu'on se place dans $\mathbb{R}^2$ ou tout polyèdre dans $\mathbb{R}^3$). La théorie de l'interpolation est à ce titre un outil de base de la *méthode des éléments finis*, introduite par Courant [47] [Cou43], qui, tout comme les méthodes spectrales, est très utilisée pour la résolution numérique des équations aux dérivées partielles (voir par exemple [CL09]).

Terminons ce chapitre par une application originale de l'interpolation de Lagrange, le *partage de clé secrète de Shamir* [48] (*Shamir's secret sharing* en anglais) [Sha79] en cryptographie. Cette méthode consiste en la répartition d'informations liées à une donnée secrète, comme une clé de chiffrement ou un mot de passe, entre plusieurs dépositaires, ces derniers ne pouvant retrouver facilement la donnée que si un nombre suffisant d'entre eux mettent en commun les partie (*shares* en anglais)s qu'ils ont reçues. Formellement, on suppose que l'on souhaite distribuer le secret en n parties et que seule la connaissance d'un nombre m, avec $m \leq n$, dit *seuil* (*threshold* en anglais), de ces parties rende aisé le recouvrement du secret. Faisant l'hypothèse que le secret est un élément s d'un *corps fini* [49], l'idée de Shamir est de choisir au hasard $m-1$ coefficients $a_1, \ldots, a_{m-1}$ dans ce corps fini et de construire le polynôme $p(x) = s + a_1 x + \cdots + a_{m-1} x^{m-1}$, les parties distribuées aux dépositaires étant les n couples $(x_i, p(x_i))$, $i = 1, \ldots, n$, composés d'éléments x_i distincts du corps fini et de leurs images par p dans ce même corps. On retrouve alors le secret à partir d'un jeu de m parties en construisant tout d'abord le polynôme d'interpolation de Lagrange de degré $m-1$ qui lui est associé et en identifiant ensuite le coefficient associé au terme de degré zéro dans ce polynôme, comme le montre l'exemple ci-dessous. Ce type de partage est dit *inconditionnellement sûr* (on parle encore de *perfect secrecy* en anglais), au sens où un attaquant ne peut récupérer aucune information sur le secret tant qu'il possède moins de m parties.

Exemple d'application du partage de clé secrète de Shamir. Illustrons le principe du partage de Shamir sur un exemple dans lequel les calculs sont effectués dans un premier temps en arithmétique entière. Supposons que le secret est $s = 1234$ et que l'on veut effectuer un partage en $n = 6$ parties sachant que la connaissance $m = 3$ parties doit être suffisante pour reconstruire le secret. On tire donc au hasard les $m-1$ coefficients $a_1 = 166$ et $a_2 = 94$ afin d'obtenir le polynôme $p(x) = 1234 + 199\,x + 94\,x^2$ permettant de générer les données à distribuer. On construit ensuite les n couples de données suivants : $(1, 1494)$, $(2, 1942)$, $(3, 2578)$, $(4, 3402)$, $(5, 4414)$ et $(6, 5614)$, chaque couple étant donné à un dépositaire différent. Considérons à présent que l'on dispose des trois parties $(2, 1942)$, $(4, 3402)$ et $(5, 4414)$ pour déterminer le secret. Les polynômes de Lagrange associés sont alors

$$l_0(x) = \frac{1}{6}(x-4)(x-5) = \frac{10}{3} - \frac{3}{2}x + \frac{1}{6}x^2,\ l_1(x) = -\frac{1}{2}(x-2)(x-5) = -5 + \frac{7}{2}x - \frac{1}{2}x^2$$
$$\text{et } l_2(x) = \frac{1}{3}(x-2)(x-4) = \frac{8}{3} - 2x + \frac{1}{3}x^2,$$

et le polynôme d'interpolation de Lagrange est

$$\Pi_2(x) = 1942\left(\frac{10}{3} - \frac{3}{2}x + \frac{1}{6}x^2\right) + 3402\left(-5 + \frac{7}{2}x - \frac{1}{2}x^2\right) + 4414\left(\frac{8}{3} - 2x + \frac{1}{3}x^2\right) = 1234 + 166\,x + 92\,x^2.$$

En se rappelant que le secret est donné par la valeur de ce dernier polynôme en $x = 0$, on trouve que $s = 1234$.

Si le procédé que l'on vient de présenter fonctionne en pratique, il présente un problème de sécurité. Un attaquant peut en effet gagner de l'information avec chaque couple de données qu'il obtient en exploitant le fait que les points du graphe du polynôme d'interpolation sur $\mathbb{N}$ se trouvent sur une courbe régulière. C'est pour cette raison qu'une arithmétique sur un corps fini à q éléments, avec $q > s$ et $q > n$, doit être utilisée pour les calculs, les points du graphe du polynôme sur un corps fini apparaissant totalement désorganisés.

47. Richard Courant (8 janvier 1888 - 27 janvier 1972) était un mathématicien germano-américain. Il est en grande partie à l'origine de la *méthode des éléments finis* utilisée pour la résolution numérique de nombreux problèmes d'équations aux dérivées partielles.

48. Adi Shamir (né le 6 juillet 1952) est un cryptologue israélien. Il est l'inventeur de la cryptanalyse différentielle et, avec Ron Rivest et Len Adleman, l'un des co-inventeurs de l'algorithme RSA. Il lui doit aussi des contributions dans divers domaines de l'informatique, notamment en théorie de la complexité algorithmique.

49. Un corps fini est un corps commutatif qui est par ailleurs fini. Dans la plupart des applications en cryptographie, on utilise un corps fini de la forme $\mathbb{Z}_q = \mathbb{Z}/q\mathbb{Z}$, avec q un nombre premier, ou une extension d'un tel corps.

Ce changement primordial modifie cependant peu la manière de faire. Il suffit de sélectionner un nombre premier q plus grand que le nombre de parties n et que chacun des coefficients a_i, $i = 0, \ldots, m-1$, et de produire les données en évaluant des couples de la forme $(x, p(x) \mod q)$ en place de $(x, p(x))$. Dans le cas de l'exemple précédent, en choisissant $q = 1613$, ceci revient à utiliser les couples de données $(1, 1494)$, $(2, 329)$, $(3, 965)$, $(4, 176)$, $(5, 1188)$ et $(6, 775)$. On observe que la valeur de l'entier q doit être connue de chacun des dépositaires et est considérée comme publique. Elle ne doit donc pas être choisie trop petite sous peine que le secret puisse être facilement deviné en testant les q valeurs qu'il peut prendre.

Références

[Ait32] A. C. AITKEN. On interpolation by iteration of proportional parts, without the use of differences. *Proc. Edinburgh Math. Soc. (2)*, 3(1):56–76, 1932. DOI: 10.1017/S0013091500013808.

[Bec00] B. BECKERMANN. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Math.*, 85(4):553–577, 2000. DOI: 10.1007/PL00005392.

[Ber13] S. N. BERNSTEIN. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Comm. Soc. Math. Kharkov*, 13(1-2) :501-517, 1913.

[Ber18] S. N. BERNSTEIN. Quelques remarques sur l'interpolation. *Math. Ann.*, 79(1-2) :1-12, 1918. DOI : 10.1007/BF01457173.

[Ber31] S. BERNSTEIN. Sur la limitation des valeurs d'un polynôme $P_n(x)$ de degré n sur tout un segment par ses valeurs en $(n+1)$ points du segment. *Bull. Acad. Sci. URSS*, 8 :1025-1050, 1931.

[Bir06] G. D. BIRKHOFF. General mean value and remainder theorems with applications to mechanical differentiation and quadrature. *Trans. Amer. Math. Soc.*, 7(1):107–136, 1906. DOI: 10.1090/S0002-9947-1906-1500736-1.

[BM97] J.-P. BERRUT and H. D. MITTELMANN. Lebesgue constant minimizing linear rational interpolation of continuous functions over the interval. *Comput. Math. Appl.*, 33(6):77–86, 1997. DOI: 10.1016/S0898-1221(97)00034-5.

[Bor05] E. BOREL. *Leçons sur les fonctions de variables réelles et les déloppements en séries de polynômes*. Gauthier-Villars, 1905.

[BP70] Å. BJÖRCK and V. PEREYRA. Solution of Vandermonde systems of equations. *Math. Comput.*, 24(112):893–903, 1970. DOI: 10.1090/S0025-5718-1970-0290541-1.

[Bru78] L. BRUTMAN. On the Lebesgue function for polynomial interpolation. *SIAM J. Numer. Anal.*, 15(4):694–704, 1978. DOI: 10.1137/0715046.

[BT04a] Z. BATTLES and L. N. TREFETHEN. An Extension of MATLAB to continuous functions and operators. *SIAM J. Sci. Comput.*, 25(5):1743–1770, 2004. DOI: 10.1137/S1064827503430126.

[BT04b] J.-P. BERRUT and L. N. TREFETHEN. Barycentric Lagrange interpolation. *SIAM Rev.*, 46(3):501–517, 2004. DOI: 10.1137/S0036144502417715.

[Cau40] A.-L. CAUCHY. Sur les fonctions interpolaires. *C. R. Acad. Sci. Paris*, 11 :775-789, 1840.

[CL09] P. CIARLET et É. LUNÉVILLE. *La méthode des éléments finis. De la théorie à la pratique. I. Concepts généraux*. Les presses de l'École Nationale Supérieure de Techniques Avancées (ENSTA), 2009.

[Cou43] R. COURANT. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49(1):1–23, 1943. DOI: 10.1090/S0002-9904-1943-07818-4.

[Cox72] M. G. COX. The numerical evaluation of *B*-splines. *IMA J. Appl. Math.*, 10(2):134–149, 1972. DOI: 10.1093/imamat/10.2.134.

[dBoo01] C. de BOOR. *A practical guide to splines*, volume 27 of *Applied mathematical sciences*. Springer, revised edition, 2001.

[dBoo72] C. de BOOR. On calculating with *B*-splines. *IMA J. Appl. Math.*, 6(1):50–62, 1972. DOI: 10.1016/0021-9045(72)90080-9.

[DHT14] T. A. DRISCOLL, N. HALE, and L. N. TREFETHEN. *Chebfun guide*. Pafnuty publications, 2014. URL: http://www.chebfun.org/docs/guide/.

[dlVal10] C.-J. de la VALLÉE POUSSIN. Sur les polynômes d'approximation et la représentation approchée d'un angle. *Bull. Cl. Sci. Acad. Roy. Belg.*, 12 :808-844, 1910.

[Dup48] M. DUPUY. Le calcul numérique des fonctions par l'interpolation barycentrique. *C. R. Acad. Sci. Paris*, 226 :158-159, 1948.

[Epp87] J. F. EPPERSON. On the Runge example. *Amer. Math. Monthly*, 94(4):329–341, 1987. DOI: 10.2307/2323093.

[Erd61] P. ERDŐS. Problems and results on the theory of interpolation. II. *Acta Math. Hungar.*, 12(1-2):235–244, 1961. DOI: 10.1007/BF02066686.

[EV80] P. ERDŐS and P. VÉRTESI. On the almost everywhere divergence of Lagrange interpolatory polynomials for arbitrary system of nodes. *Acta Math. Hungar.*, 36(1-2):71–94, 1980. DOI: 10.1007/BF01897094.

[Fab14] G. FABER. Über die interpolatorische Darstellung stetiger Funktionen. *Jahresber. Deutsch. Math.-Verein.*, 23:192–210, 1914.

[FR89] B. FISCHER and L. REICHEL. Newton interpolation in Fejér and Chebyshev points. *Math. Comput.*, 53(187):265–278, 1989. DOI: 10.1090/S0025-5718-1989-0969487-3.

[Gau75] W. GAUTSCHI. Norm estimates for inverses of Vandermonde matrices. *Numer. Math.*, 23(4):337–347, 1975. DOI: 10.1007/BF01438260.

[Her78] C. HERMITE. Sur la formule d'interpolation de Lagrange. *J. Reine Angew. Math.*, 1878(84) :70-79, 1878. DOI : 10.1515/crll.1878.84.70.

[Hig04] N. J. HIGHAM. The numerical stability of barycentric Lagrange interpolation. *IMA J. Numer. Anal.*, 24(4):547–556, 2004. DOI: 10.1093/imanum/24.4.547.

[Hig87] N. J. HIGHAM. Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems. *Numer. Math.*, 50(5):613–632, 1987. DOI: 10.1007/BF01408579.

[HM76] C. A. HALL and W. W. MEYER. Optimal error bounds for cubic spline interpolation. *J. Approx. Theory*, 16(2):105–122, 1976. DOI: 10.1016/0021-9045(76)90040-X.

[Hol57] J. C. HOLLADAY. A smoothest curve approximation. *Math. Comput.*, 11(60):233–243, 1957. DOI: 10.1090/S0025-5718-1957-0093894-6.

[IK94] E. ISAACSON and H. B. KELLER. *Analysis of numerical methods*. Dover, 1994.

[Ker80] P. KERGIN. A natural interpolation of C^K functions. *J. Approx. Theory*, 29(4):278–293, 1980. DOI: 10.1016/0021-9045(80)90116-1.

[Kuh64] H. KUHN. Ein elementarer Beweis des Weierstraßschen Approximationssatzes. *Arch. Math. (Basel)*, 15(1):316–317, 1964. DOI: 10.1007/BF01589203.

[Lag77] J. L. LAGRANGE. Leçons élémentaires sur les mathématiques données à l'École normale en 1795. Leçon cinquième. Sur l'usage des courbes dans la solution des problèmes. In J.-A. SERRET, éditeur, *Œuvres de Lagrange, tome septième*, pages 271-287. Gauthier–Villars, Paris, 1877.

[Lej57] F. LEJA. Sur certaines suites liées aux ensembles plans et leur application à la représentation conforme. *Ann. Polon. Math.*, 4 :8-13, 1957. DOI : 10.4064/ap-4-1-8-13.

[Mei02] E. MEIJERING. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proc. IEEE*, 90(3):319–342, 2002. DOI: 10.1109/5.993400.

[Nev34] E. H. NEVILLE. Iterative interpolation. *J. Indian Math. Soc.*, 20:87–120, 1934.

[Pin00] A. PINKUS. Weierstrass and approximation theory. *J. Approx. Theory*, 107(1):1–66, 2000. DOI: 10.1006/jath.2000.3508.

[PM72] T. PARKS and J. MCCLELLAN. Chebyshev approximation for nonrecursive digital filters with linear phase. *IEEE Trans. Circuit Theory*, 19(2):189–194, 1972. DOI: 10.1109/TCT.1972.1083419.

[PTK11] R. B. PLATTE, L. N. TREFETHEN, and A. B. J. KUIJLAARS. Impossibility of fast stable approximation of analytic functions from equispaced samples. *SIAM Rev.*, 53(2):308–318, 2011. DOI: 10.1137/090774707.

[Rei90] L. REICHEL. Newton interpolation at Leja points. *BIT*, 30(2):332–346, 1990. DOI: 10.1007/BF02017352.

[Rem34a] E. REMES. Sur la détermination des polynômes d'approximation de degré donné. *Comm. Soc. Math. Kharkov*, 10 :41-63, 1934.

[Rem34b] E. REMES. Sur le calcul effectif des polynômes d'approximation de Tchebichef. *C. R. Acad. Sci. Paris*, 199 :337-340, 1934.

[Rem34c] E. REMES. Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation. *C. R. Acad. Sci. Paris*, 198 :2063-2065, 1934.

[Rie16] M. RIESZ. Über einen Satz des Herrn Serge Bernstein. *Acta Math.*, 40(1):337–347, 1916. DOI: 10.1007/BF02418550.

[Riv90] T. J. RIVLIN. *Chebyshev polynomials. From approximation theory to algebra and number theory*. Wiley-Interscience, second edition edition, 1990.

[Run01] C. RUNGE. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Z. Math. Phys.*, 46:224–243, 1901.

[Sal72] H. E. SALZER. Lagrangian interpolation at the Chebyshev points $x_{n,\nu} \equiv \cos(\nu\pi/n)$, $\nu = 0(1)n$; some unnoted advantages. *Comput. J.*, 15(2):156–159, 1972. DOI: 10.1093/comjnl/15.2.156.

[SB02] J. STOER and R. BULIRSCH. *Introduction to numerical analysis*, volume 12 of *Texts in applied mathematics*. Springer, third edition, 2002. DOI: 10.1007/978-0-387-21738-3.

[Sch46a] I. J. SCHOENBERG. Contributions to the problem of approximation of equidistant data by analytic functions. Part A. On the problem of smoothing or graduation. A first class of analytic approximation formulae. *Quart. Appl. Math.*, 4(1):45–99, 1946. DOI: 10.1090/qam/15914.

[Sch46b] I. J. SCHOENBERG. Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae. *Quart. Appl. Math.*, 4(2):112–141, 1946. DOI: 10.1090/qam/16705.

[Sch66] I. J. SCHOENBERG. On Hermite-Birkhoff interpolation. *J. Math. Anal. Appl.*, 16(3):538–543, 1966. DOI: 10.1016/0022-247X(66)90160-0.

[Sha79] A. SHAMIR. How to share a secret. *Comm. ACM*, 22(11):612–613, 1979. DOI: 10.1145/359168.359176.

[Tay45] W. J. TAYLOR. Method of lagrangian curvilinear interpolation. *J. Res. Nat. Bur. Standards*, 35(2):151–155, 1945. DOI: 10.6028/jres.035.006.

[Tch54] P. L. TCHEBYCHEF. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mém. Acad. Impér. Sci. St.-Pétersbourg (Savants Étrangers)*, 7 :539-568, 1854.

[Tch59] P. L. TCHEBYCHEF. Sur les questions de minima qui se rattachent à la représentation approximative des fonctions. *Mém. Acad. Impér. Sci. St.-Pétersbourg (6)*, 7 :199-291, 1859.

[Tre00] L. N. TREFETHEN. *Spectral methods in* MATLAB. SIAM, 2000. DOI: 10.1137/1.9780898719598.

[TW91] L. N. TREFETHEN and J. A. C. WEIDEMAN. Two results on polynomial interpolation in equally spaced points. *J. Approx. Theory*, 65(3):247–260, 1991. DOI: 10.1016/0021-9045(91)90090-W.

[vdCor35] J. G. van der CORPUT. Verteilungsfunktionen. *Proc. Konink. Nederl. Akad. Wetensch.*, 38:813–821, 1935.

[Vei60] L. VEIDINGER. On the numerical determination of the best approximations in the Chebyshev sense. *Numer. Math.*, 2(1):99–105, 1960. DOI: 10.1007/BF01386215.

[War79] E. WARING. Problems concerning interpolations. *Philos. Trans. Roy. Soc. London*, 69:59–67, 1779. DOI: 10.1098/rstl.1779.0008.

[Wei85] K. WEIERSTRASS. Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen. *Sitzungsber. Preuss. Akad. Wiss. Berlin*:633–639, 789–805, 1885.

Chapitre 7

Formules de quadrature

L'évaluation d'une intégrale définie [1] de la forme

$$I(f\,;a,b)=\int_a^b f(x)\,\mathrm{d}x,$$

où $[a,b]$ est un intervalle non vide et borné de $\mathbb{R}$ et f est une fonction d'une variable réelle, continue sur $[a,b]$, à valeurs réelles, est un problème classique intervenant dans de nombreux domaines, qu'ils soient scientifiques ou non. Nous nous intéressons dans le présent chapitre à l'utilisation de *formules de quadrature* [2] (*quadrature rules* en anglais) qui approchent la valeur de l'intégrale par une somme pondérée finie de valeurs de la fonction f, et, lorsqu'elle en possède, de ses dérivées, en des points choisis. Les formules que nous considérons pour ce faire sont essentiellement de la forme

$$I_n(f\,;a,b)=\sum_{i=0}^{n}\alpha_i\,f(x_i), \tag{7.1}$$

où n est un entier naturel, les coefficients $\{\alpha_i\}_{i=0,\ldots,n}$ sont réels et les points $\{x_i\}_{i=0,\ldots,n}$ appartiennent à l'intervalle $[a,b]$, mais d'autres types de formules sont également évoqués.

Les raisons conduisant à une évaluation seulement approchée d'une intégrale comme $I(f)$ sont variées. Tout d'abord, si l'on a, en vertu du théorème fondamental de l'analyse (voir le théorème B.131), que

$$I(f\,;a,b)=F(b)-F(a),$$

où la fonction F est une primitive de f, on ne sait pas toujours, même en ayant recours à des techniques plus ou moins sophistiquées telles que le changement de variable ou l'intégration par parties, exprimer F en termes de fonctions algébriques, trigonométriques, exponentielles ou logarithmiques. Lorsque c'est toutefois le cas, l'évaluation numérique d'une telle intégrale peut encore s'avérer difficile et coûteuse en pratique, tout en n'étant réalisée qu'avec une certaine exactitude en arithmétique en précision finie et donc, au final, approchée [3]. Par ailleurs, on a vu dans la section 1.4 que l'évaluation numérique d'une intégrale par une formule de récurrence pouvait fournir un résultat catastrophique si l'on ne prenait pas garde à d'éventuels problèmes de conditionnement, ce qui nuit à la généricité de l'implémentation de la formule de récurrence considérée et exige du praticien une certaine familiarité avec la notion de stabilité numérique d'un algorithme. Enfin, le recours à une évaluation approchée est obligatoire lorsque l'intégrande est solution d'une équation fonctionnelle (une équation différentielle par exemple) que l'on ne sait pas explicitement résoudre.

1. On sait qu'une telle intégrale existe en vertu de la proposition B.130.

2. L'usage du terme « quadrature » remonte à l'Antiquité et au problème, posé par l'école pythagoricienne, de la construction à la règle et au compas d'un carré ayant la même aire qu'une surface donnée.

3. En guise d'illustration, on peut reprendre l'exemple donné en introduction du livre de Davis et Rabinowitz [DR84],

$$\int_0^t \frac{\mathrm{d}x}{1+x^4}=\frac{1}{4\sqrt{2}}\ln\left(\frac{t^2+\sqrt{2}\,t+1}{t^2-\sqrt{2}\,t+1}\right)+\frac{1}{2\sqrt{2}}\left(\arctan\left(\frac{t}{\sqrt{2}-t}\right)+\arctan\left(\frac{t}{\sqrt{2}+t}\right)\right).$$

7.1 Formules de quadrature interpolatoires

Nous limitons dans ces pages l'exposé à la classe des *formules de quadrature interpolatoires* (*interpolatory quadrature rules* en anglais), en mettant particulièrement l'accent sur les *formules de Newton–Cotes*[4], qui sont un cas particulier de formules de quadrature basées sur l'interpolation de Lagrange introduite dans le précédent chapitre.

7.1.1 Généralités

On notera tout d'abord que les points x_i et les coefficients α_i, $i = 0, \ldots, n$, dans l'expression (7.1) sont respectivement appelés *nœuds* (*nodes* en anglais) et *poids* (*weights* en anglais) de la formule de quadrature. Les nœuds de quadrature appartiennent en général à l'intervalle d'intégration $[a, b]$, mais cela n'est pas toujours le cas. EXEMPLE ? ? ?

Comme pour les problèmes d'interpolation étudiés au chapitre précédent, la précision d'une formule de quadrature pour une fonction f continue sur l'intervalle $[a, b]$ donnée se mesure notamment en évaluant l'*erreur de quadrature*

$$E_n(f\,;a,b) = I(f\,;a,b) - I_n(f\,;a,b). \tag{7.2}$$

On définit par ailleurs le *degré (algébrique) d'exactitude*[5] (*(algebraic) degree of exactness* en anglais) d'une formule de quadrature $I_n(f)$ comme le plus grand entier naturel r pour lequel

$$\forall m \in \{0, \ldots, r\},\ \forall f \in \mathbb{P}_m,\ I(f\,;a,b) = I_n(f\,;a,b).$$

Ce degré ne dépend de l'intervalle d'intégration sur lequel la formule de quadrature est considérée.

Enfin, une formule de quadrature interpolatoire est obtenue en remplaçant la fonction f dans l'intégrale par son polynôme d'interpolation de Lagrange (voir la section 6.2 du précédent chapitre) ou de Hermite ou de Birkhoff (voir la sous-section 6.2.4). Dans le premier cas, on pose ainsi

$$I_n(f\,;a,b) = \int_a^b \Pi_n f(x)\,\mathrm{d}x, \tag{7.3}$$

où $\Pi_n f$ désigne le polynôme d'interpolation de Lagrange de f associé à un ensemble de nœuds $\{x_i\}_{i=0,\ldots,n}$ donné. En vertu de la définition 6.15, on a alors, par la propriété de linéarité de l'intégrale,

$$I_n(f\,;a,b) = \int_a^b \left(\sum_{i=0}^n f(x_i)\, l_i(x) \right) \mathrm{d}x = \sum_{i=0}^n f(x_i) \int_a^b l_i(x)\,\mathrm{d}x,$$

et, en identifiant avec (7.1), on trouve que les poids de quadrature sont simplement les intégrales respectives des polynômes de Lagrange $\{l_i\}_{i=0,\ldots,n}$ sur l'intervalle $[a, b]$, c'est-à-dire

$$\alpha_i = \int_a^b l_i(x)\,\mathrm{d}x,\ i = 0, \ldots, n. \tag{7.4}$$

Pour certaines familles de formules (voir par exemple la sous-section 7.1.3), il arrive que les valeurs des nœuds et poids de quadrature soient données relativement à l'intervalle d'intégration « standard » $[-1, 1]$. Dans ce cas, on utilise le changement de variable affine

$$t \mapsto \frac{1}{2}(b-a)t + \frac{1}{2}(a+b),$$

pour obtenir les valeurs des nœuds correspondant à un intervalle borné $[a, b]$ arbitraire, et l'on multiplie les valeurs des poids par $\frac{1}{2}(b-a)$ en vertu de la formule de changement de variable dans une intégrale définie.

4. Roger Cotes (10 juillet 1682 - 5 juin 1716) était un mathématicien anglais, premier titulaire de la chaire de professeur plumien d'astronomie et de philosophie expérimentale de l'université de Cambridge. Bien qu'il ne publia qu'un article de son vivant, il apporta d'importantes contributions en calcul intégral, en théorie des logarithmes et en analyse numérique.

5. On trouve parfois aussi le terme de *degré (algébrique) de précision* (*(algebraic) degree of precision* en anglais).

Plus généralement, la formule de quadrature prendra la forme (DECIDER DE LA NOTATION)

$$I_{ap}(f;a,b)=\sum_{i=0}^{n_0}\alpha_{0,i}\,f(x_{0,i})+\sum_{i=0}^{n_1}\alpha_{1,i}\,f'(x_{1,i})+\cdots+\sum_{i=0}^{n_m}\alpha_{m,i}\,f^{(m)}(x_{m,i}). \tag{7.5}$$

COMPLETER
On a la caractérisation suivante pour les formules de quadrature interpolatoires de la forme (7.1).

Théorème 7.1 *Soit n un entier naturel. Toute formule de quadrature utilisant $n+1$ nœuds distincts et donnée par* (7.1) *est interpolatoire si et seulement si son degré d'exactitude au moins égal à n.*

DÉMONSTRATION. Soit un intervalle non vide $[a,b]$ de $\mathbb{R}$. Pour toute formule de quadrature interpolatoire à $n+1$ nœuds distincts x_i, $i=0,\ldots,n$, on déduit de l'égalité (6.25) que, pour toute fonction polynomiale f de degré inférieur ou égal à n, l'erreur de quadrature $E_n(f;a,b)$ est nulle.

Réciproquement, si le degré d'exactitude de la formule de quadrature est au moins égal à n, les poids de quadrature α_i, $i=0,\ldots,n$, doivent vérifier les relations

$$\begin{aligned}\sum_{i=0}^{n}\alpha_i &= b-a,\\ \sum_{i=0}^{n}\alpha_i\,x_i &= \frac{1}{2}(b^2-a^2),\\ &\vdots\\ \sum_{i=0}^{n}\alpha_i\,{x_i}^n &= \frac{1}{n+1}(b^{n+1}-a^{n+1}),\end{aligned} \tag{7.6}$$

qui constituent un système linéaire de $n+1$ équations à $n+1$ inconnues admettant une unique solution (le déterminant qui lui est associé étant de Vandermonde et les nœuds x_i, $i=0,\ldots,n$, étant supposés distincts). On remarque alors que la formule de quadrature (7.3) est par définition exacte pour toute fonction polynomiale de degré inférieur ou égal à n, et plus particulièrement pour tout polynôme d'interpolation associé aux nœuds x_i, $i=0,\ldots,n$. Le choix (7.4) satisfait donc chacune des équations de (7.6). □

7.1.2 Formules de Newton–Cotes

Les formules de quadrature de Newton–Cotes (*Newton–Cotes formulas* ou *Newton–Cotes quadrature rules* en anglais) sont basées sur l'interpolation de Lagrange à nœuds *équirépartis* dans l'intervalle $[a,b]$; ce sont donc des cas particuliers de formules de quadrature interpolatoires de Lagrange. Pour n un entier positif fixé, notons $x_i=x_0+ih$, $i=0,\ldots,n$, les nœuds de quadrature. On peut définir deux types de formules de Newton–Cotes :

- les formules *fermées* (*closed formulae* en anglais), pour lesquelles les extrémités de l'intervalle $[a,b]$ font partie des nœuds, c'est-à-dire $x_0=a$, $x_n=b$ et $h=\frac{b-a}{n}$ $(n\geq 1)$, et dont les règles bien connues *du trapèze* $(n=1)$ et *de Simpson* $(n=2)$ sont des cas particuliers,
- les formules *ouvertes* (*open formulae* en anglais), pour lesquelles $x_0=a+h$, $x_n=b-h$ et $h=\frac{b-a}{n+2}$ $(n\geq 0)$, auxquelles appartient la *règle du point milieu* $(n=0)$.

Une propriété intéressante de ces formules est que leurs poids de quadrature ne dépendent explicitement que de n et h et non de l'intervalle d'intégration $[a,b]$; ceux-ci peuvent donc être calculés *a priori*. En effet, en introduisant, dans le cas des formules fermées, le changement de variable

$$x=x_0+th=a+th,\ \text{avec } t\in[0,n],$$

on vérifie que, pour tout $(i,j)\in\{0,\ldots,n\}^2$, $i\neq j$, $n\geq 1$,

$$\forall t\in[0,n],\ \frac{x-x_j}{x_i-x_j}=\frac{a+th-(a+jh)}{a+ih-(a+jh)}=\frac{t-j}{i-j},$$

et donc

$$\forall i\in\{0,\ldots,n\},\ l_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{t-j}{i-j},$$

d'où l'expression suivante pour les poids de quadrature

$$\forall i \in \{0,\dots,n\},\ \alpha_i = \int_a^b l_i(x)\,\mathrm{d}x = h\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t.$$

On obtient ainsi que

$$I_n(f;a,b) = h\sum_{i=0}^{n} w_i f(x_i),\ \text{avec } w_i = \int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t.$$

En procédant de manière analogue pour les formules ouvertes, on trouve que

$$I_n(f;a,b) = h\sum_{i=0}^{n} w_i f(x_i),\ \text{avec } w_i = \int_{-1}^{n+1} \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t,$$

en posant $x_{-1} = x_0 - h = a$ et $x_{n+1} = x_0 + (n+1)h = b$. Dans le cas particulier où $n = 0$, on a $w_0 = 2$ puisque $l_0(x) = 1$.

Ajoutons par ailleurs, en vertu d'une propriété de symétrie des polynômes de Lagrange, que les poids w_i et w_{n-i} sont égaux pour $i = 0,\dots,n$ pour les formules fermées ($n \geq 1$) et ouvertes ($n \geq 0$). Pour cette raison, on ne tabule les valeurs des poids que pour $0 \leq i \leq \lfloor \frac{n}{2} \rfloor$ (voir la table 7.1), où $\lfloor \frac{n}{2} \rfloor$ désigne la partie entière de $\frac{n}{2}$.

n	w_0	w_1	w_2	w_3	
1	$\frac{1}{2}$				(règle du trapèze)
2	$\frac{1}{3}$	$\frac{4}{3}$			(règle de Simpson)
3	$\frac{3}{8}$	$\frac{9}{8}$			(règle des trois huitièmes)
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$		(règle de Boole[6] [Boo60])
5	$\frac{95}{288}$	$\frac{375}{288}$	$\frac{125}{144}$		
6	$\frac{3}{10}$	$\frac{3}{2}$	$\frac{3}{10}$	$\frac{9}{5}$	(règle de Weddle[7] [Wed54])

n	w_0	w_1	w_2	
0	2			(règle du point milieu)
1	$\frac{3}{2}$			
2	$\frac{8}{3}$	$-\frac{4}{3}$		
3	$\frac{55}{24}$	$\frac{5}{24}$		
4	$\frac{33}{10}$	$-\frac{21}{5}$	$\frac{39}{5}$	

TABLE 7.1 – Poids des formules de Newton–Cotes fermées (à gauche) et ouvertes (à droite) pour quelques valeurs de l'entier n.

On remarque la présence de poids négatifs pour certaines formules ouvertes[8], ce qui peut conduire à des instabilités dues aux erreurs d'annulation (voir la sous-section 1.4.1) lors de l'évaluation numérique de la formule. La convergence de la suite des intégrales approchées par une formule de Newton-Cotes à $n+1$ nœuds n'est par ailleurs pas assurée lorsque n tend vers l'infini, même si l'intégrande est une fonction analytique sur l'intervalle d'intégration[9] (ce comportement est à relier à la divergence de l'interpolation de Lagrange avec nœuds équirépartis pour la fonction de Runge (6.31)). Pour ces deux raisons, l'utilisation des formules de Newton–Cotes (fermées ou ouvertes) utilisant plus de huit nœuds reste délicate et est généralement déconseillée en pratique. Ainsi, si l'on souhaite améliorer la précision de l'approximation d'une intégrale obtenue par une formule de quadrature de

6. George Boole (2 novembre 1815 - 8 décembre 1864) était un mathématicien et philosophe anglais. Il est le créateur de la logique moderne, fondée sur une structure algébrique et sémantique, dont les applications se sont avérées primordiales pour la théorie des probabilités, les systèmes informatiques ou encore les circuits électroniques et téléphoniques. Il a aussi travaillé dans d'autres domaines des mathématiques, publiant notamment des traités sur les équations différentielles et les différences finies.

7. Thomas Weddle (30 novembre 1817 - 4 décembre 1853) était un mathématicien anglais, connu pour ses travaux en analyse et en géométrie.

8. La formule fermée à neuf nœuds ainsi que toutes les formules fermées à plus de onze points présentent elles aussi au moins un poids négatif.

9. On peut néanmoins montrer qu'on a convergence, en l'absence d'erreur d'arrondi, si l'intégrande est une fonction analytique dans une région suffisamment grande du plan complexe contenant cet intervalle (voir [Dav55]).

Newton–Cotes donnée, on fera plutôt appel à des formules composées (voir la section 7.2) ou encore aux *formules de Gauss* (voir les notes de fin de chapitre).

Présentons maintenant de quelques exemples importants de formules de quadrature de Newton–Cotes.

Règle du point milieu (*midpoint rule* en anglais). Cette formule, aussi appelée *règle du rectangle* [10] (*rectangle rule* en anglais), est obtenue en remplaçant dans l'intégrale la fonction f par une fonction constante dont la valeur est celle prise par f au milieu de l'intervalle $[a,b]$ (voir la figure 7.1), d'où

$$I_0(f;a,b) = (b-a) f\left(\frac{a+b}{2}\right). \tag{7.9}$$

Le poids de quadrature vaut donc $\alpha_0 = b-a$ et le nœud est $x_0 = \dfrac{a+b}{2}$.

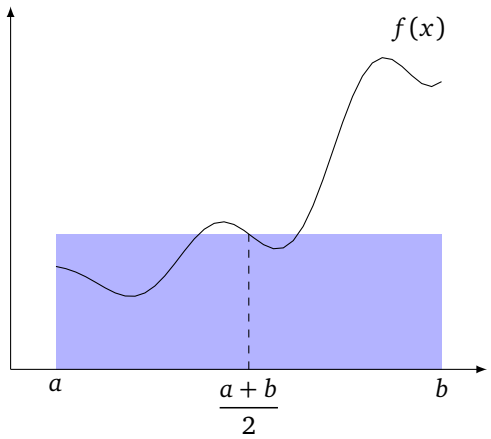


FIGURE 7.1 – Illustration de la règle du point milieu. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

En supposant la fonction f de classe $\mathscr{C}^2$ sur $[a,b]$, on peut utiliser le théorème 7.3 pour montrer que l'erreur de quadrature de cette formule vaut

$$E_0(f;a,b) = \frac{(b-a)^3}{24} f''(\eta), \text{ avec } \eta \in]a,b[.$$

Son degré d'exactitude est par conséquent égal à un.

Règle du trapèze (*trapezoidal rule* en anglais). On obtient cette formule en remplaçant dans l'intégrale la fonction f par son polynôme d'interpolation de Lagrange de degré un aux points a et b (voir la figure 7.2). Il vient alors

$$I_1(f;a,b) = \frac{b-a}{2}\left(f(a)+f(b)\right). \tag{7.10}$$

Les poids de quadrature valent $\alpha_0 = \alpha_1 = \dfrac{b-a}{2}$, tandis que les nœuds sont $x_0 = a$ et $x_1 = b$.

10. D'autres formules de quadrature interpolatoires sont basées sur un polynôme d'interpolation de Lagrange de degré nul (et donc un seul nœud de quadrature) : ce sont les règles du *rectangle à gauche*

$$I_0(f;a,b) = (b-a) f(a), \tag{7.7}$$

et *du rectangle à droite*

$$I_0(f;a,b) = (b-a) f(b), \tag{7.8}$$

dont le degré d'exactitude est égal à zéro. Elles ne font cependant pas partie des formules de Newton–Cotes.

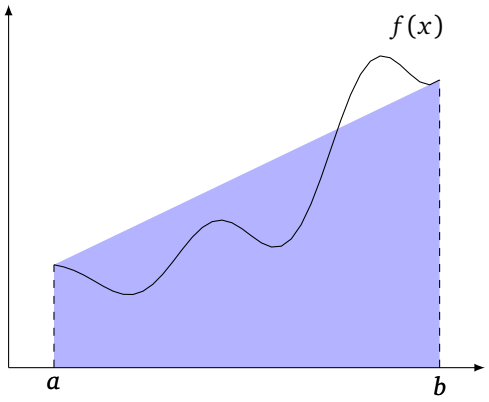


FIGURE 7.2 – Illustration de la règle du trapèze. La valeur approchée de l'intégrale $I(f\,;a,b)$ correspond à l'aire colorée en bleu.

En supposant f de classe $\mathscr{C}^2$ sur $[a,b]$, on obtient la valeur suivante pour l'erreur de quadrature

$$E_1(f\,;a,b) = -\frac{(b-a)^3}{12} f''(\xi), \text{ avec } \xi \in]a,b[.$$

et l'on en déduit que cette formule à un degré d'exactitude égal à un.

Règle de Simpson (*Simpson's rule* en anglais). Cette dernière formule est obtenue en substituant dans l'intégrale à la fonction f son polynôme d'interpolation de Lagrange de degré deux aux nœuds $x_0 = a$, $x_1 = \dfrac{a+b}{2}$ et $x_2 = b$ (voir la figure 7.3) et s'écrit

$$I_2(f\,;a,b) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{7.11}$$

Les poids de quadrature sont donnés par $\alpha_0 = \alpha_2 = \dfrac{b-a}{6}$ et $\alpha_1 = 2\,\dfrac{b-a}{3}$.

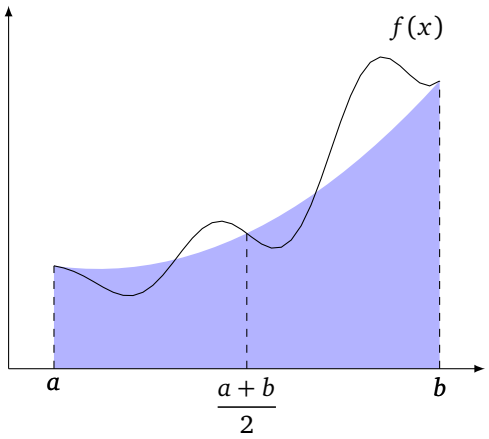


FIGURE 7.3 – Illustration de la règle de Simpson. La valeur approchée de l'intégrale $I(f\,;a,b)$ correspond à l'aire colorée en bleu.

On montre, grâce au théorème 7.3, que, si la fonction f est de classe $\mathscr{C}^4$ sur l'intervalle $[a,b]$, l'erreur de quadrature peut s'écrire

$$E_2(f\,;a,b) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \text{ avec } \xi \in]a,b[. \tag{7.12}$$

Cette formule a donc un degré d'exactitude égal à trois.

7.1.3 Formules de Fejér

Fejér [11] proposa en 1933 deux formules de quadrature interpolatoires, aujourd'hui nommées *règles de quadrature de Fejér* (*Fejér's quadrature rules* en anglais), utilisant des nœuds associés aux racines des polynômes de Chebyshev de première et de seconde espèces respectivement [Fej33]. Ces racines étant connues de manière explicite, il est possible d'obtenir des expressions, elles aussi explicites, des poids de quadrature de ces formules et de démontrer que ces derniers sont tous strictement positifs. On a plus précisément le résultat suivant.

Proposition 7.2 *Soit n un entier naturel. Les nœuds et poids de quadrature de la première règle de quadrature de Fejér à $n+1$ nœuds sur l'intervalle $[-1,1]$ sont donnés par*

$$x_i = \cos(\theta_i) \text{ et } \alpha_i = \frac{2}{n+1}\left(1 - 2\sum_{j=1}^{\lfloor \frac{n+1}{2} \rfloor} \frac{\cos(2j\,\theta_i)}{4j^2-1}\right), \text{ avec } \theta_i = \frac{2i+1}{2(n+1)}\pi,\ i = 0,\dots,n,$$

où $\lfloor \frac{n+1}{2} \rfloor$ désigne la partie entière de $\frac{n+1}{2}$. Pour la seconde règle de quadrature de Fejér, on a

$$x_i = \cos(\theta_i) \text{ et } \alpha_i = \frac{4\sin(\theta_i)}{n+2}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor+1} \frac{\sin((2j-1)\,\theta_i)}{2j-1}, \text{ avec } \theta_i = \frac{i+1}{n+2}\pi,\ i = 0,\dots,n.$$

Les poids de quadrature de ces deux formules sont strictement positifs.

DÉMONSTRATION. Les nœuds de quadrature de la premìere règle de Fejér sont les racines du polynôme de Chebyshev de première espèce de degré $n+1$, noté T_{n+1} et défini par

$$\forall x \in [-1,1],\ T_{n+1}(x) = \cos((n+1)\arccos(x)).$$

En remarquant que ce polynôme est égal, à une constante multiplicative près, au polynôme de Newton de degré $n+1$ associé aux nœuds de quadrature, on peut écrire, en se servant de l'identité (6.13), pour les polynômes de Lagrange associés à ces mêmes noeuds,

$$\forall i \in \{0,\dots,n\},\ \forall x \in \mathbb{R},\ l_i(x) = \frac{T_{n+1}(x)}{T_{n+1}{}'(x_i)(x-x_i)}.$$

Un calcul simple montrant que

$$\forall i \in \{0,\dots,n\},\ T_{n+1}{}'(x_i) = \frac{(-1)^i(n+1)}{\sin(\theta_i)},$$

on obtient, en utilisant l'identité

$$\forall i \in \{0,\dots,n\},\ 1 + 2\sum_{k=1}^{n+1} T_k(x_i)T_k(x) = -\frac{T_{n+1}(x)T_{n+2}(x_i)}{x-x_i}$$

découlant de la *formule de Christoffel* [12]*–Darboux* [13], que

$$\forall i \in \{0,\dots,n\},\ \alpha_i = \int_{-1}^{1} l_i(x)\,\mathrm{d}x = -\frac{(-1)^i\sin(\theta_i)}{(n+1)\,T_{n+2}(x_i)}\left(2 + 2\sum_{k=1}^{n+1} T_k(x_i)\int_{-1}^{1} T_k(x)\,\mathrm{d}x\right).$$

Pour le calcul des intégrales de polynômes de Chebyshev T_k, $k = 1,\dots,n+1$, on effectue un changement de variable, suivi de deux intégrations par parties successives, pour arriver à

$$\int_{-1}^{1} T_k(x)\,\mathrm{d}x = \int_0^{\pi} \cos(ky)\sin(y)\,\mathrm{d}y = \cos(k\pi) + 1 + k^2\int_0^{\pi} \cos(ky)\sin(y)\,\mathrm{d}y,$$

11. Lipót Fejér (9 février 1880 - 15 octobre 1959) était un mathématicien hongrois. Ses activités de recherche se concentrèrent sur l'analyse harmonique, et plus particulièrement les séries de Fourier, mais il publia aussi d'importants articles dans d'autres domaines des mathématiques, dont un, écrit en collaboration avec Carathéodory, sur les fonctions entières en 1907 ou un autre, issu d'un travail avec Riesz, sur les transformations conformes en 1922.

12. Elwin Bruno Christoffel (10 novembre 1829 - 15 mars 1900) était un mathématicien et physicien allemand. Il s'intéressa notamment à l'étude des transformations conformes, la théorie des invariants, la géométrie différentielle, l'analyse tensorielle, la théorie du potentiel, la physique mathématique, ainsi qu'aux polynômes orthogonaux, aux fractions continues et aux ondes de choc. Plusieurs résultats et objets mathématiques sont aujourd'hui associés à son nom.

13. Jean Gaston Darboux (14 août 1842 - 23 février 1917) était un mathématicien français. Ses travaux concernèrent l'analyse et la géométrie différentielle.

d'où

$$(1-k^2)\int_{-1}^{1} T_k(x)\,\mathrm{d}x = 1+(-1)^k.$$

On en déduit alors que

$$\int_{-1}^{1} T_k(x)\,\mathrm{d}x = \begin{cases} \dfrac{2}{1-k^2} & \text{si } k \text{ est pair,} \\ 0 & \text{si } k \text{ est impair,} \end{cases}$$

et donc, en remarquant que

$$\forall i \in \{0,\dots,n\},\ T_{n+2}(x_i) = \cos((n+2)\theta_i) = \cos\left(\theta_i + (2i+1)\frac{\pi}{2}\right) = (-1)^i \cos\left(\theta_i + \frac{\pi}{2}\right) = (-1)^{i+1}\sin(\theta_i),$$

que

$$\forall i \in \{0,\dots,n\},\ \alpha_i = \frac{2}{n+1}\left(1 + \sum_{j=1}^{\lfloor \frac{n+1}{2} \rfloor} \frac{2\,T_{2j}(x_i)}{1-(2j)^2}\right) = \frac{2}{n+1}\left(1 - 2\sum_{j=1}^{\lfloor \frac{n+1}{2} \rfloor} \frac{\cos(2j\,\theta_i)}{4j^2-1}\right).$$

Il reste à montrer que ces poids de quadrature sont strictement positifs. On a

$$\forall i \in \{0,\dots,n\},\ \sum_{j=1}^{\lfloor \frac{n+1}{2} \rfloor} \frac{\cos(2j\,\theta_i)}{4j^2-1} < \sum_{j=1}^{+\infty} \frac{1}{4j^2-1} = \lim_{k\to+\infty} \frac{1}{2}\sum_{j=1}^{k}\left(\frac{1}{2j-1} - \frac{1}{2j+1}\right) = \lim_{k\to+\infty} \frac{1}{2}\left(1 - \frac{1}{2k+1}\right) = \frac{1}{2},$$

ce qui établit clairement la propriété.

Pour la seconde règle, les nœuds sont les racines du polynôme de Chebyshev de seconde espèce de degré $n+1$, noté U_{n+1} et défini par

$$\forall \theta \in]0,\pi[,\ U_{n+1}(\cos(\theta)) = \frac{\sin((n+2)\theta)}{\sin(\theta)}.$$

En notant l_i, $i = 0,\dots,n$, les polynômes de Lagrange de degré n associés à ces nœuds de quadrature et en ayant laborieusement recours aux identités trigonométriques de linéarisation, on trouve que

$$\forall i \in \{0,\dots,n\},\ \alpha_i = \int_{-1}^{1} l_i(x)\,\mathrm{d}x = \int_0^{\pi} l_i(\cos(\theta))\sin(\theta)\,\mathrm{d}\theta = \int_0^{\pi}\left(\sum_{j=1}^{n+1} a_j \sin(j\theta)\right)\mathrm{d}\theta = 2\sum_{j=1}^{\lfloor \frac{n+2}{2} \rfloor} \frac{a_{2j-1}}{2j-1},$$

où

$$\forall j \in \{1,\dots,n+1\},\ a_j = \frac{2}{n+2}\sum_{k=1}^{n+1} l_i(\cos(\theta_k))\sin(\theta_k)\sin(j\,\theta_k) = \frac{2}{n+2}\,\sin(\theta_i)\sin(j\,\theta_i),$$

par propriété des polynômes de Lagrange. On en déduit facilement la formule donnée dans l'énoncé.

Pour prouver la stricte positivité des poids de cette règle, il suffit enfin de voir que

$$\begin{aligned}\forall i \in \{0,\dots,n\},\ 2\sin(\theta_i)\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor+1} \frac{\sin((2j-1)\,\theta_i)}{2j-1} &= \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor+1} \frac{\cos((2j-2)\,\theta_i) - \cos(2j\,\theta_i)}{2j-1} \\ &< 1 - \sum_{j=1}^{+\infty}\left(\frac{1}{2j-1} - \frac{1}{2j+1}\right) = 1 - \lim_{k\to+\infty}\left(1 - \frac{1}{2k+1}\right) = 0.\end{aligned}$$

□

On notera que les deux règles de Fejér sont des exemples de formule de quadrature ouverte : les extrémités de l'intervalle d'intégration (ici les points -1 et 1) ne sont en effet pas utilisées comme nœuds de quadrature. La seconde règle est aussi considérée comme plus pratique que la première, en raison de la disposition de ses nœuds [14].

14. Pour passer l'approximation obtenue par la formule à $n+1$ nœuds à celle donnée par la formule à $2n+1$ nœuds, on a seulement besoin d'effectuer n nouvelles évaluations de fonction pour la seconde règle de Fejér contre $2n+1$ pour la première règle.

7.1.4 Formules de Clenshaw–Curtis **

Les *formules de Clenshaw–Curtis* [CC60] constituent une version fermée de la seconde règle de Fejér. Les nœuds de quadrature de la formule sont les $n+1$ points d'extremum de T_n sur$[-1,1]$, y compris -1 et 1

formule explicite pour les poids

$$\alpha_i = \frac{c_i}{n}\left(1 - \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \frac{b_j \cos(2j\theta_i)}{4j^2-1}\right),$$

avec

$$b_j = \begin{cases} 1 & j = \frac{n}{2} \\ 2 & j < \frac{n}{2} \end{cases} \text{ et } c_i = \begin{cases} 1 & i = 0 \text{ ou } n \\ 2 & \text{sinon} \end{cases}$$

et $\alpha_i > 0$

calcul des poids par la formule explicite est coûteux ($O(n^2)$ flops) pour des valeurs élevées de l'entier n et pas stable numériquement. Gentleman [Gen72a, Gen72b] a néanmoins montré qu'on pouvait les obtenir au moyen d'une transformée en cosinus discrète, proche de la transformée de Fourier discrète. On peut alors tirer parti d'une *transformée de Fourier rapide* (*fast Fourier transform* en anglais), comme l'*algorithme de Cooley et Tukey* [CT65], pour le calcul de cette transformée de Fourier discrète, permettant de ramener très avantageusement le coût à $O(n\ln(n))$ opérations. voir l'algorithme dans [Wal06] ou la méthode présentée dans [GLR07]

A VOIR : discussion critique sur les mérites respectifs des formules de Gauss et de Clenshaw–Curtis, comparaison des vitesses de convergence des méthodes pour certains intégrands [Tre08])

VOIR AUSSI *formules de Basu* [Bas71]

7.1.5 Estimations d'erreur

Cette section est consacrée à l'établissement d'expressions permettant d'arriver à une estimation de l'erreur d'une formule de quadrature interpolatoire. L'accent est mis sur un résultat spécifique aux formules de Newton–Cotes, mais une technique de représentation *intégrale* de l'erreur, très générale, est également décrite.

Cas des formules de Newton–Cotes

Démontrons maintenant le théorème fournissant les estimations des erreurs de quadrature des règles du point milieu, du trapèze et de Simpson annoncées plus haut.

Théorème 7.3 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, n un entier positif et f une fonction de $\mathscr{C}^{n+2}([a,b])$ si n est pair, de $\mathscr{C}^{n+1}([a,b])$ si n est impair. Alors, l'erreur de quadrature pour les formules de Newton–Cotes fermées est donnée par*

$$E_n(f;a,b) = \begin{cases} \dfrac{K_n}{(n+2)!} f^{(n+2)}(\xi),\ K_n = \displaystyle\int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x < 0, & \text{si } n \text{ est pair,} \\ \dfrac{K_n}{(n+1)!} f^{(n+1)}(\xi),\ K_n = \displaystyle\int_a^b \omega_{n+1}(x)\,\mathrm{d}x < 0, & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \xi < b$, et par

$$E_n(f;a,b) = \begin{cases} \dfrac{K'_n}{(n+2)!} f^{(n+2)}(\eta),\ K'_n = \displaystyle\int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x > 0, & \text{si } n \text{ est pair,} \\ \dfrac{K'_n}{(n+1)!} f^{(n+1)}(\eta),\ K'_n = \displaystyle\int_a^b \omega_{n+1}(x)\,\mathrm{d}x > 0, & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \eta < b$, pour les formules de Newton–Cotes ouvertes.

DÉMONSTRATION. Traitons tout d'abord le cas d'une formule fermée avec n pair. En intégrant la relation (6.27) entre a et b et en posant

$$\Omega_{n+1}(x) = \int_a^x \omega_{n+1}(s)\,\mathrm{d}s,$$

on obtient, après une intégration par parties (légitime en vertu du lemme 6.18),

$$E_n(f;a,b)=\int_a^b \omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x=\Big[\Omega_{n+1}(x)[x_0,\dots,x_n,x]f\Big]_{x=a}^b-\int_a^b \Omega_{n+1}(x)\,\frac{\mathrm{d}}{\mathrm{d}x}[x_0,\dots,x_n,x]f\,\mathrm{d}x.$$

Il est clair que $\Omega_{n+1}(a)=0$; de plus, la fonction $\omega_{n+1}\left(\frac{a+b}{2}+\cdot\right)$ étant impaire [15], on a $\Omega_{n+1}(b)=0$, d'où

$$E_n(f;a,b)=-\int_a^b \Omega_{n+1}(x)[x_0,\dots,x_n,x,x]f\,\mathrm{d}x=-\int_a^b \Omega_{n+1}(x)\,\frac{f^{(n+2)}(\zeta(x))}{(n+2)!}\,\mathrm{d}x,$$

avec ζ une fonction continue à valeurs dans $]a,b[$, en utilisant respectivement (6.30) et le théorème 6.17. Enfin, comme [16] on a $\Omega_{n+1}(x)>0$ pour $a<x<b$, on déduit de la formule de la moyenne généralisée (voir le théorème B.134) que

$$E_n(f;a,b)=-\frac{f^{(n+2)}(\xi)}{(n+2)!}\int_a^b \Omega_{n+1}(x)\,\mathrm{d}x,$$

avec $a<\xi<b$, d'où, après une intégration par parties,

$$E_n(f;a,b)=\frac{f^{(n+2)}(\xi)}{(n+2)!}\int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x.$$

Dans le cas où l'entier n est impair, on note que la fonction ω_{n+1} ne change pas de signe sur l'intervalle $[b-h,b]$. On a alors, en vertu de la formule de la moyenne généralisée et de l'égalité (6.28),

$$\begin{aligned}E_n(f;a,b)&=\int_a^{b-h} \omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x+\int_{b-h}^b \omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x\\&=\int_a^{b-h} \omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x+\frac{f^{(n+1)}(\xi')}{(n+1)!}\int_{b-h}^b \omega_{n+1}(x)\,\mathrm{d}x,\end{aligned}$$

avec $a<\xi'<b$. Par les propriétés (6.16) et (6.18) des différences divisées, on peut alors écrire

$$\int_a^{b-h} \omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x=\int_a^{b-h} \omega_n(x)\,\big([x_0,\dots,x_{n-1},x]f-[x_0,\dots,x_n]f\big)\,\mathrm{d}x,$$

avec $\omega_{n+1}(x)=(x-x_n)\,\omega_n(x)$. En posant alors

$$\Omega_n(x)=\int_a^x \omega_n(s)\,\mathrm{d}s,$$

15. Ceci découle d'une propriété de la fontion polynomiale associée au $n+1^{\mathrm{e}}$ *polynôme factoriel*,

$$\pi_{n+1}(t)=t(t-1)\dots(t-n).$$

Par utilisation du changement de variable $x=x_0+th$, on a en effet

$$\omega_{n+1}(x)=h^{n+1}\,\pi_{n+1}(t).$$

Il est alors clair que les fonctions $\pi_{n+1}\left(\frac{n}{2}-\cdot\right)$ et $\pi_{n+1}\left(\frac{n}{2}+\cdot\right)$ sont polynomiales, de degré égal à $n+1$, et possèdent les mêmes $n+1$ racines $\frac{n}{2}$, $\frac{n}{2}-1,\dots,-\frac{n}{2}$; elles ne diffèrent donc que d'un facteur constant. En comparant les coefficients de leurs termes de plus haut degré, on trouve finalement

$$\pi_{n+1}\left(\frac{n}{2}+\tau\right)=(-1)^{n+1}\,\pi_{n+1}\left(\frac{n}{2}-\tau\right).$$

16. Pour établir ce résultat, il faut tout d'abord remarquer que le polynôme ω_{n+1} est de degré impair et a pour seules racines les points $a,x_1,\dots,x_{n-1},b$. On a par conséquent $\omega_{n+1}(s)>0$ pour $a<s<x_1$, d'où $\Omega_{n+1}(x)>0$ pour $a<x<x_1$. D'autre part, on a, pour $a<x+h<\frac{a+b}{2}$ avec $x\neq x_i$, $i=0,\dots,n$, et en utilisant le changement de variable $x=x_0+th$,

$$\left|\frac{\omega_{n+1}(x+h)}{\omega_{n+1}(x)}\right|=\left|\frac{\pi_{n+1}(t+1)}{\pi_{n+1}(t)}\right|=\left|\frac{(t+1)t\dots(t-n+1)}{t(t-1)\dots(t-n)}\right|=\left|\frac{t+1}{t-n}\right|=\frac{t+1}{(n+1)-(t+1)}\leq\frac{\frac{n}{2}}{(n+1)-\frac{n}{2}}=\frac{1}{1+\frac{2}{n}}<1,$$

d'où $|\omega_{n+1}(x+h)|<|\omega_{n+1}(x)|$. Ceci montre que la contribution négative de $\omega_{n+1}(s)$ à l'intégrale $\Omega_{n+1}(x)$ sur le sous-intervalle $[x_1,x_2]$ est de moindre importance que la contribution positive sur $[a,x_1]$, d'où $\Omega_{n+1}(x)>0$ pour $a<x<x_2$. Cet argument peut être répété de manière à couvrir l'ensemble de l'intervalle $\left]a,\frac{a+b}{2}\right[$ et l'antisymétrie de ω_{n+1} par rapport à $\frac{a+b}{2}$ suffit alors pour conclure.

on a, $n-1$ étant pair, $\Omega_n(a)=\Omega_n(b-h)=0$ et $\Omega_n(x)$ est strictement positive pour tout x dans $]a,b-h[$, et l'on trouve, après une intégration par parties et l'application de la formule de la moyenne généralisée,

$$\int_a^{b-h}\omega_{n+1}(x)[x_0,\dots,x_n,x]f\,\mathrm{d}x=\int_a^{b-h}\omega_n(x)[x_0,\dots,x_{n-1},x]f\,\mathrm{d}x=-\frac{f^{(n+1)}(\xi'')}{(n+1)!}\int_a^{b-h}\Omega_n(x)\,\mathrm{d}x,$$

avec $a<\xi''<b$. On a donc établi que

$$E_n(f;a,b)=-\frac{f^{(n+1)}(\xi'')}{(n+1)!}\int_a^{b-h}\Omega_n(x)\,\mathrm{d}x+\frac{f^{(n+1)}(\xi')}{(n+1)!}\int_{b-h}^{b}\omega_{n+1}(x)\,\mathrm{d}x.$$

Puisque le réel b est la plus grande racine de ω_{n+1} et que $\omega_{n+1}(x)>0$ pour $x>b$, on a $\omega_{n+1}(x)\le 0$ pour tout x dans $[b-h,b]$ et donc

$$\int_{b-h}^{b}\omega_{n+1}(x)\,\mathrm{d}x<0.$$

On sait par ailleurs que $\Omega_n(x)>0$ pour $a<x<b-h$, d'où

$$\int_a^{b-h}\Omega_n(x)\,\mathrm{d}x>0.$$

La fonction $f^{(n+1)}$ étant continue sur l'intervalle $[a,b]$, la formule de la moyenne discrète (voir le théorème B.135) implique alors l'existence d'un réel ξ strictement compris entre a et b tel que

$$E_n(f;a,b)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\left(-\int_a^{b-h}\Omega_n(x)\,\mathrm{d}x+\int_{b-h}^{b}\omega_{n+1}(x)\,\mathrm{d}x\right),$$

et l'on conclut en remarquant que

$$\int_{b-h}^{b}\omega_{n+1}(x)\,\mathrm{d}x=\Big[(b-x)\,\Omega_n(x)\Big]_{x=b-h}^{b}-\int_{b-h}^{b}\Omega_n(x)\,\mathrm{d}x=-\int_{b-h}^{b}\Omega_n(x)\,\mathrm{d}x.$$

Les résultats pour les formules ouvertes s'obtiennent exactement de la même façon que pour les formules fermées, en remplaçant les fonctions Ω_i, avec $i=n,n+1$, introduites dans les preuves par les fonctions

$$\widetilde{\Omega}_i(x)=\int_a^x\omega_i(s)\,\mathrm{d}s,\ i=n,n+1,$$

qui diffèrent des précédentes par le fait que l'on a maintenant $a<x_0$ et $x_n<b$. On montre néanmoins, par des arguments similaires à ceux précédemment invoqués, que $\widetilde{\Omega}_{n+1}(a)=\widetilde{\Omega}_{n+1}(b)=0$ et $\widetilde{\Omega}_{n+1}(x)<0$, $a<x<b$, pour tout entier n pair. □

Ce théorème montre que le degré d'exactitude d'une formule de Newton–Cotes à $n+1$ nœuds est égal à $n+1$ lorsque n est pair et n lorsque n est impair, que la formule soit fermée ou ouverte. Il est donc en général préférable d'employer une formule avec un nombre *impair* de nœuds.

Notons que l'on peut également chercher à faire apparaître la dépendance de l'erreur de quadrature par rapport au pas h et utilisant le changement de variable $x=x_0+th$. On obtient ainsi facilement le résultat suivant.

Corollaire 7.4 *Sous les hypothèses du théorème 7.3, on a les expressions suivantes pour les erreurs de quadrature respectives des formules de Newton–Cotes fermées et ouvertes*

$$E_n(f;a,b)=\begin{cases}\dfrac{M_n}{(n+2)!}\,h^{n+3}f^{(n+2)}(\xi),\ M_n=\displaystyle\int_0^n t\,\pi_{n+1}(t)\,\mathrm{d}t<0, & \textit{si } n \textit{ est pair,}\\[2ex] \dfrac{M_n}{(n+1)!}\,h^{n+2}f^{(n+1)}(\xi),\ M_n=\displaystyle\int_0^n \pi_{n+1}(t)\,\mathrm{d}t<0, & \textit{si } n \textit{ est impair,}\end{cases}$$

avec $a<\xi<b$,

$$E_n(f;a,b)=\begin{cases}\dfrac{M'_n}{(n+2)!}\,h^{n+3}f^{(n+2)}(\eta),\ M'_n=\displaystyle\int_{-1}^{n+1} t\,\pi_{n+1}(t)\,\mathrm{d}t>0, & \textit{si } n \textit{ est pair,}\\[2ex] \dfrac{M'_n}{(n+1)!}\,h^{n+2}f^{(n+1)}(\eta),\ M'_n=\displaystyle\int_{-1}^{n+1} \pi_{n+1}(t)\,\mathrm{d}t>0, & \textit{si } n \textit{ est impair,}\end{cases}$$

avec $a<\eta<b$.

Représentation intégrale de l'erreur de quadrature *

En vertu de la propriété de linéarité de l'intégrale (et de la dérivée), l'erreur de quadrature (7.2) peut être vue comme le résultat l'action d'un opérateur linéaire sur une fonction suffisamment régulière. Le résultat suivant, dû à Peano [17] [Pea13] et donné ici sous une forme générale s'adaptant à toutes les formules de quadrature interpolatoires précédemment décrites, utilise ce fait pour fournir une représentation élégante de l'erreur.

Théorème 7.5 ***(représentation intégrale de l'erreur de quadrature)*** *Soit n un entier naturel tel que $E(f) = 0$ pour tout f de $\mathbb{P}_n$. Alors, pour tout f de classe $\mathscr{C}^{n+1}([a,b])$, on a*

$$E(f;a,b) = \int_a^b f^{(n+1)}(t)K(t)\,\mathrm{d}t,$$

avec

$$K(t) = \frac{1}{n!}E(x \mapsto [(x-t)_+]^n),$$

où $(\cdot)_+$ désigne la partie positive. La fonction K est appelée le ***noyau de Peano*** *(**Peano kernel** en anglais) de l'opérateur d'erreur de la formule de quadrature.*

DÉMONSTRATION. A REPRENDRE Le reste du développement de Taylor avec reste intégral de $f(x)$ au point a est

$$\frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\,\mathrm{d}t = \frac{1}{n!}\int_a^b [(x-t)_+]^n f^{(n+1)}(t)\,\mathrm{d}t.$$

L'application de l'opérateur d'erreur de quadrature au développement conduit alors à

$$E(f;a,b) = \frac{1}{n!}E\left(x \mapsto \int_a^b [(x-t)_+]^n f^{(n+1)}(t)\,\mathrm{d}t\right),$$

la formule de quadrature considérée étant de degré d'exactitude égal à n. Pour arriver à la représentation donnée dans l'énoncé à partir de cette dernière identité, il faut pouvoir échanger l'intégrale avec E. □

On peut montrer que les noyaux de Peano associés à certaines classes de formules de quadrature sont de signe constant sur l'intervalle $[a,b]$. C'est notamment le cas pour les formules de Newton–Cotes (voir [Ste27]). Par utilisation de la formule de la moyenne généralisée (voir le théorème B.134), il vient alors

$$E(f;a,b) = f^{(n+1)}(\xi)\int_a^b K(t)\,\mathrm{d}t \text{ avec } \xi \in]a,b[.$$

On obtient par cette technique une autre preuve des résultats du théorème 7.3.

Exemple d'application de la représentation intégrale de l'erreur de quadrature. La règle de Simpson a un degré d'exactitude égal à trois (a montrer), l'application du théorème conduit par conséquent à

$$K(t) = \frac{1}{3!}\left(\int_a^b (x-t)_+^3\,\mathrm{d}x - \frac{1}{3}(a-t)_+^3 - \frac{4}{3}\left(\frac{a+b}{2}-t\right)_+^3 - \frac{1}{3}(b-t)_+^3\right).$$

Par définition de la partie positive, il vient

$$\int_a^b (x-t)_+^3\,\mathrm{d}x = \int_t^b (x-t)_+^3\,\mathrm{d}x = \frac{(b-t)^4}{4}$$

et

$$(a-t)_+^3 = 0,\ \left(\frac{a+b}{2}-t\right)_+^3 = \begin{cases} 0 & \text{si } t \geq \frac{a+b}{2} \\ \left(\frac{a+b}{2}-t\right)^3 & \text{si } t < \frac{a+b}{2} \end{cases},\ (b-t)_+^3 = (b-t)^3.$$

retrouver (7.12)

17. Giuseppe Peano (27 août 1858 - 20 avril 1932) était un mathématicien italien. Analyste puis logicien, il s'est principalement intéressé aux fondements des mathématiques, ainsi qu'à la théorie des langages.

7.2 Formules de quadrature composées

Les formules de quadrature introduites jusqu'à présent ont toutes été obtenues en substituant à l'intégrande son polynôme d'interpolation de Lagrange à nœuds équirépartis sur l'intervalle d'intégration, la valeur de l'intégrale considérée étant alors approchée par la valeur de l'intégrale du polynôme. Pour améliorer la précision de cette approximation, on est donc tenté d'augmenter le degré de l'interpolation polynomiale utilisée. Le phénomène de Runge (voir la section 6.2.3 du précédent chapitre) montre cependant qu'un polynôme d'interpolation de degré élevé peut, lorsque les nœuds d'interpolation équirépartis, fournir une approximation catastrophique d'une fonction pourtant très régulière, ce qui a des conséquences désastreuses lorsque l'on cherche, par exemple, à approcher l'intégrale

$$\int_{-5}^{5} \frac{\mathrm{d}x}{1+x^2} = 2\arctan(5) = 2,74680153389\ldots \tag{7.13}$$

par une formule de Newton–Cotes (voir la table 7.2). Il a d'ailleurs été démontré par Pólya [18] [Pól33] que les formules de Newton–Cotes ne convergent généralement pas lorsque l'entier n tend vers l'infini, même lorsque la fonction à intégrer est analytique. Ceci, allié à l'observation que les poids de quadrature n'ont pas tous le même signe à partir de $n = 2$ pour les formules ouvertes et $n = 8$ pour les formules fermées, ce qui qui pose des problèmes de stabilité numérique, conduit les praticiens à ne pas, ou peu, utiliser les formules de quadrature de Newton–Cotes dont le nombre de nœuds est supérieur ou égal à huit. Il existe d'autres formules de quadrature interpolatoires, comme les formules de Gauss, aux nœuds non équirépartis, qui ne sont pas sujettes à ce problème de divergence, mais dont le calcul des nœuds et poids de quadrature lorsque le nombre de nœuds devient important peut s'avérer coûteux (cette affirmation étant néanmoins à tempérer, voir la section 7.4).

n	$I_n(f;-5,5)$
1	5,19231
2	6,79487
3	2,08145
4	2,374
5	2,30769
6	3,87045
7	2,89899
8	1,50049
9	2,39862
10	4,6733

TABLE 7.2 – Valeur de $I_n(f;-5,5)$, obtenue par une formule de quadrature de Newton–Cotes fermée, en fonction de n pour l'approximation de l'intégrale (7.13) de la fonction de Runge $f(x) = \frac{1}{1+x^2}$. On n'observe *a priori* pas de convergence de la valeur approchée vers la valeur exacte lorsque n augmente.

Il est cependant possible construire très simplement des formules de quadrature dont la mise en œuvre est aisée et dont la précision pourra être aussi grande que souhaitée : ce sont les *formules de quadrature composées* (*composite quadrature rules* ou *compound quadrature rules* en anglais).

7.2.1 Principe

Les formules de quadrature interpolatoires composées utilisent la technique de l'interpolation polynomiale par morceaux introduite dans la section 6.3, qui consiste une interpolation polynomiale à nœuds équirépartis de bas degré sur des sous-intervalles obtenus en partitionnant l'intervalle d'intégration. On peut construire de nombreuses classes de formules de quadrature interpolatoires composées, mais nous ne présentons ici que les plus courantes, en lien avec les formules de Newton–Cotes qui viennent d'être étudiées.

18. George Pólya (Pólya György en hongrois, 13 décembre 1887 - 7 septembre 1985) était un mathématicien américain d'origine austro-hongroise. Il fit d'importantes contributions à la combinatoire, la théorie des nombres et la théorie des probabilités. Il rédigea également plusieurs ouvrages sur l'heuristique et la pédagogie des mathématiques.

Étant donné un entier m supérieur ou égal a 1, on pose

$$H = \frac{b-a}{m}$$

et l'on introduit une partition de l'intervalle $[a,b]$ en m sous-intervalles $[x_{j-1}, x_j]$, $j = 1,\dots,m$, de longueur H, avec $x_i = a + iH$, $i = 0,\dots,m$. Comme

$$I(f;a,b) = \int_a^b f(x)\,\mathrm{d}x = \sum_{j=1}^{m} \int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x,$$

il suffit d'approcher chacune des intégrales apparaissant dans le membre de droite de l'égalité ci-dessus en utilisant une formule de quadrature interpolatoire, généralement la même sur chaque sous-intervalle, pour obtenir une formule composée, conduisant à une approximation de $I(f;a,b)$ de la forme

$$I_{m,n}(f;a,b) = \sum_{j=1}^{m} I_n(f;x_{j-1},x_j) = \sum_{j=1}^{m}\sum_{i=0}^{n} \alpha_{i,j} f(x_{i,j}), \tag{7.14}$$

où les coefficients $\alpha_{i,j}$ et les points $x_{i,j}$, $i = 0,\dots,n$, désignent respectivement les poids et les nœuds de la formule de quadrature interpolatoire utilisée sur le j^{e} sous-intervalle, $j = 1,\dots,m$, de la partition de $[a,b]$. Dans les *formules de Newton–Cotes composées*, la formule de quadrature utilisée sur chaque sous-intervalle est une même formule de Newton–Cotes, fermée ou ouverte, à $n+1$ nœuds équirépartis, $n \geq 0$, et l'on a par conséquent

$$x_{i,j} = x_{j-1} + i\,h,\ i = 0,\dots,n,\ j = 1,\dots,m,$$

avec $h = \frac{H}{n}$, $n \geq 1$, pour une formule fermée, et $h = \frac{H}{n+2}$, $n \geq 0$, pour une formule ouverte, les poids de quadrature $\alpha_{i,j} = h\,w_i$ étant indépendants de j.

En notant que l'erreur de quadrature d'une formule composée, notée $E_{m,n}(f;a,b)$, se décompose de la manière suivante

$$E_{m,n}(f;a,b) = I(f;a,b) - I_{m,n}(f;a,b) = \sum_{j=1}^{m}\left(\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x - \sum_{i=0}^{n} \alpha_{i,j} f(x_{i,j})\right),$$

on obtient sans mal, grâce à l'analyse d'erreur réalisée dans la précédente section, le résultat suivant pour les formules de Newton–Cotes composées.

Théorème 7.6 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, n un entier positif et f une fonction de $\mathscr{C}^{n+2}([a,b])$ si n est pair, de $\mathscr{C}^{n+1}([a,b])$ si n est impair. Alors, en conservant les notations du corollaire 7.4, l'erreur de quadrature pour les formules de Newton–Cotes composées vaut*

$$E_{m,n}(f;a,b) = \begin{cases} \dfrac{M_n}{(n+2)!}\dfrac{b-a}{n^{n+3}} H^{n+2} f^{(n+2)}(\xi) & \text{si } n \text{ est pair,} \\ \dfrac{M_n}{(n+1)!}\dfrac{b-a}{n^{n+2}} H^{n+1} f^{(n+1)}(\xi) & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \xi < b$, pour les formules fermées et

$$E_{m,n}(f;a,b) = \begin{cases} \dfrac{M'_n}{(n+2)!}\dfrac{b-a}{(n+2)^{n+3}} H^{n+2} f^{(n+2)}(\eta) & \text{si } n \text{ est pair,} \\ \dfrac{M'_n}{(n+1)!}\dfrac{b-a}{(n+2)^{n+2}} H^{n+1} f^{(n+1)}(\eta) & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \eta < b$.

DÉMONSTRATION. Considérons le cas d'une formule de quadrature composée de Newton–Cotes fermée. Puisque la même formule de quadrature est utilisée sur chacun des sous-intervalles $[x_{j-1}, x_j]$, $j = 1,\dots,m$, il découle du corollaire 7.4, en remarquant que la constante M_n ne dépend pas de l'intervalle d'intégration, que

$$E_{m,n}(f;a,b) = \begin{cases} \dfrac{M_n}{(n+2)!} h^{n+3} \sum_{j=1}^{m} f^{(n+2)}(\xi_j) & \text{si } n \text{ est pair,} \\ \dfrac{M_n}{(n+1)!} h^{n+2} \sum_{j=1}^{m} f^{(n+1)}(\xi_j) & \text{si } n \text{ est impair,} \end{cases}$$

avec $x_{j-1} < \xi_j < x_j$, $j = 1, \ldots, m$. Par définition de H, il vient alors

$$E_{m,n}(f;a,b) = \begin{cases} \dfrac{M_n}{(n+2)!}\dfrac{b-a}{mn^{n+3}} H^{n+2} \sum_{j=1}^{m} f^{(n+2)}(\xi_j) & \text{si } n \text{ est pair,} \\ \dfrac{M_n}{(n+1)!}\dfrac{b-a}{mn^{n+2}} H^{n+1} \sum_{j=1}^{m} f^{(n+1)}(\xi_j) & \text{si } n \text{ est impair,} \end{cases}$$

dont se déduit l'estimation annoncée en appliquant le théorème de la moyenne discrète. L'estimation pour les formules ouvertes s'obtient de manière analogue. □

On déduit de ce théorème que, à n fixé, l'erreur de quadrature d'une formule de Newton–Cotes composée tend vers 0 lorsque m tend vers l'infini, c'est-à-dire lorsque H tend vers 0, ce qui assure la convergence de la valeur approchée de l'intégrale vers sa valeur exacte (voir la table 7.3). De plus, le degré d'exactitude d'une formule composée coïncide avec celui de la formule dont elle dérive. En pratique, on utilise généralement des formules de Newton–Cotes composées basées sur des formules à peu de nœuds ($n \leq 2$), ce qui garantit que tous les poids de quadrature sont positifs. À cet égard, les formules dans l'exemple qui suit sont très couramment employées.

m	$E_{m,1}(f;-5,5)$	$E_{m,2}(f;-5,5)$
1	2,36219	−4,04807
2	−2,44551	0,09649
4	−0,53901	0,12942
8	−0,03769	0,01348
16	0,00069	$9{,}08169\,10^{-5}$
32	0,00024	$4{,}54992\,10^{-8}$
64	$6{,}0182\ \ 10^{-5}$	$2{,}60675\,10^{-9}$
128	$1{,}50475\,10^{-5}$	$1{,}63011\,10^{-10}$
256	$3{,}76199\,10^{-6}$	$1{,}01887\,10^{-11}$

TABLE 7.3 – Valeurs des erreurs de quadrature des règles de quadrature du trapèze et de Simpson composées en fonction du nombre de sous-intervalles m pour l'approximation de l'intégrale (7.13) de la fonction de Runge $f(x) = \frac{1}{1+x^2}$. On constate que l'erreur de quadrature $E_{m,n}(f;-5,5)$, $n = 1, 2$, tend vers zéro lorsque m augmente.

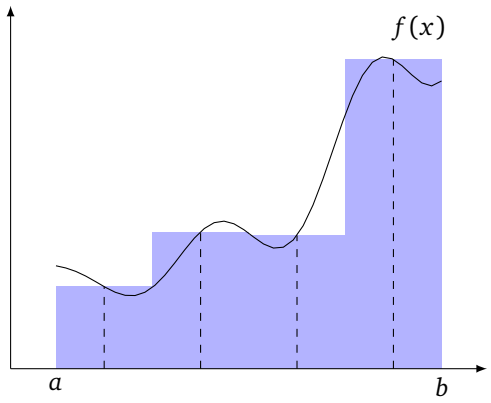


FIGURE 7.4 – Illustration de la règle du point milieu composée à quatre sous-intervalles sur l'intervalle $[a, b]$. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

Exemples de formules de Newton–Cotes composées. On présente, avec leur erreur de quadrature (obtenue via le théorème 7.6) ci-dessous trois formules Newton–Cotes composées parmi les plus utilisées. La *règle du point milieu composée* (voir la figure 7.4) fait partie des formules ouvertes et ne possède qu'un nœud de quadrature dans chaque sous-intervalle de la

partition de l'intervalle $[a,b]$. Dans ce cas, on a $h = \frac{H}{2}$ et

$$I(f;a,b) = H\sum_{j=1}^{m} f\left(a+\left(j-\frac{1}{2}\right)H\right) + \frac{b-a}{24}H^2 f''(\eta),$$

avec $\eta \in]a,b[$.

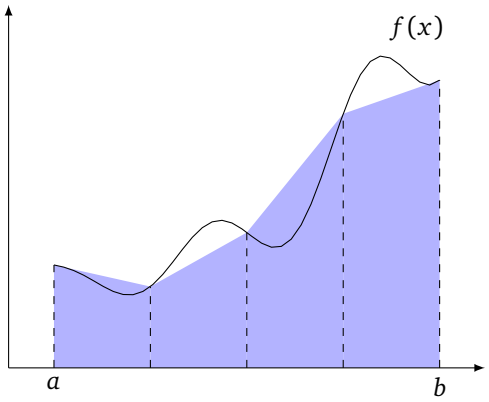


FIGURE 7.5 – Illustration de la règle du trapèze composée à quatre sous-intervalles sur l'intervalle $[a,b]$. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

La *règle du trapèze composée* (voir la figure 7.5) est une formule fermée qui a pour nœuds de quadrature les extrémités de chaque sous-intervalle. On a alors $h = H$ et

$$I(f;a,b) = \frac{H}{2}\left(f(a)+2\sum_{j=1}^{m-1} f(a+jH)+f(b)\right) - \frac{b-a}{12}H^2 f''(\xi),$$

avec $\xi \in]a,b[$. Enfin, la *règle de Simpson composée* (voir la figure 7.6) utilise comme nœuds de quadrature les extrémités et le milieu de chaque sous-intervalle, d'où $h = \frac{H}{2}$ et

$$I(f;a,b) = \frac{H}{6}\left(f(a)+2\sum_{j=1}^{m-1} f(a+jH)+4\sum_{j=1}^{m} f\left(a+\left(j-\frac{1}{2}\right)H\right)+f(b)\right) - \frac{b-a}{2880}H^4 f^{(4)}(\xi),$$

avec, là encore, $\xi \in]a,b[$.

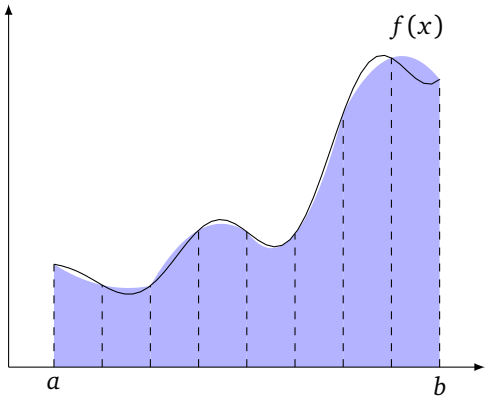


FIGURE 7.6 – Illustration de la règle de Simpson composée à quatre sous-intervalles sur l'intervalle $[a,b]$. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

On peut établir la convergence d'une formule de quadrature composée sous des hypothèses bien moins restrictives que celles du théorème 7.6. C'est l'objet du résultat suivant.

Théorème 7.7 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, f une fonction continue sur $[a,b]$, $\{x_j\}_{j=0,\dots,m}$ l'ensemble des nœuds d'une partition de $[a,b]$ en m sous-intervalles et une formule de quadrature composée de la forme* (7.14) *relativement à cette partition, de degré exactitude égal à r et dont les poids de quadrature $\alpha_{i,j}$, $i=0,\dots,n$, $j=1,\dots,m$, sont positifs. Alors, on a*

$$\lim_{m\to+\infty} I_{m,n}(f;a,b) = I(f;a,b).$$

DÉMONSTRATION. Lorsque m tend vers l'infini, les mesures des sous-intervalles de la partition de $[a,b]$ deviennent arbitrairement petites et, pour tout $\varepsilon>0$, on peut donc trouver un entier M tel que, si $m\geq M$, il existe des polynômes p_j, $j=1,\dots,m$, de degré inférieur ou égal à r tels que

$$\max_{x_{j-1}\leq x\leq x_j} \left|f(x)-p_j(x)\right| \leq \varepsilon,\ j=1,\dots,m.$$

Il vient alors, en utilisant que le degré d'exactitude de la formule de quadrature composée est r,

$$\left|\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x - \int_{x_{j-1}}^{x_j} p_j(x)\,\mathrm{d}x\right| + \left|\int_{x_{j-1}}^{x_j} p_j(x)\,\mathrm{d}x - \sum_{i=0}^{n}\alpha_{i,j}f(x_{i,j})\right|$$
$$\leq \varepsilon\left|x_j-x_{j-1}\right| + \left|\sum_{i=0}^{n}\alpha_{i,j}\left(p_j(x_{i,j})-f(x_{i,j})\right)\right| \leq \varepsilon\left|x_j-x_{j-1}\right| + \varepsilon\sum_{i=0}^{n}\left|\alpha_{i,j}\right|.$$

Par ailleurs, la formule de quadrature étant exacte pour une fonction constante et les poids de quadrature étant tous positifs, il vient

$$\sum_{i=0}^{n}\left|\alpha_{i,j}\right| = \left|x_j-x_{j-1}\right|.$$

On a par conséquent

$$\left|E_{m,n}(f;a,b)\right| \leq 2\varepsilon\,|b-a|,$$

et le résultat est démontré. □

Notons qu'en prenant $p_j(x) = f\left(\frac{x_{j-1}+x_j}{2}\right)$, $\forall x\in[x_{j-1},x_j]$, $j=1,\dots,m$, dans la preuve ci-dessus, on obtient le corollaire suivant.

Corollaire 7.8 *Sous les hypothèses du précédent théorème, on a de plus*

$$\left|I(f;a,b)-I_{m,n}(f;a,b)\right| \leq 2(b-a)\,\omega\left(f,\frac{H}{2}\right),$$

où, pour tout réel δ strictement positif,

$$\omega(f,\delta) = \sup\left\{|f(x)-f(y)| \mid (x,y)\in[a,b]^2,\ x\neq y,\ |x-y|\leq\delta\right\}$$

est le module de continuité de la fonction f.

7.2.2 Formules adaptatives **

REPRENDRE On a vu que, pour une fonction f continue, il est possible d'atteindre une précision arbitraire pour le calcul de l'intégrale $I(f)$ avec un formule de quadrature composée, simplement en prenant la longueur H suffisamment petit. Lorsque la fonction présente par endroits de brusques variations ou des oscillations, et que ces régions représentent une petite portion de l'intervalle d'intégration $[a,b]$, une réduction *uniforme* du pas H sur l'intervalle entier conduira à un coût élevé, et en partie inutile, pour la méthode. On peut donc souhaiter concentrer les évaluations de fonction là où cela est réellement nécessaire, en ajustant localement le pas et donnant ainsi lieu à une *formule de quadrature composée adaptative* (*adaptive composite quadrature rule* en anglais). Les programmes de calcul numérique d'intégrales les plus courants/robustes font appel à ce type de technique.

Premières tentatives : règle de Simpson adaptative [Kun62, McK62] (voir aussi [Lyn69])

Pour en apprendre plus sur ces techniques, qui ne sont d'ailleurs pas imparables, voir [GG00].

7.3 Évaluation d'intégrales sur un intervalle borné de fonctions particulières **

7.3.1 Fonctions périodiques **

Commençons par une expérience numérique dans laquelle la règle du trapèze composée est utilisée pour l'évaluation de l'intégrale de la fonction ... entre ... et ...

PLUSIEURS EXEMPLES DE CONVERGENCE pour des fonctions périodiques

parler de la convergence géométrique/exponentielle de la règle du trapèze composée pour une fonction analytique intégrée sur une période ou la droite réelle

Ces derniers exemples mettent en évidence un phénomène de convergence extrêmement rapide – on parle parfois de *superconvergence* – de la méthode, de loin meilleure que ce que laissent présager les résultats du théorème 7.6 (FAIRE LA VERIFICATION).

introduction abordable au phénomène dans [Wei02]

Théorème 7.9 (« formule [19] d'Euler–Maclaurin [20] ») *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, p un entier naturel et f une fonction de classe $\mathscr{C}^{p+1}$ sur $[a,b]$. Pour tout entier naturel m non nul, on a*

$$\int_a^b f(x)\,\mathrm{d}x = H\left(\frac{1}{2}f(x_0)+f(x_1)+\cdots+f(x_{m-1})+\frac{1}{2}f(x_m)\right)+\sum_{i=1}^{\lfloor p/2\rfloor} H^{2i}\frac{b_{2i}}{(2i)!}\int_a^b f^{(2i)}(x)\,\mathrm{d}x + \frac{H^{p+1}}{(p+1)!}\int_a^b \tilde{B}_{p+1}\left(\frac{x-a}{H}\right)f^{(p+1)}(x)\,\mathrm{d}x, \quad (7.15)$$

*où $H=\frac{b-a}{m}$, $x_j=a+jH$, $j=0,\dots,m$, $\tilde{B}_{p+1}$ désigne une extension périodique du **polynôme de Bernoulli** B_{p+1}, définie par*

$$\forall t\in\mathbb{R},\ \tilde{B}_j(t)=B_j(t-\lfloor t\rfloor)=\begin{cases}B_j(t) & 0\le t<1\\ \tilde{B}_j(t-1) & \textit{sinon}\end{cases}$$

*et les nombres (rationnels) $b_{2i}=B_{2i}(0)$, sont les **nombres de Bernoulli**,*

On a tout d'abord que

$$\int_a^b f(x)\,\mathrm{d}x = H\int_0^m f(a+Ht)\,\mathrm{d}t = H\sum_{j=0}^{m-1}\int_j^{j+1} f(a+Ht)\,\mathrm{d}t.$$

Pour un entier j de $\{0,\dots,m-1\}$, considérons à présent l'intégrale $\int_j^{j+1} f(a+Ht)\,\mathrm{d}t$. En utilisant le changement de variable $s=t-\lfloor t\rfloor$ sur l'intervalle $[j,j+1]$, il vient que $t=j+s$ et l'on peut alors poser $g(s)=f(a+(j+s)H)$, d'où

$$\int_j^{j+1} f(a+Ht)\,\mathrm{d}t = \int_0^1 g(s)\,\mathrm{d}s.$$

On va maintenant établir la formule pour cette dernière intégrale. 3 - on n'a plus qu'à démontrer la formule sur $[0,1]$, où $\tilde{B}_j=B_j$ 4 - relation de récurrence pour les polynômes de Bernoulli, conditions en 0 et 1 pour la détermination et conséquence sur le nombres, relation de symétrie 4 - IPP+récurrence, simplification, on trouve la formule, retour sur $[a,b)$ par sommation

à utiliser pour la convergence de la règle du trapèze

approche suivie par Poisson [21] dans S.-D. Poisson, Sur le calcul numérique des intégrales définies, Mémoires de l'Académie Royale des Sciences de l'Institut de France, 4 (1827), pp. 571–602.

application l'approximation d'une série, question de la divergence (lien avec travail de Poisson)

Article détaillé sur la règle du trapèze pour l'intégration de fonctions périodiques [TW14]

19. Cette formule fut découverte de manière indépendante par Leonhard Euler et Colin Maclaurin vers 1735. Euler l'utilisa pour le calcul de limites de séries lentement convergentes tandis que Maclaurin s'en servit pour l'évaluation approchée d'intégrales.

20. Colin Maclaurin (février 1698 - 14 juin 1746) était un mathématicien écossais. Il fit des travaux remarquables en géométrie, plus précisément dans l'étude de courbes planes, et écrivit un important mémoire sur la théorie des marées.

21. Siméon Denis Poisson (21 juin 1781 - 25 avril 1840) était un mathématicien et physicien français. Il est l'auteur de nombreux travaux, notamment sur les intégrales définies, les séries de Fourier et les probabilités en mathématiques, sur la mécanique céleste, la théorie de l'électricité et du magnétisme ainsi que celle de l'élasticité en physique.

7.3.2 Fonctions rapidement oscillantes **

On dira qu'un intégrand est *rapidement oscillant* si celui-ci présente de nombreux (typiquement plus de dix) maxima et minima locaux sur l'intervalle d'intégration. De telles fonctions interviennent par exemple lors du calcul des coefficients d'une série de Fourier, avec l'évaluation d'intégrales réelles comme

$$\int_a^b f(x)\cos(kx)\,\mathrm{d}x \text{ ou } \int_a^b f(x)\sin(kx)\,\mathrm{d}x,\ k\in\mathbb{N}.$$

difficultés particulières...
parler de la *méthode de Filon* [Fil30], en partie basée sur une formule de quadrature composée

7.4 Notes sur le chapitre

Le lecteur intéressé trouvera de nombreux détails complémentaires sur la théorie et les aspects pratiques de l'intégration numérique dans les ouvrages de référence de Davis et Rabinowitz [DR84] et de Brass et Petras [BP11].

Une question venant naturellement à l'esprit concernant les formules de quadrature est de savoir quel(s) choix judicieux de nœuds et de poids permet(tent) d'atteindre le degré d'exactitude maximal possible avec une formule à $n+1$ nœuds distincts. Une conséquence du théorème 7.1 est qu'une formule de quadrature à $n+1$ points ayant un degré d'exactitude maximal est nécessairement interpolatoire. On est donc amené à se demander s'il existe des choix de points x_i, $i=0,\dots,n$, conduisant à une formule de quadrature interpolatoire capable d'intégrer exactement tout polynôme de degré inférieur ou égal à $n+m$ pour un entier m strictement positif. Or, un résultat, dû à Jacobi [Jac26], montre que la condition

$$\forall q\in\mathbb{P}_{m-1},\ \int_a^b \omega_{n+1}(x)q(x)\,\mathrm{d}x=0, \tag{7.16}$$

où ω_{n+1} est le polynôme de Newton de degré $n+1$ (voir la sous-section 6.2.2) associé aux nœuds de quadrature, fournit une condition nécessaire [22] et suffisante [23] caractérisant une telle formule. Cette dernière impose exactement m contraintes sur les nœuds x_i, $i=0,\dots,n$, et l'on a alors forcément $m\le n+1$, faute de quoi ω_{n+1} serait orthogonal à $\mathbb{P}_{n+1}$ et par conséquent nul. Le degré d'exactitude maximal atteint par une formule de quadrature interpolatoire à $n+1$ nœuds distincts est par conséquent égal $2n+1$ et la condition (7.16) montre que ses nœuds sont les racines du polynôme ω_{n+1}, qui n'est autre, à un coefficient multiplicatif près, que le $(n+1)^{\text{e}}$ élément d'une suite de *polynômes orthogonaux relativement au produit scalaire de* $L^2([a,b])$. La formule de quadrature optimale (en termes du degré d'exactitude) ainsi obtenue porte le nom de *formule de quadrature de Gauss*, en référence à Johann Carl Friedrich Gauss qui la développa pour les besoins de ses calculs en astronomie sur les perturbations des orbites planétaires et la publia en 1814 dans un mémoire intitulé *Methodus nova integralium valores per approximationem inveniendi* présenté à la Société scientifique de Göttingen. Ces formules de quadrature ont par la suite été généralisées par Christoffel [Chr58] aux intégrales de la forme

$$I_w(f;a,b)=\int_a^b f(x)\,w(x)\,\mathrm{d}x,$$

22. En effet, le produit $\omega_{n+1}\,q$ étant un polynôme de degré inférieur ou égal à $n+1+m-1=n+m$, il est exactement intégré par la formule de quadrature et on a alors

$$\int_a^b \omega_{n+1}(x)q(x)\,\mathrm{d}x=\sum_{i=0}^n \alpha_i\,\omega_{n+1}(x_i)q(x_i)=0,$$

puisque les nœuds x_i, $i=0,\dots,n$, sont les racines de ω_{n+1}.

23. Tout polynôme p de $\mathbb{P}_{n+m}$ pouvant s'écrire sous la forme $p=\omega_{n+1}\,q+r$, où $q\in\mathbb{P}_{m-1}$ et $r\in\mathbb{P}_n$ sont respectivement le quotient et le reste de la division euclidienne de p par ω_{n+1}, on a

$$\int_a^b p(x)\,\mathrm{d}x=\int_a^b \omega_{n+1}(x)q(x)\,\mathrm{d}x+\int_a^b r(x)\,\mathrm{d}x=\int_a^b r(x)\,\mathrm{d}x,$$

en vertu de (7.16). La formule de quadrature étant interpolatoire, elle intègre exactement le polynôme r et l'on obtient alors

$$\int_a^b p(x)\,\mathrm{d}x=\sum_{i=0}^n \alpha_i\,r(x_i)=\sum_{i=0}^n \alpha_i\,(p(x_i)-\omega_{n+1}(x_i)q(x_i))=\sum_{i=0}^n \alpha_i\,p(x_i).$$

où w est une fonction positive et intégrable [24] sur $]a,b[$, appelée *fonction poids*. Les nœuds de la formule sont alors les racines des polynômes orthogonaux (voir la sous-section 6.1.3) pour le produit scalaire induit par la fonction poids et l'intervalle considérés, ce qui donne lieu à différentes familles de quadrature. Parmi les choix les plus courants, on peut citer

- les *formules de Gauss–Legendre*, pour la fonction poids $w(x)=1$ sur l'intervalle $]-1,1[$,
- les *formules de Gauss–Chebyshev* [Tch74], pour la fonction poids $w(x)=\frac{1}{\sqrt{1-x^2}}$ sur l'intervalle $]-1,1[$,
- les *formules de Gauss–Gegenbauer* [25], pour la fonction poids $w(x)=(1-x^2)^{\lambda-\frac{1}{2}}$ sur l'intervalle $]-1,1[$, avec λ un réel strictement supérieur à $-\frac{1}{2}$,
- les *formules de Gauss–Jacobi*, dont les trois précédents types de formules sont des cas particuliers, pour la fonction poids $w(x)=(1-x)^{\alpha}(1+x)^{\beta}$ sur l'intervalle $]-1,1[$, avec α et β des réels strictement supérieurs à -1,
- les *formules de Gauss–Laguerre*, pour la fonction poids $w(x)=e^{-x}$ sur l'intervalle $[0,+\infty[$,
- les *formules de Gauss–Hermite*, pour la fonction poids $w(x)=e^{-x^2}$ sur $\mathbb{R}$.

Les poids de quadrature d'une formule de Gauss sont tous strictement positifs et ses nœuds sont contenus dans l'intervalle *ouvert* $]a,b[$ (voir [Sti84]) et non uniformément répartis, comme on peut le constater sur la figure 7.7 pour les formules de Gauss–Legendre. On peut néanmoins être amené à inclure parmi les nœuds de quadrature soit l'une des deux, soit les deux extrémités de l'intervalle d'intégration, ce qui conduit respectivement aux *formules de Gauss–Radau* [26] [Rad80], dont le degré d'exactitude est égal à $2n$ pour une formule à $n+1$ points, et aux *formules de Gauss–Lobatto* [27] [Lob52], dont le degré d'exactitude vaut $2n-1$ pour une formule à $n+1$ nœuds. On a par ailleurs convergence des formules de Gauss, ainsi que de Gauss–Radau et de Gauss–Lobatto, vers l'intégrale $I_w(f)$ lorsque n tend vers l'infini pour tout intégrande f *continu* [Sti84].

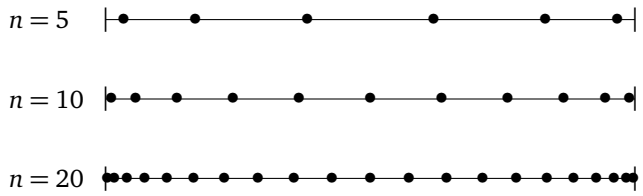


FIGURE 7.7 – Répartition sur l'intervalle $[-1,1]$ des nœuds de la formule de quadrature de Gauss–Legendre pour $n=5$, 10 et 20. On observe que les nœuds s'accumulent au voisinage des bornes de l'intervalle.

SUR LE CALCUL DES POIDS : Une méthode numérique stable de calcul des poids de formules de quadrature interpolatoires est présentée dans [KE82].

extension possibles : *formules de Gauss–Turán* [28] (utilisation des valeurs des dérivées, basée sur l'interpolation de Hermite) [Tur50] et *de Gauss–Kronrod* [29] (formule à $2n+1$ points obtenue par ajout de $n+1$ nœuds et poids à une formule à n nœuds, choisis de manière à maximiser le degré d'exactitude) [Kro65]

Mentionnons enfin la *méthode de Romberg* [30] [Rom55], qui est une méthode itérative de calcul numérique

24. L'intervalle $]a,b[$ n'étant pas forcément borné, on s'assure lorsque c'est le cas que l'intégrale ci-dessus est bien définie, au moins lorsque la fonction f est polynomiale, en requérant que tous les moments $\int_a^b x^s w(x)\,\mathrm{d}x$, $s\in\mathbb{N}$, existent et soient finis.

25. Leopold Bernhard Gegenbauer (2 février 1849 - 3 juin 1903) était un mathématicien autrichien. Surtout connu pour ses travaux en algèbre, il s'est également intéressé aux théories des fonctions et de l'intégration.

26. Jean-Charles Rodolphe Radau (22 janvier 1835 - 21 décembre 1911) était un astronome et mathématicien français d'origine allemande. Parmi ses travaux, on peut retenir deux mémoires consacrés à la réfraction, parus dans les *Annales de l'Observatoire de Paris* en 1881 et 1889, qui lui valurent chacun un prix de l'Académie des Sciences.

27. Rehuel Lobatto (6 juin 1797 - 9 février 1866) était un mathématicien hollandais. Il s'intéressa entre autres à l'intégration numérique d'équations différentielles, à une généralisation des formules de quadrature de Gauss et, pour les besoins du gouvernement hollandais, aux statistiques.

28. Paul Turán (Turán Pál en hongrois, 18 août 1910 - 26 septembre 1976) était un mathématicien hongrois. Travaillant principalement en théorie des nombres, en analyse et en théorie des graphes, il eut une longue collaboration avec son compatriote Paul Erdős, qui s'étendit sur quarante-six ans et se concrétisa par la publication de vingt-huit articles.

29. Aleksandr Semenovich Kronrod (Алекса́ндр Семёнович Кронро́д en russe, 22 octobre 1921 - 6 octobre 1986) était un mathématicien et informaticien russe, connu pour les formules de quadrature portant son nom. Il s'intéressa au calcul numérique appliqué à la physique théorique et à l'économie, ainsi qu'à l'intelligence artificielle et à la médecine.

30. Werner Romberg (16 mai 1909 - 5 février 2003) était un mathématicien allemand. Il est à l'origine d'une procédure récursive améliorant la précision du calcul d'une intégrale par la règle du trapèze composée.

d'intégrale basée sur l'application du *procédé d'extrapolation de Richardson* [RG27] pour l'accélération de la convergence de la règle du trapèze composée associée à des subdivisions dyadiques [31] successives de l'intervalle d'intégration.

NOTES : cette méthode est basée sur un développement de l'erreur par la formule d'Euler–Maclaurin,
elle nécessite le calcul du tableau des valeurs extrapolées $R(k,m)$, $0 \leq m \leq k \leq N$, dont les éléments satisfont la relation de récurrence

$$R(k,m) = \frac{1}{4^m - 1} \left(4^m R(k,m-1) - R(k-1,m-1)\right), 1 \leq m \leq k \leq N,$$

et la première colonne telle que $R(k,0) = I_{2^k,1}(f)$, $k = 0,\ldots,N$.
et l'utilisation d'un critère d'arrêt : $|R(k,k) - R(k-1,k-1)| \leq \varepsilon$

Les nœuds des formules de quadrature construites successivement de la méthode de Romberg sont équidistribués sur l'intervalle d'intégration, mais ces formules ne sont en revanche pas des formules de Newton–Cotes composées, excepté [32] pour de petites valeurs de l'entier m. En particulier, elles ne sont pas sujettes aux problèmes d'instabilité numérique mentionnés dans la section 7.1.2 lorsque m augmente. (QUESTION : les poids sont-ils toujours positifs ?)

Références

[Bas71] N. K. BASU. Evaluation of a definite integral using Tschebyscheff approximation. *Mathematica*, 13(36):13–23, 1971.

[Boo60] G. BOOLE. *A treatise on the calculus of finite differences*. Macmillan and Co., 1860.

[BP11] H. BRASS and K. PETRAS. *Quadrature theory: the theory of numerical integration on a compact interval*, volume 178 of *Mathematical surveys and monographs*. American Mathematical Society, 2011.

[CC60] C. W. CLENSHAW and A. R. CURTIS. A method for numerical integration on an automatic computer. *Numer. Math.*, 2(1):197–205, 1960. DOI: 10.1007/BF01386223.

[Chr58] E. B. CHRISTOFFEL. Über die Gaußische Quadratur und eine Verallgemeinerung derselben. *J. Reine Angew. Math.*, 1858(55):61–82, 1858. DOI: 10.1515/crll.1858.55.61.

[CT65] J. W. COOLEY and J. W. TUKEY. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19(90):297–301, 1965. DOI: 10.1090/S0025-5718-1965-0178586-1.

[Dav55] P. DAVIS. On a problem in the theory of mechanical quadratures. *Pacific J. Math.*, 5(suppl. 1):669–674, 1955. DOI: 10.2140/pjm.1955.5.669.

[DR84] P. J. DAVIS and P. RABINOWITZ. *Methods of numerical integration, Computer sciences and applied mathematics*. Academic Press, second edition, 1984.

[Fej33] L. FEJÉR. Mechanische Quadraturen mit positiven Cotesschen Zahlen. *Math. Z.*, 37(1):287–309, 1933. DOI: 10.1007/BF01474575.

[Fil30] L. N. G. FILON. On a quadrature formula for trigonometric integrals. *Proc. Roy. Soc. Edinburgh*, 49:38–47, 1930. DOI: 10.1017/S0370164600026262.

[Gen72a] W. M. GENTLEMAN. Implementing Clenshaw-Curtis quadrature, I Methodology and experience. *Comm. ACM*, 15(5):337–342, 1972. DOI: 10.1145/355602.361310.

[Gen72b] W. M. GENTLEMAN. Implementing Clenshaw-Curtis quadrature, II Computing the cosine transformation. *Comm. ACM*, 15(5):343–346, 1972. DOI: 10.1145/355602.361311.

[GG00] W. GANDER and W. GAUTSCHI. Adaptive quadrature – revisited. *BIT*, 40(1):84–101, 2000. DOI: 10.1023/A:1022318402393.

[GLR07] A. GLASER, X. LIU, and V. ROKHLIN. A fast algorithm for the calculation of the roots of special functions. *SIAM J. Sci. Comput.*, 29(4):1420–1438, 2007. DOI: 10.1137/06067016X.

[Jac26] C. G. J. JACOBI. Ueber Gauß neue Methode, die Werthe der Integrale näherungsweise zu finden. *J. Reine Angew. Math.*, 1826(1):301–308, 1826. DOI: 10.1515/crll.1826.1.301.

[KE82] J. KAUTSKY and S. ELHAY. Calculation of the weights of interpolatory quadratures. *Numer. Math.*, 40(3):407–422, 1982. DOI: 10.1007/BF01396453.

31. A DEFINIR

32. Pour $m = 0$, on retrouve en effet la règle du trapèze composée. Pour $m = 1$ et 2, les poids obtenus coïncident respectivement avec ceux de la règle de Simpson composée et de la règle de Boole composée.

[Kro65] A. S. KRONROD. *Nodes and weights of quadrature formulas. Sixteen-place tables*. Consultants Bureau, 1965.

[Kun62] G. F. KUNCIR. Algorithm 103: Simpson's rule integrator. *Comm. ACM*, 5(6):347, 1962. DOI: 10.1145/367766.368179.

[Lob52] R. LOBATTO. *Lessen over de differentiaal- en integraal-rekening. Tweede deel. Integral-rekening*. Gebroeders Van Cleef, 1852.

[Lyn69] J. N. LYNESS. Notes on the adaptive Simpson quadrature routine. *J. ACM*, 16(3):483–495, 1969. DOI: 10.1145/321526.321537.

[McK62] W. M. MCKEEMAN. Algorithm 145: Adaptive numerical integration by Simpson's rule. *Comm. ACM*, 5(12):604, 1962. DOI: 10.1145/355580.369102.

[Pea13] G. PEANO. Resto nelle formule di quadratura espresso con un integrale finito. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 22:562–569, 1913.

[Pól33] G. PÓLYA. Über die Konvergenz von Quadraturverfahren. *Math. Z.*, 37(1):264–286, 1933. DOI: 10.1007/BF01474574.

[Rad80] R. RADAU. Étude sur les formules d'approximation qui servent à calculer la valeur numérique d'une intégrale définie. *J. Math. Pures Appl. (3)*, 6 :283-336, 1880.

[RG27] L. F. RICHARDSON and J. A. GAUNT. The deferred approach to the limit. Part I. Single lattice. Part II. Interpenetrating lattices. *Philos. Trans. Roy. Soc. London Ser. A*, 226(636-646):299–361, 1927. DOI: 10.1098/rsta.1927.0008.

[Rom55] W. ROMBERG. Vereinfachte numerische Integration. *Norske Vid. Selsk. Forh. (Trondheim)*, 28(7):30–36, 1955.

[Ste27] J. F. STEFFENSEN. *Interpolation*. The Williams & Wilkins Company, 1927.

[Sti84] T. J. STIELTJES. Quelques recherches sur la théorie des quadratures dites mécaniques. *Ann. Sci. École Norm. Sup. (3)*, 1 :409-426, 1884. DOI : 10.24033/asens.245.

[Tch74] P. TCHEBICHEF. Sur les quadratures. *J. Math. Pures Appl. (2)*, 19 :19-34, 1874.

[Tre08] L. N. TREFETHEN. Is Gauss quadrature better than Clenshaw-Curtis? *SIAM Rev.*, 50(1):67–87, 2008. DOI: 10.1137/060659831.

[Tur50] P. TURÁN. On the theory of mechanical quadrature. *Acta Sci. Math. (Szeged)*, 12(A):30–37, 1950.

[TW14] L. N. TREFETHEN and J. A. C. WEIDEMAN. The exponentially convergent trapezoidal rule. *SIAM Rev.*, 56(3):385–458, 2014. DOI: 10.1137/130932132.

[Wal06] J. WALDVOGEL. Fast construction of the Fejér and Clenshaw–Curtis quadrature rules. *BIT*, 46(1):195–202, 2006. DOI: 10.1007/s10543-006-0045-4.

[Wed54] T. WEDDLE. On a new and simple rule for approximating to the area of a figure by means of seven equidistant ordinates. *Cambridge and Dublin Math. J.*, 9:79–80, 1854.

[Wei02] J. A. C. WEIDEMAN. Numerical integration of periodic functions: a few examples. *Amer. Math. Monthly*, 109(1):21–36, 2002. DOI: 10.2307/2695765.

Troisième partie

Équations différentielles et aux dérivées partielles

Dans cette dernière partie, on s'intéresse à la résolution numérique de problèmes d'équations dites *d'évolution*, c'est-à-dire de problèmes basées sur des équations différentielles ou aux dérivées partielles donc la solution dépend d'un paramètre qui, dans le cas le plus courant, représente la variable de temps.

COMPLETER :

parler des équations différentielles ordinaires

définir les équations aux dérivées partielles du second ordre avec classification (elliptique, hyperbolique, parabolique) dans le cas linéaire

Chapitre 8

Résolution numérique des équations différentielles ordinaires

Ce chapitre concerne la résolution numérique approchée d'équations, et de systèmes d'équations, différentielles ordinaires. De telles équations interviennent dans de nombreux problèmes issus de la modélisation mathématique de phénomènes physiques ou biologiques et se rencontrent par conséquent dans des disciplines aussi variées que l'ingénierie, la mécanique ou l'économie (plusieurs exemples sont donnés dans la section 8.2).

L'élaboration de techniques de résolution approchée des équations différentielles ordinaires constitue un vaste domaine d'études et de recherches depuis plus de trois siècles et notre objectif est d'en offrir au lecteur un premier aperçu. Après quelques rappels concernant les bases de la théorie des équations différentielles ordinaires, nous décrivons des méthodes de résolution numérique parmi les plus classiques et analysons leurs propriétés au moyen de techniques générales. Les notions de consistance, de stabilité et de convergence, déjà rencontrées dans ce cours et revisitées à cette occasion, occupent ici une place centrale et réapparaîtront lors de l'étude de méthodes de résolution numérique d'équations aux dérivées partielles aux chapitres 10 et 11.

8.1 Rappels sur les équations différentielles ordinaires *

Considérons une *équation différentielle ordinaire du premier ordre* [1] (*first-order ordinary differential equation* en anglais)

$$x'(t) = f(t, x(t)), \tag{8.1}$$

où x est une fonction d'une variable réelle à valeurs [2] dans $\mathbb{R}^d$ (avec d un entier supérieur ou égal à 1), que l'on cherche à déterminer, et f est une application définie et *continue* sur un ouvert D de $\mathbb{R}^{d+1}$, également à valeurs dans $\mathbb{R}^d$. Dans la plupart des cas, l'ouvert D aura la forme $I \times \Omega$, avec I un intervalle de $\mathbb{R}$, dans lequel la variable « *de temps* » t prend ses valeurs, et Ω un ouvert de $\mathbb{R}^d$, dans lequel l'« *état* » $x(t)$ prend ses valeurs.

Lorsque la fonction f est de la forme $f(t, x) = A(t)\,x + b(t)$, avec A et b des fonctions continues respectivement à valeurs dans $M_d(\mathbb{R})$ et $\mathbb{R}^d$, on dit que l'équation différentielle (8.1) est *linéaire* (*à coefficients constants* si A et b ne dépendent pas de t) et *linéaire homogène* si l'on a de plus $b \equiv 0$; elle est *non linéaire* dans les autres cas. Par ailleurs, l'équation est *autonome* lorsque la fonction f ne dépend pas de la variable t.

Exemples d'équations différentielles ordinaires du premier ordre. Parmi les équations différentielles ordinaires du premier ordre les plus célèbres de l'histoire des mathématiques, on peut mentionner l'*équation de Bernoulli*, proposée en 1695,

$$x'(t) = b_1(t)\,x(t) + b_2(t)\,x^m(t), \tag{8.2}$$

1. Le qualificatif « ordinaire » signifie que l'inconnue x de l'équation différentielle est une fonction qui ne dépend que d'une seule variable (ici, la variable t). L'équation est dite « du premier ordre » car elle ne fait intervenir que la dérivée première de x.

2. On notera que l'on a fait le choix, par souci de simplicité, de fonctions f et x à valeurs réelles, mais l'on aurait pu tout aussi bien envisager qu'elles prennent des valeurs complexes.

avec b_1 et b_2 des fonctions à valeurs réelles, en général continues, définies sur un intervalle ouvert de $\mathbb{R}$ et m un entier naturel (ou un réel, à condition que la solution soit à valeurs strictement positives) différent de 0 et 1, et l'*équation de Riccati*[3], introduite en 1720,

$$x'(t) = r_0(t) + r_1(t)\,x(t) + r_2(t)\,x^2(t), \tag{8.3}$$

avec r_0, r_1 et r_2 des fonctions à valeurs réelles, généralement continues, définies sur un intervalle ouvert de $\mathbb{R}$, telles que $r_0 \not\equiv 0$ et $r_2 \not\equiv 0$.

Il existe des équations différentielles d'ordre supérieur à un : étant donné un entier k supérieur ou égal à deux, un ouvert D de $\mathbb{R}^{kd+1}$ et une fonction f de D dans $\mathbb{R}^d$, on appelle *équation différentielle ordinaire d'ordre k* une équation de la forme

$$x^{(k)}(t) = f(t, x(t), x'(t), \ldots, x^{(k-1)}(t)), \tag{8.4}$$

l'entier k correspondant alors à l'ordre maximal des dérivées de la fonction x pouvant apparaître dans l'équation. Néanmoins, nous nous intéresserons uniquement à des équations différentielles ordinaires du premier ordre dans le reste de cette section. Les raisons de ce choix seront données plus loin.

8.1.1 Solutions

La première notion qu'il est important de préciser est celle de *solution* d'une équation différentielle ordinaire. Elle est l'objet de la définition suivante.

Définition 8.1 (solution d'une équation différentielle ordinaire) *On appelle **solution de l'équation différentielle ordinaire*** (8.1) *tout couple (J, x), avec J un intervalle de $\mathbb{R}$ et x une fonction dérivable sur J à valeurs dans $\mathbb{R}^d$ tels que l'équation* (8.1) *est satisfaite et $(t, x(t))$ appartient à D pour tout t appartenant à J.*

Exemple de solutions d'une équation différentielle ordinaire du premier ordre. On considère l'ouvert $D = \mathbb{R} \times \mathbb{R}_+^*$ et la fonction $f(t,x) = x^\alpha = \exp(\alpha \ln(x))$. On cherche alors un intervalle J de $\mathbb{R}$ et une application dérivable x de J dans $\mathbb{R}$ satisfaisant $x(t) > 0$ et $x'(t) = (x(t))^\alpha$, $\forall t \in J$. Cette dernière équation peut encore s'écrire

$$\frac{d}{dt}\left((1-\alpha)^{-1}(x(t))^{1-\alpha}\right) = 1,$$

et il existe une constante K telle que, $\forall t \in J$, $(1-\alpha)^{-1}(x(t))^{1-\alpha} = t + K$, ce qui implique que $t + K > 0$. Ainsi, pour tout $J \subset]K, +\infty[$, la fonction x donnée par

$$x(t) = ((1-\alpha)(t+K))^{1/(1-\alpha)},$$

vérifie l'équation pour tout t de J. Les solutions de l'équation sont par conséquent obtenues en faisant le choix d'un réel K, puis d'un intervalle J inclus dans $]K, +\infty[$ et en donnant enfin l'expression obtenue pour la fonction x.

On voit avec l'exemple précédent que les solutions d'une équation différentielle ordinaire peuvent être définies sur des intervalles plus ou moins grands. Cette remarque conduit à l'introduction de notions supplémentaires.

Définition 8.2 (prolongement d'une solution d'une équation différentielle ordinaire) *Soit (J, x) et $(\tilde{J}, \tilde{x})$ deux solutions de l'équation différentielle* (8.1)*. On dit que $(\tilde{J}, \tilde{x})$ est un **prolongement** (**extension** en anglais) de (J, x) si $J \subset \tilde{J}$ et $\tilde{x}_{|_J} = x$.*

Le prolongement induisant une relation d'ordre sur l'ensemble des solutions d'une équation, on a la définition et le résultat suivants.

Définition 8.3 (solution maximale d'une équation différentielle ordinaire) *On dit que le couple (J, x) est une **solution maximale** (**maximal solution** en anglais) de l'équation différentielle* (8.1) *si elle n'admet pas d'autre prolongement qu'elle-même.*

Théorème 8.4 *Toute solution de l'équation différentielle* (8.1) *se prolonge en une solution maximale (non nécessairement unique).*

DÉMONSTRATION. A ECRIRE □

Lorsque le domaine D est de la forme $D = I \times \Omega$, il est tout à fait possible que la solution maximale (J, x) d'un problème de Cauchy soit telle que J n'est pas égal à I. Ceci conduit à introduire la définition suivante.

3. Jacopo Francesco Riccati (28 mai 1676 - 15 avril 1754) était un mathématicien et physicien italien. Il s'intéressa à la résolution d'équations différentielles ordinaires par des méthodes de séparation de variables et de réduction d'ordre.

Définition 8.5 (solution globale d'une équation différentielle ordinaire) *Une solution de l'équation différentielle* (8.1) *est dite* ***globale*** *(**global** en anglais) si elle est définie sur l'ouvert I tout entier.*

On observera que toute solution globale est maximale, mais que la réciproque n'est pas vraie.

Exemple. A ECRIRE $x' = x^2$

A VOIR : obstruction au prolongement est du au fait que la solution maximale sort de tout compact quand t arrive aux bornes de J

Théorème 8.6 (« théorème des bouts ou de sortie de tout compact ») *A ECRIRE*

Donnons enfin un résultat relatif à la régularité des solutions d'une équation différentielle ordinaire du premier ordre.

Théorème 8.7 *Soit k un entier positif. Si la fonction f est de classe $\mathscr{C}^k$ (en tant que fonction de deux variables), alors toute solution de l'équation* (8.1) *est de classe $\mathscr{C}^{k+1}$ (en tant que fonction d'une variable).*

DÉMONSTRATION. Raisonnons par récurrence sur l'entier k. Si $k = 0$, la fonction f est continue. Pour toute solution (J, x) de (8.1), la fonction x est, par définition, dérivable sur J. Elle est donc continue et possède une dérivée continue sur J. Elle est par conséquent de classe $\mathscr{C}^1$ sur J.
Supposons à présent que le résultat est vrai à l'ordre $k-1$, $k \geq 1$. Toute solution de l'équation différentielle est alors au moins de classe $\mathscr{C}^k$. La fonction f étant de classe $\mathscr{C}^k$ par hypothèse de récurrence, il s'ensuit que la fonction x' est de classe $\mathscr{C}^k$ en tant que composée de fonctions de classe $\mathscr{C}^k$. On en déduit que la solution x est de classe $\mathscr{C}^{k+1}$ sur J. □

8.1.2 Problème de Cauchy

Nous avons vu plus haut qu'il existe en général une infinité de solutions d'une équation différentielle ordinaire. Cependant, dans la plupart des cas, on ne cherche pas à déterminer toutes ces solutions, mais seulement celles qui vont satisfaire une *condition initiale* (*initial condition* en anglais) prescrite. Ceci conduit à l'introduction de la notion de *problème de Cauchy* (*initial value (or Cauchy) problem* en anglais), associant une équation différentielle à une condition initiale donnée.

Définition 8.8 (« problème de Cauchy ») *Soit $(t, \eta) \in D$ donné. Résoudre le* ***problème de Cauchy*** *d'équation* (8.1) *et de* ***condition initiale*** *(**initial condition** en anglais)*

$$x(t_0) = \eta, \tag{8.5}$$

consiste à trouver une solution (J, x) de (8.1)*, telle que J contienne t_0 en son intérieur et que la fonction x satisfasse* (8.5).

Il est essentiel de remarquer que résoudre le problème de Cauchy équivaut à résoudre une *équation intégrale*, comme le montre le résultat suivant.

Lemme 8.9 *Un couple (J, x) est solution du problème de Cauchy* (8.1)-(8.5) *si et seulement si x est une fonction continue sur J, telle que $(t, x(t)) \in D$ pour tout $t \in J$, et qu'on a*

$$\forall t \in J,\ x(t) = \eta + \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s. \tag{8.6}$$

DÉMONSTRATION. A REPRENDRE
Soit une solution (J, x) du problème de Cauchy. Par application du théorème fondamental de l'analyse (voir le théorème B.131), on a

$$\forall t \in J,\ x(t) = x(t_0) + \int_{t_0}^{t} x'(s)\,\mathrm{d}s = \eta + \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s.$$

Réciproquement, considérons l'application $x : t \mapsto \eta + \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s$ définie sur J. On a tout d'abord que $x(t_0) = \eta + \int_{t_0}^{t_0} f(s, x(s))\,\mathrm{d}s = \eta$. De plus, la fonction f étant continue, l'application $t \mapsto \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s$ est de classe $\mathscr{C}^1$ et a pour dérivée $t \mapsto f(t, x(t))$. On en déduit que x est bien solution du problème de Cauchy. □

On appelle *courbe intégrale* (*integral curve* en anglais) toute courbe représentative d'une solution de l'équation différentielle ordinaire (8.1). On peut ainsi interpréter la résolution du problème de Cauchy comme la détermination d'une courbe intégrale de l'équation (8.1) passant par le point (t_0, η) associé à la condition initiale (8.5).

La notion de solution d'un problème de Cauchy étant introduite, il convient naturellement de se poser la question de l'existence d'une telle solution et, le cas échéant, de son unicité, en particulier si l'on envisage de résoudre numériquement le problème.

Le résultat que nous allons maintenant énoncer assure qu'un problème de Cauchy admet *localement* une unique solution. Pour cela, nous allons supposer que la fonction f est *localement lipschitzienne par rapport à sa seconde variable* sur le domaine D, ce qui signifie que REPRENDRE, pour tout $(t_0, \eta) \in D$, il existe une constante L strictement positive et un voisinage $V_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B(\eta, r_0)}$ de (t_0, η) dans D tels que f soit L-lipschitzienne par rapport à x sur V_0, c'est-à-dire

$$\forall (t, \eta_1) \in V_0,\ \forall (t, \eta_2) \in V_0,\ \|f(t, \eta_1) - f(t, \eta_2)\| \leq L\,\|\eta_1 - \eta_2\|. \tag{8.7}$$

où l'on désigne par $\|\cdot\|$ une norme choisie sur $\mathbb{R}^d$. Par ailleurs, la fonction f est dite *globalement lipschitzienne par rapport à sa seconde variable* sur le domaine D s'il existe $L > 0$ tellle que

$$\forall (t, \eta_1) \in D,\ \forall (t, \eta_2) \in D,\ \|f(t, \eta_1) - f(t, \eta_2)\| \leq L\,\|\eta_1 - \eta_2\|. \tag{8.8}$$

NOTE : une condition suffisante, satisfaite dans de nombreux cas pratiques, pour que f soit localement (resp. globalement) lipschitzienne est que f admette des dérivées partielles $\frac{\partial f_i}{\partial x_j}$, $1 \leq i, j \leq d$, continues sur un voisinage ouvert (resp. sur D). En faisant appel au théorème des accroissements finis, on peut alors poser $L = d \max\limits_{1 \leq i,j \leq d} \max\limits_{(t,x) \in C_0} \left| \frac{\partial f_i}{\partial x_j}(t, x) \right|$. Une condition suffisante encore plus forte est qu'elle soit de classe $\mathscr{C}^1$ sur un voisinage ouvert de (t_0, η) (resp. sur D).

On a le résultat fondamental suivant.

Théorème 8.10 (« théorème de Cauchy–Lipschitz ou de Picard [4]–Lindelöf [5] » [Lip68, Pic93, Lin94]) *Supposons que la fonction f soit continue et localement lipschitzienne par rapport à sa seconde variable. Alors, pour toute donnée (t_0, η), il existe, au voisinage de t_0, une unique solution au problème de Cauchy* (8.1)-(8.5).

DÉMONSTRATION. On peut démontrer ce résultat de diverses façons. Pour des détails sur la preuve originelle (et constructive) d'existence d'une solution, on se référera aux notes de fin de chapitre (voir la section 8.9). La démonstration proposée ici est basée sur la forme intégrale (8.6) et un argument de point fixe, dû à Picard.

Soit $C = [t_0 - T, t_0 + T] \times \overline{B(x_0, r_0)}$ avec $T \leq \min(T_0, \frac{r_0}{\sup \|f\|})$ et notons $\mathscr{F} = \mathscr{C}([t_0 - T, t_0 + T], \overline{B(x_0, r_0)})$ l'ensemble des applications continues de $[t_0 - T, t_0 + T]$ dans $\overline{B(x_0, r_0)}$, que l'on munit de la norme de la convergence uniforme. À toute fonction x de $\mathscr{F}$, associons la fonction $\phi(x)$ définie par

$$(\phi(x))(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s,\ t \in [t_0 - T, t_0 + T].$$

D'après le lemme (d'équivalence entre la résolution du problème de Cauchy (8.1)-(8.5) et celle d'une équation intégrale), la fonction x est une solution du problème de Cauchy (8.1)-(8.5) si et seulement si elle est un point fixe de l'application ϕ. Observons alors que

$$\|(\phi(x))(t) - x_0\| = \left\| \int_{t_0}^{t} f(s, x(s))\,\mathrm{d}s \right\| \leq \sup \|f\|\ |t - t_0| \leq \sup \|f\|\ T \leq r_0,$$

4. Charles Émile Picard (24 juillet 1856 - 11 décembre 1941) était un mathématicien français, également philosophe et historien des sciences. Il est l'auteur de deux difficiles théorèmes en analyse complexe et fut le premier à utiliser le théorème du point fixe de Banach dans une méthode d'approximations successives de solutions d'équations différentielles ou d'équations aux dérivées partielles.

5. Ernst Leonard Lindelöf (7 mars 1870 - 4 juin 1946) était un mathématicien finlandais. Il travailla principalement en topologie, en analyse complexe, en théorie des équations différentielles, et œuvra pour l'étude de l'histoire des mathématiques finlandaises.

d'où $\phi(x)$ appartient à $\mathscr{F}$. L'opérateur ϕ envoie donc $\mathscr{F}$ dans $\mathscr{F}$.
Nous allons maintenant montrer qu'il existe une itérée de ϕ qui est contractante. Soit deux fonctions x et y de $\mathscr{F}$. On a

$$\|(\phi(x))(t)-(\phi(y))(t)\| = \|\int_{t_0}^{t}(f(s,x(s))-f(s,y(s)))\,\mathrm{d}s\| \leq \left|\int_{t_0}^{t} L\,\|x(s)-y(s)\|\,\mathrm{d}s\right| \leq L\,|t-t_0|\,\|x-y\|.$$

De la même manière,

$$\|((\phi\circ\phi)(x))(t)-((\phi\circ\phi)(y))(t)\| \leq \left|\int_{t_0}^{t} L\,\|(\phi(x))(s)-(\phi(y))(s)\|\,\mathrm{d}s\right| \leq \frac{L^2}{2}\,|t-t_0|^2\,\|x-y\|.$$

En raisonnant par récurrence, on montre alors que

$$\|((\phi^m)(x))(t)-((\phi^m)(y))(t)\| \leq \frac{L^m}{m!}\,|t-t_0|^m\,\|x-y\|,\ m\geq 1.$$

Il s'ensuit que l'application ϕ^m est lipschitzienne de constante $\frac{L^m}{m!}\,T^m$. Puisque $\lim_{m\to+\infty}\frac{L^m}{m!}\,T^m = 0$, il existe un entier m tel que $\frac{L^m}{m!}\,T^m < 1$ et pour lequel l'application ϕ^p est contractante. $\mathscr{F}$ étant un espace de Banach (espace métrique complet), le théorème du point fixe (généralisé au cas d'une application dont une itérée est contractante) montre que ϕ admet un unique point fixe. □

Corollaire 8.11 (existence d'une solution maximale) *A VOIR On suppose que f est continue et localement lipschitzienne par rapport à x sur l'ouvert D. Pour toute donnée initiale $(t_0,\eta)\in D$, il existe une unique solution maximale du problème de Cauchy. L'intervalle de définition de cette solution est ouvert.*

DÉMONSTRATION. A ECRIRE □

Exemple de solution maximale d'un problème de Cauchy. $x' = x^2,\ x(0)=\eta\neq 0$

Lorsque l'application f satisfait seulement l'hypothèse de continuité formulée en début de section, on peut encore énoncer un résultat d'existence locale de solution, mais l'unicité de cette dernière ne peut toutefois être assurée.

Théorème 8.12 (« théorème de Peano » [Pea86, Pea90]) *A ECRIRE*

DÉMONSTRATION. A ECRIRE, rappeler le

Théorème 8.13 (« théorème d'Arzelà [6]–Ascoli [7] » [Asc84, Arz95]) *Soit I un intervalle de $\mathbb{R}$ et $\mathscr{C}(I,\mathbb{R}^d)$ l'ensemble des applications continues de I dans $\mathbb{R}^d$, que l'on munit de la norme de la convergence uniforme. Alors, une famille de $\mathscr{C}(I,\mathbb{R}^d)$ est relativement compacte (c'est-à-dire que son adhérence est compacte) si et seulement si elle est équibornée et équicontinue.*

□

Sous ces conditions cependant, il se peut qu'une équation différentielle ordinaire admettent plusieurs, voire une infinité de, solutions, comme le montre l'exemple suivant.

Exemple de solutions non uniques d'un problème de Cauchy. COMPLETER $x' = 2|x|^{1/2},\ x(0)=0$

A VOIR : Existence et unicité d'une solution globale sous une condition de Lipschitz globale par rapport à x, uniformément ou non par rapport à t : f définie sur l'ouvert $I\times\mathbb{R}^d$

$$\forall t\in I,\ \forall (x,y)\in\mathbb{R}^d\times\mathbb{R}^d,\ \|f(t,x)-f(t,y)\|\leq L\,\|x-y\|, \tag{8.9}$$

c'est par exemple le cas si f est bornée, cas des équations linéaires

A VOIR : QUESTION de la sensibilité par rapport aux données initiales

6. Cesare Arzelà (6 mars 1847 - 15 mars 1912) était un mathématicien italien. Il est passé à la postérité pour ses contributions à la théorie des fonctions d'une variable réelle, et plus particulièrement la caractérisation des suites de fonctions continues.

7. Giulio Ascoli (20 janvier 1843 - 12 juillet 1896) était un mathématicien italien. On lui doit, parmi d'autres contributions à la théorie des fonctions d'une variable réelle, la notion d'équicontinuité.

Définition 8.14 *A FAIRE : notion de stabilité (et stabilité asymptotique) des solutions d'une EDO*

A VOIR : notion de *flot* associé à l'équation différentielle ordinaire

A VOIR :

Proposition 8.15 (« inégalité de Grönwall [8] ») *Soit L une fonction positive intégrable sur l'intervalle $]t_0, t_0+T[$, K et φ deux fonctions continues sur $[t_0, t_0+T]$, K étant non décroissante. Si φ satisfait l'inégalité*

$$\forall t \in [t_0, t_0+T],\ \varphi(t) \leq K(t) + \int_{t_0}^{t} L(s)\varphi(s)\,\mathrm{d}s,$$

alors

$$\forall t \in [t_0, t_0+T],\ \varphi(t) \leq K(t)\, e^{\int_{t_0}^{t} L(s)\,\mathrm{d}s}.$$

8.1.3 Une remarque fondamentale

Nous avons fait le choix, apparemment arbitraire, de ne traiter dans les définitions et résultats précédents que le cas d'équations différentielles ordinaires du premier ordre. Ceci s'explique par le fait que toute équation différentielle ordinaire d'ordre supérieur à un peut être ramenée mécaniquement à un *système* équivalent d'équations différentielles d'ordre un en introduisant des inconnues supplémentaires.

Considérons en effet l'équation différentielle ordinaire d'ordre k (8.4). Il est clair qu'en posant

$$\boldsymbol{y} = \begin{pmatrix} x \\ x' \\ \vdots \\ x^{(k-1)} \end{pmatrix},$$

on peut réécrire (8.4) sous la forme

$$\boldsymbol{y}'(t) = F(t, \boldsymbol{y}(t)),$$

où

$$\forall (t, \boldsymbol{y}) \in D,\ F(t, \boldsymbol{y}) = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_k \\ f(t, y_1, \ldots, y_k) \end{pmatrix}.$$

En observant que la fonction F est continue et localement (resp. globalement) lipschitzienne par rapport à $\boldsymbol{y}$ si et seulement si la fonction f est continue et localement (resp. globalement) lipschitzienne par rapport à $x, x', \ldots, x^{(k-1)}$, il devient alors possible d'étudier l'existence et l'unicité de solutions de tout problème mettant en jeu l'équation (8.4) avec les seuls outils et résultats introduits pour les équations du premier ordre.

Indiquons que ce procédé n'a pas qu'un intérêt théorique : il en effet mis à profit dans les bibliothèques de programmes de résolution des équations différentielles ordinaires, qui supposent généralement que l'équation du problème à résoudre est sous la forme d'un système d'équations différentielles du premier ordre. De fait, les méthodes d'approximation numérique présentées dans ce chapitre ne concernent que les équations du premier ordre.

8.2 Exemples d'équations et de systèmes différentiels ordinaires

Comme nous l'avons écrit, les modèles mathématiques basés sur des équations différentielles ordinaires sont extrêmement courants. Dans nombre de situations concrètes, la variable t représente le temps et x est une famille

8. Thomas Hakon Grönwall (16 janvier 1877 - 9 mai 1932) était un mathématicien suédois. Il travailla principalement dans les domaines de l'analyse, de la théorie des nombres et de la physique mathématique, mais s'intéressa aussi aux applications des mathématiques en ingénierie et dans l'industrie en tant que consultant.

de paramètres décrivant l'état d'un système matériel donné. L'équation différentielle ordinaire (8.1) traduit ainsi mathématiquement la loi d'évolution du système considéré au cours du temps. Savoir résoudre un problème de Cauchy revient donc à savoir prévoir la configuration du système à tout moment alors qu'on en connaît seulement une description à un instant initial donné.

Les exemples suivants proviennent de divers champs scientifiques et présentent des problèmes pour lesquels une solution explicite du système d'équations différentielles ordinaires considéré n'est généralement pas disponible. Le recours aux méthodes numériques introduites dans ce chapitre est par conséquent incontournable. Afin de rester consistant avec les conventions employées dans la plupart des ouvrages (de physique, de mécanique, de biologie...) discutant de ces modèles, la notation différentielle de la dérivée a été adoptée dans toute cette section.

8.2.1 Problème à N corps en mécanique céleste

En mécanique classique, c'est-à-dire dans le cas où les effets de la théorie de la relativité générale peuvent être négligés [9], la résolution du *problème à N corps*, avec N un entier naturel strictement plus grand que 1, consiste en la détermination des trajectoires de N corps en interaction gravitationnelle, connaissant leurs masses ainsi que leurs positions et vitesses initiales. En vertu de la *relation fondamentale de la dynamique en translation* [10], ce problème est modélisé par un système de N équations différentielles ordinaires du second ordre dont les inconnues sont à valeurs dans $\mathbb{R}^3$,

$$\frac{\mathrm{d}^2 \boldsymbol{x}_i}{\mathrm{d}t^2} = G \sum_{\substack{j=1 \\ j \neq i}}^{N} m_j \frac{\boldsymbol{x}_j - \boldsymbol{x}_i}{\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^3}, \ i = 1, \ldots, N, \tag{8.10}$$

dans lesquelles les quantités $\boldsymbol{x}_i$ et m_i, $i = 1, \ldots, N$, sont les positions, dépendantes du temps t, et masses respectives des corps, G est la *constante universelle de gravitation* [11] et $\|\cdot\|$ désigne la norme euclidienne, que l'on complète par la donnée de deux conditions initiales

$$\boldsymbol{x}_i(0) = \boldsymbol{x}_{i0} \text{ et } \frac{\mathrm{d}\boldsymbol{x}_i}{\mathrm{d}t}(0) = \boldsymbol{v}_{i0}, \ i = 1, \ldots, N, \tag{8.11}$$

avec $\boldsymbol{x}_{j_0} \neq \boldsymbol{x}_{k0}$ pour tous entiers j et k distincts appartenant à $\{1, \ldots, N\}$.

Le problème (8.10)-(8.11) se résout facilement par la théorie de Newton lorsque $N = 2$. Dans tout autre cas, il possède, comme découvert par Sundman [12] en 1909 pour le problème à trois corps [Sun09] et généralisé par Wang en 1991 pour $N > 3$ [Wan91], une solution analytique qui se présente sous la forme d'une série infinie, qu'une très lente convergence rend malheureusement inutilisable en pratique. Il faut donc généralement faire appel à une méthode numérique pour obtenir des solutions approchées du type de celle présentée sur la figure 8.1.

8.2.2 Modèle de Lotka–Volterra en dynamique des populations

Le *modèle de Lotka* [13]*–Volterra* [14] est utilisé pour décrire la dynamique de systèmes biologiques dans lesquels un prédateur et sa proie interagissent. Faisant des hypothèses sur l'environnement et l'évolution des populations de prédateurs et de proies, à savoir que

- les proies ont une nourriture abondante et se reproduisent de manière exponentielle en l'absence de prédation,
- les prédateurs se nourrissent exclusivement de proies et ont tendance à disparaître de façon exponentielle lorsque la nourriture manque,
- le taux de prédation sur les proies est proportionnel à la fréquence de rencontre entre les prédateurs et les proies,

9. Ceci suppose que les vitesses de mouvement des corps considérés sont petites devant celle de la lumière dans le vide.

10. Il s'agit de la deuxième *loi de Newton*, que l'on énonce ainsi : *l'accélération subie par un corps de masse constante dans un référentiel galiléen est proportionnelle à la résultante des forces qu'il subit, et inversement proportionnelle à sa masse*.

11. Dans le système international d'unités, la valeur de G recommandée est $6,67428(67)\ 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$ [MTN08].

12. Karl Frithiof Sundman (28 octobre 1873 - 28 septembre 1949) était un astronome et mathématicien finlandais. Il est connu pour avoir prouvé, au moyen de méthodes analytiques de régularisation, l'existence d'une série convergente solution du problème à trois corps.

13. Alfred James Lotka (2 mars 1880 - 5 décembre 1949) était un mathématicien et statisticien américain, théoricien de la dynamique des populations.

14. Vito Volterra (3 mai 1860 - 11 octobre 1940) était un mathématicien et physicien italien. Il est surtout connu pour ses travaux sur les équations intégrales et intégro-différentielles, sur les dislocations dans les cristaux et sur la dynamique des populations.

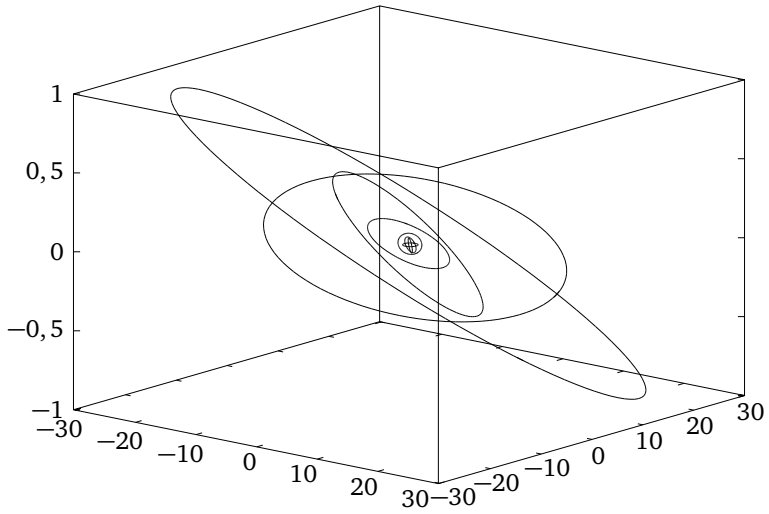


FIGURE 8.1 – Orbites des huit planètes du système solaire (Mercure, Vénus, la Terre, Mars, Jupiter, Saturne, Uranus et Neptune, le Soleil étant placé à l'origine du repère) durant une période de révolution de Neptune (soit environ 164,79 années terrestres), obtenues par résolution numérique du problème (8.10)-(8.11) pour $N = 8$ (le système de vingt-quatre équations différentielles du second ordre ayant été préalablement écrit sous la forme d'un système de quarante-huit équations différentielles du premier ordre). Les distances sont exprimées en *unités astronomiques* (1 ua = 149597870691 m).

• le taux de croissance des prédateurs est proportionnel à la quantité de nourriture à leur disposition,

celui-ci consiste en un système de deux équations différentielles ordinaires non linéaires couplées

$$\begin{cases} \dfrac{\mathrm{d}N}{\mathrm{d}t} &= N\,(\alpha - \beta\,P)\,, \\ \dfrac{\mathrm{d}P}{\mathrm{d}t} &= P\,(\gamma\,N - \delta)\,, \end{cases} \tag{8.12}$$

où la variable t désigne le temps, $N(t)$ est le nombre de proies à l'instant t, $P(t)$ est le nombre de prédateurs à l'instant t, et où les paramètres α, β, γ et δ sont respectivement le taux de reproduction des proies en l'absence de prédateurs, le taux de mortalité des proies due à la prédation, le taux de reproduction des prédateurs en fonction de la consommation de proies et le taux de mortalité des prédateurs en l'absence de proies, auquel on adjoint une condition initiale

$$N(0) = N_0,\ P(0) = P_0, \tag{8.13}$$

où N_0 et P_0 sont des constantes strictement positives représentant respectivement les effectifs de proies et de prédateurs à l'instant initial.

Ce modèle a été proposé indépendamment par Lotka, initialement pour l'étude de systèmes chimiques [Lot10], puis organiques [Lot20] (un exemple simple étant celui d'une espèce végétale et d'une espèce animale herbivore la consommant), et Volterra [Vol26], qui cherchait à fournir une explication aux fluctuations des prises de certaines espèces de poissons dans la mer Adriatique au sortir de la première guerre mondiale.

Le système d'équations autonomes (8.12) a largement été étudié d'un point de vue mathématique. On peut montrer que ses solutions sont positives, bornées, périodiques et que la fonction $H(N,P) = \gamma\,N - \delta\,\ln(N) + \beta\,P - \alpha\,\ln(P)$ est une *intégrale première* [15]. Il possède aussi deux points d'équilibre, $(0,0)$ (qui est instable) et $(\frac{\delta}{\gamma}, \frac{\alpha}{\beta})$ (qui est stable). La figure 8.2 présente un exemple de solution numérique du problème (8.12)-(8.13).

15. On appelle intégrale première d'un système d'équations différentielles toute fonction des solutions du système restant constante le long d'une trajectoire quelconque.

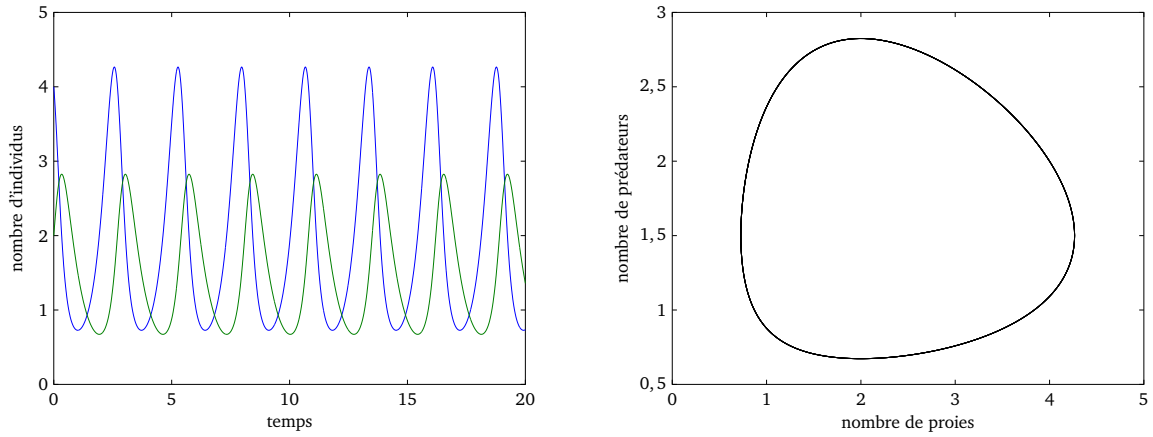


FIGURE 8.2 – Évolution des populations de proies (courbe bleue) et de prédateurs (courbe verte) au cours du temps (à gauche) et diagramme de phase (à droite) obtenus par résolution numérique du problème (8.12)-(8.13) sur l'intervalle $[0, T]$, avec $T = 20$, les valeurs des paramètres étant $\alpha = 3$, $\beta = 2$, $\gamma = 1$, $\delta = 2$, $N_0 = 4$ et $P_0 = 2$.

Le modèle de Lotka–Volterra peut être rendu plus réaliste en recourant à une loi de croissance *logistique* [Ver38], plutôt que *malthusienne*, des proies en l'absence de prédateurs (la quantité de nourriture présente dans le milieu naturel étant finie) ou en cherchant à transcrire une certaine « satiété » des prédateurs vis-à-vis des proies lorsque celles-ci sont en nombre important (un prédateur ne pouvant manger, et digérer, ou chasser qu'un nombre limité de proies) en utilisant, par exemple, une *réponse fonctionnelle de Holling de type II* [Hol59] pour la prédation. Dans ce cas, le système (8.12) se trouve modifié de la manière suivante

$$\begin{cases} \dfrac{\mathrm{d}N}{\mathrm{d}t} &= N\left(\alpha\left(1-\dfrac{N}{\kappa}\right)-\dfrac{\beta}{1+\beta\tau N}P\right), \\ \dfrac{\mathrm{d}P}{\mathrm{d}t} &= P\left(\dfrac{\gamma}{1+\beta\tau N}N-\delta\right), \end{cases}$$

où les constantes κ et τ, toutes deux strictement positives, sont respectivement la *capacité de charge* (*carrying capacity* en anglais) du milieu, qui correspond à la capacité de l'environnement à supporter la croissance des proies, et un *temps de manipulation* (*handling time* en anglais), représentant le temps moyen consacré à la consommation d'une proie.

Ce type de système d'équations différentielles se retrouve dans plusieurs modèles en économie, le plus célèbre et le plus ancien d'entre eux étant certainement le *modèle de Goodwin* [16] (*Goodwin's class struggle model* en anglais) [Goo67]. Ce dernier constitue une formalisation mathématique d'un cycle économique endogène s'expliquant principalement par le conflit de classes pour le partage de la richesse produite entre le capital et le travail dans le cadre d'un régime salarial concurrentiel.

Il se base sur les observations suivantes. En période de chômage élevé, les salaires réels progressent moins vite que la productivité du travail et la part des bénéfices dans le produit intérieur s'accroît alors, permettant une augmentation de l'investissement. La croissance et l'accumulation de capital étant soutenues, elles entraînent une reprise de la demande de travail et donc un déclin du chômage. Cette diminution redonne aux salariés un pouvoir de négociation, ce qui provoque à terme une baisse des bénéfices. Ainsi limitées dans leur capacité de financement, les entreprises réduisent leurs investissements, ralentissant l'accumulation et ramenant le taux de chômage à la hausse. Un nouveau cycle peut alors commencer. Goodwin reconnut dans cette dynamique celle des systèmes biologiques étudiés par Volterra et décrits plus haut. Nous conclurons cette sous-section en détaillant la dérivation de son modèle.

Quelques hypothèses simplificatrices sont tout d'abord faites. On considère un système économique fermé, dans lequel interagissent deux types d'agents : des travailleurs et des capitalistes. La production y résulte de l'utilisation

16. Richard M. Goodwin (24 février 1913 - 13 août 1996) était un économiste et mathématicien américain. Il est connu pour ses travaux sur les relations entre la croissance économique à long terme et les « cycle des affaires ».

de deux facteurs, qui sont le capital (*capital* en anglais), noté K, constitué d'équipements (usines, machines...) et le travail (*labour* en anglais), noté L, correspondant à la population active occupée. On distingue ainsi deux classes de revenus : ceux provenant du travail, versés sous forme de salaires (*wages* en anglais) aux travailleurs et utilisés en totalité pour l'achat de biens de consommation, et ceux issus des bénéfices (*profits* en anglais), distribués aux capitalistes et intégralement réinvestis. Enfin, on suppose que le progrès technique et l'augmentation de la population active sont constants.

En notant W la masse salariale (*wage bill* en anglais) et P le montant des bénéfices, on peut écrire que l'intégralité du produit intérieur, noté Y, est distribué en revenus,

$$\forall t > 0,\ Y(t) = W(t) + P(t),$$

avec $W = wL$, où w est le salaire nominal (*nominal wage* en anglais). On fait ensuite l'hypothèse que la des salaires est utilisée pour

$$\forall t > 0,\ W(t) = C(t),$$

où C..., et que les bénéfices sont intégralement investis,

$$\forall t > 0,\ P(t) = I(t),$$

où I... REPRENDRE La conséquence est que le marché des produits est toujours équilibré. Les cycles économiques ne peuvent donc pas provenir de déséquilibres entre l'offre et la demande globales. On voit aussi que, dans une perspective de long terme, les bénéfices financent l'accumulation du capital puisque, par définition, la variation du stock de capital est égale à l'investissement :

$$K'(t) = I(t),\ t > 0.$$

On suppose aussi que la productivité du capital est constante au cours du temps

$$\frac{Y(t)}{K(t)} = \sigma,\ t > 0$$

avec $\sigma > 0$, et que la productivité du travail croît à un taux α constant

$$\frac{Y(t)}{L(t)} = \frac{Y(0)}{L(0)}\, e^{\alpha\, t},\ t > 0 \tag{8.14}$$

avec $Y(0)$, $L(0)$ et α (: taux de croissance de la productivité) strictement positifs.
- la population active, notée N croît avec un taux démographique constant

$$N(t) = N(0)\, e^{\beta\, t},\ \forall t > 0 \tag{8.15}$$

avec $N(0) > 0$ et $\beta \geq 0$ (: taux de croissance de la population active). Le taux d'emploi, noté v est le rapport $\frac{L}{N}$ (compris entre 0 (pas de salarié) et 1 (pas de chômeur)). Son complément à 1 est le taux de chômage $u = \frac{N-L}{N}$. Le fait de distinguer l'emploi effectif de l'emploi potentiel traduit l'existence d'un "marché du travail" ou, pour mieux dire dans la théorie marxiste, d'un marché où s'échange la force de travail.

- On retient une interprétation marxienne de la *courbe/relation de Phillips* [17] [Phi58], bien connue en macroéconomie comme relation décroissante entre le taux de chômage et le taux de croissance du salaire moyen/une relation empirique décroissante entre le taux de chômage et l'inflation ou taux de croissance des salaires nominaux, noté w. Quand le taux d'emploi est faible, le pouvoir de négociation salariale des salariés est faible et la croissance du salaire moyen est réduite, voire négative. Réciproquement, quand le taux d'emploi est élevé, les salariés obtiennent des augmentations du salaire moyen. On retient la formulation linéaire

$$w'(t) = (\rho\, v(t) - \gamma)\, w(t) \tag{8.16}$$

17. Alban William Housego Phillips (18 novembre 1914 - 4 mars 1975) était un économiste néo-zélandais. Sa contribution la plus connue à la macroéconomie est la courbe portant aujourd'hui son nom. Il conçut et construisit en 1949 le *MONIAC* (acronyme de *Monetary National Income Analogue Computer* en anglais), un ordinateur analogique hydraulique utilisé pour simuler le fonctionnement de l'économie du Royaume-Uni.

avec $\gamma, \rho > 0$. Les évolutions temporelles de la masse salariale $W = wL$ se déduisent alors des évolutions conjointes de la population active occupée et du salaire moyen :

$$\frac{W'}{W} = \frac{w'}{w} + \frac{L'}{L}$$

La dynamique du modèle est commandée de façon endogène par les effets conjugués de l'accumulation du capital et de la formation des salaires. D'une part, l'évolution du stock de capital dépend des profits, qui dépendent des salaires distribués et donc des revendications salariales. D'autre part, les évolutions salariales dépendent du taux d'emploi, mais l'emploi est un facteur de production complémentaire au capital utilisé, dont les variations sont financées par les profits. Goodwin choisit de mettre en avant la dynamique conjointe du taux d'emploi et de la part des salaires dans le revenu, qu'on notera

$$\omega = \frac{W}{Y} = \frac{wL}{Y}$$

(ω : part salariale de la valeur ajoutée)

Il reste à obtenir le système différentiel exprimant les lois d'évolution du taux d'emploi et de la part des salaires dans le revenu. Par définition de ν, on a

$$\frac{\nu'}{\nu} = \frac{L'}{L} - \frac{N'}{N} = \frac{L'}{L} - \beta$$

d'après (8.15). Puisque les gains de productivité du travail sont constants, il vient d'après (8.14)

$$\frac{Y'}{Y} - \frac{L'}{L} = \alpha$$

La croissance du produit est égale au rythme d'accumulation du capital (car la productivité du capital est constante par (constance prod capital)), donc

$$\frac{K'}{K} = \frac{Y'}{Y}$$

ce qui entraîne

$$\frac{L'}{L} = \frac{K'}{K} - \alpha$$

Mais, par (constance prod capital),

$$\frac{K'}{K} = \frac{K'}{Y}\frac{Y}{K} = \sigma\,\frac{K'}{Y}$$

et $K' = P$ (l'accumulation du capital est financée par les profits, A VOIR) et de plus $P = Y - W$ (par ...) d'où

$$\frac{K'}{K} = \sigma\,\frac{K'}{Y} = \sigma\,\frac{Y-W}{Y} = \sigma\,(1-\omega)$$

et donc

$$\frac{\nu'}{\nu} = \sigma\,(1-\omega) - (\alpha+\beta)$$

d'où

$$\nu' = (\sigma\,(1-\omega) - (\alpha+\beta))\nu$$

Par définition de ω, on a

$$\frac{\omega'}{\omega} = \frac{w'}{w} + \frac{L'}{L} - \frac{Y'}{Y}$$

Par (8.16) et (8.14), on trouve $\frac{\omega'}{\omega} = \rho\,\nu(t) - (\gamma+\alpha)$, d'où

$$\omega' = (\rho\,\nu - (\alpha+\gamma))\omega$$

8.2.3 Oscillateur de van der Pol

L'*oscillateur de van der Pol* [18] est un exemple d'oscillateur dont l'évolution est gouvernée par l'équation différentielle ordinaire du second ordre suivante

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \mu(x^2-1)\frac{\mathrm{d}x}{\mathrm{d}t} + x = 0 \tag{8.17}$$

que van der Pol réalisa au moyen d'un circuit électrique composé de deux résisteurs de résistances respectives R et r, d'un condensateur de capacité électrique C, d'une bobine et d'une tétrode, la période d'oscillation étant donnée par $\mu = RC$ [vdPol26].

Dans ce système, le terme non linéaire a pour effet d'amplifier les oscillations de faible amplitude et, *a contrario*, d'atténuer celles de forte amplitude. On s'attend par conséquent à l'existence de solutions particulières périodiques stables, que des solutions issues de conditions aux limites « voisines » approchent asymptotiquement (on parle dans ce cas de *cycles limites*).

On a représenté sur les figures 8.3 et 8.4 une solution du problème, obtenue par résolution numérique, pour deux régimes distincts, déterminés par la valeur du paramètre μ, l'un faiblement amorti ($\mu = 0,1$) et l'autre amorti ($\mu = 3,5$).

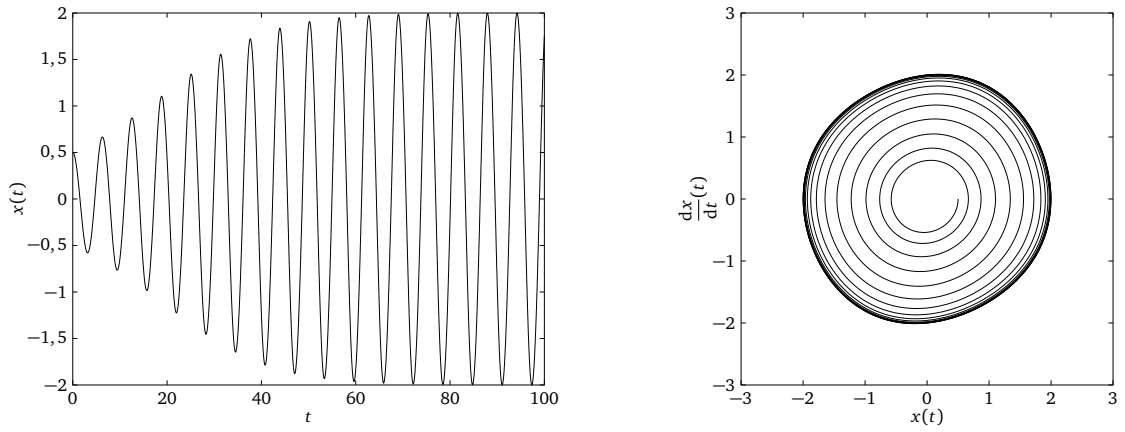


FIGURE 8.3 – Évolution au cours du temps (à gauche) et portrait de phase (à droite) de la solution de l'équation (8.17) vérifiant la condition initiale $x(0) = 0,5$, $\dfrac{\mathrm{d}x}{\mathrm{d}t}(0) = 0$, dans le cas faiblement amorti ($\mu = 0,1$).

Dans le premier cas, les oscillations sont quasi-sinusoïdales et le cycle limite vers lequel la solution converge a pratiquement la forme d'un cercle dans l'espace des phases. Pour le second cas, l'évolution de la solution fait apparaître des variations lentes de l'amplitude entrecoupées de changements soudains, ce qui donne lieu à des oscillations dites *de relaxation*.

8.2.4 Modèle SIR de Kermack–McKendrick en épidémiologie

Un *modèle SIR* (acronyme anglais pour *Susceptible-Infected-Recovered*) est un modèle compartimental d'évolution d'une maladie infectieuse au sein une population donnée au cours du temps. Il considère que la population est divisée en trois *compartiments*, représentant chacun un état possible d'un individu face à la maladie : le compartiment S des individus susceptibles de contracter la maladie, le compartiment I des individus infectés et le compartiment R des individus ayant guéri et ainsi acquis une immunité face à la maladie [19], une personne passant d'un compartiment à l'autre selon le schéma

$$S \longrightarrow I \longrightarrow R.$$

18. Balthasar van der Pol (27 janvier 1889 - 6 octobre 1959) était un physicien expérimentateur hollandais. Principalement intéressé par la propagation des ondes radioélectriques, la théorie des circuits électriques et la physique mathématique, ses travaux sur les oscillations non-linéaires connurent un regain d'intérêt dans les années 1980 à la faveur de la théorie du chaos.

19. Dans le cas d'une maladie mortelle, les individus décédés du fait de l'infection peuvent être inclus dans ce dernier compartiment.

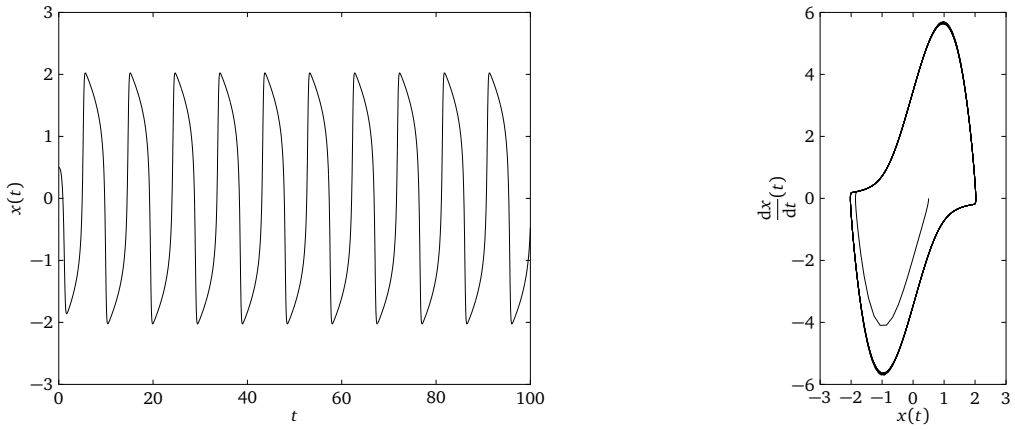


FIGURE 8.4 – Évolution au cours du temps (à gauche) et portrait de phase (à droite) de la solution de l'équation (8.17) vérifiant la condition initiale $x(0) = 0,5$, $\dfrac{\mathrm{d}x}{\mathrm{d}t}(0) = 0$, dans le cas amorti ($\mu = 3,5$).

Des règles, spécifiant dans quelles mesures et proportions les passages ci-dessus s'opèrent, complètent le modèle.

Le *modèle de Kermack–McKendrick* [KM27] est l'un des premiers modèles de ce type. Il a été proposé en 1927 pour expliquer l'augmentation et la diminution rapides du nombre de patients infectés observées lors d'épidémies de peste (à Londres en 1665 et à Bombay en 1896) et de choléra (à Londres en 1865) et suppose que la population est de taille fixée (ce qui se justifie lorsque l'épidémie se déroule sur une courte échelle de temps) et homogène (aucune structure d'âge, de genre, spatiale ou sociale n'est considérée), que la période d'incubation (c'est-à-dire le délai entre la contamination et l'apparition des premiers symptômes de la maladie) est instantanée, et que la durée durant laquelle un individu infecté est contagieux correspond à celle durant laquelle celui-ci est malade.

D'un point de vue mathématique, ce modèle se traduit par un système de trois équations différentielles ordinaires non linéaires couplées, à savoir

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}S}{\mathrm{d}t} & = & -r\,SI, \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} & = & r\,SI - a\,I, \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} & = & a\,I, \end{array} \right. \tag{8.18}$$

ainsi que la donnée d'une condition initiale,

$$S(0) = S_0,\ I(0) = I_0,\ R(0) = R_0, \tag{8.19}$$

avec généralement $S_0 > 0$, $I_0 > 0$, $R_0 = 0$, dans lesquels la variable t désigne le temps (l'instant initial $t = 0$ correspondant au début de l'épidémie), $S(t)$, $I(t)$ et $R(t)$ sont les nombres respectifs de personnes appartenant à chacun des compartiments à l'instant t, r est le taux d'infection et a est le taux de guérison. On note que le fait que la population reste stable au cours du temps est une propriété intrinsèque du modèle puisque, en additionnant les trois équations du système, il vient

$$\frac{\mathrm{d}S}{\mathrm{d}t} + \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{\mathrm{d}R}{\mathrm{d}t} = 0,$$

ce qui implique que $S(t) + I(t) + R(t) = N$ pour tout $t \geq 0$, où $N = S_0 + I_0 + R_0$ est le nombre total d'individus.

Bien que simple, ce modèle possède un certain nombre de propriétés qualitatives intéressantes. Parmi celles-ci, on peut mentionner l'existence d'un nombre, égal à

$$\mathscr{R}_0 = \frac{rS_0}{a}$$

et appelé le *taux de reproduction de base*, gouvernant l'évolution de la solution de ces équations. Heuristiquement, cette quantité représente le nombre moyen attendu de nouveaux cas d'infection, engendrés par un individu infectieux avant sa guérison (ou sa mort), dans une population entièrement constituée d'individus susceptibles. Le « *théorème du seuil* », énoncé par Kermack et McKendrick, fournit alors un critère pour décider de la propagation ou non d'une maladie infectieuse donnée au sein d'une population donnée. Si $\mathscr{R}_0 < 1$, chaque personne ayant contracté la maladie en infecte en moyenne moins d'une autre, conduisant à une disparition totale des malades de la population après quelques temps ; en revanche, si $\mathscr{R}_0 > 1$, chaque cas d'infection produit plusieurs cas secondaires et l'on assiste au développement d'une épidémie.

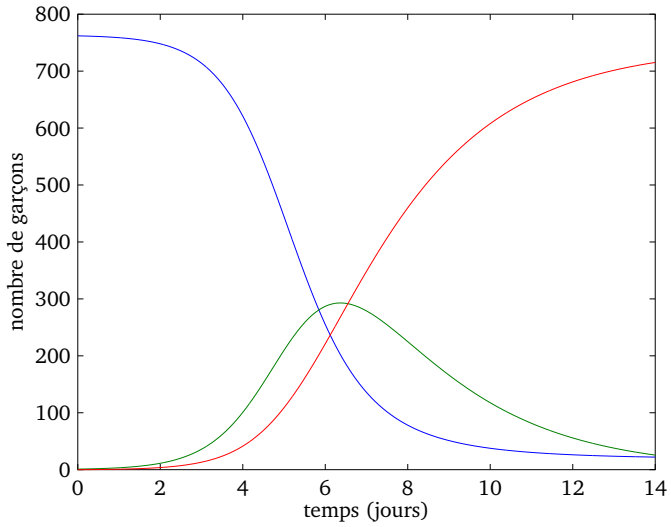


FIGURE 8.5 – Solution numérique du problème (8.18)-(8.19) dont les valeurs des paramètres $S_0 = 762$, $I_0 = 1$, $R_0 = 0$, $r = 0,00218$ et $a = 0,44036$ ont été obtenues par un étalonnage du modèle effectué à partir des données d'une épidémie de grippe dans une école de garçons parues dans la revue médicale britannique *The Lancet* le 4 mars 1978 (voir [Mur02], chapitre 10). Les courbes de couleur bleue, verte et rouge représentent les évolutions respectives des nombres de sujets des catégories S, I et R au cours du temps.

On notera que le modèle de Kermack–McKendrick peut être modifié de diverses manières afin de mieux rendre compte des caractéristiques d'une maladie (transmission indirecte, vecteurs multiples, différents niveaux d'infectiosité...) ou de la structure complexe d'une population (hétérogénéité d'âge, répartition géographique, démographie...) considérée. Un exemple est le *modèle SEIR* (la lettre E étant l'initiale du mot anglais *exposed*), dans lequel un compartiment a été introduit pour traduire le fait qu'un individu susceptible exposé à la maladie n'est généralement pas immédiatement capable de la transmettre, mais seulement après une certaine période de latence. Il est aussi possible un temps moyen d'immunisation au-delà duquel une personne est de nouveau susceptible d'être infectée, quittant ainsi le compartiment R pour réintégrer le compartiment S, ce qui donne lieu à des modèles de type SIRS ou SEIRS.

Dans tous les cas, des simulations numériques, comme celle présentée sur la figure 8.5, permettent d'explorer la gamme de comportements générés par les équations qui composent le modèle et d'améliorer la compréhension de ce dernier, participant ainsi à la définition de stratégies de vaccination ou d'isolement par mise en quarantaine des malades.

8.2.5 Modèle de Lorenz en météorologie

Le *modèle de Lorenz*[20] [Lor63] fut introduit pour rendre compte, de manière déterministe et idéalisée, des interactions entre l'atmosphère terrestre et l'océan, et plus particulièrement des courants convectifs. Il se présente

20. Edward Norton Lorenz (23 mai 1917 - 16 avril 2008) était un mathématicien et météorologue américain, pionnier de la théorie du chaos. Il découvrit la notion d'attracteur étrange et introduisit l'« effet papillon », une expression qui résume de manière métaphorique le problème de la prédictibilité en météorologie.

sous la forme d'un système de trois équations différentielles ordinaires du premier ordre,

$$\left\{\begin{array}{rcl} \dfrac{\mathrm{d}x_1}{\mathrm{d}t} & = & \sigma\,(x_2 - x_1)\,, \\ \dfrac{\mathrm{d}x_2}{\mathrm{d}t} & = & x_1\,(r - x_3) - x_2, \\ \dfrac{\mathrm{d}x_3}{\mathrm{d}t} & = & x_1 x_2 - b\,x_3, \end{array}\right. \tag{8.20}$$

complété d'une condition initiale

$$x_1(0) = 0,\ x_2(0) = 1,\ x_3(0) = 0, \tag{8.21}$$

dont les inconnues x_1, x_2 et x_3 sont des quantités respectivement proportionnelles à l'intensité des mouvements de convection, à l'écart de température entre les courants ascendants et descendants et à la distorsion du profil vertical de température par rapport à un profil linéaire, et où $\sigma = 10$ est le *nombre de Prandtl*[21], $r = 28$ est un *nombre de Rayleigh* réduit et $b = \frac{3}{8}$ est un paramètre associé à la géométrie du problème.

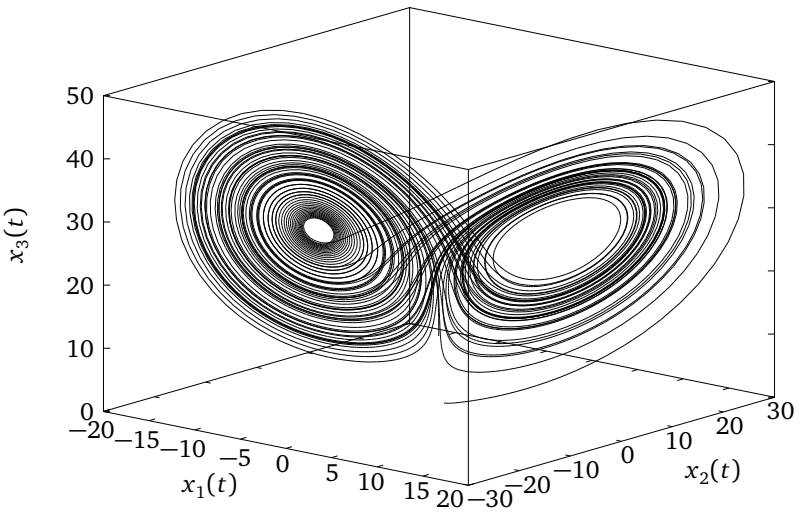


FIGURE 8.6 – Représentation dans l'espace des phases $(x_1(t), x_2(t), x_3(t))$ de l'attracteur étrange du problème de Lorenz obtenu par résolution numérique du problème (8.20)-(8.21) sur l'intervalle $[0, T]$, avec $T = 75$.

Pour les valeurs indiquées des données, issues de considérations physiques, la résolution numérique du problème permit à Lorenz de constater la sensibilité du système dynamique face à des variations de la condition initiale et d'observer que les orbites calculées semblent s'accumuler, pour presque tout choix de condition initiale[22], sur un ensemble compact de structure compliquée, que l'on qualifie d'*attracteur étrange* du fait du comportement chaotique exhibé par les trajectoires (voir les figures 8.6 et 8.7).

21. Ludwig Prandtl (4 février 1875 - 15 août 1953) était un ingénieur et physicien allemand. Il a apporté d'importantes contributions à la mécanique des fluides, notamment en développant les bases mathématiques des principes de l'aérodynamique des écoulements subsoniques et transoniques, ainsi qu'en décrivant le phénomène de couche limite et en mettant en évidence son importance pour l'étude de la traînée.

22. Le système (8.20) possède en effet trois points fixes lorsque $r > 1$, $(0,0,0)$, $(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1)$ et $(\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1)$, qui sont de plus instables pour la valeur choisie $r = 28$.

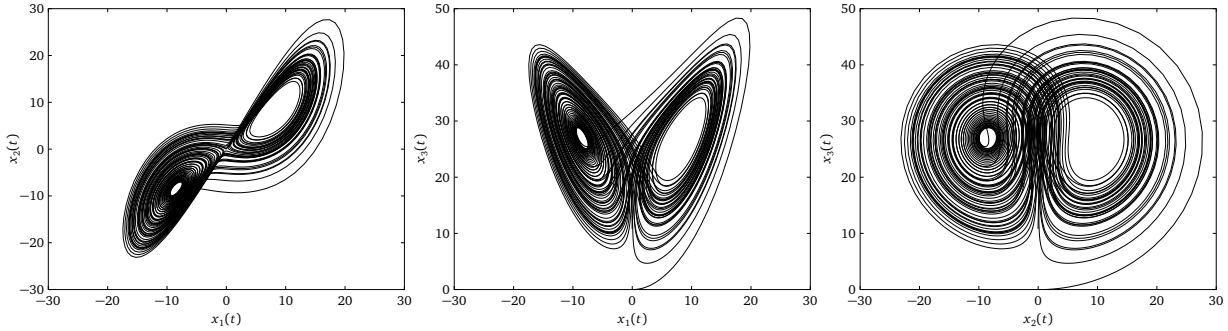


FIGURE 8.7 – Trois portraits de phase de l'attracteur étrange du problème de Lorenz obtenu par résolution numérique du problème (8.20)-(8.21) sur l'intervalle $[0, T]$, avec $T = 75$.

8.2.6 Problème de Robertson en chimie

Le *problème de Robertson* [Rob66] décrit la cinétique d'une réaction chimique autocatalytique [23] mettant en jeu trois espèces chimiques, ici appelées A, B et C, et dont l'équation bilan globale est $A \to C$. Le mécanisme de la réaction peut être décomposé en trois réactions élémentaires, la première,

$$A \to B,$$

étant lente (sa constante de vitesse vaut $k_1 = 0,04\,\mathrm{s}^{-1}$) et décrivant la formation du catalyseur B à partir du réactif A, la deuxième,

$$B + B \to B + C,$$

étant très rapide ($k_2 = 3\ 10^7\,\mathrm{mol}^{-1}\,\mathrm{m}^3\,\mathrm{s}^{-1}$) et correspondant à la formation du produit C de la réaction, et la troisième,

$$B + C \to A + C,$$

étant rapide ($k_3 = 10^4\,\mathrm{mol}^{-1}\,\mathrm{m}^3\,\mathrm{s}^{-1}$) et exprimant la disparition du catalyseur par recomposition. En notant $x_A = [A]$, $x_B = [B]$ et $x_C = [C]$ les concentrations molaires (exprimées en $\mathrm{mol\,m}^{-3}$) respectives des espèces chimiques intervenant dans la réaction, les lois de la cinétique chimique conduisent à modéliser mathématiquement ce problème par le système d'équations différentielles ordinaires suivant

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}x_A}{\mathrm{d}t} & = & -k_1\,x_A + k_3\,x_B x_C, \\ \dfrac{\mathrm{d}x_B}{\mathrm{d}t} & = & k_1\,x_A - k_2\,{x_B}^2 - k_3\,x_B x_C, \\ \dfrac{\mathrm{d}x_C}{\mathrm{d}t} & = & k_2\,{x_B}^2, \end{array} \right. \tag{8.22}$$

que l'on complète d'une condition initiale traduisant le fait que seul le réactif A est présent en début de réaction,

$$x_A(0) = 1,\ x_B(0) = 0,\ x_C(0) = 0. \tag{8.23}$$

La figure 8.8 présente la solution numérique du problème (8.22)-(8.23). On voit que le réactif A est entièrement transformé en produit C et que le catalyseur B disparaît en temps long, ce qui est conforme aux observations des chimistes.

On constate également que la concentration du catalyseur croît de manière très abrupte aux premiers instants de la réaction pour ensuite diminuer lentement et régulièrement. Ce comportement est typique des problèmes de cinétique chimique dans lesquels les différences entre les constantes de vitesse des réactions élémentaires sont grandes. Cette phase transitoire rapide pose des difficultés à certaines méthodes numériques lors du calcul d'une solution approchée. Le système (8.22) constitue à ce titre un exemple de système dit « *raide* » (voir la section 8.7).

23. Une réaction chimique est dite *autocatalytique* si l'un de ses propres produits de réaction est un catalyseur, c'est-à-dire une substance influant sur la vitesse de transformation chimique, pour elle.

8.3 Méthodes numériques de résolution

Exception faite de quelques cas particuliers [24], on ne sait généralement pas donner une forme explicite à la solution d'un problème de Cauchy. Pour cette raison, il est en pratique courant d'approcher cette solution numériquement, soit au moyen d'une méthode analytique, dans laquelle l'approximation de la solution prend généralement la forme d'une série tronquée, soit par une *méthode de discrétisation* (*discrete variable method* en anglais), qui cherche à approcher la solution en un nombre fini de points d'un intervalle donné. C'est à la description de quelques méthodes de ce second type qu'est consacrée la présente section.

Soulignons tout d'abord que l'exposé qui va suivre ne concerne que le cas d'équations différentielles *scalaires*. Si l'adaptation des méthodes présentées à des systèmes d'équations différentielles est relativement directe, l'extension de certains des résultats obtenus dans la section 8.4 ne l'est pas toujours. Les différences ou difficultés les plus marquantes seront soulignées à l'occasion, notamment dans la sous-section 8.4.6 consacrée aux spécificités des systèmes.

On supposera également que la fonction f est définie et continue sur $D = [t_0, t_0 + T] \times \mathbb{R}$, telle qu'il existe une constante L strictement positive telle que

$$\forall t \in [t_0, t_0 + T],\ \forall (x, x_*) \in \mathbb{R}^2,\ |f(t,x) - f(t,x_*)| \leq L\ |x - x_*|\,, \tag{8.24}$$

garantissant que la solution du problème de Cauchy existe et est unique sur $[t_0, t_0 + T]$.

A VOIR : vrai pour les équations différentielles linéaires. On peut discuter de la légitimité de cette hypothèse, mais c'est la condition qui intervient naturellement si l'on veut qu'il existe une solution en temps arbitrairement long (sur $[t_0, t_0 + T]$ avec T arbitrairement grand)

REPRENDRE Plus précisément, nous considérons la résolution numérique du problème de Cauchy (8.1)-(8.5) sur un intervalle $[t_0, t_0 + T]$. Pour cela, une subdivision de $[t_0, t_0 + T]$ en N sous-intervalles $[t_n, t_{n+1}]$, $n = 0, \ldots, N-1$, est réalisée, l'entier N étant destiné à tendre vers l'infini. L'ensemble des points $\{t_n\}_{0 \leq n \leq N}$ est appelé une *grille de discrétisation* (*discretization grid* en anglais) et le scalaire $h_n = t_{n+1} - t_n$, $0 \leq n \leq N-1$, est la longueur du *pas de discrétisation* (*discretization step* en anglais) au point de grille t_n. La *finesse* de la grille se mesure par la quantité

$$h = \max_{0 \leq n \leq N-1} h_n, \tag{8.25}$$

et lorsque $h_0 = h_1 = \cdots = h_{N-1} = h = \frac{T}{N}$, la grille est dite *uniforme*.

L'idée des méthodes de discrétisation est de construire une suite de valeurs [25] $(x_n)_{0 \leq n \leq N}$ approchant aux points de grille la solution x du problème de Cauchy considéré, c'est-à-dire telle que

$$x_n \approx x(t_n),\ 0 \leq n \leq N,$$

24. Pour les équations différentielles ordinaires du premier ordre, on peut par exemple citer les cas des systèmes d'équations linéaires à coefficients constants, des systèmes de dimension deux dont on connaît une intégrale première non triviale, de certaines équations *à variables séparées*, des équations *homogènes* ou encore des équations de Bernoulli (8.2) et de Riccati (8.3).

25. Cette suite sera à valeurs vectorielles si l'on s'intéresse à un système d'équations différentielles ordinaires.

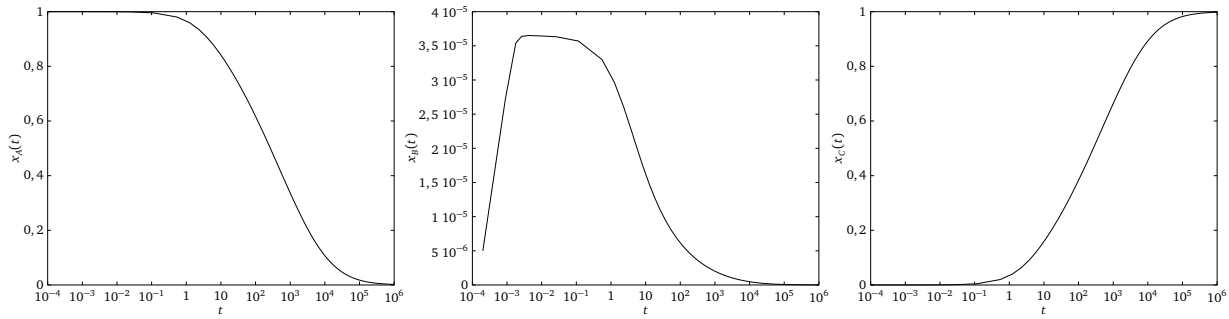


FIGURE 8.8 – Évolution sur l'intervalle $[0, T]$, avec $T = 10^6$ s, des concentrations des espèces chimiques A, B et C obtenues par résolution numérique du problème (8.22)-(8.23).

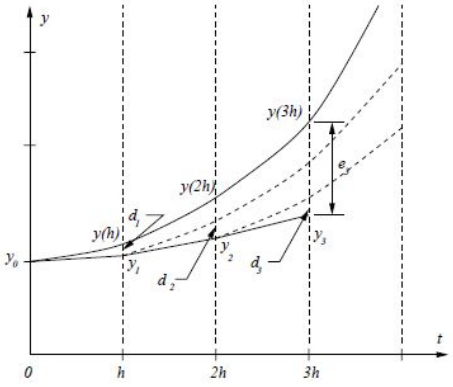


Figure 2.1.1: Exact solution and Euler approximation of $y' = f(t, y)$ with local and global errors shown.

FIGURE 8.9 – DESSIN du type ci-dessus illustrant la construction des approximations par la méthode d'Euler

en un sens qu'il nous faudra préciser. Une approximation continue x_h de la fonction x est alors obtenue par une interpolation linéaire par morceaux (voir la section 6.3) des valeurs x_n, $0 \leq n \leq N$, calculées aux point de la grille de discrétisation, c'est-à-dire

$$\forall n \in \{0, \ldots, N-1\},\ \forall t \in [t_n, t_{n+1}[,\ x_h(t) = x_n + \frac{x_{n+1} - x_n}{h_n}(t - t_n).$$

On peut essentiellement [26] distinguer deux classes de méthodes de discrétisation pour la résolution des équations différentielles ordinaires. Celle des *méthodes à un pas* (où *à pas séparés*) est caractérisée par le fait que, pour tout $n \geq 0$, la valeur approchée x_{n+1} de la solution au point t_{n+1} fournie par la méthode ne dépend que de la valeur x_n calculée à l'étape précédente et ne fait par conséquent intervenir qu'un seul pas de discrétisation. Au contraire, les *méthodes à pas multiples* (également dites *à pas liés*) font appel aux approximations de la solution en un certain nombre d'instants antérieurs t_i, avec $0 \leq i \leq n$, pour déterminer la valeur approchée x_{n+1}.

De manière à pouvoir introduire naturellement plusieurs premières notions fondamentales relatives à l'étude des méthodes, nous allons tout d'abord nous intéresser à l'exemple historique, et particulièrement didactique, de la *méthode d'Euler*.

8.3.1 La méthode d'Euler

La méthode d'Euler (*the forward Euler method* en anglais) est la plus ancienne et certainement la plus simple des méthodes numériques d'approximation des équations différentielles ordinaires du premier ordre. Elle est définie par la relation de récurrence, ou *schéma* (*scheme* en anglais),

$$x_{n+1} = x_n + h_n f(t_n, x_n),\ n = 0, \ldots, N-1, \tag{8.26}$$

la valeur x_0 étant donnée.

En supposant que l'approximation de la solution du problème (8.1)-(8.5) exactement connue en un point t_n, c'est-à-dire que l'on a $x_n = x(t_n)$, on observe que cette méthode revient simplement à construire une approximation de la solution en $t_{n+1} = t_n + h_n$ en confondant sur l'intervalle $[t_n, t_{n+1}]$ la courbe intégrale $x(t)$ issue de $(t_n, x(t_n))$ avec sa tangente en ce même point (voir la figure 8.9). Partant d'une donnée initiale x_0, ce procédé est appliqué sur chaque sous-intervalle de la subdivision de façon à obtenir, par récurrence, des approximations x_n de la solution aux points de grille t_n, $1 \leq n \leq N$.

Une seconde interprétation de la méthode est offerte en considérant l'équation intégrale

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(t, x(t))\,\mathrm{d}t. \tag{8.27}$$

26. Une telle distinction s'avère toutefois quelque peu artificielle. Nous verrons en effet dans la section 8.4 que l'analyse de plusieurs méthodes à un pas « classiques » peut être réalisée dans le cadre de la théorie développée pour des méthodes à pas multiples particulières.

satisfaite par toute solution de l'équation (8.1) sur $[t_n, t_{n+1}]$ en vertu du théorème fondamental de l'analyse (voir le théorème B.131). On voit alors clairement que l'approximation x_{n+1} est obtenue en remplaçant dans l'intégrale la fonction $f(\cdot, x(\cdot))$ par une fonction constante ayant pour valeur $f(t_n, x(t_n))$, ce qui revient encore à approcher l'intégrale présente dans (8.27) par une formule de quadrature interpolatoire, introduite au chapitre 7, qui n'est autre que la règle du rectangle à gauche (considérer la formule (7.7)).

On ne manquera pas remarquer que l'on aurait pu tout aussi bien choisir une autre formule de quadrature pour l'évaluation de l'intégrale, comme la règle du rectangle à droite (voir la formule (7.8)) ou encore celle du trapèze (voir la formule (7.10)), ce qui aurait respectivement donné lieu à la *méthode d'Euler implicite* (*the backward Euler method* en anglais)

$$x_{n+1} = x_n + h_n f(t_{n+1}, x_{n+1}),\ n = 0, \ldots, N-1, \tag{8.28}$$

ou à la *méthode de la règle du trapèze* (*trapezoidal rule* en anglais)

$$x_{n+1} = x_n + \frac{h_n}{2}\left(f(t_n, x_n) + f(t_{n+1}, x_{n+1})\right),\ n = 0, \ldots, N-1. \tag{8.29}$$

Il est cependant important de souligner que ces deux modifications sont lourdes de conséquences en pratique. En effet, pour déterminer la valeur x_{n+1} à partir de celle de x_n, il faut résoudre une équation *a priori* non linéaire. Pour cette raison, les méthodes définies par les relations de récurrence (8.28) et (8.29) sont qualifiées d'*implicites*, alors celle basée sur (8.26) est dite *explicite*. On peut d'ores et déjà noter que l'utilisation d'une méthode implicite pose des questions d'existence et d'unicité d'un point de vue théorique et de mise en œuvre d'un point de vue calculatoire, mais, de manière assez typique en analyse numérique, l'effort supplémentaire demandé par rapport à l'emploi d'une méthode explicite se trouve compensé par le renforcement de certaines propriétés. Nous reviendrons en détail sur ces points plus loin.

Évidemment, de telles méthodes de discrétisation ne sont intéressantes que si elles permettent d'approcher numériquement la solution du problème (8.1)-(8.5), et ceci d'autant mieux que la grille de discrétisation est fine (voir la figure 8.10). En effet, lorsque le paramètre h tend vers 0, le nombre de points de grille tend vers l'infini et la grille elle-même tend vers l'intervalle $[t_0, t_0 + T]$. On s'attend alors à ce que l'approximation fournie par la méthode tende vers la solution du problème et l'on dit que la méthode *converge*. Ceci se traduit mathématiquement par le fait que l'*erreur globale* de la méthode $x_{n+1} - x(t_{n+1})$ au point t_{n+1}, $n = 0, \ldots, N-1$, tend vers zéro lorsque h tend vers zéro, sous réserve que $x_0 = x(t_0)$ (éventuellement à la limite). Nous allons maintenant prouver que c'est le cas pour la méthode d'Euler.

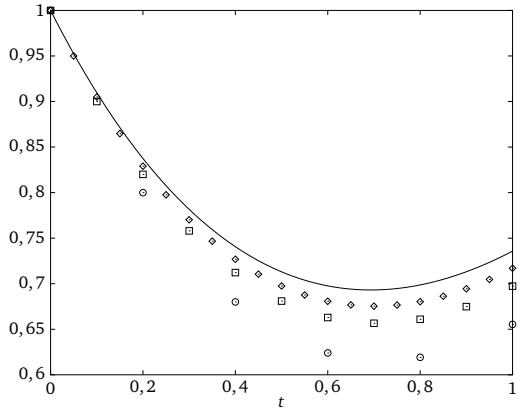


FIGURE 8.10 – Illustration de la convergence de la méthode d'Euler pour la résolution du problème de Cauchy d'équation $x'(t) = t - x(t)$ et de condition initiale $x(0) = 1$, sur l'intervalle $[0, 1]$. La courbe représente le graphe de la solution du problème, $x(t) = t - 1 + 2e^{-t}$, tandis que les familles de points désignés par les symboles $\circ$, $\square$ et $\diamond$ sont obtenues au moyen d'une résolution numérique par la méthode d'Euler sur des grilles de discrétisation uniformes dont les longueurs de pas respectives sont $h = 0,2$, $h = 0,1$ et $h = 0,05$.

En utilisant la définition (8.26) de la méthode et l'équation différentielle (8.1), on peut écrire, pour $n =$

$0,\dots,N-1$,

$$\begin{aligned} x_{n+1}-x(t_{n+1}) &= x_n + h_n f(t_n,x_n) - x(t_{n+1}) \\ &= x_n - x(t_n) + x(t_n) + h_n x'(t_n) - h_n f(t_n,x(t_n)) + h_n f(t_n,x_n) - x(t_{n+1}). \end{aligned}$$

En vertu de la condition de Lipschitz (8.24) satisfaite par f, il vient alors

$$|x_{n+1}-x(t_{n+1})| \leq (1+h_nL)\,|x_n-x(t_n)| + |\tau_{n+1}|\,,\ n=0,\dots,N-1, \tag{8.30}$$

où l'on a introduit l'*erreur de troncature locale* de la méthode au point t_{n+1} en posant

$$\tau_{n+1} = x(t_{n+1}) - x(t_n) - h_n\,x'(t_n),\ n=0,\dots,N-1.$$

La solution x étant de classe $\mathscr{C}^1$ sur l'intervalle $[t_0,t_0+T]$, on sait, par application du théorème des accroissements finis (voir le théorème B.113), qu'il existe, pour tout entier n compris entre 0 et $N-1$, un réel ζ_n appartenant à l'intervalle $]t_n,t_{n+1}[$ tel que

$$\tau_{n+1} = h_n\left(x'(\zeta_n)-x'(t_n)\right).$$

On a par conséquent

$$\sum_{n=0}^{N-1}|\tau_{n+1}| = \sum_{n=0}^{N-1} h_n\left|x'(\zeta_n)-x'(t_n)\right| \leq \omega(x',h)\sum_{n=0}^{N-1}h_n = \omega(x',h)\,T,$$

où $\omega(x',\cdot)$ désigne le module de continuité de la fonction x', qui est uniformément continue sur l'intervalle $[t_0,t_0+T]$ en vertu du théorème de Heine (voir le théorème B.95), ce qui montre que

$$\lim_{h\to 0}\sum_{n=0}^{N-1}|\tau_{n+1}| = 0. \tag{8.31}$$

Cette propriété de consistance de la méthode d'Euler est une condition nécessaire à sa convergence.

Revenons à la majoration de l'erreur de la méthode. En utilisant récursivement (8.30), on arrive à

$$|x_{n+1}-x(t_{n+1})| \leq \left(\prod_{i=0}^{n}(1+h_iL)\right)|x_0-x(t_0)| + \sum_{i=0}^{n}\left(\prod_{j=i+1}^{n}(1+h_jL)\right)|\tau_{i+1}|\,, n=0,\dots,N-1.$$

En remarquant que $\prod_{i=0}^{n}(1+h_iL) \leq e^{L\sum_{i=0}^{n}h_i}$ et que $\sum_{i=0}^{n}h_i \leq \sum_{i=0}^{N-1}h_i = T$, $n=0,\dots,N-1$, on obtient finalement que

$$|x_{n+1}-x(t_{n+1})| \leq e^{LT}\left(|x_0-x(t_0)| + \sum_{i=0}^{n}|\tau_{i+1}|\right), n=0,\dots,N-1,$$

cette dernière inégalité caractérisant la *stabilité* de la méthode. On déduit alors la *convergence* de la méthode de la propriété de consistance (8.31) si l'on a par ailleurs

$$\lim_{h\to 0}|x_0-x(t_0)|\,.$$

On peut établir une estimation plus précise de la vitesse à laquelle la méthode converge lorsque la longueur du pas de discrétisation tend vers zéro en supposant que la fonction f est de classe $\mathscr{C}^1$. Dans ce cas, la solution x est de classe $\mathscr{C}^2$ et l'on trouve [27], en utilisant la formule de Taylor–Lagrange (voir le théorème B.116),

$$|x_{n+1}-x(t_{n+1})| \leq e^{LT}\left(|x_0-x(t_0)| + \frac{MT}{2}h\right), n=0,\dots,N-1,$$

27. On a en effet que

$$\forall n\in\{0,\dots,N-1\},\ \tau_{n+1} = \frac{{h_n}^2}{2}x''(\xi_n),\ \text{avec } \xi_n\in]t_n,t_{n+1}[,$$

dont on déduit alors que

$$\sum_{n=0}^{N-1}|\tau_{n+1}| \leq \frac{1}{2}h\sum_{n=0}^{N-1}h_n\left|x''(\xi_n)\right| \leq \frac{M}{2}h\sum_{n=0}^{N-1}h_n = \frac{MT}{2}h,$$

en posant $M = \max_{t\in[t_0,t_0+T]}\left|x''(t)\right|$.

où M est une constante positive majorant la fonction $|x''|$ sur l'intervalle $[t_0, t_0+T]$. Si $|x_0 - x(t_0)| = O(h)$, on trouve que $|x_n - x(t_n)| = O(h)$, $n = 0, \ldots, N$. On dit alors que (la convergence de) la méthode est *d'ordre un*.

L'analyse que nous venons d'effectuer amène plusieurs remarques. La première est que la convergence de la méthode d'Euler repose sur les propriétés fondamentales de consistance et de stabilité ; il en sera de même pour toutes les méthodes que nous présenterons dans ce chapitre. La seconde est que la méthode d'Euler n'est pas très précise et contraint à employer une grille de discrétisation fine si l'on souhaite que l'erreur globale soit petite, ce qui a des répercussions sur le coût de calcul de la méthode.

Il est possible de construire des méthodes de discrétisation dont la précision est meilleure. Pour cela, on recense essentiellement trois [28] manières de faire. On peut tout d'abord utiliser plus d'une évaluation de la fonction f à chaque étape pour obtenir la valeur approchée de la solution (voir la sous-section 8.3.2) ou bien faire dépendre cette valeur de plus d'une valeur précédemment calculée (voir la sous-section 8.3.3). Enfin, on peut également se servir d'évaluations des dérivées de la fonction f, lorsque cette dernière est suffisamment régulière, dans le schéma de la méthode (voir la sous-section 8.3.4).

COMPLETER la preuve si f définie sur $D = [t_0, t_0+T] \times \Omega$ avec Ω un ouvert de $\mathbb{R}$ (on doit justifier l'existence de la suite $(x_n)_n$ en montrant que $(t_n, x_n) \in D$, $\forall n$)

Dans le cas où la fonction f est de classe $\mathscr{C}^1$: l'intervalle $[t_0, t_0+T]$ étant compact et la solution x du problème de Cauchy étant continue, le graphe de x possède un voisinage compact K inclus dans D (construction : pour $t \in [t_0, t_0+T]$, on choisit $\zeta(t) > 0$ tel que la boule de centre $(t, x(t))$ et de rayon $\zeta(t)$ soit contenue dans D. Le graphe de x étant compact, on peut extraire un nombre fini des boules précédentes qui recouvrent le graphe, la réunion de l'adhérence de ces boules fournissant un voisinage convenable). En utilisant alors l'inégalité de stabilité eqref, on choisit $|x_0 - x(t_0)|$ et h suffisamment petits pour que

$$\forall n,\ \forall y \in \left[x_n - e^{LT}\left(|x_0 - x(t_0)| + \frac{MT}{2}h\right), x_n + e^{LT}\left(|x_0 - x(t_0)| + \frac{MT}{2}h\right)\right],\ (t_n, y) \in K.$$

On alors par un raisonnement par récurrence que $\forall n$, x_n est bien défini et satisfait l'inégalité eqref. Ceci est en effet trivial pour $n = 0$ compte tenu des choix faits. Par ailleurs, si x_n satisfait eqref, alors cette inégalité est suffisante pour assurer l'existence de x_{n+1}, ce qui permt d'effectuer les calculs aboutissant au fait que x_{n+1} satisfait eqref au rang $n+1$.

8.3.2 Méthodes de Runge–Kutta

Les *méthodes de Runge–Kutta* [29] [Run95, Kut01] visent à étendre à la résolution d'équations différentielles ordinaires l'usage des techniques de calcul approché d'intégrales que sont les formules de quadrature interpolatoires. Elles font pour cela appel à de multiples évaluations de la fonction f, en des point obtenus par substitutions successives (pour les méthodes explicites), sur chaque sous-intervalle de la grille de discrétisation, cet « échantillonnage » de la dérivée de la courbe intégrale recherchée permettant de réaliser l'intégration numérique approchée de cette dernière au moyen d'une somme pondérée des valeurs recueillies. Pour illustrer cette idée, donnons un premier exemple.

Exemple de méthode de Runge–Kutta explicite. Considérons une modification de la méthode d'Euler proposée par Runge dans [Run95]. Supposons que l'on connaisse la valeur $x(t_n)$ et que l'on cherche à calculer une approximation x_{n+1} d'ordre deux de $x(t_{n+1})$. Pour cela, il semble naturel d'utiliser une formule de quadrature comme la règle du point milieu (voir la formule (7.9)), dont on sait que le degré d'exactitude est égal à un, en place de la formule du rectangle à gauche (dont le degré d'exactitude vaut zéro). Cette dernière nécessite cependant d'évaluer la fonction f au point $(t_n + \frac{h_n}{2}, x(t_n + \frac{h_n}{2}))$, sachant que la seule valeur à disposition est $x_n = x(t_n)$... L'idée est de se servir de la méthode d'Euler, sur l'intervalle $[t_n, t_n + \frac{h_n}{2}]$, pour approcher cette valeur inconnue. On obtient ainsi la *méthode d'Euler modifiée*, parfois appelée *méthode du point milieu explicite*, dont le schéma s'écrit

$$x_{n+1} = x_n + h_n f\left(t_n + \frac{h_n}{2}, x_n + \frac{h_n}{2} f(t_n, x_n)\right), n = 0, \ldots, N-1. \tag{8.32}$$

28. On trouve également dans la littérature d'autres classes de méthodes combinant ces trois approches (voir la section 8.9 en fin de chapitre).

29. Martin Wilhelm Kutta (3 novembre 1867 - 25 décembre 1944) était un mathématicien allemand. Il développa avec Carl Runge une méthode de résolution numérique des équations différentielles aujourd'hui très utilisée. En aérodynamique, son nom est associé à une condition permettant de déterminer la circulation autour d'un profil d'aile et, par suite, d'en déduire la portance.

On voit que deux évaluations successives de la fonction f sont nécessaires pour faire avancer d'un pas la solution numérique et que la méthode sacrifie la dépendance linéaire (entre x_{n+1} et x_n d'une part et $f(t_n, x_n)$ et/ou $f(t_{n+1}, x_{n+1})$ d'autre part) qui existe dans les méthodes d'Euler ou de la règle du trapèze. En contrepartie, on peut montrer que cette méthode est effectivement d'ordre deux.

En toute généralité, une méthode de Runge–Kutta à s niveaux pour la résolution du problème de Cauchy (8.1)-(8.5) est définie par

$$x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i\, k_i,\ k_i = f\Big(t_n + c_i\, h_n, x_n + h_n \sum_{j=1}^{s} a_{ij} k_j\Big),\ i = 1, \ldots, s,\ n = 0, \ldots, N-1, \tag{8.33}$$

la valeur x_0 étant donnée. Une telle méthode est donc entièrement caractérisée par la donnée des coefficients $\{a_{ij}\}_{1\leq i,j\leq s}$, $\{b_i\}_{1\leq i\leq s}$ et $\{c_i\}_{1\leq i\leq s}$, que l'on a coutume de présenter, depuis la publication de l'article [But64a], dans le *tableau de Butcher*[30] (*Butcher tableau* en anglais) suivant

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{22} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss} \\ \hline & b_1 & b_2 & \ldots & b_s \end{array} \tag{8.34}$$

que l'on peut encore écrire, en introduisant la matrice carrée A d'ordre s et les vecteurs $\boldsymbol{b}$ et $\boldsymbol{c}$,

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^T \end{array}.$$

Dans la suite, nous supposerons[31] que les méthodes sont telles que leurs coefficients vérifient les conditions suivantes

$$c_i = \sum_{j=1}^{s} a_{ij},\ i = 1, \ldots, s, \tag{8.35}$$

qui garantissent que tous les points en lesquels la fonction f est évaluée sont des approximations du premier ordre de la solution, simplifiant ainsi grandement l'écriture des conditions d'ordre que doit satisfaire une méthode de Runge–Kutta d'ordre élevé.

On observe que si $a_{ij} = 0$, $1 \leq i \leq j \leq s$, chacune des quantités k_i s'exprime uniquement en fonction de valeurs k_j, $1 \leq j < i \leq s$, déjà connues et la méthode de Runge–Kutta est dite *explicite* (*explicit Runge–Kutta method* en anglais, *ERK* en abrégé). Si ce n'est pas le cas, la méthode est dite *implicite*[32] (*implicit Runge–Kutta method* en anglais, *IRK* en abrégé) et la quantité x_{n+1} est alors définie de manière unique sous condition (voir l'inégalité (8.44)). La méthode d'Euler sous sa forme explicite (resp. implicite), définie par (8.26) (resp. (8.28)), est l'exemple le plus simple[33] de méthode de Runge–Kutta explicite (resp. implicite).

30. John Charles Butcher (né le 31 mars 1933) est un mathématicien néo-zélandais spécialisé dans l'étude des méthodes de résolution numérique des équations différentielles ordinaires.

31. Il s'avère possible d'obtenir des méthodes explicites, à deux ou trois niveaux et d'ordre maximal, en se passant de ces hypothèses (voir [Oli75]).

32. En se référant au tableau de Butcher (8.34), on voit qu'une méthode de Runge–Kutta est explicite si et seulement si la matrice A est strictement triangulaire inférieure. Si A est seulement triangulaire inférieure, c'est-à-dire si $a_{ij} = 0$ pour tout couple (i, j) de $\{1, \ldots, s\}^2$ tel que $i < j$ mais qu'au moins l'un des coefficients a_{ii}, $1 \leq i \leq s$, est non nul, la méthode est *semi-implicite* (*semi-implicit Runge–Kutta method* en anglais). On la qualifie alors en anglais de *diagonally implicit Runge–Kutta method* (*DIRK* en abrégé) ou, si tous les éléments diagonaux a_{ii}, $i = 1, \ldots, s$, sont identiques, de *singly-diagonally implicit Runge–Kutta method* (*SDIRK* en abrégé). Elle est implicite dans tout autre cas.

33. Vue comme une méthode de Runge–Kutta à un niveau, cette méthode a en effet pour tableau de Butcher associé

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \text{ (resp. } \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}\text{)}$$

à sa version explicite (resp. implicite).

Construction d'une méthode de Runge–Kutta explicite pour une équation scalaire

Jusqu'aux travaux de Butcher dans les années 1960, seules les méthodes de Runge–Kutta explicites étaient considérées et la dérivation de leurs coefficients était une tâche fastidieuse, demandant de nombreux calculs et d'autant plus ardue que l'ordre demandé pour la méthode est élevé. La technique généralement utilisée pour ce faire consiste à faire correspondre, jusqu'à l'ordre souhaité, le développement de Taylor au point t_n d'une solution régulière du problème de Cauchy en t_{n+1} avec celui d'une approximation numérique au même point, notée $\tilde{x}_{n+1}$, fournie par la méthode en supposant que

$$x_n = x(t_n) \tag{8.36}$$

(on parle d'*hypothèse localisante*). Nous allons maintenant illustrer cette approche en construisant des méthodes de Runge–Kutta explicites de un à trois niveaux et d'ordre identique au nombre de niveaux.

D'après la définition (8.33) et les conditions (8.35), le schéma d'une méthode de Runge–Kutta explicite à trois niveaux peut s'écrire, sous l'hypothèse localisante (8.36),

$$\begin{array}{rcl} \tilde{x}_{n+1} & = & x(t_n) + h_n\,(b_1\,k_1 + b_2\,k_2 + b_3\,k_3), \\ k_1 & = & f(t_n, x(t_n)), \\ k_2 & = & f(t_n + h_n\,c_2, x(t_n) + h_n\,c_2 k_1), \\ k_3 & = & f(t_n + h_n\,c_3, x(t_n) + h_n\,(c_3 - a_{32})k_1 + h_n\,a_{32}k_2). \end{array} \tag{8.37}$$

Par ailleurs, en supposant la fonction f suffisamment régulière ainsi qu'en utilisant et dérivant de façon répétée l'équation (8.1), on obtient le développement de Taylor de $x(t_{n+1})$ suivant

$$\begin{aligned} x(t_{n+1}) = x(t_n) &+ h_n f(t_n, x(t_n)) \\ &+ \frac{{h_n}^2}{2}\left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n))\frac{\partial f}{\partial x}(t_n, x(t_n))\right) \\ &+ \frac{{h_n}^3}{6}\left(\frac{\partial^2 f}{\partial t^2}(t_n, x(t_n)) + 2f(t_n, x(t_n))\frac{\partial^2 f}{\partial t\,\partial x}(t_n, x(t_n)) + (f(t_n, x(t_n)))^2\frac{\partial^2 f}{\partial x^2}(t_n, x(t_n))\right. \\ &\left. + \left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n))\frac{\partial f}{\partial x}(t_n, x(t_n))\right)\frac{\partial f}{\partial x}(t_n, x(t_n))\right) + O({h_n}^4). \end{aligned} \tag{8.38}$$

Il reste à effectuer des développements de similaires pour les quantités k_2 et k_3 dans le schéma (8.37). On trouve

$$\begin{aligned} k_2 = f(t_n, x(t_n)) &+ h_n\,c_2\left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + k_1\frac{\partial f}{\partial x}(t_n, x(t_n))\right) \\ &+ \frac{{h_n}^2}{2}{c_2}^2\left(\frac{\partial^2 f}{\partial t^2}(t_n, x(t_n)) + 2k_1\frac{\partial^2 f}{\partial t\,\partial x}(t_n, x(t_n)) + {k_1}^2\frac{\partial^2 f}{\partial x^2}(t_n, x(t_n))\right) + O({h_n}^3), \end{aligned}$$

et

$$\begin{aligned} k_3 = f(t_n, x(t_n)) &+ h_n\left(c_3\frac{\partial f}{\partial t}(t_n, x(t_n)) + ((c_3 - a_{32})k_1 + a_{32}k_2)\,\frac{\partial f}{\partial x}(t_n, x(t_n))\right) \\ &+ \frac{{h_n}^2}{2}\left({c_3}^2\frac{\partial^2 f}{\partial t^2}(t_n, x(t_n)) + 2c_3\,((c_3 - a_{32})k_1 + a_{32}k_2)\,\frac{\partial^2 f}{\partial t\,\partial x}(t_n, x(t_n))\right. \\ &\left. + ((c_3 - a_{32})k_1 + a_{32}k_2)^2\,\frac{\partial^2 f}{\partial x^2}(t_n, x(t_n))\right) + O({h_n}^3). \end{aligned}$$

En substituant ces expressions dans (8.37) et en ne conservant que les termes d'ordre inférieur ou égal à trois en h_n, on obtient finalement

$$\begin{aligned} \tilde{x}_{n+1} = x(t_n) &+ h_n\,(b_1 + b_2 + b_3)\,f(t_n, x(t_n)) \\ &+ {h_n}^2\,(b_2c_2 + b_3c_3)\left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n))\frac{\partial f}{\partial x}(t_n, x(t_n))\right) \\ &+ \frac{{h_n}^3}{2}\left(\left(b_2{c_2}^2 + b_3{c_3}^2\right)\left(\frac{\partial^2 f}{\partial t^2}(t_n, x(t_n)) + 2f(t_n, x(t_n))\frac{\partial^2 f}{\partial t\,\partial x}(t_n, x(t_n)) + (f(t_n, x(t_n)))^2\frac{\partial^2 f}{\partial x^2}(t_n, x(t_n))\right)\right. \\ &\left. + 2b_3c_2a_{32}\left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n))\frac{\partial f}{\partial x}(t_n, x(t_n))\right)\frac{\partial f}{\partial x}(t_n, x(t_n))\right) + O({h_n}^4). \end{aligned} \tag{8.39}$$

Il faut à présent essayer de faire coïncider les termes de ce dernier développement avec ceux de (8.38) en fonction du nombre de niveaux de la méthode. Dans le cas d'une méthode à un niveau, on pose $b_2 = b_3 = 0$ et on a simplement

$$\tilde{x}_{n+1} = x(t_n) + h_n b_1 f(t_n, x(t_n)).$$

Il vient alors $b_1 = 1$, ce qui conduit à une unique méthode de Runge-Kutta explicite à un niveau, qui n'est autre que la méthode d'Euler (8.26) introduite dans la sous-section précédente.

Pour construire une méthode à deux niveaux, on pose $b_3 = 0$ et le développement (8.39) devient alors

$$\begin{aligned}\tilde{x}_{n+1} = x(t_n) + h_n (b_1 + b_2) f(t_n, x(t_n)) + {h_n}^2 b_2 c_2 \left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n)) \frac{\partial f}{\partial x}(t_n, x(t_n)) \right) \\ + \frac{{h_n}^3}{2} b_2 {c_2}^2 \left(\frac{\partial^2 f}{\partial t^2}(t_n, x(t_n)) + 2 f(t_n, x(t_n)) \frac{\partial^2 f}{\partial t \partial x}(t_n, x(t_n)) + (f(t_n, x(t_n)))^2 \frac{\partial^2 f}{\partial x^2}(t_n, x(t_n)) \right) + O({h_n}^4).\end{aligned}$$

Les premiers termes de ce développement sont ceux de (8.38) si l'on impose que

$$\begin{aligned} &b_1 + b_2 = 1, \\ &b_2 c_2 = \tfrac{1}{2}. \end{aligned} \tag{8.40}$$

Ce système de deux équations à trois inconnues possède une famille infinie de solutions dépendantes d'un paramètre, illustrant le fait que les méthodes de Runge–Kutta explicites de nombre de niveaux et d'ordre donnés ne sont généralement pas définies de manière unique. On remarque également qu'aucune solution ne conduit à une méthode d'ordre plus haut que deux.

Exemples de méthode de Runge–Kutta explicite à deux niveaux et d'ordre deux. Deux solutions particulières du système (8.40) conduisent à des méthodes de Runge–Kutta connues. Ce sont respectivement la méthode d'Euler modifiée (8.32), de tableau de Butcher associé [34]

$$\begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

et la *méthode de Heun* [35] *d'ordre deux* [Run95, Heu00]

$$x_{n+1} = x_n + \frac{h_n}{2} \left(f(t_n, x_n) + f\left(t_n + h_n, x_n + h_n f(t_n, x_n)\right) \right), \tag{8.41}$$

de tableau de Butcher associé

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

En procédant de la même façon pour une méthode à trois niveaux, on arrive aux quatre conditions suivantes

$$\begin{aligned} &b_1 + b_2 + b_3 = 1, \\ &b_2 c_2 + b_3 c_3 = \tfrac{1}{2}, \\ &b_2 {c_2}^2 + b_3 {c_3}^2 = \tfrac{1}{3}, \\ &b_3 c_2 a_{32} = \tfrac{1}{6}, \end{aligned} \tag{8.42}$$

faisant intervenir six inconnues. Les solutions de ce système forment une famille infinie dépendant de deux paramètres, aucune ne menant à une méthode d'ordre plus haut que trois.

34. Dans la suite, on omet de noter la partie triangulaire supérieure identiquement nulle de la matrice A dans les tableaux de Butcher des méthodes de Runge–Kutta explicites.

35. Karl Heun (3 avril 1859 - 10 janvier 1929) était un mathématicien allemand, connu pour ses travaux sur les équations différentielles.

Exemples de méthode de Runge–Kutta explicite à trois niveaux et d'ordre trois. Parmi les solutions du système (8.42) deux conduisent à des méthodes connues qui sont la *méthode de Heun d'ordre trois* [Heu00], de tableau de Butcher associé

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3} & & \\ \frac{2}{3} & 0 & \frac{2}{3} & \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array},$$

et la *méthode de Kutta d'ordre trois* [Kut01], de tableau de Butcher associé

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}.$$

L'obtention de méthodes d'ordre supérieur est un travail laborieux, le nombre de conditions augmentant rapidement avec l'ordre : il faut en satisfaire 8 pour atteindre l'ordre 4, 37 pour l'ordre 6, 200 pour l'ordre 8, 1205 pour l'ordre 10... Cette manière de faire n'est donc pas viable [36] pour la dérivation de méthodes d'ordre élevé. De plus, elle ne s'applique pas au cas des systèmes d'équations différentielles ordinaires, pour lesquels la fonction f est à valeurs vectorielles. C'est en réalité dans un cadre algébrique, celui de la *théorie de Butcher*, que la structure générale des conditions d'ordre des méthodes de Runge–Kutta se trouve révélée. Nous renvoyons le lecteur intéressé aux notes de fin de chapitre pour des références sur ce sujet.

Exemples de méthode de Runge–Kutta explicite à quatre niveaux et d'ordre quatre. Une méthode de Runge–Kutta explicite d'ordre quatre, due à Kutta [Kut01], extrêmement populaire [37], au point d'être appelée *« la » méthode de Runge–Kutta*, et généralisant la règle de Simpson (voir la formule (7.11)) est celle donnée par le tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}. \tag{8.43}$$

Une autre méthode d'ordre quatre, également découverte par Kutta, généralisant la règle des trois huitièmes (voir la table 7.1), a pour tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{3} & \frac{1}{3} & & & \\ \frac{2}{3} & -\frac{1}{3} & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array}.$$

On trouvera dans l'article [Ral62] des choix de coefficients minimisant l'erreur de troncature locale pour des méthodes explicites à deux, trois et quatre niveaux d'ordre maximal.

Méthodes de Runge–Kutta implicites

Une première question naturelle se posant pour une méthode de Runge–Kutta implicite est celle de l'existence des quantités k_i, $i = 1, \dots, s$, solutions du système (8.33), la matrice carrée A définissant une méthode implicite n'étant pas strictement triangulaire inférieure et la fonction f étant *a priori* non linéaire. En supposant que cette dernière satisfait la condition de Lipschitz globale [38] (8.24) par rapport à la variable x, on montre que les conditions

36. On pourra consulter l'article [Hut'56] sur l'obtention de méthodes de Runge–Kutta explicites d'ordre six pour s'en convaincre.

37. Dans [Lam91], cette popularité historique est imputée aux valeurs des coefficients de la matrice A et du vecteur $\mathbf{c}$ associés à la méthode, qui facilitaient très probablement les évaluations de la fonction f sur un calculateur mécanique.

38. REPRENDRE Si la condition de Lipschitz est seulement satisfaite dans un voisinage de la condition initiale, des restrictions additionnelles doivent être faites sur les pas de discrétisation pour s'assurer que les points en lesquels on évalue la fonction f appartiennent à ce voisinage. L'unicité des valeurs k_i, $i = 1, \dots, s$, est alors de nature locale.

du théorème 5.9 sont vérifiées dès que les longueurs des pas de la grille de discrétisation satisfont

$$h_n < \frac{1}{L\,\|A\|_\infty},\ n = 0,\dots,N-1, \tag{8.44}$$

et les réels k_i, $i = 1,\dots,s$, sont alors définis de manière unique. On notera que cette restriction sur la finesse de la grille peut être particulièrement sévère pour la résolution d'un système raide, la constante de Lipschitz L étant souvent très grande dans ce cas.

Il découle de ces considérations que les méthodes de Runge–Kutta implicites sont bien plus coûteuses en temps de calcul et bien plus difficiles à implémenter que leurs analogues explicites, les valeurs k_i, $i = 1,\dots,s$, devant généralement être toutes calculées concurremment et non plus successivement à chaque étape, ce qui induit un effort de calcul important que nous allons détailler.

Réécrivons tout d'abord (8.33) sous la forme

$$x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i\, f(t_n + c_i\, h_n, x_n + X_i),\ n = 0,\dots,N-1, \tag{8.45}$$

$$X_i = x_n + h_n \sum_{j=1}^{s} a_{ij}\, f(t_n + c_j\, h_n, x_n + X_j),\ i = 1,\dots,s,\ n = 0,\dots,N-1. \tag{8.46}$$

On peut alors résoudre à chaque étape le système d'équations non linéaires (8.46) par la méthode des approximations successives introduite la section 5.4.1, dont la relation de récurrence est ici

$$X_i^{(k+1)} = x_n + h_n \sum_{j=1}^{s} a_{ij}\, f(t_n + c_j\, h_n, x_n + X_j^{(k)}),\ i = 1,\dots,s,\ k \geq 0, \tag{8.47}$$

et qui est convergente pour tout choix de valeurs d'initialisation $X_i^{(0)}$, $i = 1,\dots,s$, si la condition (8.44) est satisfaite. Chacune des itérations (8.47) demande s évaluations de la fonction f, $s(s+1)d$ multiplication et $s^2 d$ additions si l'on résout un système de d équations différentielles ordinaires scalaires. Une fois les valeurs X_i obtenues, $i = 1,\dots,s$, l'approximation de la solution au point t_{n+1} est calculée au moyen de la formule (8.45), ce qui nécessite [39] encore s évaluations de la fonction f, $sd+1$ multiplications et $(s-1)d+1$ additions.

Les méthodes de Runge–Kutta implicites étant en pratique quasiment exclusivement réservées au traitement des systèmes raides, les itérations de point fixe (8.47) ne convergent dans de nombreux cas que pour des longueurs de pas de discrétisation excessivement petites et l'on a alors avantage à recourir à la méthode de Newton–Raphson. Dans le cas d'un système de d équations différentielles ordinaires scalaires, l'application de la méthode à l'équation (8.46) conduit à la relation de récurrence

$$\begin{pmatrix} I_d - h_n a_{11} \frac{\partial f}{\partial x}(t_n + c_1 h_n, x_n + X_1^{(k)}) & \dots & -h_n a_{1s} \frac{\partial f}{\partial x}(t_n + c_s h_n, x_n + X_s^{(k)}) \\ \vdots & \ddots & \vdots \\ -h_n a_{s1} \frac{\partial f}{\partial x}(t_n + c_1 h_n, x_n + X_1^{(k)}) & \dots & I_d - h_n a_{ss} \frac{\partial f}{\partial x}(t_n + c_s h_n, x_n + X_s^{(k)}) \end{pmatrix} \begin{pmatrix} X_1^{(k+1)} - X_1^{(k)} \\ \vdots \\ X_s^{(k+1)} - X_s^{(k)} \end{pmatrix}$$
$$= \begin{pmatrix} x_n + h_n \sum_{j=1}^{s} a_{1j} f(t_n + c_j h_n, x_n + X_j^{(k)}) - X_1^{(k)} \\ \vdots \\ x_n + h_n \sum_{j=1}^{s} a_{sj} f(t_n + c_j h_n, x_n + X_j^{(k)}) - X_s^{(k)} \end{pmatrix},\ k \geq 0, \tag{8.48}$$

qui implique de résoudre un système linéaire de taille sd à chaque itération, ce qui demande de l'ordre de $\frac{2}{3}(sd)^3$ opérations arithmétiques pour la factorisation de la matrice du système et de l'ordre de $(sd)^2$ opérations pour la résolution des des deux systèmes triangulaires obtenus (on renvoie au chapitre 2 pour plus de détails). On diminue généralement ce coût par une modification de la méthode de Newton consistant à remplacer chacune des matrices jacobiennes $\frac{\partial f}{\partial x}(t_n + c_i h_n, x_n + Y_i^{(k)})$, $i = 1,\dots,s$, $k \geq 0$, par une approximation J, indépendante du niveau et de l'itération, un choix courant étant

$$J = \frac{\partial f}{\partial x}(t_n, x_n). \tag{8.49}$$

39. Lorsque la matrice A de la méthode est inversible, il est possible d'éviter ces évaluations en voyant que $x_{n+1} = x_n + \sum_{i=1}^{s} d_i X_i$, avec $\boldsymbol{d}^T = \boldsymbol{b}^T A^{-1}$.

La relation (8.48) devient alors

$$(I_s \otimes I_d - h_n A \otimes J) \begin{pmatrix} {X_1}^{(k+1)} - {X_1}^{(k)} \\ \vdots \\ {X_s}^{(k+1)} - {X_s}^{(k)} \end{pmatrix} = \boldsymbol{e} \otimes x_n + h_n (A \otimes I_s) \begin{pmatrix} f(t_n + c_1 h_n, x_n + {X_1}^{(k)}) \\ \vdots \\ f(t_n + c_s h_n, x_n + {X_s}^{(k)}) \end{pmatrix} - \begin{pmatrix} {X_1}^{(k)} \\ \vdots \\ {X_s}^{(k)} \end{pmatrix}, \quad k \geq 0, \quad (8.50)$$

avec $\boldsymbol{e}$ un vecteur à s composantes toutes égales à 1 et où l'on a noté $A \otimes B$ le produit de Kronecker de deux matrices A et B. La matrice du système linéaire (8.50) restant identique au cours des itérations, on n'a besoin d'effectuer qu'une seule factorisation LU en début d'étape, suivie de la résolution de deux systèmes linéaires triangulaires à chaque itération. Si cette modification ne diminue en rien les valeurs des quantités X_i, $i = 1, \ldots, s$, obtenues à convergence, elle a néanmoins pour effet de ralentir cette convergence. Ajoutons qu'il y a moyen de diminuer le coût de la factorisation de la matrice $(I_s \otimes I_d - h_n A \otimes J)$, pour le rendre proportionnel à sd^3 opérations lorsque toutes les valeurs propres de A sont distinctes (car on est alors ramené à la résolution de s systèmes de taille d, en s'appuyant sur une réduction sous une *forme canonique de Jordan* (voir [But76]).

Passons à présent à la construction proprement dite de méthodes de Runge–Kutta implicites. Celle-ci s'avère beaucoup plus aisée que pour leurs homologues explicites si l'on s'appuie sur les conditions algébriques, liant les coefficients $\{a_{ij}\}_{1 \leq i,j \leq s}$ $\{b_i\}_{1 \leq i \leq s}$ et $\{c_i\}_{1 \leq i \leq s}$, suivantes

$$\begin{aligned} B(p) &: \sum_{i=1}^{s} b_i {c_i}^{k-1} = \frac{1}{k}, \ k = 1, \ldots, p, \\ C(\eta) &: \sum_{j=1}^{s} a_{ij} {c_j}^{k-1} = \frac{{c_i}^k}{k}, \ i = 1, \ldots, s, \ k = 1, \ldots, \eta, \\ D(\zeta) &: \sum_{i=1}^{s} a_{ij} b_i {c_i}^{k-1} = \frac{b_j}{k} (1 - {c_j}^k), \ j = 1, \ldots, s, \ k = 1, \ldots, \zeta. \end{aligned}$$

La condition $B(p)$ signifie simplement que la formule de quadrature interpolatoire ayant pour nœuds les coefficients c_i et pour poids les coefficients b_i, $i = 1, \ldots, s$, est d'ordre p sur l'intervalle $[0, 1]$. L'importance des deux conditions restantes tient au résultat suivant, dû à Butcher [But64a].

Théorème 8.16 *Si les coefficients d'une méthode de Runge–Kutta à s niveaux satisfont les conditions $B(p)$, $C(\eta)$ et $D(\zeta)$ avec $p \leq \eta + \zeta + 1$ et $p \leq 2\eta + 2$, cette méthode est d'ordre p.*

On peut ainsi obtenir des méthodes implicites à s niveaux et d'ordre $p = 2s$ en se basant sur les formules de quadrature de Gauss–Legendre (voir la section 7.4) pour la détermination des coefficients $\{b_i\}_{1 \leq i \leq s}$ et $\{c_i\}_{1 \leq i \leq s}$, ce qui revient à satisfaire la condition $B(2s)$, et en cherchant à vérifier les conditions $C(s)$ et $D(s)$ pour trouver les coefficients $\{a_{ij}\}_{1 \leq i,j \leq s}$ (voir [Kun61, But64a]).

Exemples de méthode de Runge–Kutta implicite basée sur une formule de Gauss–Legendre. Pour $s \leq 5$, les coefficients de ces méthodes s'expriment en termes de radicaux. Pour les méthodes à un, deux ou trois niveaux, on a les tableaux de Butcher associés suivants

- $s = 1$, $p = 2$

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array},$$

- $s = 2$, $p = 4$

$$\begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array},$$

- $s = 3$, $p = 6$

$$\begin{array}{c|ccc} \frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{10-3\sqrt{15}}{45} & \frac{25-6\sqrt{15}}{180} \\ \frac{1}{2} & \frac{10+3\sqrt{15}}{72} & \frac{2}{9} & \frac{10-3\sqrt{15}}{72} \\ \frac{5+\sqrt{15}}{10} & \frac{25+6\sqrt{15}}{180} & \frac{10+3\sqrt{15}}{45} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}.$$

La méthode à un niveau est appelée la *méthode du point milieu implicite*.

À ce premier type de méthodes s'ajoutent les méthodes basées sur les formules de quadrature de Gauss–Radau et de Gauss–Lobatto, pour lesquelles la fonction f est évaluée respectivement en l'une ou l'autre et aux deux extrémités du sous-intervalle d'intégration courant à chaque étape. Imposer que $c_1 = 0$ conduit aux méthodes de Gauss–Radau I, que $c_s = 1$ à celles de Gauss–Radau II (l'ordre atteint dans ces deux cas étant $p = 2s-1$), que $c_1 = 0$ et $c_s = 1$ à celles de Gauss–Lobatto III (dont l'ordre est $p = 2s-2$, avec $s \geq 2$). Plusieurs sous-familles [40] de méthodes existent selon les conditions algébriques imposées pour la construction de la matrice A, certains choix se traduisant par l'annulation de coefficients matriciels et rendant par conséquent les méthodes obtenues plus efficaces d'un point calculatoire, mais aussi moins adaptées au traitement des systèmes raides (voir la sous-section 8.4.5).

Exemples de méthode de Runge–Kutta implicite basée sur une formule de Gauss–Radau. Les tableaux de Butcher associés aux différentes familles de méthodes de Runge–Kutta implicites à un, deux ou trois niveaux associées aux formules de quadrature de Gauss–Radau sont respectivement

- $s = 1$, $p = 1$

1	1
	1

Gauss–Radau IIA

- $s = 2$, $p = 3$

0	0	0
$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	$\frac{1}{4}$	$\frac{3}{4}$

Gauss–Radau I

0	$\frac{1}{4}$	$-\frac{1}{4}$
$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{4}$	$\frac{3}{4}$

Gauss–Radau IA

$\frac{1}{3}$	$\frac{1}{3}$	0
1	1	0
	$\frac{3}{4}$	$\frac{1}{4}$

Gauss–Radau II

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

Gauss–Radau IIA

- $s = 3$, $p = 5$

0	0	0	0
$\frac{6-\sqrt{6}}{10}$	$\frac{9+\sqrt{6}}{75}$	$\frac{24+\sqrt{6}}{120}$	$\frac{168-73\sqrt{6}}{600}$
$\frac{6+\sqrt{6}}{10}$	$\frac{9-\sqrt{6}}{75}$	$\frac{168+73\sqrt{6}}{600}$	$\frac{24-\sqrt{6}}{120}$
	$\frac{1}{9}$	$\frac{16+\sqrt{6}}{36}$	$\frac{16-\sqrt{6}}{36}$

Gauss–Radau I

0	$\frac{1}{9}$	$\frac{-1-\sqrt{6}}{18}$	$\frac{-1+\sqrt{6}}{18}$
$\frac{6-\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{88-43\sqrt{6}}{360}$
$\frac{6+\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+43\sqrt{6}}{360}$	$\frac{88-7\sqrt{6}}{360}$
	$\frac{1}{9}$	$\frac{16+\sqrt{6}}{36}$	$\frac{16-\sqrt{6}}{36}$

Gauss–Radau IA

$\frac{4-\sqrt{6}}{10}$	$\frac{24-\sqrt{6}}{120}$	$\frac{24-11\sqrt{6}}{120}$	0
$\frac{4+\sqrt{6}}{10}$	$\frac{24+11\sqrt{6}}{120}$	$\frac{24+\sqrt{6}}{120}$	0
1	$\frac{6-\sqrt{6}}{12}$	$\frac{6+\sqrt{6}}{12}$	0
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

Gauss–Radau II

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

Gauss–Radau IIA

La méthode de Gauss–Radau IIA à un niveau n'est autre que la méthode d'Euler implicite. On observe que le premier (resp. dernier) niveau des méthodes de Gauss–Radau I (resp. II) est explicite.

Exemples de méthode de Runge–Kutta implicite basée sur une formule de Gauss–Lobatto. Les tableaux de Butcher associés aux différentes familles de méthodes de Runge–Kutta implicites à deux ou trois niveaux associées aux formules de quadrature de Gauss–Lobatto sont respectivement

40. En plus des familles de Gauss–Radau I (satisfaisant les conditions $a_{1j} = 0$, $j = 1,\ldots,s$, $B(2s-1)$ et $C(s)$, ce qui implique que $D(s-1)$ est vérifiée) et II (conditions $a_{is} = 0$, $i = 1,\ldots,s$, $B(2s-1)$ et $D(s)$, impliquant $C(s-1)$) et de Gauss-Lobatto III (conditions $a_{is} = 0$ et $a_{1j} = 0$, $i,j = 1,\ldots,s$, $B(2s-2)$ et $C(s-1)$, impliquant $D(s-1)$) construites dans [But64b], mentionnons celles de Gauss–Radau IA (conditions $B(2s-1)$ et $D(s)$, impliquant $C(s-1)$) et IIA (conditions $B(2s-1)$ et $C(s)$, impliquant $D(s-1)$), de Gauss–Lobatto IIIA (conditions $B(2s-2)$ et $C(s)$, impliquant $D(s-2)$) et IIIB (conditions $B(2s-2)$ et $D(s)$, impliquant $C(s-2)$), introduites dans [Ehl69], ainsi que celle de Gauss–Lobatto IIIC (conditions $B(2s-2)$ et $C(s-1)$, impliquant $D(s-1)$), introduite dans [Chi71].

• $s=2$, $p=2$

0	0	0
1	1	0
	$\frac{1}{2}$	$\frac{1}{2}$

Gauss–Lobatto III

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Gauss–Lobatto IIIA

0	$\frac{1}{2}$	$-\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Gauss–Lobatto IIIC

• $s=3$, $p=4$

0	0	0	0
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0
1	0	1	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Gauss–Lobatto III

0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Gauss–Lobatto IIIA

0	$\frac{1}{6}$	$-\frac{1}{6}$	0
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$	0
1	$\frac{1}{6}$	$\frac{5}{6}$	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Gauss–Lobatto IIIB

0	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Gauss–Lobatto IIIC

On observe que les premier et dernier niveaux des méthodes de Gauss–Lobatto III sont explicites. On remarque également qu'il n'existe pas de méthode de Gauss–Lobatto IIIB à deux niveaux. La méthode de Gauss–Lobatto IIIA n'est autre que la méthode de la règle du trapèze (8.29).

Indiquons que certaines de ces méthodes peuvent aussi être interprétées comme des *méthodes de collocation*. Appliquée à la résolution numérique du problème de Cauchy (8.1)-(8.5), une méthode de collocation consiste à poser à chaque étape

$$x_{n+1} = p(t_n + h_n),\ n = 0, \dots, N-1, \tag{8.51}$$

où p désigne l'unique polynôme de degré s satisfaisant

$$p(t_n) = x_n,\ p'(t_n + c_j h_n) = f(t_n + c_j h_n, P(t_n + c_j h_n)),\ j = 1, \dots, s. \tag{8.52}$$

Il a été démontré dans [Wri70] qu'un tel procédé correspond à une méthode de Runge–Kutta implicite. En effet, en notant $k_j = p'(t_n + c_j h_n)$, $j = 1, \dots, s$, on a, en utilisant les résultats sur l'interpolation de Lagrange et les notations du chapitre 6, que

$$p'(t_n + c\, h_n) = \sum_{j=1}^{s} k_j\, l_j(c),\ c \in [0,1],$$

avec

$$l_j(c) = \prod_{\substack{k=1 \\ k \neq j}}^{s} \frac{c - c_k}{c_j - c_k},\ j = 1, \dots, s.$$

En intégrant cette identité entre 0 et c_i, $i = 1, \dots, s$, on obtient

$$p(t_n + c_i h_n) - p(t_n) = h_n \sum_{j=1}^{s} k_j \left(\int_0^{c_i} l_j(c)\, \mathrm{d}c \right),\ i = 1, \dots, s,$$

d'où, en posant

$$a_{ij} = \int_0^{c_i} l_j(c)\, \mathrm{d}c,\ i, j = 1, \dots, s, \tag{8.53}$$

et en utilisant les conditions (8.52),

$$k_j = p'(t_n + c_j h_n) = f(t_n + c_j h_n, p(t_n + c_j h_n)) = f\left(t_n + c_j h_n, x_n + h_n \sum_{j=1}^{s} a_{ij}\, k_j\right),\ j = 1, \dots, s.$$

En intégrant l'identité entre 0 et 1, il vient

$$p(t_n + h_n) - p(t_n) = h_n \sum_{j=1}^{s} k_j \left(\int_0^{1} l_j(c)\, \mathrm{d}c \right),$$

dont on déduit finalement, en posant $b_j = \int_0^1 l_j(c)\,\mathrm{d}c$, $j = 1,\dots,s$, et en se servant de (8.51) et (8.52),

$$x_{n+1} = x_n + h_n \sum_{j=1}^{s} k_j\, b_j.$$

Réciproquement, une méthode de Runge–Kutta implicite dont les coefficients c_i, $i = 1,\dots,s$, sont distincts et d'ordre au moins s sera une méthode de collocation si elle fournit une solution exacte lorsque $f(t,x) = p(t)$, pour tout polynôme p de degré inférieur ou égal à $s-1$, ce qui revient à demander que

$$\sum_{j=1}^{s} a_{ij}\, p(c_j) = \int_0^{c_i} p(t)\,\mathrm{d}t,\ i = 1,\dots,s, \tag{8.54}$$

ce qui équivaut à

$$\sum_{j=1}^{s} a_{ij}\, {c_j}^r = \frac{{c_i}^{r+1}}{r+1},\ r = 0,\dots,s-1,\ i = 1,\dots,s,$$

ces conditions s'avérant nécessaires [41]. On peut ainsi vérifier que les méthodes de Gauss–Legendre, de Gauss–Radau de type IIA et de Gauss–Lobatto de type IIIA sont des méthodes de collocation.

Nous concluons cette section en évoquant brièvement des méthodes de Runge–Kutta semi-implicites, pour lesquelles les matrices de coefficients $\{a_{ij}\}_{1\le i,j\le s}$ sont triangulaires inférieures. Ceci a pour conséquence que le système d'équations non linéaires (8.46) associé à de telles méthodes peut être résolu de manière séquentielle, via une méthode de descente par blocs, entraînant une réduction substantielle du coût de calcul engendré par rapport à une méthode de Runge–Kutta implicite. Dans le cas des méthodes *DIRK* (pour *diagonally implicit Runge–Kutta* en anglais) [Ale77], on impose aux coefficients diagonaux a_{ii}, $i = 1,\dots,s$, d'être tous identiques, ce qui conduit à une diminution supplémentaire du coût de résolution lorsque l'on utilise la modification (8.50) de la méthode de Newton–Raphson, les blocs diagonaux à factoriser étant les mêmes. Un inconvénient majeur de ces méthodes est que leur construction semble difficile pour des ordres élevés.

Exemples de méthode DIRK. Pour $s = 2$ et 3, il existe une unique méthode DIRK à s niveaux et d'ordre $s+1$ possédant la propriété d'être *A-stable* (voir la définition 8.44). Ces méthodes sont respectivement données par les tableaux de Butcher

- $s = 2$, $p = 3$

$$\begin{array}{c|cc} \frac{1}{2}+\frac{1}{2\sqrt{3}} & \frac{1}{2}+\frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{2}-\frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2}+\frac{1}{2\sqrt{3}} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array},$$

- $s = 3$, $p = 4$

$$\begin{array}{c|ccc} \frac{1+\mu}{2} & \frac{1+\mu}{2} & 0 & 0 \\ \frac{1}{2} & -\frac{\mu}{2} & \frac{1+\mu}{2} & 0 \\ \frac{1-\mu}{2} & 1+\mu & -(1+2\mu) & \frac{1+\mu}{2} \\ \hline & \frac{1}{6\mu^2} & 1-\frac{1}{3\mu^2} & \frac{1}{6\mu^2} \end{array},$$

avec $\mu = \frac{2\sqrt{3}}{3}\cos\left(\frac{\pi}{18}\right)$.

Enfin, les méthodes *SIRK* (pour *singly implicit Runge–Kutta* en anglais), introduites dans [Nør76], conservent l'idée d'une matrice dont le spectre est constitué d'une unique valeur propre réelle de multiplicité s tout en abandonnant une forme triangulaire inférieure. Ce sont donc des méthodes de Runge–Kutta complètement implicites, mais dont le coût de résolution effectif rivalise avec celui des méthodes DIRK si l'on utilise la méthode de Newton modifiée (8.50) en conjonction avec la réduction sous forme canonique de Jordan mentionnée plus haut. Contrairement aux méthodes DIRK, des méthodes SIRK d'ordre arbitrairement élevé et possédant de bonnes propriétés de stabilité peuvent simplement être obtenues comme des méthodes de collocation.

41. Les coefficients a_{ij}, $1 \le i, j \le s$, donnés par (8.53) satisfont en effet les conditions (8.54) par définition du polynôme d'interpolation de Lagrange.

Exemple de méthode SIRK. Un exemple de méthode SIRK à deux niveaux et d'ordre trois est celui donné par le tableau de Butcher suivant

$$\begin{array}{c|cc} (2-\sqrt{2})\mu & \frac{(4-\sqrt{2})\mu}{4} & \frac{(4-3\sqrt{2})\mu}{4} \\ (2+\sqrt{2})\mu & \frac{(4+3\sqrt{2})\mu}{4} & \frac{(4-\sqrt{2})\mu}{4} \\ \hline & \frac{4(1+\sqrt{2})\mu-\sqrt{2}}{8\mu} & \frac{4(1-\sqrt{2})\mu+\sqrt{2}}{8\mu} \end{array},$$

avec $\mu = \frac{3\pm\sqrt{3}}{6}$.

8.3.3 Méthodes à pas multiples linéaires

Les *méthodes à pas multiples linéaires* (*linear multistep methods* en anglais) adoptent une philosophie inverse de celles des méthodes de Runge–Kutta pour améliorer la précision de l'approximation qu'elles calculent, au sens où celles-ci font appel à q valeurs précédemment calculées de la solution approchée, l'entier q, $q \geq 1$, étant le *nombre de pas* de la méthode, pour faire avancer à chaque étape la résolution numérique du problème.

Cette classe de méthodes possède des liens étroits avec l'interpolation de Lagrange, présentée au chapitre 6, la dérivation de méthodes se faisant suivant deux approches, basées respectivement sur l'intégration et la différentiation de polynômes d'interpolation de Lagrange particuliers.

Principe

Étant donné un entier $q \geq 1$, une méthode à q pas linéaire est définie par le schéma

$$x_{n+1} = \sum_{i=0}^{q-1} a_i\, x_{n-i} + h \sum_{i=0}^{q} b_i\, f(t_{n-i+1}, x_{n-i+1}),\ n \geq q-1, \tag{8.55}$$

dans lequel les valeurs $f(t_i, x_i)$, $i = n-q+1, \ldots, n+1$, interviennent de manière linéaire [42], ce qui n'était pas le cas pour les méthodes de Runge–Kutta. Ceci suggère d'écrire la relation de récurrence (8.55) sous la forme générale d'*équation aux différences* linéaire,

$$\sum_{i=0}^{q} \alpha_i\, x_{n+i} = h \sum_{i=0}^{q} \beta_i\, f(t_{n+i}, x_{n+i}),\ n \geq 0, \tag{8.56}$$

en posant $a_i = -\frac{\alpha_{q-1-i}}{\alpha_q}$, $i = 0, \ldots, q-1$, et $b_i = \frac{\beta_{q-i}}{\alpha_q}$, $i = 0, \ldots, q$, la valeur du coefficient α_q étant fixée. Lorsque le coefficient β_q est nul, la méthode est explicite, elle est implicite sinon.

Le lecteur attentif aura remarqué que la longueur du pas de discrétisation dans la formule (8.56) ne dépend pas de l'entier n, alors que c'était le cas pour toutes les méthodes à un pas introduites auparavant. De fait, on a ici supposé que la grille de discrétisation était uniforme. Faire cette hypothèse simplificatrice n'est pas anodin : en son absence, les coefficients α_i et β_i, $i = 0, \ldots, q$, dépendent en effet de l'entier n et varient donc à chaque étape. Ceci rend l'implémentation de ces méthodes complexe dans l'optique d'une adaptation de la longueur du pas de discrétisation (voir néanmoins la fin de la section 8.6 pour la présentation d'approches possibles).

Une seconde hypothèse classique que l'on fera est de supposer que les coefficients (maintenant constants) α_i et β_i satisfont aux conditions [43]

$$\alpha_q = 1,\ |\alpha_0| + |\beta_0| \neq 0. \tag{8.57}$$

Lorsque la méthode à pas multiples linéaire est implicite, on est amené à résoudre à chaque étape une équation (ou, le cas échéant, un système d'équations) généralement non linéaire,

$$x_{n+q} = h\beta_q\, f(t_{n+q}, x_{n+q}) + \sum_{i=0}^{q-1} (h\beta_i\, f(t_{n+i}, x_{n+i}) - \alpha_i\, x_{n+i}). \tag{8.58}$$

42. Si cette observation justifie le qualificatif donné à ces méthodes, on notera cependant que la relation (8.55) n'est généralement pas linéaire par rapport aux approximations numériques x_i, $i = n-q+1, \ldots, n+1$, la fonction f pouvant être non linéaire.

43. La première condition interdit que l'on puisse définir une méthode d'une infinité de manières (celle-ci restant inchangée lorsque l'on multiplie (8.56) par une constante non nulle) en normalisant un coefficient particulier, alors que la seconde vise simplement à interdire l'écriture d'une méthode à q pas sous la forme d'une méthode en théorie à $q+1$ pas. Sans cette dernière hypothèse, il serait par exemple possible de définir la méthode d'Euler explicite en posant $x_{n+2} - x_{n+1} = h_{n+1}\, f(t_{n+1}, x_{n+1})$, donnant ainsi l'illusion d'une méthode à deux pas.

La fonction f satisfaisant la condition de Lipschitz globale (8.24) par rapport à la variable x, il découle du théorème 5.9 que cette équation possède une unique solution si la longueur h du pas de discrétisation est telle que

$$h < \frac{1}{L\,|\beta_q|}, \tag{8.59}$$

et que l'on peut approcher numériquement cette solution par la méthode des approximations successives : étant donnée une valeur $x_{n+q}^{(0)}$ arbitraire, la valeur x_{n+q} est la limite de la suite $(x_{n+q}^{(k)})_{k\in\mathbb{N}}$ définie par la relation de récurrence

$$x_{n+q}^{(k+1)} = h\beta_q f(t_{n+q}, x_{n+q}^{(k)}) + \sum_{i=0}^{q-1} (h\beta_i f(t_{n+i}, x_{n+i}) - \alpha_i x_{n+i}),\ k \geq 0. \tag{8.60}$$

Dans de nombreux cas, la condition (8.59) ne s'avère pas restreignante, en raison de contraintes sur le pas plus sévères imposées par la précision voulue sur la solution numérique. Comme on l'a déjà vu dans la sous-section 8.3.2, ceci n'est malheureusement plus vrai dans le cas particulier des systèmes raides, pour lesquels $L \gg 1$. Il faut dans ce cas abandonner la méthode des approximations successives pour la remplacer par la méthode de Newton–Raphson. C'est généralement une modification de cette dernière, similaire à celle proposée pour les méthodes de Runge–Kutta implicites et visant à réduire le nombre de factorisations LU effectuées à chaque étape, qui est implémentée en pratique, la relation de récurrence associée prenant alors la forme

$$\left(I_m - h\beta_q \frac{\partial f}{\partial x}(t_{n+q}, x_{n+q}^{(0)})\right)\left(x_{n+q}^{(k+1)} - x_{n+q}^{(k)}\right) = -x_{n+q}^{(k)} + h\beta_q f(t_{n+q}, x_{n+q}^{(k)}) + \sum_{i=0}^{q-1} (h\beta_i f(t_{n+i}, x_{n+i}) - \alpha_i x_{n+i}),\ k \geq 0. \tag{8.61}$$

Les méthodes à pas multiples linéaires implicites sont donc bien plus coûteuses en temps de calcul et plus complexes à mettre en œuvre que leurs analogues explicites. Il est cependant possible de diminuer le nombre d'itérations de point fixe (8.60) (ou (8.61)) nécessaires en choisissant judicieusement la valeur $x_{n+q}^{(0)}$. On peut, par exemple, effectuer une étape d'une méthode à pas multiples linéaire explicite de même ordre et poursuivre les itérations de point fixe de la méthode implicite à partir de la valeur obtenue ; c'est le type de stratégie retenue par les *méthodes de prédiction-correction* présentées dans la section 8.5.

Enfin, il est important de noter que, pour démarrer, une méthode définie par la relation de récurrence (8.56) nécessite de connaître q valeurs approchées de la solution aux temps t_i, $i = 0, \ldots, q-1$. Or, seule la valeur de la solution à l'instant t_0 est fournie par la condition initiale du problème de Cauchy. Il faut donc avoir recours à une autre méthode pour calculer les $q-1$ valeurs d'initialisation faisant défaut. En pratique, cette *procédure de démarrage* est généralement accomplie au moyen d'une méthode de Runge–Kutta d'ordre supérieur ou égal à celui de la méthode à pas multiples linéaire considérée.

Avant de passer à la présentation de quatre familles de méthodes à pas multiples linéaires, décrites schématiquement dans le tableau 8.1, concluons cette brève introduction en définissant les *polynômes caractéristiques premier*

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j,\ \forall z \in \mathbb{C}, \tag{8.62}$$

et *second*

$$\sigma(z) = \sum_{j=0}^{q} \beta_j z^j,\ \forall z \in \mathbb{C}, \tag{8.63}$$

associés à toute méthode à pas multiples linéaire définie par (8.56), sur lesquels repose en grande partie l'analyse théorique réalisée dans la section 8.4. On déduit de (8.57) que le degré de ρ toujours égal à q. En revanche, celui de σ égal à q si la méthode est implicite et strictement inférieur à q si elle est explicite.

Exemples de polynômes caractéristiques associés aux méthodes à un pas déjà introduites. Dans le cas de la méthode d'Euler explicite (resp. implicite) définie par (8.26) (resp. (8.28)), nous avons $\rho(z) = z - 1$ et $\sigma(z) = 1$ (resp. $\rho(z) = z - 1$ et $\sigma(z) = z$). Pour la méthode de la règle du trapèze, définie par (8.29), il vient $\rho(z) = z - 1$ et $\sigma(z) = \frac{1}{2}(z + 1)$. Chacune de ces méthodes, pourtant présentées comme des méthodes à un pas, sont bien des exemples particuliers de méthodes à pas multiples linéaires.

	Adams–Bashforth		Adams–Moulton		Nyström		Milne–Simpson généralisée		BDF	
i	α_i	β_i	α_i	β_i	α_i	β_i	α_i	β_i	α_i	β_i
q	∘		∘	∘	∘		∘	∘	∘	∘
$q-1$	∘	∘	∘	∘		∘		∘	∘	
$q-2$		∘		∘	∘	∘	∘	∘	∘	
$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$	
n		∘		∘		∘		∘	∘	

TABLE 8.1 – Représentation schématique des indices de support des coefficients de différentes classes de méthodes à q pas linéaires.

Méthodes d'Adams

Bien qu'étant les méthodes à pas multiples linéaires les plus anciennes, les *méthodes d'Adams*[44] restent très utilisées et sont présentes dans bon nombre de codes de résolution numérique de systèmes d'équations différentielles ordinaires non raides basés sur des paires de prédicteur-correcteur (voir la section 8.5).

Pour comprendre leur dérivation, supposons que l'on dispose d'approximations $x_n, \ldots, x_{n+q-1}$, $n \geq 0$, de la solution du problème de Cauchy (8.1)-(8.5) aux points $t_n, \ldots, t_{n+q-1}$. En considérant la forme intégrale du problème de Cauchy sur l'intervalle $[t_{n+q-1}, t_{n+q}]$,

$$x(t_{n+q}) = x(t_{n+q-1}) + \int_{t_{n+q-1}}^{t_{n+q}} f(t, x(t))\,\mathrm{d}t, \tag{8.64}$$

et en utilisant une technique voisine de celle des formules de quadrature interpolatoires (voir le chapitre 7), on voit qu'une approximation de $x(t_{n+q})$ peut être obtenue assez naturellement en substituant dans (8.64) x_{n+q-1} à $x(t_{n+q-1})$ et en remplaçant la fonction dans l'intégrale soit par le polynôme d'interpolation de Lagrange de degré $q-1$ associé aux couples $(t_i, f(t_i, x_i))$, $i = n, \ldots, n+q-1$, pour une méthode explicite à q pas, soit par celui de degré q associé aux couples $(t_i, f(t_i, x_i))$, $i = n, \ldots, n+q$, pour une méthode implicite à q pas. Les méthodes correspondantes, dites méthodes d'Adams, sont consistantes par construction et d'ordre maximal relativement au nombre de pas considérés. Elles s'écrivent, en utilisant la forme générale (8.56) des méthodes à pas multiples linéaires,

$$x_{n+q} - x_{n+q-1} = h \sum_{i=0}^{q} \beta_i f(t_{n+i}, x_{n+i}),\ n \geq 0. \tag{8.65}$$

On observe que l'on a $\rho(z) = z^q - z^{q-1}$, ce qui correspond au choix le plus simple possible de premier polynôme caractéristique compte tenu des hypothèses (8.57). De plus, la détermination des coefficients β_j, $i = 0, \ldots, q$, peut être considérablement facilitée en ayant recours à la forme de Newton du polynôme d'interpolation de Lagrange vue au chapitre 6.

Pour le voir, posons $[t_i]f = f(t_i, x_i)$, $i = n, \ldots, n+q-1$. Dans le cas des méthodes d'Adams explicites, encore dites *méthodes d'Adams–Bashforth*[45], on a $\beta_q = 0$ et, en utilisant (6.11), il vient

$$\begin{aligned}\Pi_{q-1}(t) = [t_{n+q-1}]f + (t - t_{n+q-1})[t_{n+q-1}, t_{n+q-2}]f + (t - t_{n+q-1})(t - t_{n+q-2})[t_{n+q-1}, t_{n+q-2}, t_{n+q-3}]f \\ + \cdots + (t - t_{n+q-1})(t - t_{n+q-2}) \ldots (t - t_{n-1})[t_{n+q-1}, t_{n+q-2}, \ldots, t_n]f.\end{aligned}$$

Lorsque le pas de discrétisation est supposé constant et de longueur égale à h, on peut réécrire Π_{q-1} sous la

44. John Couch Adams (5 juin 1819 - 21 janvier 1892) était un mathématicien et astronome britannique. Son fait le plus célèbre fut de prédire l'existence de la planète Neptune, dont il calcula en 1845, indépendamment de Le Verrier, la position en étudiant les irrégularités du mouvement d'Uranus.

45. Francis Bashforth (8 janvier 1819 - 12 février 1912) était un mathématicien anglais. Son intérêt pour la balistique le conduisit à effectuer plusieurs séries d'expériences avec un chronographe de son invention dans le but de comprendre l'effet de la résistance de l'air sur la trajectoire d'un projectile.

forme compacte suivante, apparentée à la *formule de Gregory*[46]*–Newton régressive*,

$$\Pi_{q-1}(t) = \sum_{i=0}^{q-1} \frac{\nabla^i f_{n+q-1}}{i!\, h^i} \prod_{j=0}^{i} (t - t_{n+q-1-j}), \tag{8.66}$$

en introduisant les *différences finies régressives* définies par récurrence,

$$\nabla^0 f_i = [t_i]f,\ \nabla^m f_i = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1},\ m \geq 1.$$

En intégrant entre t_{n+q-1} et t_{n+q}, on trouve alors

$$x_{n+q} = x_{n+q-1} + h \sum_{i=0}^{q-1} \gamma_i \nabla^i f_{n+q-1},$$

où

$$\gamma_i = \frac{1}{i!} \int_0^1 \prod_{j=0}^{i} (u+j)\, \mathrm{d}u = \int_0^1 \binom{u+i-1}{i} \mathrm{d}u,$$

par définition de la généralisation du coefficient binomial, $\binom{z}{k} = \frac{z(z-1)(z-2)\ldots(z-k+1)}{k!}$, $\forall z \in \mathbb{C}$, $\forall k \in \mathbb{N}$.
Les valeurs des premiers coefficients γ_i sont les suivantes[47]

$$\gamma_0 = 1,\ \gamma_1 = \frac{1}{2},\ \gamma_2 = \frac{5}{12},\ \gamma_3 = \frac{3}{8},\ \gamma_4 = \frac{251}{70},\ \gamma_5 = \frac{95}{288},\ \gamma_6 = \frac{19087}{60480}.$$

Une fois ces constantes calculées, les méthodes d'Adams–Bashforth sont explicitement obtenues en ré-exprimant les différences finies régressives en termes des valeurs f_i. Pour les premières valeurs de q, ceci conduit aux formules :

- $q = 1 : x_{n+1} = x_n + h f(t_n, x_n)$ (méthode d'Euler explicite),
- $q = 2 : x_{n+1} = x_n + h \left(\frac{3}{2} f(t_n, x_n) - \frac{1}{2} f(t_{n-1}, x_{n-1}) \right)$,
- $q = 3 : x_{n+1} = x_n + h \left(\frac{23}{12} f(t_n, x_n) - \frac{16}{12} f(t_{n-1}, x_{n-1}) + \frac{5}{12} f(t_{n-2}, x_{n-2}) \right)$,
- $q = 4 : x_{n+1} = x_n + h \left(\frac{55}{24} f(t_n, x_n) - \frac{59}{24} f(t_{n-1}, x_{n-1}) + \frac{37}{24} f(t_{n-2}, x_{n-2}) - \frac{9}{24} f(t_{n-3}, x_{n-3}) \right)$.

En vertu des théorèmes 8.25 et 8.31, ces méthodes explicites à q pas sont d'ordre q et stables, elles sont par conséquent optimales au sens du théorème 8.32.

Pour les méthodes d'Adams implicites, encore appelées les *méthodes d'Adams–Moulton*[48], c'est le polynôme de degré q interpolant les valeurs $f_n, \ldots, f_{n+q}$ aux nœuds $t_n, \ldots, t_{n+q}$ qu'il faut considérer. En supposant que le pas de discrétisation est uniforme, on a cette fois

$$x_{n+q} = x_{n+q-1} + h \sum_{i=0}^{q} \gamma_i^* \nabla^i f_{n+q}, \tag{8.67}$$

avec

$$\gamma_i^* = \frac{1}{i!} \int_0^1 \prod_{j=0}^{i} (u+j-1)\, \mathrm{d}u = \int_0^1 \binom{u+i-2}{i} \mathrm{d}u.$$

46. James Gregory (novembre 1638 - octobre 1675) était un mathématicien et astronome écossais. Il publia un descriptif d'un des premiers modèles de télescope à miroir secondaire concave, sans néanmoins parvenir à le construire, et découvrit les développements en série de plusieurs fonctions trigonométriques.

47. Pour les calculer, on utilise le fait que la relation de récurrence suivante existe entre ces coefficients (voir, par exemple, [HNW93] pour une démonstration)

$$\gamma_0 = 1,\ \gamma_m + \sum_{i=1}^{m} \frac{1}{i+1} \gamma_{m-i} = 1,\ m \geq 1.$$

48. Forest Ray Moulton (29 avril 1872 - 7 décembre 1952) était un astronome américain. Il est, avec Thomas Chrowder Chamberlin, l'instigateur d'une hypothèse selon laquelle la formation d'une planète serait due à l'accrétion de corps célestes plus petits appelés *planétésimaux*.

On en déduit[49] les valeurs des premiers coefficients γ_i^*

$$\gamma_0^*=1,\ \gamma_1^*=-\frac{1}{2},\ \gamma_2^*=-\frac{1}{12},\ \gamma_3^*=\frac{1}{24},\ \gamma_4^*=-\frac{19}{720},\ \gamma_5^*=-\frac{3}{160},\ \gamma_6^*=-\frac{863}{60480}.$$

Les méthodes d'Adams–Moulton pour les premières valeurs de l'entier q sont alors :

- $q=1 : x_{n+1}=x_n+h\left(\frac{1}{2}f(t_{n+1},x_{n+1})+\frac{1}{2}f(t_n,x_n)\right)$ (méthode de la règle du trapèze),
- $q=2 : x_{n+1}=x_n+h\left(\frac{5}{12}f(t_{n+1},x_{n+1})+\frac{8}{12}f(t_n,x_n)-\frac{1}{12}f(t_{n-1},x_{n-1})\right)$,
- $q=3 : x_{n+1}=x_n+h\left(\frac{9}{24}f(t_{n+1},x_{n+1})+\frac{19}{24}f(t_n,x_n)-\frac{5}{24}f(t_{n-1},x_{n-1})+\frac{1}{24}f(t_{n-2},x_{n-2})\right)$,
- $q=4$:

$$x_{n+1}=x_n+h\left(\frac{251}{720}f(t_{n+1},x_{n+1})+\frac{646}{720}f(t_n,x_n)-\frac{264}{720}f(t_{n-1},x_{n-1})+\frac{106}{720}f(t_{n-2},x_{n-2})-\frac{19}{720}f(t_{n-3},x_{n-3})\right).$$

Formellement, on observe qu'il est possible de construire une formule d'Adams–Moulton avec $q=0$, choix pour lequel on retrouve la méthode d'Euler implicite (8.28), qui est une méthode à un pas... Pour cette raison, nous considérerons dans la suite que l'entier q est strictement positif, afin d'éviter toute confusion liée à l'existence de *deux* méthodes d'Adams–Moulton à un pas.

Les méthodes d'Adams–Moulton à q pas sont d'ordre $q+1$ et stables ; elles sont donc optimales lorsque le nombre de pas est impair.

Méthodes de Nyström

Une autre famille de méthodes à pas multiples linéaires est celle proposée par Nyström dans [Nys25]. Ces méthodes sont obtenues de manière similaire aux méthodes d'Adams–Bashforth, mais en se basant sur la relation intégrale

$$x(t_{n+q})=x(t_{n+q-2})+\int_{t_{n+q}}^{t_{n+q-2}} f(t,x(t))\,\mathrm{d}t,\ n\geq 1, \tag{8.68}$$

en lieu de (8.64), ce qui revient à faire le choix

$$\rho(z)=z^q-z^{q-2}$$

pour le premier polynôme caractéristique. Le pas de discrétisation étant supposé de longueur constante, on obtient dans ce cas, en utilisant l'expression (8.66) de Π_{q-1} et en procédant comme précédemment,

$$x_{n+q}=x_{n+q-2}+h\sum_{i=0}^{q-1}\kappa_i\,\nabla^i f_{n+q},$$

où

$$\kappa_i=(-1)^i\int_{-1}^{1}\binom{-s}{i}\mathrm{d}s.$$

Les valeurs des premiers coefficients κ_i sont les suivantes

$$\kappa_0=2,\ \kappa_1=0,\ \kappa_2=\frac{1}{3},\ \kappa_3=\frac{1}{3},\ \kappa_4=\frac{29}{90},\ \kappa_5=\frac{14}{45},\ \kappa_6=\frac{1139}{3780},$$

Nyström recommandant dans son article leur usage plutôt que celui des coefficients γ_i des méthodes d'Adams–Bashforth, les calculs étant effectués à l'époque par des calculateurs humains et non des machines.

49. La relation de récurrence entre ces coefficients est

$$\gamma_0^*=1,\ \gamma_m^*+\sum_{i=1}^{m}\frac{1}{i+1}\gamma_{m-i}^*=0,\ m\geq 1.$$

Exemple de méthode de Nyström. Pour $q = 2$ (et aussi $q = 1$, le coefficient κ_1 étant nul), la méthode de Nyström s'écrit

$$x_{n+1} = x_{n-1} + 2h f(t_n, x_n),$$

et est parfois appelée la *méthode de la règle du point milieu* par analogie avec la formule de quadrature du même nom.

Généralisations de la méthode de Milne–Simpson

Ce groupe de méthodes est constitué des analogues implicites des méthodes de Nyström que nous venons de décrire. Pour une grille uniforme, il vient par conséquent

$$x_{n+q} = x_{n+q-2} + h \sum_{i=0}^{q} \kappa_i^* \nabla^i f_{n+q},$$

où

$$\kappa_i^* = (-1)^i \int_{-2}^{0} \binom{-s}{i} \mathrm{d}s,$$

les valeurs des premiers coefficients κ_i^* étant

$$\kappa_0^* = 2,\ \kappa_1^* = -2,\ \kappa_2^* = \frac{1}{3},\ \kappa_3^* = 0,\ \kappa_4^* = -\frac{1}{90},\ \kappa_5^* = -\frac{1}{90},\ \kappa_6^* = -\frac{37}{3780}.$$

Exemple de la méthode de Milne–Simpson. Pour $q = 2$, on trouve une méthode dont la règle de quadrature associée n'est autre que celle de Simpson (voir la formule (7.11)),

$$x_{n+1} = x_{n-1} + h\left(\frac{1}{3} f(t_{n+1}, x_{n+1}) + \frac{4}{3} f(t_n, x_n) + \frac{1}{3} f(t_{n-1}, x_{n-1})\right). \tag{8.69}$$

Cette méthode à deux pas, introduite par Milne dans [Mil26], est d'ordre quatre, ce qui la rend optimale au sens du théorème 8.32.

Méthodes utilisant des formules de différentiation rétrograde

Chacune des familles de méthodes que nous venons de présenter sont basées sur une équation intégrale satisfaite par toute solution de l'équation différentielle (8.1) et tirent parti de l'intégration numérique d'un polynôme d'interpolation de la fonction $f(\cdot, x(\cdot))$. Une approche « duale » consiste à directement approcher au point t_{n+q} la dérivée de la solution par celle d'un polynôme d'interpolation de Lagrange associé aux approximations des valeurs x aux points $t_n + q, \ldots, t_n$. Pour obtenir les coefficients de telles méthodes, il faut ainsi considérer le polynôme d'interpolation de Lagrange associé aux valeurs $x_{n+q}, \ldots, x_n$. En faisant appel à la formule de Gregory–Newton régressive dans d'une grille uniforme de pas h, il vient

$$\sum_{i=1}^{q} \frac{1}{i} \nabla^i x_{n+q} = h f(t_{n+q}, x_{n+q}), \tag{8.70}$$

ce qui conduit [50] aux formules suivantes pour les premières valeurs de q :

- $q = 1 : x_{n+1} - x_n = h f(t_{n+1}, x_{n+1})$ (méthode d'Euler implicite),
- $q = 2 : \frac{3}{2} x_{n+1} - 2 x_n + \frac{1}{2} x_{n-1} = h f(t_{n+1}, x_{n+1})$,
- $q = 3 : \frac{11}{6} x_{n+1} - 3 x_n + \frac{3}{2} x_{n-1} - \frac{1}{3} x_{n-2} = h f(t_{n+1}, x_{n+1})$,
- $q = 4 : \frac{25}{12} x_{n+1} - 4 x_n + 3 x_{n-1} - \frac{4}{3} x_{n-2} + \frac{1}{4} x_{n-3} = h f(t_{n+1}, x_{n+1})$.

50. Étant directement issues de (8.70), les formules données ne satisfont pas la condition de normalisation $\alpha_q = 1$ imposée par (8.57). On se référera au tableau 8.2 pour les valeurs des coefficients normalisés.

Sous les hypothèses (8.57), on observe que l'on a, pour une méthode à q pas,

$$\rho(z) = z^q \left((1 - z^{-1}) + \cdots + \frac{1}{q}(1 - z^{-1})^q \right)$$

et

$$\sigma(z) = \beta_q z^q,$$

ce qui correspond au choix le plus simple possible pour σ pour une méthode implicite. Une telle méthode est d'ordre q.

q	α_0	α_1	α_2	α_3	α_4	α_5	α_6	β_q
1	-1	1						1
2	$\frac{1}{3}$	$-\frac{4}{3}$	1					$\frac{2}{3}$
3	$-\frac{2}{11}$	$\frac{9}{11}$	$-\frac{18}{11}$	1				$\frac{6}{11}$
4	$\frac{3}{25}$	$-\frac{16}{25}$	$\frac{36}{25}$	$-\frac{48}{25}$	1			$\frac{12}{25}$
5	$-\frac{12}{137}$	$\frac{75}{137}$	$-\frac{200}{137}$	$\frac{300}{137}$	$-\frac{300}{137}$	1		$\frac{60}{137}$
6	$\frac{10}{147}$	$-\frac{72}{147}$	$\frac{225}{147}$	$-\frac{400}{147}$	$\frac{450}{147}$	$-\frac{360}{147}$	1	$\frac{60}{147}$

TABLE 8.2 – Coefficients normalisés des méthodes BDF à q pas, $1 \leq q \leq 6$.

L'introduction de ces méthodes à pas multiples linéaires implicites dites *BDF* (acronyme anglais de *backward differentiation formula*) remonte à l'article de Curtiss et Hirschfelder [CH52]. Si elles sont généralement négligées au profit de méthodes implicites comme les méthodes d'Adams–Moulton, plus précises à nombre de pas égal, leur stabilité supérieure (voir la sous-section 8.4.5) les rendent particulièrement attractives lorsque le système du problème à résoudre est raide (voir la section 8.7).

8.3.4 Méthodes basées sur des développements de Taylor

Nous concluons cette section sur la présentation d'une classe de méthodes utilisant non seulement les valeurs de la dérivée première de la solution recherchée, mais aussi celles de ses dérivées d'ordre supérieur. Celle-ci est s'appuie sur un développement de Taylor de la solution, exprimé en termes des dérivées « totales » de la fonction f, qui doit donc présenter une certaine régularité.

Supposons que f soit infiniment dérivable. C'est alors aussi le cas pour la solution du problème de Cauchy (8.1)-(8.5) et l'on a

$$x''(t) = \frac{\partial f}{\partial t}(t, x(t)) + f(t, x(t)) \frac{\partial f}{\partial x}(t, x(t)) = f^{(1)}(t, x(t)),$$

$$x'''(t) = \frac{\partial^2 f}{\partial t^2}(t, x(t)) + 2 f(t, x(t)) \frac{\partial^2 f}{\partial t \partial x}(t, x(t)) + (f(t, x(t)))^2 \frac{\partial^2 f}{\partial x^2}(t, x(t)) = f^{(2)}(t, x(t)),$$

et cetera. En notant plus généralement

$$x^{(k+1)}(t) = f^{(k)}(t, x(t)),\ k \geq 1.$$

et en posant $f^{(0)}(t, x) = f(t, x)$, le développement de Taylor à l'ordre p, avec p un entier strictement positif, de la solution de l'équation au point $t_{n+1} = t_n + h_n$ autour du point t_n s'écrit

$$x(t_{n+1}) = x(t_n) + \sum_{k=1}^{p} \frac{{h_n}^k}{k!} f^{(k-1)}(t_n, x(t_n)) + O({h_n}^{p+1}),\ n \geq 0,$$

et l'on en déduit de manière directe le schéma suivant

$$x_{n+1} = x_n + \sum_{k=1}^{p} \frac{{h_n}^k}{k!} f^{(k-1)}(t_n, x_n),\ n \geq 0.$$

La méthode obtenue, parfois appelée *méthode de Taylor*, est par construction d'ordre p. Pour $p = 1$, on retrouve la méthode d'Euler et son interprétation en termes d'approximation de la courbe intégrale par sa tangente en un point sur un pas de discrétisation.

Ces méthodes ne sont pas sans inconvénient en pratique. Il faut en effet supposer que la fonction est *a priori* très régulière et une telle méthode aura toutes les chances d'être inutilisable en pratique si ce n'est pas le cas. De plus, pour une méthode d'ordre p, on doit être en mesure de pouvoir évaluer les dérivées de f jusqu'à l'ordre $p-1$. Or, même lorsque des expressions analytiques sont disponibles, leur complexité croît très rapidement avec l'ordre de dérivation, y compris pour des fonctions simples [51], ce qui rend la mise en œuvre difficile. De fait, l'usage des méthodes de Taylor est rarement recommandé pour p plus grand que deux, mais il existe néanmoins des approches sophistiquées (voir l'article [BWZ71] et les références qu'il contient) pour construire des programmes générant et évaluant *automatiquement* les développements de Taylor requis par la méthode pour certaines formes particulières de fonctions (fonctions rationnelles, fonctions trigonométriques, etc.).

8.4 Analyse des méthodes

Lors de l'étude de la méthode d'Euler dans la sous-section 8.3.1, nous avons introduit les notions de consistance, de stabilité, de convergence et d'ordre d'une méthode numérique fournissant une approximation de la solution du problème de Cauchy (8.1)-(8.5). L'objectif de cette section est de formaliser les définitions données au sein d'un cadre mathématique permettant d'effectuer l'analyse *a priori* de chacune des méthodes présentées et de déterminer dans quelle mesure les solutions des problèmes discrets qui leur sont associés convergent (dans un sens qui sera précisé) vers la solution du problème de Cauchy lorsque le pas de discrétisation tend vers zéro.

Nous allons traiter de manière très générale le cas des méthodes à un pas, qui inclut notamment les méthodes de Runge–Kutta, ce procédé d'analyse pouvant être facilement étendu aux méthodes à pas multiples. Nous ne nous en tenons cependant pas à cet aspect en affinant les résultats obtenus dans le cas les méthodes à pas multiples linéaires au moyen de la théorie associée aux équations aux différences linéaires, dont les grandes lignes sont rappelées dans la prochaine sous-section.

8.4.1 Rappels sur les équations aux différences linéaires *

Soit un entier $q \geq 1$. On appelle équation aux différences linéaire (scalaire) d'ordre q toute équation de la forme

$$\alpha_q\, u_{n+q} + \alpha_{q-1}\, u_{n+q-1} + \cdots + \alpha_0\, u_n = \varphi_{n+q},\ n = 0, \ldots, \tag{8.71}$$

dans laquelle les coefficients α_i, $i = 0, \ldots, q$, supposés réels et tels que $\alpha_0 \alpha_q \neq 0$, peuvent éventuellement dépendre de l'entier n et le scalaire φ_{n+q} est donné pour tout valeur de n. Une équation aux différences linéaire est dite *à coefficients constants* lorsque les coefficients sont indépendants de n et *homogène* si son membre de droite est nul pour tout $n \geq 0$.

Une telle équation admet une solution dès qu'on lui adjoint q conditions initiales spécifiant les valeurs des q premiers termes de la suite $\{u_n\}_{n\geq 0}$, puisque l'on déduit de (8.71) et de l'hypothèse sur les coefficients de l'équation que

$$u_{n+q} = -\frac{1}{\alpha_q} \sum_{i=0}^{q-1} \alpha_i\, u_{n+i} + \frac{\varphi_{n+q}}{\alpha_q},\ n \geq 0.$$

Cette solution est unique, la solution de l'équation aux différences linéaire homogène

$$\alpha_q\, u_{n+q} + \alpha_{q-1}\, u_{n+q-1} + \cdots + \alpha_0\, u_n = 0,\ n \geq 0, \tag{8.72}$$

de valeurs initiales identiquement nulles étant la solution triviale.

Considérons à présent les solutions de l'équation homogène (8.72). En raison de la linéarité de l'équation, ces solutions forment un espace vectoriel et un ensemble de q solutions linéairement indépendantes [52] est appelé un

51. Le lecteur est invité à vérifier cette affirmation avec la fonction $f(t,x) = t^2 + x^2$.

52. On dit que des solutions $\{u_n^{(1)}\}_{n\geq 0}, \{u_n^{(2)}\}_{n\geq 0}, \ldots, \{u_n^{(r)}\}_{n\geq 0}$ de l'équation (8.72) sont *linéairement indépendantes* si le fait d'avoir $\mu_1\, u_n^{(1)} + \mu_1\, u_n^{(1)} + \cdots + \mu_r\, u_n^{(r)} = 0$ pour toute valeur de l'entier n implique que $\mu_1 = \mu_2 = \cdots = \mu_r = 0$.

ensemble fondamental de solutions de l'équation homogène, toute solution pouvant en effet s'exprimer sous la forme d'une combinaison linéaire des éléments de cet ensemble.

Lorsque l'équation est à coefficients constants, il est possible d'expliciter un ensemble fondamental de solutions. Pour cela, on introduit le polynôme caractéristique associé à l'équation

$$\rho(z) = \alpha_q z^q + \alpha_{q-1} z^{q-1} + \cdots + \alpha_0,$$

dont on note les racines ξ_i, $i = 0, \ldots, q-1$. Lorsque celles-ci sont simples (et donc distinctes), l'ensemble des suites des suites linéairement indépendantes $\{{\xi_i}^n\}_{n\geq 0}$, $i = 0, \ldots, q-1$, est un ensemble fondamental de solutions, car on a

$$\alpha_q {\xi_i}^{n+q} + \alpha_{q-1} {\xi_i}^{n+q-1} + \cdots + \alpha_0 {\xi_i}^n = {\xi_i}^n \rho(\xi_i) = 0,\ n \geq 0,\ i = 0, \ldots, q-1.$$

Si au moins une racine est de multiplicité plus grande que un, on peut encore définir un ensemble fondamental de solutions. Pour le voir, supposons que le scalaire ξ soit une racine de multiplicité m du polynôme ρ, avec $m > 1$. Dans ce cas, on remarque que ξ est aussi un zéro de multiplicité m de la fonction $g(z) = z^n \rho(z)$, avec n un entier naturel arbitraire. Les $m-1$ premières dérivées de g s'annulent donc en $z = \xi$ et l'on a

$$\begin{aligned}
\alpha_q {\xi_i}^{n+q} + \alpha_{q-1} {\xi_i}^{n+q-1} + \cdots + \alpha_0 {\xi_i}^n &= 0, \\
\alpha_q (n+q) {\xi_i}^{n+q-1} + \alpha_{q-1} (n+q-1) {\xi_i}^{n+q-2} + \cdots + \alpha_0 n {\xi_i}^{n-1} &= 0, \\
&\vdots \\
\alpha_q (n+q)(n+q-1)\ldots(n+q-m+2) {\xi_i}^{n+q-m+1} + \cdots + \alpha_0 n(n-1)\ldots(n-m+2) {\xi_i}^{n-m+1} &= 0
\end{aligned}$$

ce qui revient à dire que les suites $\{n\,\xi^n\}_{n\geq 0}, \ldots, \{n(n-1)\ldots(n-m+2)\,\xi^n\}_{n\geq 0}$ sont des solutions de l'équation aux différences linéaire homogène. Les $m-1$ suites « manquantes » de l'ensemble fondamental précédemment construit sont alors obtenues à partir de combinaisons linéaires de ces solutions $\{{\xi_0}^n\}_{n\geq 0}$, à savoir $\{n\,{\xi_0}^n\}_{n\geq 0}, \ldots, \{n^{m_0-1}\,{\xi_0}^n\}_{n\geq 0}$.

Plus généralement, en supposant que le polynôme caractéristique associé à l'équation possède r, avec $r \leq q$, racines distinctes ξ_i, $i = 0, \ldots, r-1$, de multiplicités respectives m_i, $i = 0, \ldots, q'-1$, toute solution de (8.72) peut s'écrire

$$u_n = \sum_{j=0}^{r-1} \left(\sum_{i=0}^{m_j - 1} \nu_{ij}\, n^i \right) {\xi_j}^n,\ n \geq 0, \tag{8.73}$$

où les coefficients ν_{ij} sont déterminés par les q conditions initiales imposées. On notera que si une racine de ρ est complexe, une autre racine se trouve être son conjugué, le polynôme étant par hypothèse à coefficients réels. Ces deux racines complexes peut alors être utilisées pour former une paire de solutions réelles de l'équation homogène.

Une fois connue la forme générale des solutions de l'équation homogène, toute solution de l'équation aux différences linéaire non homogène (8.71) s'obtient en ajoutant une solution particulière quelconque de l'équation non homogène à une solution de l'équation homogène dont les coefficients ont été ajustés de façon à ce que la somme des deux satisfasse les q conditions initiales du problème. A REPRENDRE Une telle solution particulière peut être trouvée en résolvant l'équation (8.71) avec des valeurs initiales identiquement nulles et représentée par des solutions de l'équation homogène via un *principe de Duhamel*[53] *discret*.

Théorème 8.17 *Soit* $\left\{\{u_n^{(k)}\}_{n\geq 0}\right\}_{k=0,\ldots,q-1}$ *un ensemble fondamental de solutions de l'équation aux différences* (8.72) *dont les éléments satisfont respectivement les conditions initiales*

$$u_n^{(k)} = \delta_{nk},\ n = 0, \ldots, q-1,\ k = 0, \ldots, q-1,$$

où δ_{ik} *désigne le symbole de Kronecker. La solution de l'équation* (8.71) *s'écrit alors*

$$u_n = \sum_{i=0}^{q-1} u_i\, u_n^{(i)} + \frac{1}{\alpha_q} \sum_{j=q}^{n} \varphi_j\, u_{n-j+q-1}^{(q-1)},\ n = 0, 1 \ldots$$

DÉMONSTRATION. A ECRIRE □

A VOIR : dire un mot sur le cas à coefficients non constants ?

53. Jean-Marie Constant Duhamel (5 février 1797 - 29 avril 1872) était un mathématicien et physicien français. Il est l'auteur de travaux sur les équations aux dérivées partielles modélisant la propagation de la chaleur dans un solide, les cordes vibrantes ou encore la vibration de l'air dans des tubes.

8.4.2 Ordre et consistance

Nous allons maintenant étudier la manière dont la solution calculée par les méthodes que nous avons présentées approchent la solution d'un problème de Cauchy bien posé.

Cas des méthodes à un pas

On notera tout d'abord que toute relation de récurrence définissant une méthode à un pas explicite [54] peut s'écrire

$$x_{n+1} = x_n + h_n \Phi_f(t_n, x_n; h_n),\ n = 0, \ldots, N-1, \tag{8.74}$$

la fonction Φ_f étant parfois appelée la *fonction d'incrément* de la méthode. Dans toute la suite, nous supposons que cette fonction est *continue par rapport à ses trois arguments*.

Exemples de fonction d'incrément de méthode à un pas explicite. Pour la méthode d'Euler, la fonction d'incrément est simplement $\Phi_f(t,x;h) = f(t,x)$. Pour la méthode d'Euler modifiée définie par (8.32), il vient $\Phi_f(t,x;h) = \frac{1}{2} f\left(t+\frac{h}{2}, x+\frac{h}{2} f(t,x)\right)$, et l'on trouve $\Phi_f(t,x;h) = \frac{1}{2}\left(f(t,x) + f(t+h, x+h f(t,x))\right)$ pour la méthode de Heun de schéma (8.41).

Afin d'étudier la convergence d'une méthode de la forme (8.74), il faut en premier lieu s'intéresser à l'erreur commise à chaque étape de la récurrence. Comme nous l'avons vu dans la sous-section 8.3.1, une telle mesure peut se faire par l'intermédiaire de l'erreur de troncature locale associée à la méthode.

Définition 8.18 (erreur de troncature locale d'une méthode à un pas) *Pour tout entier n tel que $0 \leq n \leq N-1$, l'**erreur de troncature locale** au point t_{n+1} d'une méthode à un pas de la forme* (8.74) *est définie par*

$$\tau_{n+1} = \tau(t_{n+1}, x; h_n) = x(t_{n+1}) - x(t_n) - h_n \Phi_f(t_n, x(t_n); h_n), \tag{8.75}$$

où la fonction x désigne une solution de l'équation (8.1).

On remarque que l'erreur de troncature locale n'est autre que le résidu [55] obtenu en insérant une solution exacte de l'équation (8.1) en place d'une solution approchée dans la relation de récurrence (8.74) définissant la méthode. On peut alors légitimement se demander en quel sens ce résidu rend compte de l'erreur produite à chaque étape par la méthode numérique et, *a fortiori*, quel est son rapport avec l'erreur globale, qui est la seule erreur important réellement en pratique.

En explicitant la fonction d'incrément, on peut montrer que l'erreur de troncature locale de la méthode au point t_{n+1} est essentiellement [56] égale à l'*erreur locale* $x(t_{n+1}) - \tilde{x}_{n+1}$ de la méthode en ce même point, où $\tilde{x}_{n+1}$ désigne l'approximation fournie par le schéma de la méthode sous l'hypothèse (8.36), dite localisante, c'est-à-dire

$$\tilde{x}_{n+1} = x(t_n) + h_n \Phi_f(t_n, x(t_n); h_n),\ n = 0, \ldots, N-1.$$

Par exemple, on a pour la méthode d'Euler

$$\tau_{n+1} = x(t_{n+1}) - \left(x(t_n) + h_n f(t_n, x(t_n))\right),\ n = 0, \ldots, N-1,$$

ce qui correspond bien à la définition de l'erreur locale donnée plus haut. On voit avec cette interprétation que, si les erreurs de troncature locales n'ont *a priori* pas de rapport direct avec elle, leur propagation et leur accumulation au cours de la résolution numérique contribuent de manière complexe à l'erreur globale. En ce sens, l'erreur de troncature locale gouverne l'évolution de l'erreur globale, motivant la définition suivante.

54. Pour une méthode à un pas implicite, il faut considérer une relation générale de la forme

$$x_{n+1} = x_n + h_n \Phi_f(t_n, x_{n+1}, x_n; h_n),\ n = 0, \ldots, N-1.$$

Par exemple, la méthode de la règle du trapèze, définie par (8.29), a pour fonction d'incrément $\Phi_f(t_n, x_{n+1}, x_n; h_n) = \frac{1}{2}\left(f(t_n + h_n, x_{n+1}) + f(t_n, x_n)\right)$.

55. Certains auteurs définissent parfois l'erreur de troncature comme ce résidu divisé par la longueur du pas de discrétisation au point considéré.

56. Il y a égalité lorsque la méthode est explicite et égalité à une constante multiplicative près lorsque la méthode est implicite (voir par exemple le lemme 8.23 pour les méthodes à pas multiples linéaires).

Définition 8.19 (consistance d'une méthode à un pas) *Une méthode à un pas de la forme* (8.74) *est dite* ***consistante*** *avec l'équation différentielle* (8.1) *si l'on a*

$$\lim_{h\to 0} \frac{1}{h_n}\,\tau_{n+1} = 0,\ n = 0,\dots,N-1,$$

où τ_{n+1} *désigne l'erreur de troncature locale de la méthode au point* t_{n+1}*, définie par* (8.75).

En d'autres mots, une méthode est consistante si les erreurs de troncature locale aux points de la grille sont des infiniment petits en h^2 (notation $O(h^2)$) lorsque la longueur maximale des pas de discrétisation tend vers zéro.

On peut vérifier qu'une méthode à un pas est consistante en utilisant le résultat suivant.

Théorème 8.20 (condition nécessaire et suffisante de consistance d'une méthode à un pas) *Une méthode numérique à un pas de la forme* (8.74) *est consistante avec l'équation différentielle* (8.1) *si et seulement si*

$$\forall (t,x) \in [\,t_0, t_0+T\,] \times \mathbb{R},\ \Phi_f(t,x;0) = f(t,x). \tag{8.76}$$

DÉMONSTRATION. La condition (8.76) est nécessaire. En effet, si la méthode est consistante, alors, pour toute solution x de classe $\mathscr{C}^1$ de l'équation (8.1), on a

$$\lim_{h\to 0} \frac{1}{h_n}\,\tau_{n+1} = \lim_{h\to 0}\left(\frac{1}{h_n}\int_{t_n}^{t_{n+1}} f(s,x(s))\,\mathrm{d}s - \Phi_f(t_n,x(t_n);h_n)\right) = 0,\ 0 \le n \le N-1.$$

Pour tout t dans $[\,t_0, t_0+T\,]$ et toute grille de discrétisation (non nécessairement uniforme) de pas h, il existe un indice $n(h)$ tel que $t_{n(h)} \le t < t_{n(h)+1}$ et l'on a $\lim_{h\to 0} t_{n(h)} = t$. Il vient alors que

$$\lim_{h\to 0} \Phi_f(t_{n(h)}, x(t_{n(h)}); h_{n(h)}) = \lim_{h\to 0} \frac{1}{h_{n(h)}} \int_{t_{n(h)}}^{t_{n(h)+1}} f(s,x(s))\,\mathrm{d}s,$$

d'où $\Phi_f(t,x(t);0) = f(t,x(t))$. On peut par ailleurs toujours choisir une valeur initiale η en t_0 pour que le couple $(t,x(t))$ prenne une valeur arbitraire dans $[\,t_0, t_0+T\,] \times \mathbb{R}$, d'où le résultat.

Montrons maintenant que la condition est aussi suffisante. On a

$$\begin{aligned}
\frac{1}{h_n}\,\tau_{n+1} &= \frac{1}{h_n}\,\big(x(t_{n+1}) - x(t_n) - h_n\,\Phi_f(t_n,x(t_n),h_n)\big)\\
&= \frac{1}{h_n}\int_{t_n}^{t_{n+1}} \big(f(s,x(s)) - \Phi_f(t_n,x(t_n);h_n)\big)\,\mathrm{d}s\\
&= \frac{1}{h_n}\int_{t_n}^{t_{n+1}} \big(\Phi_f(s,x(s);0) - \Phi_f(t_n,x(t_n);h_n)\big)\,\mathrm{d}s,
\end{aligned}$$

d'où

$$\frac{1}{h_n}\,|\tau_{n+1}| \le \max_{t_n\le s\le t_{n+1}} \big|\Phi_f(s,x(s);0) - \Phi_f(t_n,x(t_n);h_n)\big|.$$

Le membre de droite de cette inégalité tend vers 0 avec h uniformément en n ; la méthode est donc consistante. □

Exemples de méthode à un pas consistante. On a déjà vu que $\Phi_f(t,x;h) = f(t,x)$ pour la méthode d'Euler, ce qui montre une seconde[57] fois que cette méthode est consistante. Pour une méthode de Runge–Kutta à s niveaux, on a, en toute généralité (voir (8.33)), $\Phi_f(t,x;h) = \sum_{i=1}^s b_i\,f(t+c_i\,h, x+h\sum_{j=1}^s a_{ij}k_j)$. On trouve alors que $\Phi_f(t,x;0) = f(t,x)\sum_{i=1}^s b_i$ et la méthode est par conséquent consistante si et seulement si $\sum_{i=1}^s b_i = 1$, ce que l'on l'avait constaté lors de la construction de méthodes de Runge–Kutta explicites à un, deux et trois niveaux dans la sous-section 8.3.2.

Pour avoir une estimation de la précision de l'approximation offerte par une méthode, il faut savoir à quelle vitesse ses erreurs de troncature locales tendent vers zéro avec la longueur du pas de discrétisation, cette information correspondant à la notion d'ordre d'une méthode.

57. On a en effet déjà prouvé que c'était le cas dans la sous-section 8.3.1.

Définition 8.21 (ordre d'une méthode à un pas) *Une méthode à un pas de la forme* (8.74) *est dite* ***d'ordre*** *p, avec p un entier naturel, si*

$$\tau_{n+1} = O(h^{p+1}),\ n = 0, \ldots, N-1,$$

lorsque h tend vers zéro, pour toute solution suffisamment régulière de l'équation (8.1). *Elle est dite* ***exactement d'ordre*** *p si le nombre p est le plus grand entier pour lequel la relation ci-dessus est satisfaite.*

Il découle des définitions 8.19 et 8.21 qu'une méthode d'ordre supérieur ou égal à un est consistante.

On peut affiner la notion de précision d'une méthode donnée ci-dessus en introduisant la définition suivante.

Définition 8.22 (fonction d'erreur principale d'une méthode à un pas) *On appelle* ***fonction d'erreur principale*** *d'une méthode à un pas de la forme* (8.74) *la fonction ψ continue et non identiquement nulle telle que*

$$\tau_{n+1} = \psi(t_n, x(t_n))\, h^{p+1} + O(h^{p+2}),\ n = 0, \ldots, N-1,$$

pour toute solution x suffisamment régulière de l'équation (8.1).

On voit que la fonction d'erreur principale caractérise le terme d'ordre dominant dans l'erreur de troncature locale, c'est-à-dire l'entier p apparaissant dans sa définition donnant l'ordre exact de la méthode. Elle rend explicite, lorsque la longueur du pas de discrétisation est suffisamment petite, le comportement de l'erreur de troncature.

Exemples de fonction d'erreur principale d'une méthode à un pas. Pour la méthode d'Euler, il vient, en effectuant un développement de Taylor de $x(t_{n+1})$ au point t_n au second ordre, la solution étant supposée de classe $\mathscr{C}^2$,

$$\tau_{n+1} = x(t_n) + h_n f(t_n, x(t_n)) + \frac{{h_n}^2}{2}\left(\frac{\partial f}{\partial t}(t_n, x(t_n)) + f(t_n, x(t_n))\,\frac{\partial f}{\partial x}(t_n, x(t_n))\right) + O({h_n}^3) - (x(t_n) + h_n f(t_n, x(t_n))),$$

d'où

$$\psi(t,x) = \frac{1}{2}\left(\frac{\partial f}{\partial t}(t,x) + f(t,x)\,\frac{\partial f}{\partial x}(t,x)\right).$$

Pour la méthode de Heun définie par (8.41), on doit pousser le développement un cran plus loin (voir (8.38)), et donc supposer que la solution est de classe $\mathscr{C}^3$, pour obtenir

$$\psi(t,x) = -\frac{1}{12}\left(\frac{\partial^2 f}{\partial t^2}(t,x) + 2f(t,x)\,\frac{\partial^2 f}{\partial t \partial x}(t,x) + (f(t,x))^2\,\frac{\partial^2 f}{\partial x^2}(t,x)\right) + \frac{1}{6}\left(\frac{\partial f}{\partial t}(t,x)\frac{\partial f}{\partial x}(t,x) + f(t,x)\left(\frac{\partial f}{\partial x}(t,x)\right)^2\right).$$

Pour une méthode de Taylor d'ordre p, on trouve (en reprenant la notation introduite dans la sous-section 8.3.4)

$$\psi(t,x) = \frac{1}{(p+1)!}\, f^{(p)}(t,x).$$

La détermination de l'ordre maximal atteint par une méthode de Runge–Kutta explicite de nombre de niveaux fixé peut se faire par la technique présentée dans la sous-section 8.3.2 pour la construction effective de telles méthodes. Cette approche est néanmoins très technique, car l'expression de la fonction d'erreur principale associée se complique au fur et à mesure que le nombre de niveaux augmente. En notant $p^*(s)$ l'ordre maximal d'une méthode de Runge–Kutta explicite vue comme une fonction du nombre s de niveaux, on sait depuis les travaux de Kutta [Kut01] que $p^*(s) = s$ pour $1 \leq s \leq 4$. Les méthodes explicites d'ordre supérieur nécessitent systématiquement plus de niveaux que l'ordre atteint. Plus précisément, il a été démontré par Butcher, au moyen d'une approche algébrique et pour un problème scalaire (voir [But65]), que

$$\begin{aligned} &p^*(s) = s-1 \text{ pour } 5 \leq s \leq 7,\\ &p^*(s) = s-2 \text{ pour } 8 \leq s \leq 9,\\ &p^*(s) \leq s-2 \text{ pour } s \geq 10. \end{aligned}$$

Enfin, on a vu dans la sous-section 8.3.2 que l'ordre maximal d'une méthode de Runge–Kutta implicite à s niveaux était égal à $2s$.

Cas des méthodes à pas multiples linéaires *

Pour une méthode à pas multiples linéaire de la forme (8.56) et une grille de discrétisation de pas de longueur uniforme, l'erreur de troncature locale prend la forme

$$\tau_{n+q} = \sum_{i=0}^{q} (\alpha_i\, x(t_{n+i}) - h\beta_i\, f(t_{n+i}, x(t_{n+i}))) = \sum_{i=0}^{q} \big(\alpha_i\, x(t_{n+i}) - h\beta_i\, x'(t_{n+i})\big),\ n = 0, \ldots, N-q.$$

Il est dans ce cas commode d'introduire l'*opérateur aux différences* associé à la méthode, que l'on définit, pour toute fonction arbitraire z de classe $\mathscr{C}^1$ sur l'intervalle $[t_0, t_0 + T]$, par

$$\mathscr{L}(z(t), h) = \sum_{i=0}^{q} \big(\alpha_i\, z(t + i\,h) - h\beta_i\, z'(t + i\,h)\big), \tag{8.77}$$

l'erreur de troncature locale s'écrivant alors $\tau_{n+q} = \mathscr{L}(x(t_n), h)$, $n = 0, \ldots, N-q$. On peut voir cet opérateur comme un opérateur linéaire agissant sur toute fonction différentiable.

En introduisant l'*opérateur de décalage à gauche* [58] T_h et en abusant quelque peu des notations [59], on peut réécrire (8.77) de manière compacte en termes de l'opérateur T_h et des polynômes caractéristiques associés à la méthode :

$$\mathscr{L}(z(t), h) = (\rho(T_h) - h\,\sigma(T_h))\, z(t).$$

Lemme 8.23 (lien entre l'erreur de troncature locale et l'erreur locale d'une méthode à pas multiples linéaire) *Soit un problème de Cauchy* (8.1)-(8.5) *pour lequel la fonction f est continûment différentiable et une méthode à pas multiples linéaire de la forme* (8.56) *utilisée pour sa résolution. On a la relation suivante entre l'erreur de troncature locale et l'erreur locale de la méthode*

$$\tau_{n+q} = \left(\alpha_q - h\beta_q\, \frac{\partial f}{\partial x}(t_{n+q}, \eta_{n+q})\right)\big(x(t_{n+q}) - \tilde{x}_{n+q}\big), n = 0, \ldots, N-q,$$

où $\tilde{x}_{n+q}$ est l'approximation fournie par la méthode en supposant que $x_{n+i} = x(t_{n+i})$, $i = 0, \ldots, q-1$, et η_{n+q} est un réel strictement compris entre $x(t_{n+q})$ et $\tilde{x}_{n+q}$.

DÉMONSTRATION. En utilisant l'hypothèse localisante dans (8.56), il vient

$$\sum_{i=0}^{q-1} (\alpha_i\, x(t_{n+i}) - h\beta_i\, f(t_{n+i}, x(t_{n+i}))) + \alpha_q\, \tilde{x}_{n+q} - h\beta_q\, f(t_{n+q}, \tilde{x}_{n+q}) = 0,$$

d'où

$$\mathscr{L}(x(t_n), h) = \alpha_q\, \big(x(t_{n+q}) - \tilde{x}_{n+q}\big) - h\beta_q\, \big(f(t_{n+q}, x(t_{n+q})) - f(t_{n+q}, \tilde{x}_{n+q})\big).$$

Le résultat découle alors du théorème des accroissement finis (voir le théorème B.113). □

Supposons à présent la fonction z infiniment différentiable. En effectuant des développements de Taylor au point t de $z(t + i\,h)$ et $z'(t + i\,h)$, $i = 0, \ldots, q$, dans (8.77) et en regroupant les termes, on obtient

$$\mathscr{L}(z(t), h) = C_0\, z(t) + C_1 h\, z'(t) + \cdots + C_k h^k\, z^{(k)}(t) + \ldots \tag{8.78}$$

où

$$\begin{aligned} C_0 &= \sum_{i=0}^{q} \alpha_i, \\ C_1 &= \sum_{i=0}^{q} (i\,\alpha_i - \beta_i), \\ C_k &= \sum_{i=0}^{q} \left(\frac{i^k}{k!}\,\alpha_i - \frac{i^{k-1}}{(k-1)!}\,\beta_i\right),\ k \geq 2. \end{aligned}$$

REPRENDRE Ceci conduit à la définition suivante.

58. Cet opérateur associe à toute fonction z continue d'une variable réelle la fonction $T_h z = z(\cdot + h)$.

59. On décide de noter les composées multiples de l'opérateur T_h avec lui-même de la façon suivante : ${T_h}^2 = T_h \circ T_h$, ${T_h}^3 = T_h \circ {T_h}^2$, etc.

Définition 8.24 (caractérisation de l'ordre d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire est* ***d'ordre*** *p, avec p un entier naturel, si l'opérateur aux différences, défini par* (8.77)*, qui lui est associé est tel que l'on a* $C_0 = C_1 = \cdots = C_p = 0$ *et* $C_{p+1} \neq 0$ *dans le développement* (8.78)*, la constante* C_{p+1} *étant alors appelée la* ***constante d'erreur principale*** *de la méthode.*

Observons que cette définition est légitime. En effet, le développement de Taylor (8.78) est vrai pour toute fonction suffisamment régulière et donc, en particulier, toute solution suffisamment régulière du problème. D'autre part, le calcul montre que le premier coefficient non nul, C_{p+1} est *indépendant* du point t en lequel le développement est effectué (ce qui n'est en revanche pas le cas des coefficients suivants).

En reliant les expressions des constantes C_k, $k \geq 0$, aux valeurs des polynômes ρ, σ et ρ', on obtient la caractérisation analytique suivante de la consistance et de l'ordre d'une méthode à pas multiples linéaire.

Théorème 8.25 (conditions nécessaires et suffisantes de consistance et d'ordre d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire de la forme* (8.56) *est consistante si et seulement si ses polynômes caractéristiques premier et second satisfont*

$$\rho(1) = 0 \text{ et } \rho'(1) = \sigma(1). \tag{8.79}$$

Elle est d'ordre p si et seulement si l'on a de plus

$$\frac{\rho(z)}{\sigma(z)} = \ln(z) + O\left((z-1)^{p+1}\right). \tag{8.80}$$

dans un voisinage du point $z = 1$.

DÉMONSTRATION. Par définition, une méthode à pas multiples linéaire est consistante si elle est au moins du premier ordre, c'est-à-dire si $C_0 = C_1 = 0$. On déduit les conditions (8.79), qui ne sont qu'une réécriture des expressions de ces constantes en termes des polynômes caractéristiques de la méthode.

A FINIR □

On vient de mettre en évidence un lien entre l'ordre d'une méthode à pas multiples linéaire et l'ordre d'approximation de la fonction ln par la fonction rationnelle $\frac{\rho}{\sigma}$ au voisinage du point $z = 1$.

REPRENDRE/FINIR

Détermination de l'ordre de convergence des méthodes d'Adams. compte tenu de leur construction, au moins q explicite ($q+1$ implicite), détermination de la constante d'erreur principale indique que c'est l'ordre exact

constante de la méthode de la règle du trapèze ? ? ?

Résumé dans le tableau 8.3

nom de la méthode à q pas	ordre	constante d'erreur principale
Adams–Bashforth	q	γ_q
Adams–Moulton	$q+1$	γ^*_{q+1}
Nyström ($q=2$)	2	$\frac{1}{6}$
Nyström ($q>2$)	q	$\frac{\kappa_q}{2}$
Milne–Simpson ($q=2$)	4	$-\frac{1}{180}$
Milne–Simpson ($q>3$)	$q+1$	$\frac{\kappa^*_{q+1}}{2}$
BDF	q	$-\frac{1}{q+1}$

TABLE 8.3 – Ordre et constante d'erreur principale de différentes méthodes à q pas linéaires.

8.4.3 Zéro-stabilité *

La zéro-stabilité d'une méthode de résolution numérique d'un problème de Cauchy bien posé caractérise le comportement de son schéma vis-à-vis de l'accumulation de perturbations lorsque le pas de discrétisation tend vers zéro. Cette propriété de la méthode assure que cette dernière n'est pas trop sensible aux erreurs de représentation des données ou d'arrondi en arithmétique en précision finie et qu'elle fournit une approximation effectivement calculable de la solution. On peut la voir comme un avatar, spécifique à résolution approchée des équations différentielles ordinaires, de la notion de stabilité numérique discutée dans la section 1.5.2.

on veut que la solution numérique du problème avec une donnée perturbée reste proche de la solution avec donnée non perturbée

REPRENDRE/DEPLACER on va voir que l'on peut décider du fait qu'une méthode est zéro-stable ou non en considérant simplement son application à la résolution d'un problème trivial dont l'équation est $x'(t) = 0$ d'où le nom du concept de stabilité introduit dans la définition ref / stabilité dans le cas limite $h \to 0$ d'où le nom (aussi appelée stabilité de Dahlquist)

Cas des méthodes à un pas

Définition 8.26 (zéro-stabilité d'une méthode à un pas) *On dit qu'une méthode à un pas de la forme* (8.74) *pour la résolution de l'équation différentielle ordinaire* (8.1) *est* ***zéro-stable*** *s'il existe une constante strictement positive C, indépendante de la longueur des pas de discrétisation, telle que, pour toutes suites $(x_n)_{0\leq n\leq N}$ et $(y_n)_{0\leq n\leq N}$ définies respectivement par*

$$x_{n+1} = x_n + h_n\,\Phi_f(t_n, x_n; h_n),\ n = 0,\dots,N-1,$$

et

$$y_{n+1} = y_n + h_n\left(\Phi_f(t_n, y_n; h_n) + \varepsilon_n\right),\ n = 0,\dots,N-1,$$

les initialisations x_0 et y_0 et les perturbations ε_n, $n = 0,\dots,N-1$, étant donnés, on a, pour tout h suffisamment petit,

$$\max_{0\leq n\leq N} |x_n - y_n| \leq C\left(|x_0 - y_0| + \max_{0\leq i\leq N-1} |\varepsilon_i|\right). \tag{8.81}$$

DONNER EXPLICATIONS... A VOIR CONDITION SUR h

Théorème 8.27 (condition suffisante de zéro-stabilité d'une méthode à un pas) *Une méthode à un pas de la forme* (8.74) *pour la résolution de l'équation différentielle ordinaire* (8.1) *est zéro-stable s'il existe une constante strictement positive Λ telle que l'on a*

$$\forall t \in [t_0, t_0+T],\ \forall (x,y) \in \mathbb{R}^2,\ \forall h \in [0, h_0]\ (\textbf{\textit{ou}}\ \mathbb{R}_+),\ \left|\Phi_f(t,x;h) - \Phi_f(t,y;h)\right| \leq \Lambda\,|x-y|\,. \tag{8.82}$$

La preuve de ce théorème utilise le résultat suivant, que l'on peut voir comme une version discrète de l'inégalité de Grönwall (voir la proposition 8.15).

Lemme 8.28 *Soit une suite réelle $(e_n)_{0\leq n\leq N}$ dont les termes satisfont*

$$e_{n+1} \leq a_n\,e_n + b_n,\ n = 0,\dots,N-1, \tag{8.83}$$

avec $(a_n)_{0\leq n\leq N-1}$ et $(b_n)_{0\leq n\leq N-1}$ des suites réelles avec $a_n > 0$. On a alors [60]

$$e_n \leq \left(\prod_{i=0}^{n-1} a_i\right) e_0 + \sum_{j=0}^{n-1}\left(\prod_{k=j+1}^{n-1} a_k\right) b_j,\ n = 0,\dots,N.$$

DÉMONSTRATION. On observe tout d'abord que

$$\left(\prod_{i=0}^{n} a_i\right) e_0 + \sum_{j=0}^{n}\left(\prod_{k=j+1}^{n} a_k\right) b_j = a_n\left(\left(\prod_{i=0}^{n-1} a_i\right) e_0 + \sum_{j=0}^{n-1}\left(\prod_{k=j+1}^{n-1} a_k\right) b_j\right) + b_n,\ n = 0,\dots,N-1.$$

60. On adopte ici la convention usuelle qu'un produit « vide » a pour valeur 1 et qu'une somme « vide » a pour valeur 0.

En soustrayant cette égalité à l'inégalité (8.83), on trouve

$$e_{n+1}-\left(\left(\prod_{i=0}^{n}a_i\right)e_0+\sum_{j=0}^{n}\left(\prod_{k=j+1}^{n}a_k\right)b_j\right)\leq a_n\left(e_n-\left(\left(\prod_{i=0}^{n-1}a_i\right)e_0+\sum_{j=0}^{n-1}\left(\prod_{k=j+1}^{n-1}a_k\right)b_j\right)\right),\ n=0,\dots,N-1.$$

Pour $n=0$, le membre de gauche de cette inégalité devient $e_1-(a_0\,e_0+b_0)$, cette quantité étant négative en vertu de (8.83). Un raisonnement par récurrence permet alors de montrer alors qu'on a plus généralement

$$e_n-\left(\left(\prod_{i=0}^{n-1}a_i\right)e_0+\sum_{j=0}^{n-1}\left(\prod_{k=j+1}^{n-1}a_k\right)b_j\right)\leq 0,\ n=0,\dots,N.$$

□

DÉMONSTRATION DU THÉORÈME 8.27. Considérons les suites $(x_n)_{0\leq n\leq N}$ et $(y_n)_{0\leq n\leq N}$ de la définition 8.26. Leur différence satisfait la relation de récurrence

$$y_{n+1}-x_{n+1}=y_n-x_n+h_n\left(\Phi_f(t_n,y_n;h_n)-\Phi_f(t_n,x_n;h_n)\right)+h_n\,\varepsilon_n,\ n=0,\dots,N-1.$$

Il découle alors de l'inégalité triangulaire et la condition de Lipschitz (8.82) que

$$|y_{n+1}-x_{n+1}|\leq(1+\Lambda h_n)\,|y_n-x_n|+h_n\,|\varepsilon_n|,\ n=0,\dots,N-1.$$

En utilisant le lemme 8.28 en posant $e_n=|y_n-x_n|$, $n=0,\dots,N$, $a_n=1+\Lambda h_n$ et $b_n=h_n\,|\varepsilon_n|$, $n=0,\dots,N-1$, tout en remarquant que

$$\prod_{k=j+1}^{n-1}(1+\Lambda h_k)\leq\prod_{k=0}^{n-1}(1+\Lambda h_k)\leq\prod_{k=0}^{N-1}(1+\Lambda h_k)\leq\prod_{k=0}^{N-1}e^{\Lambda h_k}=e^{\Lambda\sum_{k=0}^{N-1}h_k}=e^{\Lambda T},\ j=0,\dots,n-1,$$

on arrive à

$$|y_n-x_n|\leq e^{\Lambda T}\left(|y_0-x_0|+\sum_{j=0}^{n-1}h_j\,\left|\varepsilon_j\right|\right),\ n=0,\dots,N,$$

d'où

$$|y_n-x_n|\leq e^{\Lambda T}\left(|y_0-x_0|+T\max_{1\leq j\leq n-1}\left|\varepsilon_j\right|\right),\ n=0,\dots,N,$$

qui conduit à la condition de zéro-stabilité avec $C=e^{\Lambda T}\max\{1,T\}$. □

Les fonctions d'incrément de toutes les méthodes à un pas utilisées en pratique satisfont une condition de Lipschitz par rapport à x si cela est le cas pour la fonction f définissant l'équation différentielle ordinaire, la constante Λ du théorème 8.27 pouvant alors s'exprimer en fonction de la constante L de la condition (8.24), comme le montrent les exemples suivants.

Exemples de méthode à un pas zéro-stable. Sous l'hypothèse que la fonction f est lipschitzienne, on obtient immédiatement la zéro-stabilité de la méthode d'Euler (8.26) puisque l'on a $\Phi_f(t,x;h)=f(t,x)$, d'où $\Lambda=L$. De la même manière, pour la méthode d'Euler modifiée (8.32), on a $\Phi_f(t,x;h)=f\left(t+\frac{h}{2},x+\frac{h}{2}f(t,x)\right)$ et il vient $\Lambda\leq L\left(1+\frac{1}{2}hL\right)$. Enfin, pour la méthode de Runge–Kutta « classique » résumée dans le tableau (8.43), on obtient, après quelques majorations,

$$\left|\Phi_f(t,x;h)-\Phi_f(t,y;h)\right|\leq\frac{L}{6}\left(1+2\left(1+\frac{1}{2}hL\right)+2\left(1+\frac{1}{2}hL+\frac{1}{4}(hL)^2\right)+\left(1+hL+\frac{1}{2}(hL)^2+\frac{1}{4}(hL)^3\right)\right)|x-y|,$$

d'où $\Lambda\leq L\left(1+\frac{1}{2}hL+\frac{1}{6}(hL)^2+\frac{1}{24}(hL)^3\right)$.

FAIRE UNE REMARQUE sur le fait que la constante de stabilité $C=e^{\Lambda T}\max\{1,T\}$ devient très grande lorsque T ou L (et donc Λ) sont grands. Ce résultat de stabilité de la solution numérique n'est donc pas d'une réelle utilité lorsqu'il s'agit d'étudier la sensibilité par rapport à des perturbations en temps long (T grand) ou quand le système est raide (L grand).

Cas des méthodes à pas multiples linéaires

Définition 8.29 (zéro-stabilité d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire de la forme* (8.56) *pour la résolution de l'équation différentielle ordinaire* (8.1) *est* ***zéro-stable*** *s'il existe une constante $C > 0$, indépendante de la longueur des pas de la grille de discrétisation, telle que, pour toutes suite x_n et y_n définies respectivement par*

$$x_{n+q} = h \sum_{i=0}^{q} \beta_i f(t_{n+i}, x_{n+i}) - \sum_{i=0}^{q-1} \alpha_i x_{n+i},\ n = 0, \ldots, N-q,$$

et

$$y_{n+q} = h \left(\sum_{i=0}^{q} \beta_i f(t_{n+i}, y_{n+i}) + \varepsilon_{n+q} \right) - \sum_{i=0}^{q-1} \alpha_i y_{n+i},\ n = 0, \ldots, N-1,$$

x_i, y_i $i = 0, \ldots, q-1$, et ε_{n+q} étant donnés, on ait, pour tout h suffisamment petit,

$$\max_{0 \le n \le N} |x_n - y_n| \le C \left(\max_{0 \le i \le q-1} |x_i - y_i| + \max_{0 \le i \le N-1} |\varepsilon_i| \right). \tag{8.84}$$

zéro-stabilité d'une méthode multipas linéaire $\Leftrightarrow$ les solutions de l'équation aux différences sous-jacente sont bornées, ce qui est lié aux racines du premier polynôme caractéristique ρ et plus précisément à leur localisation dans le plan complexe

Définition 8.30 (« condition de racine ») *On dit qu'une méthode à pas multiples linéaire de la forme* (8.56) *pour la résolution du problème* (8.1)-(8.5) *satisfait la* ***condition de racine*** *si toutes les racines de son premier polynôme caractéristique ρ ont un module inférieur ou égal à l'unité et que toutes les racines de module égal à un sont simples.*

NOTE : Si la méthode est consistante ($\rho(1) = 0$) alors 1 est racine du polynôme, c'est la racine principale. les $q-1$ racines restantes sont dites "spurieuses" et proviennent de la représentation approchée d'un système différentiel du premier ordre par un système aux différences d'ordre q. On note que toute méthode à un pas consistante satisfait automatiquement la condition de racine.

REPRENDRE Comme pour la consistance, il est donc possible de caractériser analytiquement la stabilité d'une méthode multipas. On a la zéro-stabilité si les racines du polynôme satisfont la condition suivante.

Théorème 8.31 (condition nécessaire et suffisante de zéro-stabilité d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire est zéro-stable si et seulement si elle vérifie la condition de racine.*

DÉMONSTRATION. Pour montrer que la condition est nécessaire, nous allons raisonner par l'absurde, en supposant que la méthode est stable et que la condition énoncée dans la définition 8.30 est violée. L'inégalité (8.84) étant satisfaite pour tout problème de Cauchy bien posé, elle l'est en particulier pour le problème d'équation différentielle $x'(t) = 0$ et de condition initiale $x(0) = 0$, dont la solution est identiquement nulle. Pour cette équation, le schéma d'une méthode à pas multiples linéaire de la forme (8.56) s'écrit

$$\sum_{i=0}^{q} \alpha_i x_{n+i} = 0, n \ge 0.$$

Notons ξ_i, $i = 1, \ldots, r$, avec $r \le q$, les racines du polynôme ρ ; il existe alors une racine ξ_j dont le module est strictement plus grand que l'unité, ou bien de module égal à l'unité mais de multiplicité strictement supérieure à un. On sait d'après les résultats rappelés dans la sous-section 8.4.1 que la suite A COMPLETER/REPRENDRE

$$x_n = \begin{cases} {\xi_j}^n & \text{si } \xi_j \in \mathbb{R} \\ (\xi_j + \overline{\xi_j})^n & \text{si } \xi_j \in \mathbb{C} \setminus \mathbb{R} \end{cases}$$

dans le premier cas ou

$$x_n = \begin{cases} n\,{\xi_j}^n & \text{si } \xi_j \in \mathbb{R} \\ (\xi_j + \overline{\xi_j})^n & \text{si } \xi_j \in \mathbb{C} \setminus \mathbb{R} \end{cases}$$

est solution...

La preuve que la condition est également suffisante est longue et technique et nous l'omettons pour cette raison. □

On voit une nouvelle fois avec ce résultat que la vérification d'une propriété cruciale d'une méthode à pas multiples linéaire se ramène à une question algébrique par l'utilisation de l'analyse complexe. En effet, si l'on a

montré dans la sous-section précédente que la méthode est consistante si la fonction rationnelle $\frac{\rho}{\sigma}$ approche la fonction ln à l'ordre deux au voisinage de $z = 1$, on vient de prouver qu'elle est zéro-stable si les zéros de son premier polynôme caractéristique ρ sont tous contenus dans le disque unité et simples s'ils appartiennent au cercle unité.

Exemple de méthode à pas multiples linéaire instable. La méthode explicite à deux pas définie par le schéma

$$x_{n+2} = 3\,x_{n+1} - 2\,x_n + h\left(\frac{1}{2}f(t_{n+1}, x_{n+1}) - \frac{3}{2}f(t_n, x_n)\right),\ n = 0, \ldots, N-2, \tag{8.85}$$

a pour polynômes caractéristiques $\rho(z) = z^2 - 3z + 2$ et $\sigma(z) = \frac{1}{2}z - \frac{3}{2}$. Cette méthode est consistante, car $\rho(1) = 0$ et $\rho'(1) = -1 = \sigma(1)$, et d'ordre deux, car $C_2 = 0$ et $C_3 \neq 0$. Elle ne vérifie cependant pas la condition de racine, les racines de ρ étant 1 et 2, et n'est donc pas zéro-stable. La figure 8.11 illustre le phénomène d'instabilité observé en pratique, qui se traduit par une augmentation rapide de l'erreur globale de la méthode lorsque l'on diminue la longueur du pas de discrétisation.

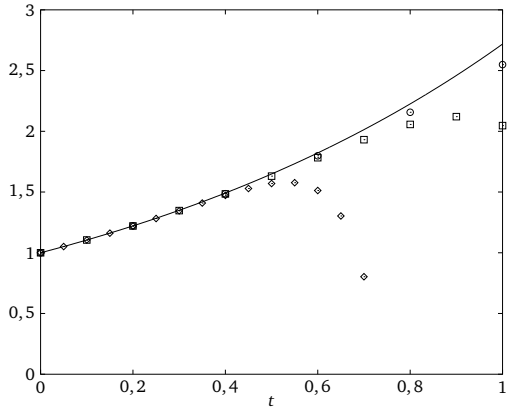


FIGURE 8.11 – Illustration de l'instabilité de la méthode à pas multiples linéaire définie par (8.85) lors de la résolution du problème de Cauchy d'équation $x'(t) = x(t)$ et de condition initiale $x(0) = 1$, sur l'intervalle $[0, 1]$. La courbe représente la solution du problème, $x(t) = e^x$, et les points marqués des symboles $\circ$, $\square$ et $\diamond$ les valeurs numériques $\{x_n\}_{0 \leq n \leq N}$ obtenues sur des grilles de discrétisation uniformes de longueurs de pas respectives $h = 0,2$, $h = 0,1$ et $h = 0,05$.

Le théorème 8.31 permet de prouver très simplement la stabilité de plusieurs des familles de méthodes à pas multiples linéaires introduites dans la sous-section 8.3.3. Dans le cas des méthodes d'Adams, il vient $\rho(z) = z^q - z^{q-1} = z^{q-1}(z-1)$, $q \geq 1$ et la condition de racine est donc satisfaite. Il en va de même pour les méthodes de Nÿstrom et de Milne–Simpson généralisées, pour lesquelles on a $\rho(z) = z^q - z^{q-2} = z^{q-2}(z^2-1)$. L'analyse de zéro-stabilité des méthodes BDF n'est pas aussi n'est pas aussi immédiate, car leurs premiers polynômes caractéristiques ne sont pas de forme triviale. Il apparaît que ces dernières méthodes ne sont stables que pour de valeurs de q inférieures ou égales à six, un fait observé numériquement dès les années 1950 [MC53] mais seulement prouvé de façon rigoureuse [61] près d'une vingtaine d'années plus tard [Cry72, CM75].

Nous terminons cette sous-section par un résultat important, montrant que l'ordre maximal atteint par une méthode à pas multiples linéaire à q pas sera de loin inférieur à la valeur théoriquement possible [62] de $2q$ si l'on souhaite qu'elle soit zéro-stable.

Théorème 8.32 (« première barrière de Dahlquist [63] » [Dah56]) *Il n'existe pas de méthode à q pas linéaire zéro-stable dont l'ordre est supérieur à $q+1$ si q est impair et à $q+2$ si q est pair. Par ailleurs, si la méthode est explicite, son ordre ne peut être plus grand que q.*

61. Une preuve très courte et élégante de l'instabilité des méthodes BDF pour $q \geq 7$ est donnée dans [HW83].

62. Pour définir une méthode à q pas linéaire, on a $2(q+1)$ coefficients α_i et β_i, $i = 0, \ldots, q$, à choisir, parmi lesquels on pose $\alpha_q = 1$ pour satisfaire (8.57). On a donc $2q+1$ paramètres libres (seulement $2q$ pour une méthode explicite car $\beta_q = 0$ dans ce cas) alors que l'on a $p+1$ équations linéaires à vérifier pour que la méthode soit d'ordre p. Par conséquent, l'ordre le plus élevé que l'on peut atteindre est $2q$ si la méthode est implicite et $2q-1$ si elle est explicite.

63. Germund Dahlquist (16 janvier 1925 - 8 février 2005) était un mathématicien suédois, principalement connu pour ses contributions à l'analyse numérique des méthodes de résolution des équations différentielles.

DÉMONSTRATION. A ECRIRE □

A VOIR on peut mentionner la théorie des *étoiles d'ordre* (*order stars* en anglais) de Wanner, Hairer et Nørsett [WHN78] ici, car il existe une preuve de ce résultat reposant dessus

REPRENDRE Il découle de ce théorème qu'une méthode à q pas zéro-stable et d'ordre $q+2$ est *optimale*. On peut montrer que toutes les racines spurieuses du premier polynôme caractéristique d'une telle méthode se trouvent sur le cercle unité, ce qui pose d'autres problèmes de stabilité (voir la sous-section 8.4.5).

8.4.4 Convergence

Munis des propriétés de consistance et de zéro-stabilité, nous pouvons maintenant prouver que les méthodes numériques convergent vers la solution du problème de Cauchy.

Cas des méthodes à un pas

Définition 8.33 (convergence d'une méthode à un pas) *On dit qu'une méthode numérique de la forme* (8.74) *est convergente si, pour tout problème de Cauchy* (8.1)-(8.5) *satisfaisant les hypothèses du théorème 8.10, on a*

$$\lim_{h\to 0}\left(\max_{0\le n\le N}|x_n - x(t_n)|\right) = 0,$$

dès que

$$\lim_{h\to 0}|x_0 - x(t_0)| = 0.$$

On remarque que cette définition impose que la convergence des suites d'approximations construites le schéma sous condition de convergence de l'initialisation x_0. EXPLIQUER en faisant une REMARQUE SUR LES ERREURS D'ARRONDI SUR MACHINE

Théorème 8.34 *Si une méthode à un pas de la forme* (8.74) *est consistante et zéro-stable, alors elle est convergente*

DÉMONSTRATION. Soit x la solution du problème (8.1)-(8.5). En appliquant l'inégalité de stabilité (8.81) aux suites $(x_n)_{0\le n\le N}$ et $(x(t_n))_{0\le n\le N}$, qui satisfont respectivement (8.74) et (8.75), il vient

$$\max_{0\le n\le N}|x_n - x(t_n)| \le C\left(|x_0 - x(t_0)| + \max_{0\le n\le N-1}\left|\frac{\tau_{n+1}}{h_n}\right|\right).$$

On déduit alors la convergence de la méthode de la condition de consistance. □

Corollaire 8.35 *Si la fonction d'incrément d'une méthode à un pas est continue par rapport à ses variables et satisfait une condition de Lipschitz par rapport à x, la méthode est convergente si elle est consistante.*

A VOIR On peut montrer que cette condition suffisante est également nécessaire. On va généraliser ces résultats pour les méthodes à pas multiples linéaires (auxquelles appartiennent des méthodes à un pas comme la méthode d'Euler).

Cas des méthodes à pas multiples linéaires

Pour les méthodes à pas multiples linéaires, la définition de la convergence doit être adaptée pour tenir compte de l'initialisation particulière requise par ces méthodes.

Définition 8.36 (convergence d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire de la forme* (8.56) *est dite* ***convergente*** *si, pour tout problème de Cauchy* (8.1)-(8.5) *satisfaisant les hypothèses du théorème 8.10, on a*

$$\lim_{h\to 0}\left|x_{\lceil (t-t_0)/h\rceil} - x(t)\right| = 0,\ \forall t\in[t_0, t_0+T],$$

dès que les valeurs d'initialisation x_i, $i=0,\dots,q-1$, satisfont

$$\lim_{h\to 0}|x_i - x(t_0+ih)| = 0,\ i=0,\dots,q-1.$$

Nous allons maintenant donner une condition nécessaire de convergence pour les méthodes à pas multiples linéaires. Nous aurons besoin de deux lemmes techniques, le premier servant à démontrer le second.

Lemme 8.37 *Supposons que le polynôme $\rho(z) = \alpha_q z^q + \alpha_{q-1} z^{q-1} + \cdots + \alpha_0$ satisfasse la condition de racines de la définition 8.30. Alors, les coefficients γ_i, $i \geq 0$, du développement*

$$\frac{1}{\alpha_0 z^q + \alpha_1 z^{q-1} + \cdots + \alpha_q} = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \ldots$$

sont bornés, i. e.

$$\Gamma = \sup_{i=0,1,\ldots} |\gamma_i| < +\infty.$$

DÉMONSTRATION. Posons $\hat{\rho}(z) = z^q \rho(z^{-1})$. Les racines du polynôme réciproque $\hat{\rho}$ sont les inverses des racines de ρ et la fonction qui a z associe $\frac{1}{\hat{\rho}(z)}$ est donc holomorphe dans le disque unité ouvert $\{z \in \mathbb{C} \,|\, |z| < 1\}$. Les racines de ρ de module égal à 1, notées $\xi_1, \xi_2, \ldots, \xi_m$ étant simples, les pôles de $\frac{1}{\hat{\rho}}$ de module égal à 1 sont d'ordre un et il existe des coefficients $A_1, A_2, \ldots, A_m$ tels que la fonction

$$g(z) = \frac{1}{\hat{\rho}(z)} - \frac{A_1}{z - {\xi_1}^{-1}} - \frac{A_1}{z - {\xi_2}^{-1}} - \cdots - \frac{A_1}{z - {\xi_m}^{-1}},$$

est aussi holomorphe dans le disque unité fermé $\{z \in \mathbb{C} \,|\, |z| \leq 1\}$. Cette dernière fonction est donc développable en série entière,

$$g(z) = \sum_{n=0}^{+\infty} a_n z^n,$$

les coefficients a_n, donnés par $a_n = \frac{1}{2i\pi} \int_{|z|=1} \frac{g(z)}{z^{n+1}} \, \mathrm{d}z$, $n \in \mathbb{N}$, étant bornés indépendamment de n. De la même manière, on peut développer en série chacun des éléments simples $\frac{A_i}{z - {\xi_i}^{-1}}$, $i = 1, \ldots, m$, les coefficients des développements associés étant également bornés, ce qui achève la preuve. □

Le second lemme concerne la croissance des solutions de l'équation aux différences non homogène suivante

$$\alpha_q z^{n+q} + \alpha_{q-1} z^{n+q-1} + \cdots + \alpha_0 z^n = h \left(\beta_q z^{n+q} + \beta_{q-1} z^{n+q-1} + \cdots + \beta_0 z^n \right) + \lambda_n, \; n = 0, \ldots, N-q. \tag{8.86}$$

Lemme 8.38 *Supposons que le polynôme $\rho(z) = \alpha_q z^q + \alpha_{q-1} z^{q-1} + \cdots + \alpha_0$ satisfasse la condition de racines et soit B^*, β et Λ des constantes positives telles que*

$$\sum_{i=0}^{q} |\beta_i| \leq B^*, \; |\beta_i| \leq \beta, \; |\lambda_i| \leq \Lambda, \; i = 0, \ldots, N,$$

et soit $0 \leq h < \left|\alpha_q\right| \beta^{-1}$. Alors toute solution de (8.86) *pour laquelle $|z_i| \leq Z$, $i = 0, \ldots, q-1$,*

$$|z_n| \leq K^* \, e^{nhL^*}, \; n = 0, \ldots, N,$$

où $L^* = \Gamma^* B^*$, $K^* = \Gamma^* (N\Lambda + (\sum_{i=0}^{q} |\alpha_i|) Zq)$, $\Gamma^* = \frac{\Gamma}{1 - h \left|\alpha_q\right|^{-1} \beta}$.

DÉMONSTRATION. A ECRIRE □

Théorème 8.39 (« théorème d'équivalence de Dahlquist » [Dah56]) *Une condition nécessaire et suffisante pour qu'une méthode à pas multiples linéaire pour la résolution de* (8.1)-(8.5) *soit convergente est qu'elle soit consistante et zéro-stable.*

DÉMONSTRATION. Commençons par montrer que la convergence de la méthode implique sa stabilité. Si la méthode est convergente, elle l'est en particulier lorsqu'on l'utilise pour la résolution du problème de Cauchy d'équation $x'(t) = 0$ et de condition initiale $x(t_0) = 0$, ayant pour solution $x(t) = 0$. Dans ce cas, le schéma de la méthode se résume à l'équation aux différences linéaire

$$\sum_{i=0}^{q} \alpha_i x_{n+i} = 0, \; n = 0, \ldots, N-q. \tag{8.87}$$

Raisonnons par l'absurde et supposons que la méthode est instable. En vertu du théorème 8.31, ceci signifie que l'équation (8.87) a pour solution particulière la suite $\{u_n\}_{n=0,\ldots,N}$, avec $u_n = \xi^n$, $|\xi| > 1$, ou bien $u_n = n \, x i^n$, $|\xi| = 1$, selon que le polynôme

caractéristique ρ associé à la méthode possède une racine ξ de module plus grand que 1 ou bien de module égal à 1 et de multiplicité supérieure à 1. Dans n'importe lequel de ces cas de figure, la suite définie par $x_n = \sqrt{h}\,u_n$, $n = 0, \ldots, N$, est telle que ses q premiers termes tendent vers 0 quand h tend vers 0. En revanche, pour tout t fixé tel que $t_0 < t \leq t_0 + T$, on a que $|x_n|$, avec $nh = t - t_0 > 0$, tend vers l'infini quand h tend vers 0, ce qui est en contradiction avec le fait que la méthode est convergente.

Prouvons maintenant que la convergence de la méthode implique sa consistance. Pour cela, considérons tout d'abord son application pour la résolution du problème d'équation $x'(t) = 0$ et de condition initiale $x(t_0) = 1$, qui a pour solution $x(t) = 1$, le schéma de la méthode étant une nouvelle fois (8.87). En supposant que les valeurs d'initialisation sont exactes, la convergence de la méthode implique que les termes de la suite $\{x_n\}_{n=0,\ldots,N}$ convergent vers 1 lorsque h tend vers 0 et l'on en déduit que $\rho(1) = \sum_{i=0}^{q} \alpha_i = 0$. Considérons ensuite la résolution du problème d'équation $x'(t) = 0$ et de condition initiale $x(t_0) = 1$, dont la solution est $x(t) = t - t_0$. Le schéma de la méthode s'écrit alors $\sum_{i=0}^{q} \alpha_i\, x_{n+i} = h \sum_{i=0}^{q} \beta_i$, $n = 0, \ldots, N-q$. Une solution particulière de cette équation est donnée par la suite $\left\{ \frac{\sigma(1)}{\rho'(1)}\, hn \right\}_{n=0,\ldots,N}$. En effet, puisque $\rho(1) = 0$, on vérifie que

$$\frac{\sigma(1)}{\rho'(1)}\, h \sum_{i=0}^{q} \alpha_i\, (n+i) - h \sum_{i=0}^{q} \beta_i = \frac{\sigma(1)}{\rho'(1)}\, h \big(n\,\rho(1) + \rho'(1) \big) - h\,\sigma(1) = \frac{\sigma(1)}{\rho'(1)}\, h\,\rho'(1) - h\,\sigma(1) = 0,\ n = 0, \ldots, N-q.$$

D'autre part, les q premiers termes de cette suite tendant vers 0 lorsque h tend vers 0, on déduit de l'hypothèse de convergence de la méthode que, pour tout t fixé tel que $t_0 \leq t \leq t_0 + T$, $\frac{\sigma(1)}{\rho'(1)}\, hn$, avec $nh = t - t_0$, tend vers hn lorsque h tend vers 0, et par conséquent $\rho'(1) = \sigma(1)$. On en conclut que la méthode est consistante par le théorème 8.25.

REPRENDRE Montrons enfin que la consistance et la zéro-stabilité de la méthode impliquent sa convergence. Soit un problème de Cauchy eqref dont la fonction f satisfait les hypothèses du théorème ref et la valeur initiale arbitraire. Considérons la suite de valeurs $\{x_n\}$ solution de l'équation aux différences eqref (schema) obtenue à partir de q valeurs d'initialisation $x_0, \ldots, x_{q-1}$ satisfaisant l'hypothèse eqref (à définir). Pour tout n..., on a d'une part, en vertu du théorème des accroissements finis

$$x(t_{n+i}) = x(t_n) + ih\, x'(\eta_i),\ i = \ldots$$

avec $x_n < \eta_i < x_{n+i}$, et d'autre part, la fonction x' étant continue sur l'intervalle fermé $[t_0, t_0 + T]$,

$$x'(t_{n+i}) = x'(t_n) + \theta_i\, \omega(x', ih),\ i = \ldots$$

avec $\omega(x', \varepsilon) =$, $|\theta_i| < 1$, et, en utilisant eqref,

$$x(t_{n+i}) = x(t_n) + ih\,(x'(t_n) + \theta_i'\, \omega(x', ih)),\ i = \ldots$$

avec $\left|\theta_i'\right| < 1$. La méthode étant consistante, on a $\sum_{i=0}^{q} \alpha_i = 0$ et $\sum_{i=0}^{q} (i\alpha_i - \beta_i) = 0$ et il vient alors

$$|\mathscr{L}(x(t_n), h)| \leq \left(\sum_{i=0}^{q} i\,|\alpha_i| + |\beta_i| \right) \omega(x', qh).$$

Par ailleurs,

$$\sum_{i=0}^{q} \alpha_i\, (x_{n+i} - x(t_{n+i})) - h \sum_{i=0}^{q} \beta_i\, (f(t_{n+i}, x_{n+i}) - f(t_{n+i}, x(t_{n+i}))) = \theta_n h \left(\sum_{i=0}^{q} i\,|\alpha_i| + |\beta_i| \right) \omega(x', qh),$$

avec $|\theta_n| \leq 1$. La fonction f vérifiant une condition de Lipschitz, on peut appliquer le lemme 8.38 avec $z_n = x_n - x(t_n)$, $Z = \max\limits_{i=0,\ldots,q-1} |x_i - x(t_0 + i\,h)|$ $\Lambda = \left(\sum_{i=0}^{q} |\alpha_i| \right) \omega(x', qh)h$, $N = \frac{T}{h}$ et $B^* = L \sum_{i=0}^{q} |\beta_i|$. Il s'ensuit que

$$|x_n - x(t_n)| \leq \Gamma^* \left[\left(\sum_{i=0}^{q} i\,|\alpha_i| \right) \max_{i=0,\ldots,q-1} |x_i - x(t_0 + i\,h)| + (t_n - t_0) \left(\sum_{i=0}^{q} i\,|\alpha_i| + |\beta_i| \right) \omega(x', qh) \right] e^{(t_n - t_0) L \Gamma^* \sum_{i=0}^{q} |\beta_i|}$$

... La fonction x' étant uniformément continue sur $[t_0, t_0 + T]$ par le théorème de Heine, $\omega(x', qh)$ tend vers 0 lorsque h tend vers 0, ce qui entraîne, avec les hypothèses sur les valeurs d'initialisation, la convergence de la méthode. □

On peut symboliquement résumer ce résultat dans la devise

consistance + stabilité = convergence

qui s'avère être d'une portée très générale (on ne manquera pas de le comparer avec le théorème 10.27, relatif à la résolution numérique d'équations aux dérivées partielles linéaires, dans le chapitre 10).

8.4.5 Stabilité absolue

La zéro-stabilité introduite dans une précédente sous-section n'est pas la seule notion de stabilité utile à qui s'intéresse à la résolution numérique d'un problème de Cauchy. Celle-ci caractérise en effet le comportement de la méthode considérée lorsque la longueur du pas de discrétisation tend vers zéro, la taille de l'intervalle sur lequel on effectue la résolution étant fixée. En pratique cependant, on ne peut effectuer qu'un nombre fini d'opérations, et c'est par conséquent des pas de longueur strictement non nulle que l'on emploie. Il faut alors s'assurer que l'accumulation des erreurs introduites par la méthode à chaque étape ne croît pas de manière incontrôlée avec le nombre d'étapes effectuées.

On étudie pour cela le comportement *asymptotique* de la solution approchée lorsque le nombre de pas tend vers l'infini, la longueur h définie par (8.25) étant fixée. Puisque ce comportement en temps long dépend à la fois de la méthode numérique et du problème que l'on cherche à résoudre, on convient de se restreindre à un problème de Cauchy *modèle*, basé sur une équation différentielle scalaire linéaire à coefficient constant et homogène, l'*équation test de Dahlquist* (*Dahlquist's test equation* en anglais),

$$x'(t) = \lambda\, x(t),\ t > 0,\ \text{et } x(0) = 1, \tag{8.88}$$

où le scalaire λ est un nombre complexe tel que $\mathrm{Re}(\lambda) < 0$. La solution de ce problème étant $x(t) = e^{\lambda\, t}$, elle satisfait, compte tenu de l'hypothèse sur λ,

$$\lim_{t \to +\infty} |x(t)| = 0.$$

Il semble alors naturel de chercher à ce que la solution obtenue par une méthode numérique appliquée à la résolution du problème (8.88) vérifie une propriété similaire, ce qui conduit à la définition suivante [64].

Définition 8.40 (stabilité absolue) *Une méthode numérique est dite* ***absolument stable*** *si l'approximation* $(x_n)_{n\in\mathbb{N}}$ *construite pour la résolution du problème* (8.88) *est telle que*

$$\lim_{n \to +\infty} |x_n| = 0, \tag{8.89}$$

pour des valeurs données du pas $h > 0$ *et du coefficient complexe* λ*, avec* $\mathrm{Re}(\lambda) < 0$.

La *région de stabilité absolue* d'une méthode numérique est le sous-ensemble du plan complexe

$$\mathscr{S} = \{h\lambda \in \mathbb{C} \,|\, \text{la condition (8.89) est satisfaite}\}.$$

La détermination de cette dernière est une étape indispensable de l'étude d'applicabilité de la méthode à la résolution de systèmes d'équations différentielles raides (voir la section 8.7).

Nous allons à présent examiner les propriétés de stabilité absolue des différentes méthodes introduites dans la section 8.3. Dans tout le reste de cette sous-section, nous supposerons pour simplifier que la grille de discrétisation utilisée est uniforme.

Cas des méthodes à un pas

L'utilisation d'une méthode à un pas pour la résolution approchée du problème (8.88) mène à la relation de récurrence

$$x_{n+1} = R(h\lambda)\, x_n,\ n \geq 0,$$

dans laquelle R est la *fonction de stabilité* de la méthode. On en déduit immédiatement qu'une méthode à un pas est absolument stable si et seulement si

$$|R(h\lambda)| < 1.$$

64. On trouve parfois dans la littérature un énoncé différent de celui de la définition 8.40, qui exige simplement que la suite $(x_n)_{n\in\mathbb{N}}$ soit *bornée*. On dit alors que la méthode est *faiblement stable*.

Détermination de la région de stabilité absolue de quelques méthodes à un pas. Considérons la méthode d'Euler sous sa forme explicite. Nous avons

$$x_{n+1} = (1+h\lambda)\, x_n,\ n \geq 0,$$

soit encore $R(h\lambda) = (1+h\lambda)$. La région de stabilité absolue de cette méthode est donc la boule ouverte $B(-1,1)$. Pour la méthode d'Euler implicite, il vient

$$R(h\lambda) = (1-h\lambda)^{-1}$$

et la région de stabilité absolue correspondante est le complémentaire dans $\mathbb{C}$ de la boule fermée $\overline{B(1,1)}$. Enfin, pour la méthode de la règle du trapèze, on trouve

$$R(h\lambda) = \frac{2+h\lambda}{2-h\lambda}.$$

La région de stabilité absolue correspondante est alors le demi-plan complexe tel que $\mathrm{Re}(z) < 0$. On l'a représentée, ainsi que celles des deux méthodes précédentes, sur la figure 8.12.

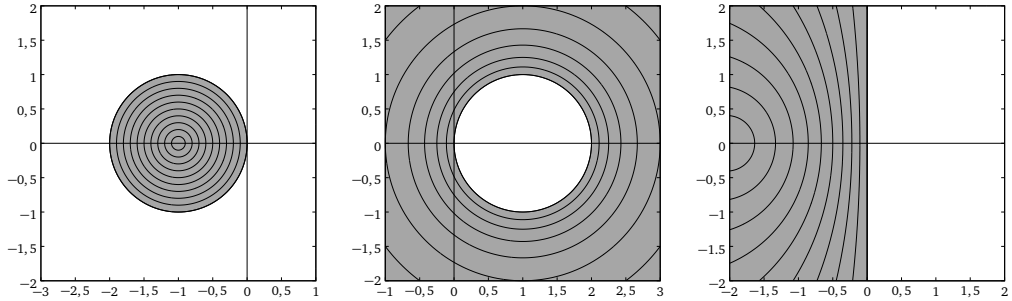


FIGURE 8.12 – Régions de stabilité absolue (en gris) pour quelques méthodes à un pas : la méthode d'Euler explicite, la méthode d'Euler implicite et la méthode de la règle du trapèze (de gauche à droite).

Dans le cas d'une méthode de Runge–Kutta, l'application de la méthode à la résolution du problème (8.88) conduit à la relation de récurrence

$$x_{n+1} = x_n + h \sum_{i=1}^{s} b_i\, k_i,\ k_i = \lambda\,(x_n + h \sum_{j=1}^{s} a_{ij} k_j),\ i = 1,\ldots,s,\ n \geq 0,$$

que l'on peut encore écrire, en introduisant le vecteur $\boldsymbol{k}$ de composantes k_i, $i = 1,\ldots,s$,

$$x_{n+1} = x_n + h\, \boldsymbol{b}^T \boldsymbol{k},\ (I_s - h\lambda A)\, \boldsymbol{k} = \lambda x_n\, \boldsymbol{e}. \tag{8.90}$$

On en déduit alors que

$$x_{n+1} = \left(1 + h\lambda\, \boldsymbol{b}^T\, (I_s - h\lambda A)^{-1}\, \boldsymbol{e}\right) x_n,\ n \geq 0,$$

d'où

$$R(h\lambda) = 1 + h\lambda\, \boldsymbol{b}^T (I_s - h\lambda A)^{-1} \boldsymbol{e}. \tag{8.91}$$

Une autre forme de la fonction de stabilité est obtenue en faisant appel à la règle de Cramer (voir la proposition A.163) pour la résolution du système linéaire de $s+1$ équations (8.90) et s'écrit

$$R(h\lambda) = \frac{\det(I_s - h\lambda\,(A + \boldsymbol{e}\boldsymbol{b}^T))}{\det\,(I_s - h\lambda A)} \tag{8.92}$$

Ces deux expressions sont complémentaires : il parfois plus facile de travailler avec l'une que l'autre et vice versa. Lorsque la méthode de Runge–Kutta est explicite, il est clair que $\det\,(I_s - h\lambda A) = 1$. On déduit alors de (8.92) que la fonction de stabilité est polynomiale et la région de stabilité absolue est nécessairement bornée.

Détermination des régions de stabilité absolue des méthodes de Runge–Kutta explicites. On considère une méthode de Runge–Kutta explicite à s niveaux et d'ordre p, avec $p \leq s$. On a vu plus haut que la fonction de stabilité d'une telle méthode était polynomiale et de degré au plus égal à s. On sait par ailleurs, la méthode étant d'ordre p, que la valeur $\tilde{x}_{n+1}$ donnée par la méthode sous l'hypothèse localisante $x_n = x(t_n)$ coïncide avec la somme des $p+1$ premiers termes du développement de Taylor

$$x(t_{n+1}) = x(t_n) + h\lambda\, x(t_n) + \frac{1}{2}(h\lambda)^2 x(t_n) + \cdots + \frac{1}{p!}(h\lambda)^p x(t_n) + O(h^{p+1})$$

de la solution exacte du problème (8.88). Ceci implique, lorsque [65] $s = p$, que

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \cdots + \frac{1}{p!}(h\lambda)^p. \tag{8.93}$$

Pour $1 \leq s \leq 4$, *toutes* les méthodes de Runge–Kutta explicites à s niveaux d'ordre maximal ont donc la même fonction de stabilité. Notons qu'un raisonnement similaire montre que la fonction de stabilité d'une méthode de Taylor d'ordre p est donnée par (8.93). Des exemples des régions de stabilité absolue correspondantes sont représentées sur la figure 8.13. On observe qu'elles sont bornées (ce qui était prévisible puisque les fonctions de stabilité de ces méthodes sont polynomiales), mais que leur taille augmente avec l'ordre.

Si [66] la méthode est d'ordre $p < s$, la fonction de stabilité est de la forme

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \cdots + \frac{1}{p!}(h\lambda)^p + \sum_{j=p+1}^{s} \gamma_j (h\lambda)^j,$$

où les scalaires γ_j, $j = p+1, \ldots, s$, sont des fonctions des coefficients de la méthode, que l'on peut déterminer en identifiant les termes en puissance de $h\lambda$, correspondants dans (8.91). Par exemple, pour $p = s-1$, on a, en posant $\boldsymbol{d} = (I_s - h\lambda A)^{-1}\boldsymbol{e}$,

$$\begin{array}{lcccccccccc}
d_1 & = & 1 & & & & & & & & \\
d_2 & = & 1 & + & h\lambda a_{21} d_1 & & & & & & \\
d_2 & = & 1 & + & h\lambda a_{31} d_1 & + & h\lambda a_{32} d_2 & & & & \\
\vdots & & \vdots & & \vdots & & \vdots & & \ddots & & \\
d_s & = & 1 & + & h\lambda a_{s1} d_1 & + & h\lambda a_{s2} d_2 & + & \cdots & + & h\lambda a_{ss-1} d_{s-1}
\end{array}$$

et l'on trouve, après substitution et en utilisant les conditions (8.35),

$$\gamma_s = b_s a_{ss-1} a_{s-1s-2} \ldots a_{32} c_2.$$

Cette technique de calcul se généralise à tout ordre et permet en particulier d'obtenir les fonctions de stabilité des *méthodes de Runge–Kutta emboîtées* introduites dans la sous-section 8.6.1.

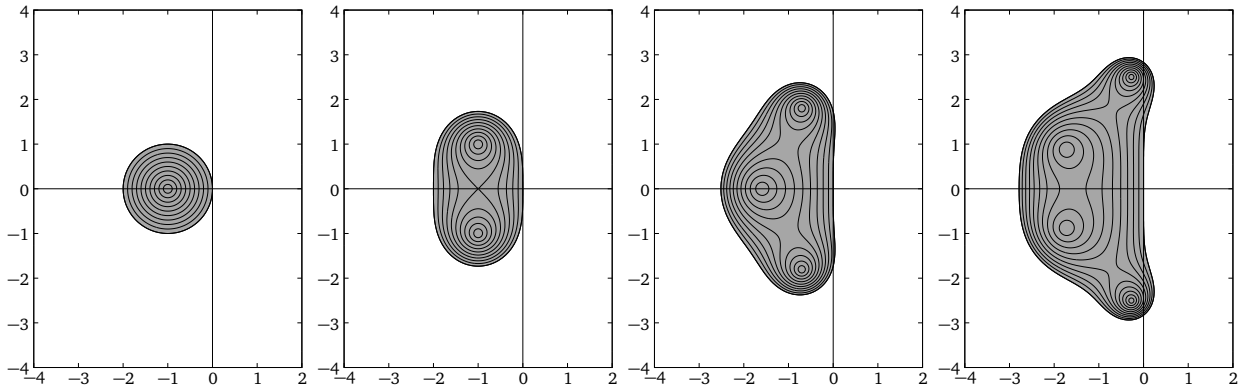


FIGURE 8.13 – Régions de stabilité absolue (en gris) pour les méthodes de Runge–Kutta explicites à s niveaux, $s = 1, \ldots, 4$.

Pour une méthode de Runge–Kutta implicite ou semi-implicite, la quantité $\det(I_s - h\lambda A)$ est elle-même une fonction polynomiale de $h\lambda$ et la fonction de stabilité est donc une fonction rationnelle. Il est dans ce cas tout à fait possible que la condition de stabilité absolue soit encore satisfaite lorsque $|h\lambda|$ tend vers l'infini, conduisant alors à une région de stabilité absolue non bornée.

65. On rappelle que ceci n'arrive que si $1 \leq s \leq 4$ (voir la sous-section 8.4.3).
66. C'est toujours le cas dès que $s > 4$ (voir une nouvelle fois la sous-section 8.4.3).

Détermination de la région de stabilité absolue d'une méthode de Runge–Kutta implicite. fonction pour la méthode implicite de Gauss-Legendre pour $s = 2$

$$R(h\lambda) = \frac{1 + \frac{h\lambda}{2} + \frac{(h\lambda)^2}{12}}{1 - \frac{h\lambda}{2} + \frac{(h\lambda)^2}{12}}.$$

Cette fraction rationnelle est l'*approximant de Padé*[67] *de type* $(2,2)$ de la fonction exponentielle[68] COMPLETER (voir la sous-section 8.7.2 pour plus de détails)

Cas des méthodes à pas multiples linéaires *

Si l'on utilise une méthode à pas multiple linéaire pour la résolution du problème (8.88), on aboutit à la relation de récurrence

$$\sum_{i=0}^{q} \alpha_i \, x_{n+i} - h\lambda \sum_{i=0}^{q} \beta_i \, x_{n+i} = 0, \; n \geq 0,$$

qui est une équation aux différences linéaires ayant pour polynôme caractéristique associé $\Pi_{h\lambda}(z) = \rho(z) - h\lambda\,\sigma(z)$. Grâce à ce dernier, on peut, comme cela est le cas pour la zéro-stabilité, caractériser la stabilité absolue d'une méthode à pas multiples linéaire en termes d'une condition de racine.

Théorème 8.41 (condition nécessaire et suffisante de stabilité absolue d'une méthode à pas multiples linéaire) *Une méthode à pas multiples linéaire est absolument stable pour une valeur donnée de λ si et seulement si toutes les racines de $\Pi_{h\lambda}$ sont de module inférieur ou égal à l'unité et que celles de module égal à un sont simples.*

DÉMONSTRATION. A ECRIRE (ou admise ?) □

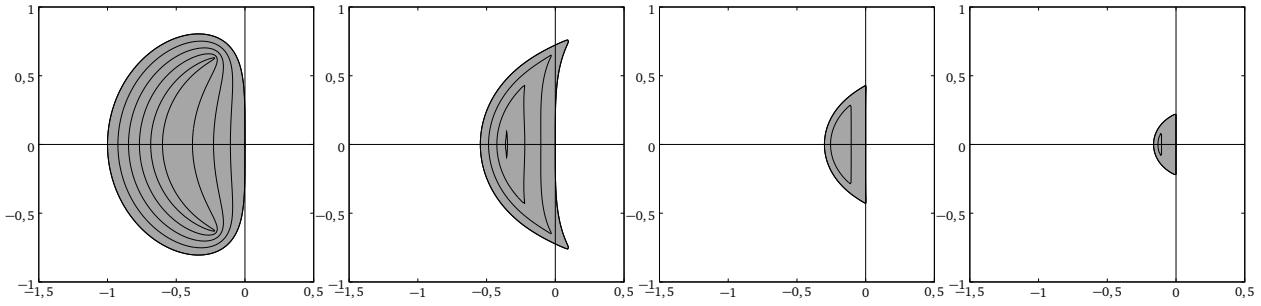


FIGURE 8.14 – Régions de stabilité absolue (en gris) pour les méthodes d'Adams–Bashforth d'ordre deux à cinq (de gauche à droite).

On a respectivement représenté sur les figures 8.14, 8.15 et 8.16 les régions de stabilité absolue des méthodes d'Adams–Bashforth de deux à cinq pas (la méthode à un pas correspondante étant la méthode d'Euler explicite (voir la figure 8.12)), d'Adams–Moulton de deux à quatre pas (la méthode à un pas correspondante étant la méthode de la règle du trapèze (voir la figure 8.12)) et BDF de deux à six pas (la méthode à un pas correspondante étant la méthode d'Euler implicite (voir la figure 8.12)).

Au vu de ces figures, il peut sembler que les régions de stabilité des méthodes implicites sont plus grandes que celles des méthodes explicites correspondantes, et que la taille de la région a tendance à diminuer lorsque l'ordre de la méthode augmente. Si la première conjecture est vraie, la seconde est en général fausse, un contre-exemple étant celui des régions de stabilité absolue des méthodes de Runge–Kutta explicites représentées sur la figure 8.13. On observera que les régions de stabilité absolue des méthodes BDF sont non bornées.

Pour déterminer graphiquement les régions de stabilité absolue d'une méthode à pas multiples linéaires, on peut exploiter la condition de racine du théorème 8.41 en cherchant les racines du polynôme $\Pi_{h\lambda}$ pour $h\lambda$ prenant

67. Henri Eugène Padé (17 décembre 1863 - 9 juillet 1953) était un mathématicien français. Il est surtout connu pour son développement d'une méthode d'approximation des fonctions analytiques par des fonctions rationnelles.

68. REPRENDRE On dit qu'une fonction rationnelle est l'approximant de Padé de type (m,n) de la fonction exponentielle si son numérateur a un degré borné m, son dénominateur a un degré borné n et que $e^z - r(z) = O(z^{m+n+1})$, $z \to 0$.

ses valeurs en des points d'une grille dans le plan complexe et en traçant les lignes de niveaux du module de la racine de plus grand module. Cette manière de faire est cependant coûteuse en calculs. Une autre technique, très simple à mettre en œuvre, se fonde sur le fait que les racines d'un polynôme sont des fonctions continues de ses coefficients. Elle consiste à déterminer la frontière de la région de stabilité par le tracé de la *courbe du lieu des racines* (*root locus curve* en anglais), d'équation

$$h\lambda = \frac{\rho(e^{\mathrm{i}\theta})}{\sigma(e^{\mathrm{i}\theta})},\ 0 \leq \theta \leq 2\pi.$$

Cette courbe possède la propriété qu'exactement une racine du polynôme $\Pi_{h\lambda}$ touche le cercle unité en chacun de ses points. Il s'ensuit que la frontière de la région de stabilité est un sous-ensemble de cette courbe (car les autres zéros du polynôme peuvent se trouver soit à l'intérieur du disque unité, soit en dehors). On décide alors si chacune des composantes connexes du plan complexe ainsi obtenues appartient ou pas à la région de stabilité absolue de la méthode en calculant les racines de $\Pi_{h\lambda}$ pour une (et une seule) valeur de $h\lambda$ qu'elle contient.

Exemples de tracé de courbe du lieu des racines. La figure 8.17 présente les courbes du lieu des racines de la méthode d'Adams–Bashforth à quatre pas, de la méthode de Nÿstrom à trois pas et la méthode de Milne–Simpson généralisée à quatre pas. Ce sont trois exemples de méthodes pour lesquelles la frontière de la région de stabilité absolue est strictement incluse dans la courbe du lieu des racines associée. Pour la méthode d'Adams–Bashforth, la comparaison avec la région de stabilité trouvée sur la figure 8.14 indique en effet que les deux lobes appartenant au demi-plan contenant les nombres complexes à partie réelle positive ne font pas partie du domaine de stabilité absolue de la méthode. Ceci est confirmé par le fait que, pour une valeur de $h\lambda$ choisie arbitrairement dans l'un ou l'autre de ces ensembles, le polynôme $\Pi_{h\lambda}$ possède deux racines de module plus grand que l'unité. On montre de cette manière que la région de stabilité absolue d'un méthode de Nÿstrom est $]-\mathrm{i},\mathrm{i}[$ pour $q = 1$ ou 2, $\{0\}$ pour $q \geq 3$ (ce qui correspond au minimum possible pour une méthode zéro-stable en vertu du théorème 8.31), et que celle d'une méthode de Milne–Simpson généralisée est $]-\mathrm{i}\sqrt{3},\mathrm{i}\sqrt{3}[$ pour $q = 2$ ou 3 et $\{0\}$ pour $q \geq 4$.

Notons que l'on n'est parfois intéressé que par la détermination de l'*intervalle de stabilité absolue* $\mathscr{S} \cap \mathbb{R}$ de la méthode (voir la sous-section 8.7.2). Dans ce cas, le paramètre $h\lambda$ étant réel, le polynôme $\Pi_{h\lambda}$ est à coefficients réels et dire qu'il satisfait la condition de racine du théorème 8.41 signifie encore que c'est un *polynôme de Schur*, c'est-à-dire un polynôme dont toutes les racines sont contenues à l'intérieur du disque unité, ce que l'on peut vérifier grâce au *critère de Routh*[69]*–Hurwitz*[70] en faisant appel à une transformation conforme. Ce critère algébrique affirmant en effet qu'un polynôme de degré q à coefficients réels $a_q\, z^q + a_{q-1}\, z^{q-1} + \cdots + a_1\, z + a_0$ a toutes ses

69. Edward John Routh (20 janvier 1831 - 7 juin 1907) était un mathématicien anglais. Il fit beaucoup pour systématiser la théorie mathématique de la mécanique et introduisit plusieurs idées essentielles au développement de la théorie moderne du contrôle des systèmes.

70. Adolf Hurwitz (26 mars 1859 - 18 novembre 1919) était un mathématicien allemand. Il démontra plusieurs résultats fondamentaux sur les courbes algébriques, dont son théorème sur les automorphismes, et s'intéressa à la théorie des nombres.

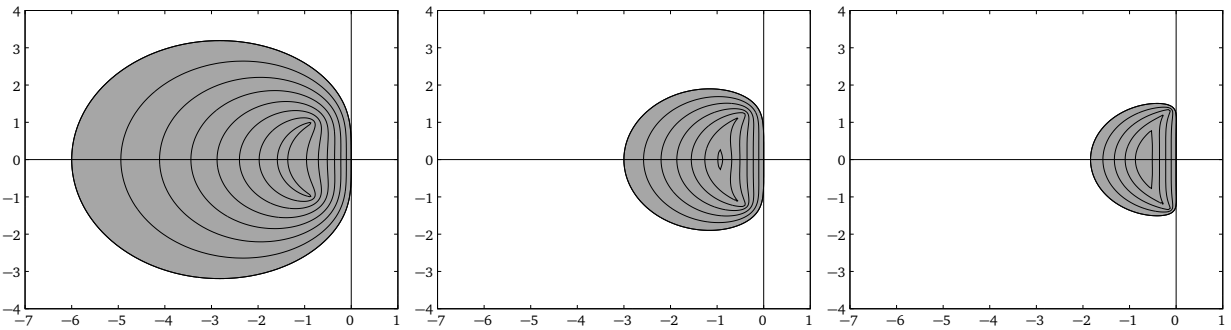


FIGURE 8.15 – Régions de stabilité absolue (en gris) pour les méthodes d'Adams–Moulton d'ordre trois à cinq (de gauche à droite).

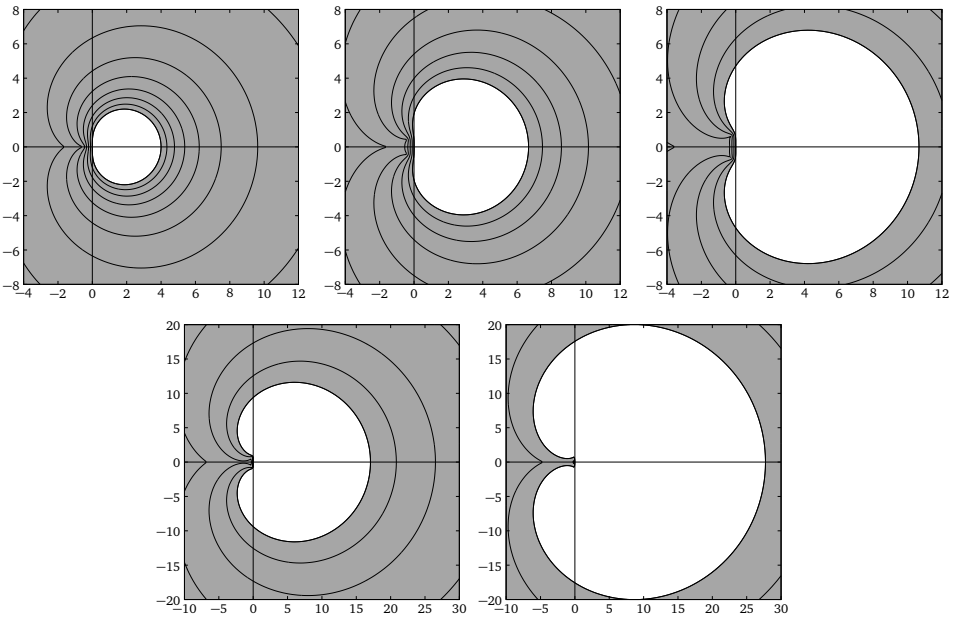


FIGURE 8.16 – Régions de stabilité absolue (en gris) pour les méthodes BDF d'ordre deux à six (de gauche à droite et de haut en bas).

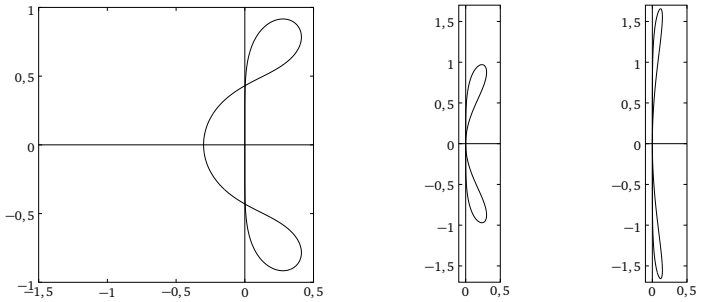


FIGURE 8.17 – Courbes du lieu des racines des méthodes d'Adams–Bashforth à quatre pas (à gauche), de Nÿstrom à trois pas (au milieu) et de Milne–Simpson généralisée à quatre pas (à droite).

racines de partie réelle strictement négative si et seulement si tous les mineurs principaux de la matrice d'ordre q

$$\begin{pmatrix} a_{q-1} & a_{q-3} & a_{q-5} & a_{q-7} & \dots & 0 \\ a_q & a_{q-2} & a_{q-4} & a_{q-6} & \dots & 0 \\ 0 & a_{q-1} & a_{q-3} & a_{q-5} & \dots & 0 \\ 0 & a_q & a_{q-2} & a_{q-4} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & a_0 \end{pmatrix}$$

sont strictement positifs, on doit faire en sorte de ramener le disque unité au demi-plan complexe de partie réelle négative pour être en mesure de l'utiliser, ce que l'on fait en condidérant le polynôme $(1-z)^q \, \Pi_{h\lambda}\left(\frac{1+z}{1-z}\right)$.

EXEMPLE ?

Nous reviendrons sur l'importance de la propriété de stabilité absolue des méthodes dans la section 8.7 consacrée à la résolution des systèmes d'équations différentielles raides.

8.4.6 Cas des systèmes d'équations différentielles ordinaires

Si l'extension au cas des systèmes d'équations différentielles ordinaires du premier ordre des différentes méthodes présentées et de leur analyse est relativement directe, il faut mentionner que certains des résultats établis en considérant une unique équation scalaire peuvent néanmoins différer.

C'est en particulier le cas pour l'ordre maximal atteint par les méthodes de Runge–Kutta explicites. La théorie de Butcher montre que les conditions d'ordre satisfaites par les coefficients de ces méthodes sont, en général, moins restrictives dans le cas scalaire que dans le cas vectoriel pour les méthodes d'ordre supérieur ou égal à cinq. De fait, il existe des méthodes dont l'ordre est cinq lorsqu'elles sont appliquées à la résolution d'une équation différentielle scalaire mais seulement quatre pour un système de plusieurs équations. Un exemple de méthode à six niveaux présentant cette particularité est donné dans [But95].

La transposition de la définition de stabilité absolue au cas de la résolution d'un système d'équations différentielles se fait en considérant tout simplement un système linéaire homogène à coefficients constants

$$\boldsymbol{x}'(t) = A\boldsymbol{x}(t), \tag{8.94}$$

avec A une matrice d'ordre d, $d \geq 2$, dont les valeurs propres λ_i, $i = 1, \ldots, d$, sont distinctes et de partie réelle négative. La solution générale d'un tel système s'écrit

$$\boldsymbol{x}(t) = \sum_{i=1}^{d} c_i \, e^{\lambda_i t} \, \boldsymbol{v_i},$$

où les c_i, $i = 1, \ldots, d$, sont des constantes arbitraires et les vecteurs $\boldsymbol{v_i}$, $i = 1, \ldots, d$, sont des vecteurs propres respectivement associés aux valeurs propres λ_i, $i = 1, \ldots, d$, de la matrice A, et satisfait

$$\lim_{t \to +\infty} \|\boldsymbol{x}(t)\| = 0$$

pour tout choix d'une norme $\|\cdot\|$ sur $\mathbb{C}^d$. Exiger qu'une méthode soit absolument stable dans ce contexte revient alors à demander à ce que l'approximation de toute solution du système (8.94) qu'elle génère pour une valeur de h donnée soit telle que

$$\lim_{n \to +\infty} \|\boldsymbol{x}_n\| = 0.$$

Les valeurs propres de la matrice A étant distinctes, celle-ci est diagonalisable et il existe une matrice inversible P telle que $\Lambda = P^{-1}AP$, avec $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. En réécrivant de manière équivalente le système (8.94) sous la forme d'un système d'équations différentielles *découplées*,

$$\boldsymbol{y}'(t) = \Lambda \boldsymbol{y}(t),$$

dans lequel on a posé $\boldsymbol{y}(t) = P^{-1}\boldsymbol{x}(t)$, on voit que la théorie développée dans le cas scalaire suffit effectivement pour traiter le cas des systèmes d'équations différentielles ordinaires.

8.5 Méthodes de prédiction-correction

Nous avons déjà évoqué les difficultés pratiques rencontrées lors de l'utilisation d'une méthode à pas multiples linéaire implicite, liées à la résolution numérique à chaque étape de l'équation (8.58), généralement non linéaire, par une méthode des approximations successives. Bien que l'on puisse garantir, en prenant une grille de discrétisation suffisamment fine, que la suite définie par la relation de récurrence (8.60) sera convergente pour toute initialisation arbitraire, on ne sait cependant pas prédire combien d'itérations – et donc combien d'évaluations de la fonction f – seront nécessaires pour atteindre une précision voulue. Cette incertitude sur le coût de calcul *a priori* des méthodes implicites rend l'emploi de ces dernières délicat dans certains domaines d'applications, comme à l'intérieur d'un *système temps réel embarqué* [71]. On peut évidemment chercher à rendre cette étape moins coûteuse en fournissant une initialisation raisonnable, mais cela ne permettra pas de « contrôler » le nombre d'itérations de

71. Il s'agit d'un système électronique et informatique chargé de contrôler un procédé en opérant avec un temps de réponse adapté à l'évolution de ce procédé. L'antiblocage de sécurité des freins (*Antiblockiersystem* en allemand), qui équipe de nombreux véhicules, est un exemple d'un tel système.

point fixe réellement effectuées. L'idée des méthodes de prédiction-correction repose sur la prise en compte de ces deux dernières remarques, en tirant parti d'une méthode explicite, qualifiée de *prédicteur*, pour obtenir une approximation $x_{n+q}^{(0)}$ de la valeur x_{n+q}, solution de l'équation (8.58), recherchée, dont on se sert pour effectuer un nombre fixé à l'avance d'itérations de point fixe associées à une méthode implicite, alors appelée le *correcteur*.

Dans toute la suite, nous distinguerons le correcteur du prédicteur en attachant des astérisques à tout paramètre s'y rapportant, comme son nombre de pas q^*, son ordre p^* ou encore ses coefficients α_i^*, $i = 0, \ldots, q^*$, et β_i^*, $i = 0, \ldots, q^*$. L'implémentation d'une méthode de prédiction-correction comporte plusieurs phases, que nous allons maintenant décrire. En supposant que les approximations x_{n+i}, $i = \min(0, q-q^*), \ldots, q-1$, ont été calculées aux étapes précédentes (ou font partie des valeurs de démarrage de la méthode) et que les quantités $f_{n+i} = f(t_{n+i}, x_{n+i})$, $i = \min(0, q-q^*), \ldots, q-1$, sont également connues, la *prédiction* consiste en l'obtention de la valeur $x_{n+q}^{(0)}$, donnée par

$$x_{n+q}^{(0)} = \sum_{i=0}^{q-1} (h\beta_i f_{n+i} - \alpha_i x_{n+i}). \tag{8.95}$$

Suit une *évaluation* de la fonction f utilisant cette approximation,

$$f(t_{n+q}, x_{n+q}^{(0)}), \tag{8.96}$$

qui permet alors une *correction*

$$x_{n+q}^{(1)} = h\beta_{q^*}^* f(t_{n+q}, x_{n+q}^{(0)}) + \sum_{i=0}^{q^*-1} \left(h\beta_i^* f_{n+q-q^*+i} - \alpha_i^* x_{n+q-q^*+i}\right). \tag{8.97}$$

Un moyen mnémotechnique pour décrire les diverses mises en œuvre possibles à partir de ces trois phases est de désigner ces dernières respectivement par les lettres P, E et C, l'ordre des lettres indiquant leur enchaînement dans la méthode. Par exemple, à chaque étape de la résolution, une méthode de type PEC effectue, dans cet ordre, les calculs (8.95), (8.96) et (8.97), suivis des affectations $x_{n+q} = x_{n+q}^{(1)}$ et $f_{n+q} = f(t_{n+q}, x_{n+q}^{(0)})$. Remarquons qu'on aurait pu choisir d'utiliser la valeur $x_{n+q}^{(1)}$ pour mettre à jour f_{n+q} en effectuant une nouvelle évaluation de la fonction f,

$$f_{n+q} = f(t_{n+q}, x_{n+q}^{(1)}),$$

ou même d'utiliser cette évaluation pour faire une seconde itération de correction,

$$x_{n+q}^{(2)} = h\beta_{q^*}^* f(t_{n+q}, x_{n+q}^{(1)}) + \sum_{i=0}^{q^*-1} \left(h\beta_i^* f_{n+q-q^*+i} - \alpha_i^* x_{n+q-q^*+i}\right),$$

et alors poser $x_{n+q} = x_{n+q}^{(2)}$, donnant ainsi respectivement lieu aux modes PECE et PECEC de la méthode de prédiction-correction. On regroupe l'ensemble des modes ainsi formés par des combinaisons de ces deux procédés sous la notation condensée $\mathrm{P(EC)}^{\mu}\mathrm{E}^{1-\tau}$, avec μ un entier naturel [72] et $\tau = 0$ ou 1.

Exemple de méthode de prédiction-correction. On peut voir la méthode de Heun, définie par la relation (8.41), comme une méthode de prédiction-correction en mode PECE, dans laquelle le prédicteur est la méthode d'Euler explicite,

$$x_{n+1}^{(0)} = x_n + h f(t_n, x_n),$$

et le correcteur est la méthode de la règle du trapèze,

$$x_{n+1} = x_n + \frac{h}{2}\left(f(t_{n+1}, x_{n+1}^{(0)}) + f(t_n, x_n)\right) = x_n + \frac{h}{2}\left(f(t_n + h, x_n + h f(t_n, x_n)) + f(t_n, x_n)\right).$$

Passons à présent à l'étude des méthodes de prédiction-correction. Compte tenu de leur fonctionnement, on conçoit facilement que l'erreur de troncature locale des schémas obtenus combine l'erreur de troncature locale du prédicteur avec celle du correcteur de manière plus ou moins évidente selon le mode considéré. Nous considérerons ici que le mode en question est $\mathrm{P(EC)}^{\mu}\mathrm{E}^{1-\tau}$, avec $\mu \geq 1$ et $\tau = 0$ ou 1.

72. On observera que le cas $\mu = 0$ (mode PE) correspond à l'emploi de la méthode explicite, alors le cas limite $\mu = +\infty$ (mode $\mathrm{P(EC)}^{+\infty}$) correspond à celui de la méthode implicite.

Supposons que la solution x du problème de Cauchy est de classe $\mathscr{C}^{\max(p,p^*)+1}$. Pour le prédicteur, nous avons, en reprenant les notations utilisées dans la sous-section 8.4.2,

$$\mathscr{L}(x(t_n),h)=\sum_{i=0}^{q}\alpha_i\,x(t_{n+i})-h\sum_{i=0}^{q-1}\beta_i\,f(t_{n+i},x(t_{n+i}))=C_{p+1}\,h^{p+1}x^{(p+1)}(t_n)+O(h^{p+2}).$$

En additionnant la seconde égalité à (8.95) sous l'hypothèse localisante

$$x_{n+i}=x(t_{n+i}),\ i=\min(0,q-q^*),\dots,q-1, \tag{8.98}$$

il vient

$$x(t_{n+q})-\tilde{x}_{n+q}^{(0)}=C_{p+1}\,h^{p+1}x^{(p+1)}(t_n)+O(h^{p+2}). \tag{8.99}$$

Pour le correcteur, nous avons

$$\mathscr{L}^*(x(t_{n+q-q^*}),h)=\sum_{i=0}^{q^*}\alpha_i^*\,x(t_{n+q-q^*+i})-h\sum_{i=0}^{q^*}\beta_i^*\,f(t_{n+q-q^*+i},x(t_{n+q-q^*+i}))$$
$$=C_{p^*+1}^*\,h^{p^*+1}x^{(p^*+1)}(t_n)+O(h^{p^*+2}),$$

et, en additionnant la seconde égalité à (8.60) pour $k\le\mu-1$ sous l'hypothèse localisante (8.98) et en utilisant le théorème des accroissements finis (voir le théorème B.113), on trouve

$$x(t_{n+q})-\tilde{x}_{n+q}^{(k+1)}=h\beta_{q^*}^*\frac{\partial f}{\partial x}(t_{n+q},\eta_k)(x(t_{n+q})-\tilde{x}_{n+q}^{(k)})+C_{p^*+1}^*\,h^{p^*+1}x^{(p^*+1)}(t_n)+O(h^{p^*+2}),$$
$$k=0,\dots,\mu-1, \tag{8.100}$$

où η_k est un point intérieur du segment joignant $x(t_{n+q})$ à $\tilde{x}_{n+q}^{(k)}$. Pour pouvoir poursuivre, il nous faut discuter en fonction des valeurs relatives des entiers p et p^*.

Si $p\ge p^*$, on obtient en reportant (8.99) dans (8.100) pour $k=0$

$$x(t_{n+q})-\tilde{x}_{n+q}^{(1)}=C_{p^*+1}^*\,h^{p^*+1}x^{(p^*+1)}(t_n)+O(h^{p^*+2}).$$

En reportant cette nouvelle égalité dans (8.100) pour $k=1$, il vient

$$x(t_{n+q})-\tilde{x}_{n+q}^{(2)}=C_{p^*+1}^*\,h^{p^*+1}x^{(p^*+1)}(t_n)+O(h^{p^*+2}).$$

En réitérant ce procédé, on trouve finalement que

$$x(t_{n+q})-\tilde{x}_{n+q}^{(\mu)}=C_{p^*+1}^*\,h^{p^*+1}x^{(p^*+1)}(t_n)+O(h^{p^*+2}). \tag{8.101}$$

Par conséquent, l'ordre et l'erreur locale de troncature principale de la méthode de prédiction-correction sont ceux du correcteur pour tout valeur de l'entier μ.

Si $p=p^*-1$, on obtient cette fois pour $k=0$

$$x(t_{n+q})-\tilde{x}_{n+q}^{(1)}=h^{p^*+1}\left(\beta_{q^*}^*\frac{\partial f}{\partial x}(t_{n+q},\eta_k)\,C_{p^*}\,x^{(p^*)}(t_n)+C_{p^*+1}\,x^{(p^*+1)}(t_n)\right)+O(h^{p^*+2}).$$

On voit que, dans le cas $\mu=1$, l'ordre de la méthode est bien celui du correcteur, mais l'erreur locale de troncature principale diffère. En revanche, si $\mu\ge 2$, on retrouve, en effectuant des substitutions successives, une erreur locale de troncature principale identique à celle du correcteur.

Si $p=p^*-2$, on a

$$x(t_{n+q})-\tilde{x}_{n+q}^{(1)}=\beta_{q^*}^*\frac{\partial f}{\partial x}(t_{n+q},\eta_k)\,C_{p^*-1}\,h^{p^*}x^{(p^*-1)}(t_n)+O(h^{p^*+1}),$$

et l'ordre de la méthode est inférieur à celui du correcteur si $\mu=1$. Pour $\mu=2$, on retrouve l'ordre du correcteur mais une erreur locale de troncature principale différente,

$$x(t_{n+q})-\tilde{x}_{n+q}^{(2)}=h^{p^*+1}\left(\left(\beta_{q^*}^*\frac{\partial f}{\partial x}(t_{n+q},\eta_k)\right)^2C_{p^*-1}^*\,x^{(p^*-1)}(t_n)+C_{p^*+1}\,x^{(p^*+1)}(t_n)\right)+O(h^{p^*+2}),$$

alors que l'ordre et l'erreur sont ceux du schéma correcteur dès que $\mu \geq 3$.

La tendance est donc claire : l'ordre d'une méthode de prédiction-correction dépend à la fois de l'écart entre les ordres du prédicteur et du correcteur et du nombre d'étapes de correction (nous laissons le soin au lecteur de vérifier que les modes P(EC)$^\mu$ et P(EC)$^\mu$E ont toujours un ordre et une erreur de troncature locale principale identiques). On peut résumer ces constatations en énonçant le résultat suivant.

Proposition 8.42 *Soit une méthode de prédiction-correction en mode P(EC)$^\mu$ ou P(EC)$^\mu$E, avec $\mu \geq 1$, basée sur une paire de méthodes à pas multiples linéaires d'ordre p pour le prédicteur et p^* pour le correcteur.*
Si $p \geq p^*$ (ou si $p < p^*$ et $\mu > p^* - p$), la méthode a le même ordre et la même erreur de troncature locale principale que le correcteur.
Si $p < p^*$ et $\mu = p^* - p$, la méthode et le correcteur ont le même ordre mais des erreurs de troncature locales principales différentes.
Enfin, si $p < p^*$ et $\mu < p^* - p$, la méthode est d'ordre $p + \mu < p^*$.

Lorsque $p = p^*$, il est particulièrement intéressant de remarquer qu'il vient, en soustrayant (8.101) à (8.99),

$$\tilde{x}_{n+q}^{(\mu)} - \tilde{x}_{n+q}^{(0)} = \left(C_{p+1} - C_{p+1}^*\right) h^{p+1} x^{(p+1)}(t_n) + O(h^{p+2}),$$

ce qui fournit, en utilisant de nouveau (8.101), l'estimation

$$\frac{C_{p+1}^*}{C_{p+1} - C_{p+1}^*} \left(x_{n+q}^{(\mu)} - x_{n+q}^{(0)}\right) \tag{8.102}$$

pour l'erreur de troncature locale principale de la méthode de prédiction-correction. Cet estimateur d'erreur, dû à Milne [Mil26], est aisément calculable et d'un coût pratiquement nul. Il est un outil essentiel pour l'adaptation du pas de discrétisation des méthodes à pas multiples linéaires utilisées comme prédicteur-correcteur (voir la sous-section 8.6.2). On remarquera que l'hypothèse localisante a été abandonnée dans (8.102), l'erreur de troncature principale constituant une mesure acceptable de la précision de la méthode.

On voit donc qu'il y a un avantage à ce que les méthodes d'une paire de prédicteur-correcteur soient du même ordre, ce qui signifie que le prédicteur aura, en général, plus de pas que pour le correcteur. Dans ce cas, on a coutume de poser que le nombre de pas de la méthode de prédiction-correction est égal à celui du prédicteur et de lever la condition $|\alpha_0| + |\beta_0| \neq 0$ sur les coefficients du correcteur [73]. Les *méthodes d'Adams–Bashforth–Moulton* (ABM en abrégé), utilisées dans de nombreux codes de résolution numérique de systèmes d'équations différentielles ordianires non raides, sont basées sur ce principe, la paire d'une méthode d'Adams–Bashforth–Moulton à q pas (et d'ordre q) étant composée d'une méthode d'Adams–Bashforth à q pas comme prédicteur et d'une méthode d'Adams–Moulton à $q-1$ pas comme correcteur.

Notons que, au vu des estimations d'erreur obtenues plus haut, une méthode de prédiction-correction n'a d'intérêt que si le correcteur est plus précis que le prédicteur. Dans le cas d'une paire de méthodes de même ordre p, ceci se traduit par le fait qu'il faut que

$$\left|C_{p+1}^*\right| < \left|C_{p+1}\right|.$$

Ceci est effectivement vrai pour les méthodes d'Adams–Bashforth–Moulton, pour lesquelles on a $C_{p+1} = \gamma_p$ et $C_{p+1}^* = \gamma_p^*$ (voir le tableau 8.3), puisque l'on peut montrer, en utilisant les définitions données dans la sous-section 8.3.3, que

$$\left|\gamma_p^*\right| < \frac{\gamma_p}{p-1},\ p \geq 2.$$

Terminons l'analyse des méthodes de prédiction-correction en examinant leurs propriétés de stabilité absolue. Pour cela, déterminons le polynôme de stabilité de la méthode en mode P(EC)$^\mu$E$^{1-\tau}$, pour $\mu \geq 1$ et $\tau = 0$ ou 1, en appliquant celle-ci à la résolution du problème de Cauchy linéaire (8.88).

La phase de prédiction s'écrit dans ce cas

$$x_{n+q}^{(0)} = \sum_{i=0}^{q-1} \left(h\lambda\beta_i\, x_{n+i}^{(\mu-\tau)} - \alpha_i\, x_{n+i}^{(\mu)}\right), \tag{8.103}$$

73. On relira à ce titre les remarques faites sur les hypothèses (8.57).

alors que celle de correction devient

$$x_{n+q}^{(k+1)} = h\lambda\beta_{q^*}^* \, x_{n+q}^{(k)} + \sum_{i=0}^{q^*-1} \left(h\lambda\beta_i^* \, x_{n+q-q^*+i}^{(\mu-\tau)} - \alpha_i^* \, x_{n+q-q^*+i}^{(\mu)} \right), \; k = 0, \ldots, \mu-1. \tag{8.104}$$

En soustrayant deux à deux les relations successives de (8.104), on obtient

$$x_{n+q}^{(k+1)} - (1 + h\lambda\beta_{q^*}^*) \, x_{n+q}^{(k)} + h\lambda\beta_{q^*}^* \, x_{n+q}^{(k-1)} = 0, \; k = 1, \ldots, \mu-1.$$

En considérant cette dernière relation comme une équation aux différences à coefficients constants pour les quantités $\{x_{n+q}^{(k)}\}_{0\le k\le\mu}$, on peut exprimer tout $x_{n+q}^{(k)}$, $1 \le k \le \mu-1$ en fonction de $x_{n+q}^{(0)}$ et $x_{n+q}^{(\mu)}$. On trouve

$$x_n^{(k)} = \frac{(h\lambda\beta_{q^*}^*)^k (1 - (h\lambda\beta_{q^*}^*)^{\mu-k})}{1 - (h\lambda\beta_{q^*}^*)^\mu} \, x_{n+q}^{(0)} + \frac{1 - (h\lambda\beta_{q^*}^*)^k}{1 - (h\lambda\beta_{q^*}^*)^\mu} \, x_{n+q}^{(\mu)}, \; k = 0, \ldots, \mu.$$

Pour $k = \mu-1$, il vient

$$x_n^{(\mu-1)} = \frac{(h\lambda\beta_{q^*}^*)^{\mu-1} (1 - h\lambda\beta_{q^*}^*)}{1 - (h\lambda\beta_{q^*}^*)^\mu} \, x_{n+q}^{(0)} + \frac{1 - (h\lambda\beta_{q^*}^*)^{\mu-1}}{1 - (h\lambda\beta_{q^*}^*)^\mu} \, x_{n+q}^{(\mu)}$$

Cette dernière expression permet d'éliminer $x_{n+q}^{(0)}$ de (8.103) et l'on obtient, en utilisant que $\alpha_q = 1$,

$$\begin{aligned}(1 - (h\lambda\beta_{q^*}^*)^\mu) \, x_{n+q}^{(\mu-1)} = (h\lambda\beta_{q^*}^*)^{\mu-1} (1 - h\lambda\beta_{q^*}^*)(h\lambda) \sum_{i=0}^{q-1} \beta_i \, x_{n+i}^{(\mu-\tau)} + (1 - (h\lambda\beta_{q^*}^*)^\mu) \, x_{n+q}^{(\mu)} \\ - (h\lambda\beta_{q^*}^*)^{\mu-1} (1 - h\lambda\beta_{q^*}^*) \sum_{i=0}^{q} \alpha_i \, x_{n+i}^{(\mu)}\end{aligned} \tag{8.105}$$

L'entier τ ne prenant que les valeurs 0 ou 1, la relation ci-dessus ne fait intervenir que les quantités $x_i^{(\mu-1)}$ et $x_i^{(\mu)}$. Une seconde relation portant sur ces quantités est obtenue en prenant $k = \mu-1$ dans (8.104) et en utilisant que $\alpha_q = 1$

$$h\lambda\beta_{q^*}^* \, x_{n+q}^{(\mu-1)} = \sum_{i=0}^{q^*} \alpha_i^* \, x_{n+q-q^*+i}^{(\mu)} - (h\lambda) \sum_{i=0}^{q^*-1} \beta_i^* \, x_{n+q-q^*+i}^{(\mu-\tau)}. \tag{8.106}$$

Pour $\tau = 0$, il suffit de multiplier (8.105) par $\frac{h\lambda\beta_{q^*}^*}{1-(h\lambda\beta_{q^*}^*)^\mu}$ et d'identifier avec (8.106) pour trouver

$$\sum_{i=0}^{q^*} \alpha_i^* \, x_{n+q-q^*+i}^{(\mu)} - (h\lambda) \sum_{i=0}^{q^*} \beta_i^* \, x_{n+q-q^*+i}^{(\mu)} + \frac{(h\lambda\beta_{q^*}^*)^\mu (1 - h\lambda\beta_{q^*}^*)}{1 - (h\lambda\beta_{q^*}^*)^\mu} \left(\sum_{i=0}^{q} \alpha_i \, x_{n+i}^{(\mu)} - (h\lambda) \sum_{i=0}^{q-1} \beta_i \, x_{n+i}^{(\mu)} \right) = 0.$$

En désignant par ρ et σ les polynômes caractéristiques du prédicteur, par ρ^* et σ^* ceux du correcteur, on a obtenu le polynôme de stabilité absolue

$$\Pi_\lambda(z) = \rho^*(z) - (h\lambda)\,\sigma^*(z) + \frac{(h\lambda\beta_{q^*}^*)^\mu (1 - h\lambda\beta_{q^*}^*)}{1 - (h\lambda\beta_{q^*}^*)^\mu} \, (\rho(z) - (h\lambda)\,\sigma(z)), \tag{8.107}$$

pour une méthode de prédiction-correction en mode P(EC)$^\mu$E.

Pour $\tau = 1$, la dérivation est beaucoup moins aisée. Cependant, en étendant naturellement l'application de l'opérateur de décalage à gauche T_h, introduit dans la sous-section 8.4.2, aux suites de valeurs aux points de la grille, les relations (8.105) et (8.106) se réécrivent respectivement dans ce cas

$$\begin{aligned}\left((1 - (h\lambda\beta_{q^*}^*)^\mu) \, {T_h}^q - (h\lambda\beta_{q^*}^*)^{\mu-1} (1 - (h\lambda\beta_{q^*}^*)^\mu)(h\lambda)\,\sigma(T_h) \right) x_n^{(\mu-1)} \\ = \left((1 - (h\lambda\beta_{q^*}^*)^\mu) \, {T_h}^q - (h\lambda\beta_{q^*}^*)^{\mu-1} (1 - (h\lambda\beta_{q^*}^*)^\mu)(h\lambda)\,\rho(T_h) \right) x_n^{(\mu)}\end{aligned}$$

et

$$(h\lambda)\,\sigma(T_h)\,x_{n+q-q^*}^{(\mu-1)} = \rho(T_h)\,x_{n+q-q^*}^{(\mu)}.$$

L'élimination conduit alors au polynôme de stabilité absolue suivant

$$\Pi_\lambda(z) = \beta_{q^*}^*\, z^{q^*}\,(\rho^*(z) - (h\lambda)\,\sigma^*(z)) + \frac{(h\lambda\beta_{q^*}^*)^\mu(1-h\lambda\beta_{q^*}^*)}{1-(h\lambda\beta_{q^*}^*)^\mu}\,(\rho(z)\sigma^*(z) - \rho^*(z)\sigma(z)) \tag{8.108}$$

pour une méthode de prédiction-correction en mode P(EC)$^\mu$.

On constate que les deux polynômes obtenus sont essentiellement des perturbations d'ordre $(h\lambda)^{\mu+1}$ du polynôme de stabilité du correcteur $\rho^*(z) - (h\lambda)\,\sigma^*(z)$, le polynôme d'une méthode en mode P(EC)$^\mu$E possédant une structure plus simple (c'est une combinaison linéaire des polynômes de stabilité absolue du prédicteur et du correcteur) que celui de la même méthode en mode P(EC)$^\mu$. En se rappelant que $|\lambda| \leq L$, avec L la constante de Lipschitz de la condition (8.24) et pour h suffisamment petit, on voit de plus qu'ils convergent effectivement vers celui du correcteur lorsque μ tend vers l'infini (le facteur z^{q^*} dans (8.108) n'ayant pas d'influence à la limite).

La figure 8.18 présente les régions de stabilité absolue de méthodes d'Adams–Bashforth–Moulton utilisées selon différents modes. En comparant cette figure avec les figures 8.14 et 8.15, on note, comme on pouvait s'y attendre, qu'une méthode d'Adams–Bashforth–Moulton en mode PEC est *a priori* moins stable que la méthode d'Adams–Bashforth (mode PE) correspondante et qu'une méthode d'Adams–Moulton (mode P(EC)$^{+\infty}$) est toujours plus stable que la méthode d'Adams–Bashforth–Moulton, tous modes confondus.

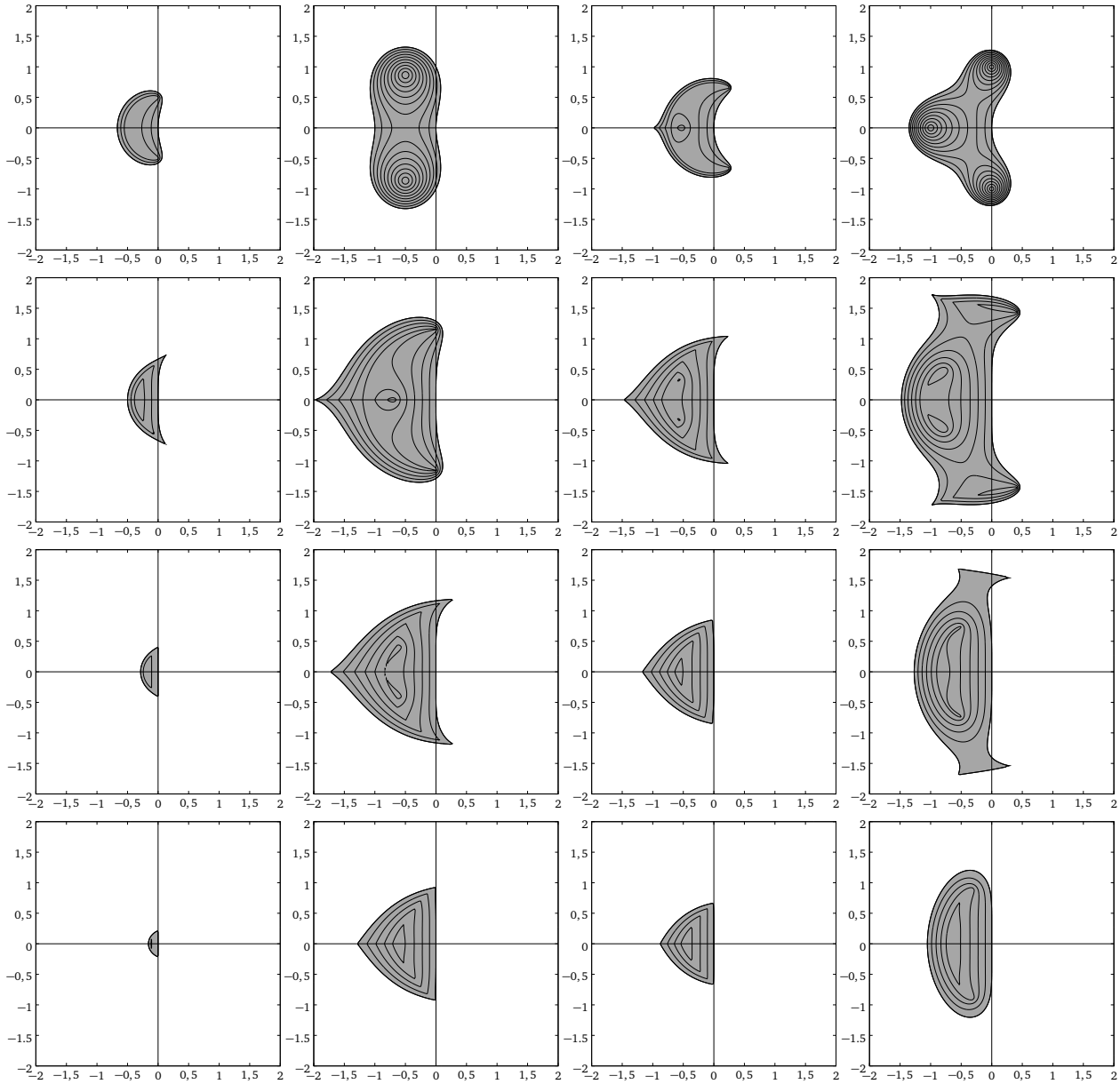


FIGURE 8.18 – Régions de stabilité absolue (en gris) pour différents modes (de gauche à droite : PEC, PECE, P(EC)^2, P(EC)^{2}E) des méthodes d'Adams–Bashforth–Moulton d'ordre un à quatre (de haut en bas).

On retiendra des résultats de cette section que si l'ordre et l'erreur de troncature locale principale d'une méthode de prédiction-correction basée sur des méthodes à pas multiples linéaires sont généralement ceux de son correcteur, ce n'est en revanche pas le cas pour son polynôme de stabilité, qui prend d'ailleurs une forme très différente selon que $\tau = 0$ ou 1.

8.6 Techniques pour l'adaptation du pas de discrétisation

L'erreur locale d'une méthode convergente diminuant avec la longueur du pas discrétisation, on peut imaginer implémenter dans un code de résolution numérique d'équations différentielles ordinaires un mécanisme d'adaptation du pas garantissant que, à chaque itération du schéma définissant la méthode employée, l'erreur commise ne dépasse pas une tolérance prescrite par l'utilisateur. La grille de discrétisation et l'approximation numérique de la

solution sont alors générées concurremment par le programme.

Si la mise en œuvre de cette idée est beaucoup plus simple pour une méthode à pas que pour une méthode à pas multiples pour des raisons déjà évoquées dans la sous-section 8.3.3, on est dans les deux cas confronté au problème pratique de l'estimation efficace de l'erreur locale. Pour le résoudre, on a en général recours à des estimateurs *a posteriori* de l'erreur locale de troncature (et plus précisément de la fonction d'erreur principale), car les estimateurs *a priori* ont souvent des formes compliquées et sont inemployables en pratique [74]. De tels estimateurs sont essentiellement fondés sur la comparaison de deux approximations distinctes de la solution du problème et peuvent être construits de diverses manières.

Indiquons enfin que l'adaptation du pas peut également servir à la détection d'un comportement singulier de la solution du problème (une explosion en temps fini par exemple) ou encore, lorsqu'une méthode explicite est utilisée, à déterminer si le système du problème que l'on cherche à résoudre est raide ou non (voir la section 8.7).

8.6.1 Cas des méthodes à un pas

Une première approche possible est celle dérivant du procédé d'extrapolation de Richardson [RG27], qui est une technique générale d'accélération de convergence applicable à nombre de méthodes numériques [75]. Pour la présenter, nous allons supposer que l'on utilise une méthode à un pas d'ordre p (typiquement une méthode de Runge–Kutta) avec un pas de discrétisation de longueur h, fournissant une approximation de la solution notée x_h. On peut écrire l'erreur de troncature locale au point t_{n+1} sous la forme

$$\tau_{n+1} = x(t_{n+1}) - (\tilde{x}_h)_{n+1} = \psi(t_n, x(t_n))\, h^{p+1} + O(h^{p+2}), \tag{8.109}$$

où la valeur $(\tilde{x}_h)_{n+1}$ est obtenue sous l'hypothèse localisante $(x_h)_n = x(t_n)$ et ψ désigne la fonction d'erreur principale de la méthode, généralement inconnue.

Admettons à présent que l'on dispose d'une seconde approximation numérique de la solution, notée x_{2h}, calculée par la même méthode, mais avec un pas de discrétisation de longueur $2h$. Sous l'hypothèse localisante $(x_{2h})_{n-1} = x(t_{n-1})$ et en effectuant un développement de Taylor au premier ordre de $\psi(t_{n-1}, x(t_{n-1}))$ au point t_n, on obtient que l'erreur locale au point t_{n+1} relative à cette approximation s'écrit

$$x(t_{n+1}) - (\tilde{x}_{2h})_{n+1} = \psi(t_{n-1}, x(t_{n-1}))\,(2h)^{p+1} + O(h^{p+2}) = \psi(t_n, x(t_n))\,(2h)^{p+1} + O(h^{p+2}).$$

En soustrayant cette égalité à (8.109), il vient

$$(\tilde{x}_{2h})_{n+1} - (\tilde{x}_h)_{n+1} = \left(1 - 2^{p+1}\right) \psi(t_n, x(t_n))\, h^{p+1} + O(h^{p+2}),$$

d'où, par substitution de l'expression trouvée pour la fonction d'erreur principale dans (8.109),

$$\tau_{n+1} = \left(1 - 2^{p+1}\right)^{-1} \left((\tilde{x}_{2h})_{n+1} - (\tilde{x}_h)_{n+1}\right) + O(h^{p+2}). \tag{8.110}$$

En abandonnant les hypothèses localisantes, on voit que l'on a obtenu un estimateur de l'erreur de troncature locale effectivement calculable. Si la valeur estimée est alors inférieure en valeur absolue à une tolérance fixée ε, l'erreur est jugée acceptable et l'on passe à l'étape suivante (si la valeur absolue est inférieure à $2^{-(p+1)}\varepsilon$, la longueur du pas est même généralement doublée). Si ce n'est pas la cas, on réitère le calcul d'estimation d'erreur en divisant cette fois par deux la longueur du pas de discrétisation.

Ce procédé fonctionne bien en pratique, mais il entraîne une importante augmentation du volume des calculs effectués à chaque étape. Pour une méthode de Runge–Kutta explicite à s niveaux, on a *a priori* besoin de $s-1$ évaluations supplémentaires de la fonction f (la valeur k_1 utilisée pour le calcul de $(x_{2h})_{n+1}$ ayant été évaluée lors du calcul de $(x_h)_n$ à l'étape précédente). De plus, en cas de réduction de la longueur du pas, la valeur approchée de la solution au point de grille courant doit être recalculée.

Une alternative à l'utilisation du procédé de Richardson est de faire appel à deux méthodes d'ordre différents p et $\widehat{p}$ (avec typiquement $\widehat{p} = p+1$), de fonctions d'incrément respectives Φ_f et $\widehat{\Phi}_f$, et d'utiliser la différence entre les valeurs approchées de la solution fournies par ces méthodes,

$$h\left(\widehat{\Phi}_f(t_n, x_n; h) - \Phi_f(t_n, x_n; h)\right),$$

74. Dans le cas d'une méthode de Runge–Kutta par exemple, toute estimation de l'erreur locale en un point nécessite plus d'évaluations de la fonction f que n'en demande la méthode pour le calcul de la valeur approchée de la solution en ce point.

75. Elle est par exemple à la base de la méthode de Romberg, mentionnée dans le chapitre 7.

comme estimation de l'erreur de troncature locale. Cette technique peut être rendue particulièrement efficace lorsqu'elle combine deux méthodes de Runge–Kutta, respectivement à s et $\widehat{s}$ niveaux (avec $s \leq \widehat{s}$) et d'ordre p et $\widehat{p}$ (avec $p < \widehat{p}$), *emboîtées*, c'est-à-dire pour lesquelles les quantités k_i coïncident pour $i = 1, \ldots, s$, car l'estimateur d'erreur est alors donné par la quantité

$$h \left(\sum_{i=1}^{s} (\widehat{b_i} - b_i) \, k_i + \sum_{i=s+1}^{\widehat{s}} \widehat{b_i} \, k_i \right),$$

dont le calcul demande un total de $\widehat{s}$ (contre $\widehat{s} + s$ *a priori*) évaluations de la fonction f.

Les méthodes de Runge–Kutta emboîtées (*embedded Runge–Kutta methods* en anglais) se caractérisent donc par la donnée d'un seul jeu de coefficients $\{a_{ij}\}_{1 \leq i,j \leq \widehat{s}}$ et de coefficients $\{c_i\}_{1 \leq i \leq \widehat{s}}$ et de deux jeux de coefficients $\{b_i\}_{1 \leq i \leq s}$ et $\{\widehat{b_i}\}_{1 \leq i \leq \widehat{s}}$, et donnent lieu à des tableaux de Butcher augmentés de la forme

$$\begin{array}{c|c} \boldsymbol{c} & A \\ \hline & \boldsymbol{b}^T \\ & \widehat{\boldsymbol{b}}^T \end{array}.$$

Lorsque $\widehat{s} = s + 1$, il est possible d'éviter une évaluation (lorsque la longueur du pas courant est acceptée) en faisant coïncider la valeur de $k_{\widehat{s}}$ avec celle de k_1 à l'étape suivante (voir par exemple les méthodes définies par les tableaux (8.111) et (8.112)), les méthodes étant dans ce cas désignées dans la littérature anglo-saxonne par l'acronyme *FSAL* (pour *first same as last* en anglais).

L'un des premiers essais d'incorporation d'un procédé d'emboîtement dans une méthode de Runge–Kutta explicite semble dû à Merson [Mer57]. Il repose sur l'utilisation d'une méthode à cinq niveaux d'ordre quatre, de tableau

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{3} & \frac{1}{3} & & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & \\ \frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & \\ 1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & \\ \hline & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \end{array},$$

pour laquelle Merson proposa d'estimer l'erreur de troncature locale par la quantité

$$\frac{h}{30} \left(-2\,k_1 + 9\,k_3 - 8\,k_4 + k_5 \right),$$

ce qui revient à comparer la solution approchée fournie par la méthode définie ci-dessus avec celle dont le tableau de Butcher est le suivant

$$\begin{array}{c|ccccc} 0 & & & & & \\ \frac{1}{3} & \frac{1}{3} & & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & \\ \frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & \\ 1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & \\ \hline & \frac{1}{10} & 0 & \frac{3}{10} & \frac{2}{5} & \frac{1}{5} \end{array}.$$

On a cependant vu dans la sous-section 8.4.2 que l'ordre maximal atteint par une méthode à cinq niveaux ne pouvait être qu'inférieur ou égal à quatre, il n'est par conséquent pas possible que la dernière méthode soit d'ordre cinq et l'on peut à juste raison penser que l'estimateur ainsi construit n'est pas valide[76]. Si la *méthode de Merson* n'entre donc pas dans le cadre exposé plus haut, elle n'en fût pas moins historiquement importante et ouvrit la voie au développement des méthodes de Runge–Kutta emboîtées.

76. On trouve de fait que l'ordre de la seconde méthode vaut généralement trois, mais qu'il est égal à cinq lorsque le système d'équations différentielles à résoudre est linéaire à coefficients constants.

Il découle néanmoins de la précédente considération qu'une méthode de Runge–Kutta explicite emboîtée d'ordre quatre requiert au moins six niveaux. C'est le cas de la *méthode d'England* $(4,5)$ [Eng69], résumée dans le tableau augmenté suivant

$$\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{2} & \frac{1}{2} & & & & & \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & & & & \\
1 & 0 & -1 & 2 & & & \\
\frac{2}{3} & \frac{7}{27} & \frac{10}{27} & 0 & \frac{1}{27} & & \\
\frac{1}{5} & \frac{28}{625} & -\frac{1}{5} & \frac{546}{625} & \frac{54}{625} & -\frac{378}{625} & \\
\hline
 & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} & 0 & 0 \\
 & \frac{1}{24} & 0 & 0 & \frac{5}{48} & \frac{27}{56} & \frac{125}{336}
\end{array}.$$

Un intérêt de cette méthode est que les coefficients b_5 et b_6 de la méthode d'ordre quatre sont nuls, ce qui fait que seulement quatre niveaux sont nécessaires si aucune estimation de l'erreur n'est requise. Une autre méthode emboîtée d'ordre quatre particulièrement populaire est la *méthode de Fehlberg* $(4,5)$, de tableau de Butcher augmenté

$$\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{4} & \frac{1}{4} & & & & & \\
\frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\
\frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\
1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\
\frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\
\hline
 & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\
 & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55}
\end{array}.$$

Cette méthode n'est qu'un exemple d'une classe de paires de schémas, introduite par Fehlberg [Feh69], dont les ordres et les nombres de niveaux respectifs sont donnés dans le tableau 8.4 et dont les coefficients sont choisis pour que la valeur absolue du coefficient de l'erreur de troncature principale de la méthode d'ordre p soit la plus petite possible.

p	s	$\hat{s}$
3	4	5
4	5	6
5	6	8
6	8	10
7	11	13
8	15	17

TABLE 8.4 – Ordre et nombres de niveaux respectifs des « paires de Fehlberg ».

En effet, dans les méthodes de Fehlberg, tout comme dans la méthode d'England, c'est la méthode d'ordre le plus bas qui est utilisée pour la construction effective de la solution approchée. Dans d'autres méthodes emboîtées, c'est celle d'ordre le plus élevé qui sert au calcul de cette solution et dont le coefficient de l'erreur de troncature principale est optimisé, ce qui les rend particulièrement appropriées pour l'adaptation du pas par extrapolation locale. Parmi celles-ci, on peut citer la *méthode de Dormand–Prince* $(5,4)$ [DP80], qui est une méthode *FSAL* à sept

niveaux d'ordre cinq, de tableau

$$\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}, \tag{8.111}$$

la *méthode de Bogacki–Shampine* $(3,2)$ [BS89], méthode *FSAL* à quatre niveaux d'ordre trois et de tableau

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{3}{4} & 0 & \frac{3}{4} & & \\
1 & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & \\
\hline
 & \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & 0 \\
 & \frac{7}{24} & \frac{1}{4} & \frac{1}{3} & \frac{1}{8}
\end{array}, \tag{8.112}$$

ou encore la *méthode de Cash–Karp* $(5,4)$ [CK90], méthode à six niveaux d'ordre cinq et de tableau

$$\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & \\
\frac{3}{5} & \frac{3}{10} & -\frac{9}{10} & \frac{6}{5} & & & \\
1 & -\frac{11}{54} & \frac{5}{2} & -\frac{70}{27} & \frac{35}{27} & & \\
\frac{7}{8} & \frac{1631}{55296} & \frac{175}{512} & \frac{575}{13824} & \frac{44275}{110592} & \frac{253}{4096} & \\
\hline
 & \frac{2825}{27648} & 0 & \frac{18575}{48384} & \frac{13525}{55296} & \frac{277}{14336} & \frac{1}{4} \\
 & \frac{37}{378} & 0 & \frac{250}{621} & \frac{125}{594} & 0 & \frac{512}{1771}
\end{array}.$$

La présentation des tableaux de Butcher augmentés n'étant pas toujours standardisée, nous avons pour chacune des précédentes méthodes fait correspondre la première ligne de coefficients b_i, $i = 1,\dots,s$, du tableau à la méthode d'ordre inférieur, le couple d'entiers apparaissant dans la dénomination de la méthode emboîtée précisant les ordres et rôles respectifs de chaque méthode individuelle. Ainsi, pour la méthode de Dormand–Prince $(5,4)$, on comprend que la méthode de Runge–Kutta d'ordre cinq fournit l'approximation à adopter en fin d'étape et que celle d'ordre quatre ne sert que pour estimer l'erreur.

On peut déterminer les fonctions de stabilité des méthodes de Runge–Kutta emboîtées au moyen des techniques employées pour les méthodes de Runge–Kutta pour lesquelles $p < s$ et décrites dans la sous-section 8.4.5.

Détermination des régions de stabilité absolue de quelques méthodes de Runge–Kutta emboîtées. La fonction de stabilité associée à une méthode de Runge–Kutta emboîtée est celle de la méthode fournissant l'approximation (et non de celle utilisée pour l'estimation d'erreur). Pour quelques-unes des méthodes présentées, on trouve

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4 + \frac{1}{144}(h\lambda)^5$$

pour la méthode de Merson ($s = 5$ et $p = 4$),

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4$$

pour la méthode d'England $(4,5)$ ($s=4$ et $p=4$, on observe que cette fonction est bien celle des méthodes de Runge–Kutta à quatre niveaux d'ordre quatre, donnée par (8.93)),

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4 + \frac{1}{104}(h\lambda)^5$$

pour la méthode de Fehlberg $(4,5)$ ($s=5$ et $p=4$), et

$$R(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4 + \frac{1}{120}(h\lambda)^5 + \frac{1}{600}(h\lambda)^6$$

pour la méthode de Dormand–Prince $(5,4)$ ($s=6$ et $p=5$). Les régions de stabilité absolue correspondantes sont représentées sur la figure 8.19. On note que la région pour la méthode Dormand–Prince $(5,4)$ est, de manière surprenante, une union d'ensembles disjoints.

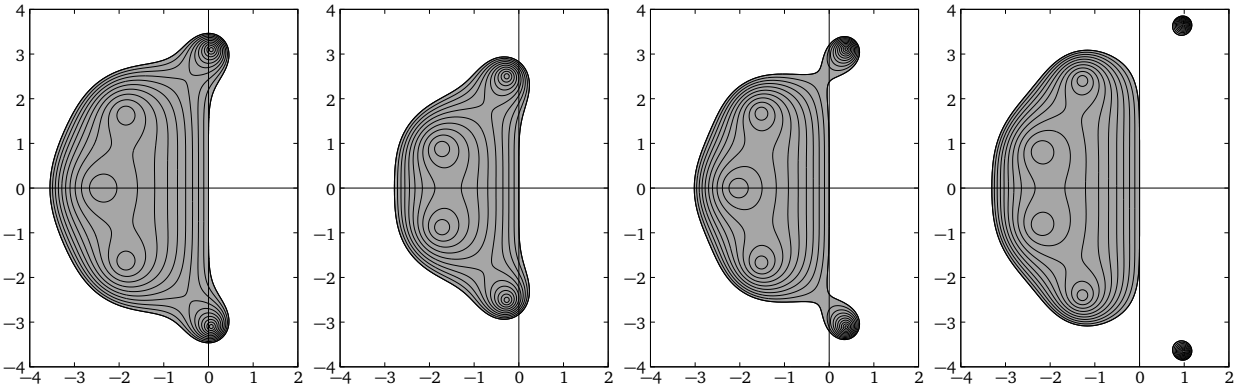


FIGURE 8.19 – Régions de stabilité absolue (en gris) pour des méthodes de Merson, d'England $(4,5)$, de Fehlberg $(4,5)$ et de Dormand–Prince $(5,4)$ (de gauche à droite).

8.6.2 Cas des méthodes à pas multiples linéaires *

Pour les méthodes à pas multiples linéaires, toute la difficulté d'adaptation du pas de discrétisation réside dans le changement de longueur du pas, puisque l'on a vu dans la section 8.5 que l'on dispose avec l'estimée de Milne (8.102) d'un estimateur d'erreur de troncature locale principale extrêmement peu coûteux et applicable à certaines classes de méthodes de prédiction-correction à pas multiples linéaires, dont font par exemple partie les méthodes d'Adams–Bashforth–Moulton. Dans le cas de ces dernières, il existe deux manières efficaces d'implémenter un mécanisme d'adaptation du pas.

Supposons que l'on travaille avec une méthode d'ordre q, ayant servi à obtenir une approximation x_n de la solution du problème au point t_n, avec $n \geq q$, et que la longueur du pas de discrétisation à utiliser pour le calcul de x_{n+1} soit modifiée de h à θh, où θ est un réel strictement positif. Si l'on veut continuer à utiliser la méthode telle quelle, on a besoin de connaître q valeurs approchées antérieures de la solution aux points équidistants $t_n, t_n - \theta h, t_n - 2\theta h, \ldots, t_n - (q-1)\theta h$, ce qui demande une interpolation des valeurs à disposition.

COMPLETER : nécessite d'interpoler pour obtenir donner les données en $t_n - \alpha h$, etc. voir [Kro73] lorsque la longueur du pas ne peut être de doublée ou divisée par deux à chaque étape, autre implémentation en exploitant le lien avec les méthodes de Nordsieck pour reformuler le problème

sinon : pas d'interpolation (on utilise les données déjà obtenues) mais les coefficients de la méthode deviennent variables lorsque le pas varie, voir [Ces61]

8.7 Systèmes raides

Nous consacrons cette avant-dernière section à la résolution numérique des équations, ou plus généralement des systèmes d'équations, différentielles ordinaires dits *raides*. Précédemment, nous avons en effet justifié à plusieurs

reprises l'emploi de méthodes de type implicite, autrement considérées comme coûteuses en temps de calcul, par le fait qu'elles étaient dans ce cas indispensables.

Si tout praticien des méthodes numériques possède une idée intuitive des phénomènes regroupés derrière le concept de *raideur*, en donner une définition mathématique correcte n'est pas une tâche aisée. Bien que l'on observe souvent les mêmes faits caractéristiques sur le plan numérique, les raisons amenant à qualifier un système d'équations différentielles de système raide peuvent être diverses. Néanmoins, l'approche la plus commune de ces difficultés se fait par le biais d'une théorie linéaire, en lien avec la notion de stabilité absolue introduite dans la sous-section 8.4.5. Avant d'aborder celle-ci, commençons par illustrer la problématique rencontrée en pratique au moyen de deux exemples.

8.7.1 Deux expériences numériques

Le premier problème que nous traitons, issu de [Mol08], est celui de la propagation d'une flamme de diffusion, c'est-à-dire de la détermination de la zone au sein de laquelle une réaction de combustion se produit entre deux réactants – un combustible et un comburant – séparés. En guise d'exemple, on peut penser à la combustion d'un solide, comme une allumette ou une bougie, en se rappelant que, juste après son allumage, la flamme augmente rapidement de volume jusqu'à atteindre une taille critique qu'elle conserve une fois que la quantité de dioxygène consommée en son intérieur s'est équilibrée avec celle disponible à sa surface. On peut modéliser de manière grossière [77] ce phénomène en considérant que la flamme a la forme d'une boule dont le rayon à l'instant t, noté $r(t)$ est solution du problème de Cauchy

$$r'(t) = r(t)^2(1-r(t)),\ t \geq 0,\ \text{avec } r(0) = \delta. \tag{8.113}$$

Le réel δ, supposé strictement positif et « petit », est le rayon de la boule à l'instant initial. On est intéressé par la détermination numérique de la solution du problème sur un intervalle de temps de longueur inversement proportionnelle à à la valeur de δ, qui va s'avérer être un paramètre critique vis-à-vis de la raideur de l'équation. On notera qu'il est ici possible de résoudre analytiquement le problème en faisant appel à la *fonction W de Lambert* [78] (voir [CGH$^+$96]). En effet, l'équation différentielle ordinaire étant à variables séparables, il trouve, après intégration, l'équation implicite

$$\frac{1}{r(t)} + \ln\left(\frac{1}{r(t)-1}\right) = \frac{1}{\delta} + \ln\left(\frac{1}{\delta-1}\right) - t,$$

dont la solution s'écrit

$$r(t) = \frac{1}{W(a\,e^{a-t}) + 1}$$

où l'on a posé $a = \frac{1}{\delta} - 1$, la fonction W satisfaisant $W(z)\,e^{W(z)} = z$ pour tout nombre complexe z.

On a représenté sur la figure 8.20 la solution du problème sur l'intervalle $\left[0, \frac{2}{\delta}\right]$ pour deux valeurs distinctes du paramètre δ. On observe que la solution croît lentement jusqu'à environ $\frac{1}{2\delta}$ pour alors atteindre très brusquement (relativement à l'échelle de temps considéré pour chaque cas) la valeur 1 qu'elle conserve ensuite.

77. On trouvera dans la sous-section 11.1.3 un modèle, basé sur une équation aux dérivées partielles plutôt qu'une équation différentielle ordinaire, plus fidèle à la dynamique de ce phénomène.

78. Jean-Henri Lambert (Johann Heinrich Lambert en allemand, 26 août 1728 - 25 septembre 1777) était un mathématicien, physicien, astronome et philosophe suisse. Auteur prolifique, on lui doit notamment l'introduction des fonctions hyperboliques en trigonométrie ou la première preuve de l'irrationalité de π, l'invention de plusieurs systèmes de projection cartographique en géographie, ainsi que des travaux fondateurs en photométrie.

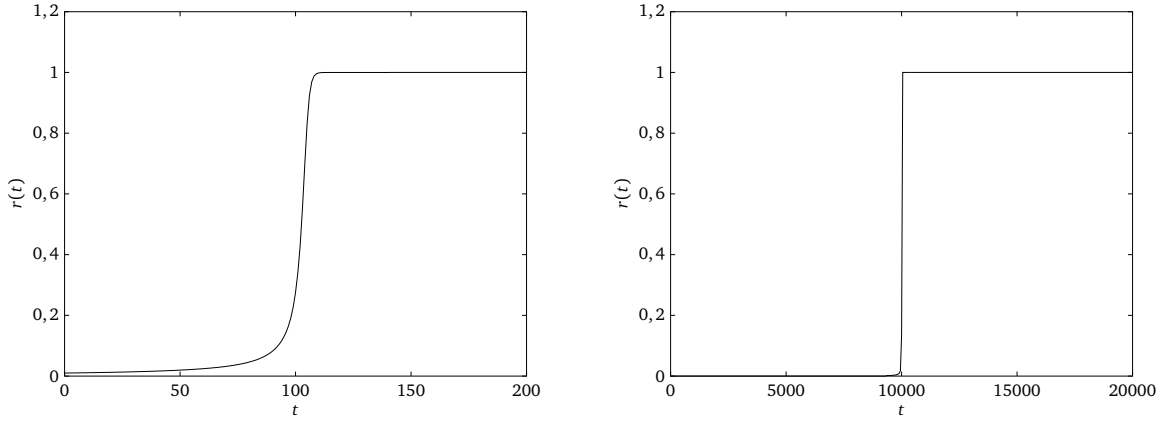


FIGURE 8.20 – Solution du problème (8.113) sur l'intervalle $\left[0, \frac{2}{\delta}\right]$, pour $\delta = 10^{-2}$ (à gauche) et $\delta = 10^{-4}$ (à droite).

Pour la résolution numérique de ce problème, on a utilisé deux méthodes à un pas, l'une adaptée à la résolution de systèmes raides, l'autre non, couplées à un mécanisme d'adaptation du pas de discrétisation, présentes dans le logiciel MATLAB (voir [SR97]). Celles-ci sont d'une part une méthode de Runge–Kutta emboîtée basée sur la paire $(5, 4)$ de Dormand–Prince (voir le tableau de Butcher (8.111)), implémentée dans la fonction `ode45`, et d'autre part une *méthode de Rosenbrock*[79] *modifiée* d'ordre deux, disponible dans la fonction `ode23s`. Cette dernière méthode peut être vue comme une généralisation d'une méthode de Runge–Kutta semi-implicite, ne requérant pas de résolution d'un système d'équations non linéaires, mais l'évaluation d'une approximation du jacobien de la fonction f et la résolution d'un système linéaire, à chaque étape (voir par exemple [Zed90] pour une présentation). Les résultats obtenus avec les deux valeurs de δ précédemment utilisées sont présentés sur les figures 8.21 et 8.22. Dans les deux cas, la tolérance fixée pour l'erreur relative utilisée pour adapter le pas est égale à 10^{-6}.

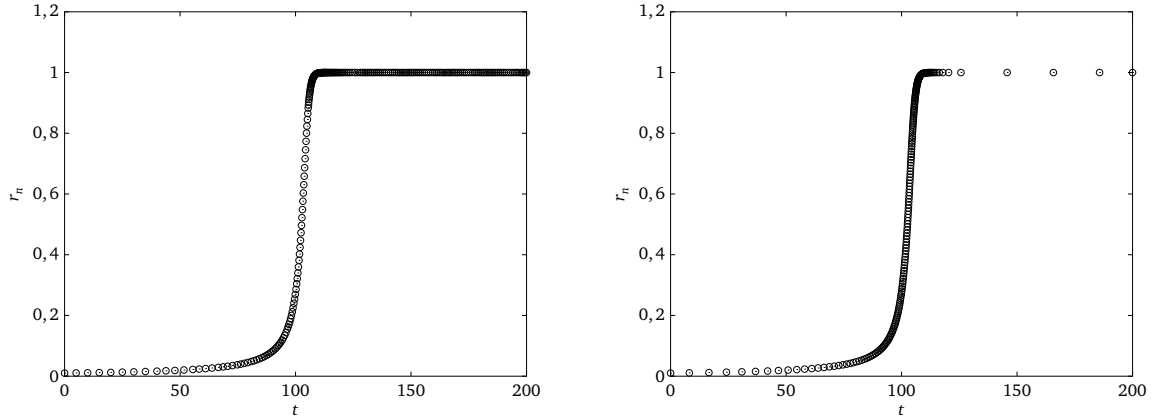


FIGURE 8.21 – Solutions numériques approchées du problème (8.113) sur l'intervalle $\left[0, \frac{2}{\delta}\right]$, pour $\delta = 10^{-2}$, respectivement obtenues par une méthode de Runge–Kutta emboîtée (à gauche) et une méthode de Rosenbrock modifiée (à droite), combinées à une stratégie d'adaptation du pas de discrétisation.

Pour $\delta = 10^{-2}$, les deux méthodes ont un comportement de manière similaires, utilisant respectivement 248 pas et 204 pas pour la seconde, cette disparité pouvant être imputée à une erreur de troncature locale parfois plus faible pour une méthode implicite. En revanche, pour $\delta = 10^{-4}$, l'équation est raide et les conséquences sur le

79. Howard Harry Rosenbrock (16 décembre 1920 - 21 octobre 2011) était un mathématicien anglais, spécialiste de la théorie du contrôle des systèmes. On lui doit l'introduction de méthodes numériques pour la résolution de problèmes d'optimisation non linéaire et de systèmes d'équations différentielles.

plan numérique sont immédiatement observables : la première méthode a maintenant besoin de 12224 pas pour résoudre le problème contre 231 pas pour la seconde. La différence fondamentale de comportement de la méthode explicite ici utilisée (la méthode implicite n'étant pas affectée) pour la résolution du problème (8.113) atteste de la raideur de ce dernier lorsque la donnée initiale est petite et l'intervalle d'intégration suffisamment grand.

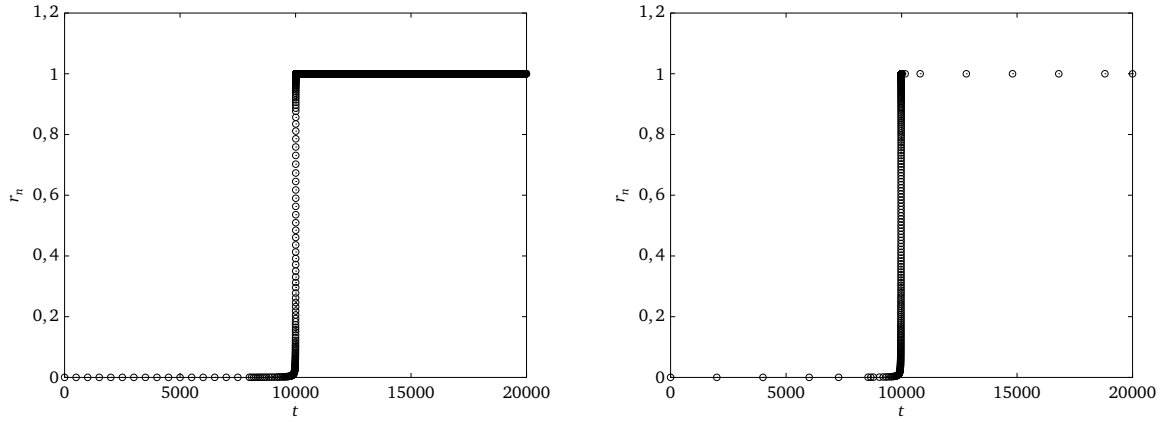


FIGURE 8.22 – Solutions numériques approchées du problème (8.113) sur l'intervalle $\left[0, \frac{2}{\delta}\right]$, pour $\delta = 10^{-4}$, respectivement obtenues par une méthode de Runge–Kutta emboîtée (à gauche) et une méthode de Rosenbrock modifiée (à droite), combinées à une stratégie d'adaptation du pas de discrétisation.

Le second exemple, tiré de [Lam91], considère la résolution sur l'intervalle $[0, 10]$ des deux problèmes de Cauchy suivants

$$\boldsymbol{x}'(t) = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \boldsymbol{x}(t) + \begin{pmatrix} 2\sin(t) \\ 2\left(\cos(t) - \sin(t)\right) \end{pmatrix}, \ \boldsymbol{x}(0) = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \tag{8.114}$$

et

$$\boldsymbol{x}'(t) = \begin{pmatrix} -2 & 1 \\ 998 & -999 \end{pmatrix} \boldsymbol{x}(t) + \begin{pmatrix} 2\sin(t) \\ 999\left(\cos(t) - \sin(t)\right) \end{pmatrix}, \ \boldsymbol{x}(0) = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \tag{8.115}$$

ayant tout deux la même solution, donnée par

$$\boldsymbol{x}(t) = 2\,e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} \sin(t) \\ \cos(t) \end{pmatrix},$$

et représentée sur la figure 8.23.

Pour la résolution numérique de ces problèmes, on a de nouveau utilisé une méthode de Runge–Kutta emboîtée basée sur la paire $(5, 4)$ de Dormand–Prince, ainsi qu'une méthode à pas multiples linéaire, issue d'une modification [80] des méthodes BDF, dite *NDF* (pour *numerical differentiation formula* en anglais) [Klo71], implémentée dans la fonction `ode15s` de MATLAB avec des mécanismes d'adaptation du pas discrétisation et de variation de l'ordre (de un à cinq) de la méthode. Si la résolution du problème (8.114) a nécessité 100 pas avec la première et 41 avec la seconde, il a en revanche fallu effectuer respectivement 12060 et 48 pas pour résoudre le problème (8.115).

80. Cette modification consiste en l'ajout d'un terme à la relation de récurrence (8.70), conduisant au schéma suivant

$$\sum_{i=1}^{q} \frac{1}{i} \nabla^i x_{n+q} = h f(t_{n+q}, x_{n+q}) + \kappa \left(\sum_{j=1}^{q} \frac{1}{j} \right) (x_{n+q} - x_{n+q}^{(0)}),$$

dans lequel

$$x_{n+q}^{(0)} = \sum_{i=0}^{q} \frac{1}{i} \nabla^i x_{n+q-1}$$

est un prédicteur et κ est un paramètre. Cette « correction » ne diminue pas l'ordre de la méthode et la constante d'erreur principale associée vaut $-\frac{1}{q+1} - \kappa \left(\sum_{j=1}^{q} \frac{1}{j} \right)$. Dans [SR97], la valeur du paramètre κ est choisie de manière à rendre la méthode plus précise que la méthode BDF correspondante tout en limitant la diminution de l'angle de A(α)-stabilité.

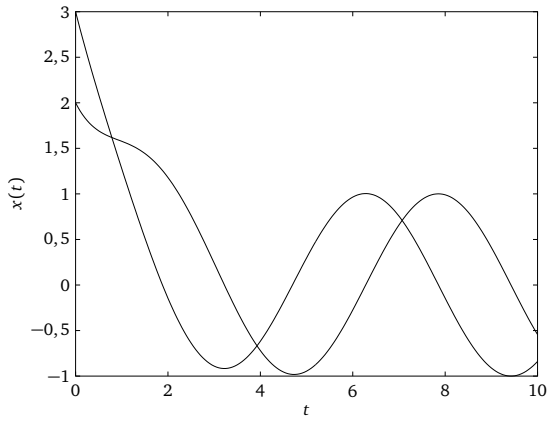


FIGURE 8.23 – Solution des problèmes (8.114) et (8.115) sur l'intervalle $[0, 10]$.

Les solutions de ces problèmes étant identiques, c'est le fait que l'un de ces deux systèmes différentiels, pourtant de même nature, soit raide et que l'autre non qui explique de tels écarts de comportement entre une méthode explicite et une méthode implicite de résolution.

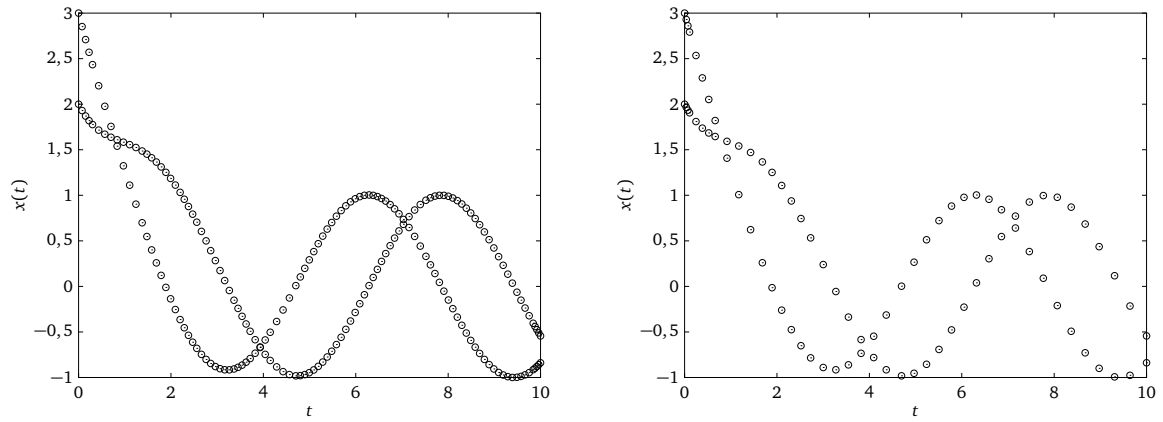


FIGURE 8.24 – Solutions numériques approchées du problème (8.114) sur l'intervalle $[0, 10]$ respectivement obtenues par une méthode de Runge–Kutta emboîtée (à gauche) et une méthode de Rosenbrock modifiée (à droite), combinées à une stratégie d'adaptation du pas de discrétisation.

8.7.2 Différentes notions de stabilité pour la résolution des systèmes raides *

Comme l'adjectif « raide » tend à le faire entendre (le terme de *raideur* désignant également le coefficient mesurant la résistance d'un corps à une déformation élastique), la résolution numérique d'un système raide impose à certaines méthodes des restrictions sur la longueur du pas de discrétisation bien plus sévères que la précision demandée ne l'exigerait. Les problèmes raides étant omniprésents dans les problèmes issus d'applications, avec pour exemples le problème de propagation de flamme (8.113), l'évolution de l'oscillateur de van der Pol dans un régime fortement amorti ($\mu \gg 1$) modélisée par l'équation (8.17) ou encore le problème de Robertson de système (8.22), il est de toute première importance de savoir les résoudre efficacement.

Il est toutefois difficile d'appréhender le phénomène sous-jacent d'un point de vue mathématique, les valeurs propres de la matrice jacobienne $\frac{\partial f}{\partial x}$, la dimension du système différentiel, la régularité de la solution, la valeur

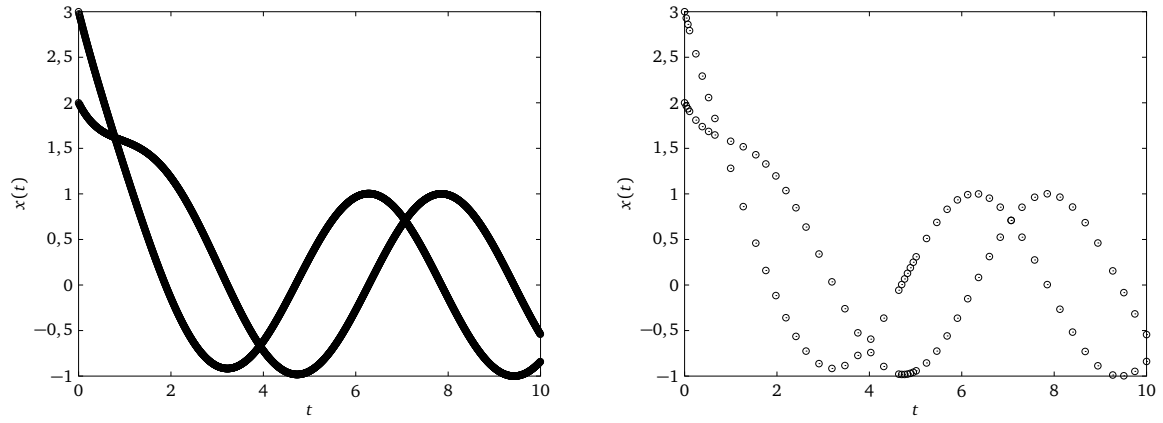


FIGURE 8.25 – Solutions numériques approchées du problème (8.115) sur l'intervalle $[0, 10]$ respectivement obtenues par une méthode de Runge–Kutta emboîtée (à gauche) et une méthode de Rosenbrock modifiée (à droite), combinées à une stratégie d'adaptation du pas de discrétisation..

initiale ou la taille de l'intervalle d'intégration pouvant toutes jouer un rôle. On reconnaît néanmoins dans bon nombre de cas que les problèmes de stabilité constatés en pratique sont dûs à l'existence de phases de transition rapide de la solution.

Considérons les problèmes du second exemple de la sous-section 8.7.1 et tentons de fournir une explication de la différence conséquente du nombre de pas utilisés pour leur résolution numérique par une méthode explicite à l'aune de la théorie de stabilité absolue des méthodes introduite dans la sous-section 8.4.5 (voir la section 8.4.6 pour son extension au cas des systèmes). Pour cela, notons tout d'abord que les solutions générales des systèmes différentiels de ces deux problèmes sont respectivement

$$\boldsymbol{x}(t) = c_1\, e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\, e^{-3\,t} \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} \sin(t) \\ \cos(t) \end{pmatrix}$$

et

$$\boldsymbol{x}(t) = c_1\, e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\, e^{-1000\,t} \begin{pmatrix} 1 \\ -998 \end{pmatrix} + \begin{pmatrix} \sin(t) \\ \cos(t) \end{pmatrix},$$

les valeurs des constantes arbitraires c_1 et c_2 étant fixées par la donnée d'une condition initiale. Ces solutions sont toutes deux composées d'une partie dite *transitoire*, tendant vers zéro lorsque t tend vers l'infini, et d'une partie *stationnaire*, qui est par conséquent observée en temps long (voir la figure 8.23).

Si l'on est intéressé par la simulation de la partie stationnaire de la solution, on se trouve confronté aux exigences de stabilité absolue de la méthode numérique utilisée, qui impliquent pour les problèmes considérés que les réels $-h$ et $-3h$ (resp. $-h$ et $-1000h$) appartiennent à l'intervalle de stabilité absolue de la méthode pour le problème (8.114) (resp. (8.115)). L'intervalle de stabilité de la paire $(5, 4)$ de Dormand–Prince étant approximativement $]-3, 0[$ (voir la figure 8.19), la contrainte sur la longueur du pas pour la résolution numérique du problème (8.115) est drastique, puisque l'on demande que $h < 0,003$, alors que l'on a $h < 1$ pour le problème (8.114). En revanche, les méthodes NDF d'ordre un à cinq n'imposent aucune restriction de ce type, leurs intervalles de stabilité respectifs étant tous égaux à $]-\infty, 0[$.

Pour un système différentiel linéaire à coefficients constants et non homogène de dimension d dont la matrice possède des valeurs propres complexes λ_i, $i = 1, \ldots, d$, ayant toutes une partie réelle strictement négative, posons

$$\mathrm{Re}(\overline{\lambda}) \leq \mathrm{Re}(\lambda_i) \leq \mathrm{Re}(\underline{\lambda}) < 0,\ i = 1, \ldots, d.$$

La situation est alors la suivante. Pour toute méthode de résolution de région de stabilité bornée, la longueur du pas de discrétisation sera d'autant plus petite que la valeur $|\mathrm{Re}(\overline{\lambda})|$ est grande, cette contrainte persistant après que la composante en question soit devenue négligeable dans la solution [81] du fait de sa décroissance rapide. Le temps

81. Ceci est encore vrai si la condition initiale est telle que la composante associée à la valeur propre $\overline{\lambda}$ n'est pas présente dans la solution du problème de Cauchy, comme c'est le cas pour les problèmes (8.114) et (8.115).

de simulation pour atteindre un régime stationnaire est lui d'autant plus long que le valeur $\left|\mathrm{Re}(\underline{\lambda})\right|$ est petite. Ceci conduit à proposer une caractérisation de la raideur par l'introduction d'un *quotient de raideur*

$$\frac{|\mathrm{Re}(\overline{\lambda})|}{\left|\mathrm{Re}(\underline{\lambda})\right|},$$

un système étant considéré comme raide lorsque le quotient est grand devant l'unité. Cette première définition n'est cependant pas sans inconvénient. En effet, si $\left|\mathrm{Re}(\underline{\lambda})\right| \simeq 0$, le quotient de raideur peut être très grand sans pour autant que $|\mathrm{Re}(\overline{\lambda})|$ le soit et que le système soit « réellement » raide (au sens où la longueur du pas ne se trouve pas contrainte par l'exigence de stabilité absolue). D'autre part, on voit qu'elle ne concerne que des systèmes d'équations différentielles linéaires à coefficients constants et se révèle complètement inadaptée [82] pour le traitement des systèmes non linéaires ou même simplement linéaires mais à coefficients non constants.

On peut néanmoins formuler la définition suivante, issue de [Lam91] et largement empirique, qui condense de manière pragmatique ce que l'on observe généralement en pratique.

Définition 8.43 (système raide) *Un système d'équations différentielles ordinaires est dit* ***raide*** *sur un intervalle d'intégration s'il force une méthode numérique dont la région de stabilité absolue est de taille finie à utiliser un pas de discrétisation excessivement petit compte tenu de la régularité de la solution exacte.*

Une première approche pour le choix de méthode adaptées à la résolution numérique des systèmes raides repose sur l'explication du phénomène dans le cas linéaire développée plus haut. Elle consiste à exiger que la méthode soit absolument stable pour toute valeur du nombre complexe λ, tel que $\mathrm{Re}(\lambda) < 0$, dans le problème (8.88) et la longueur h du pas de la grille de discrétisation, ce qu'on résume dans le concept suivant, initialement introduit pour les méthodes à pas multiples linéaires.

Définition 8.44 (A-stabilité d'une méthode [Dah63]) *Une méthode numérique pour l'approximation de la solution de* (8.88) *est dite* ***A-stable*** *si*

$$\mathscr{S} \cap \mathbb{C}_- = \mathbb{C}_-, \tag{8.116}$$

où $\mathscr{S}$ désigne la région de stabilité absolue de la méthode et $\mathbb{C}_- = \{z \in \mathbb{C} \,|\, \mathrm{Re}(z) < 0\}$.

Un rapide retour sur les diverses figures présentant les régions de stabilité absolue de méthodes dans la sous-section 8.4.5 et les sections 8.5 et 8.6 permet de voir que les méthodes d'Euler implicite, de la règle du trapèze et BDF à deux et trois pas sont A-stables.

Nous avons relevé dans la sous-section 8.4.5 une relation entre la fonction de stabilité de certaines méthodes de Runge–Kutta implicites et les approximants de Padé de la fonction exponentielle. Ce lien peut être exploité pour démontrer qu'une méthode à un pas est A-stable.

Théorème 8.45 (condition suffisante de A-stabilité d'une méthode à un pas) *Si la fonction de stabilité associée à une méthode à un pas est l'approximant de Padé de type (m,m) ou $(m+1,m)$ de la fonction exponentielle, $m = 0, 1, 2, \ldots$, alors cette méthode est A-stable.*

DÉMONSTRATION. A ECRIRE □

DONNER application : méthodes de Runge–Kutta implicites car, pour tout s, il existe une méthode A-stable et d'ordre $2s$.

Pour les méthodes à pas multiples linéaires, on a la condition nécessaire de A-stabilité suivante.

Théorème 8.46 (condition nécessaire de A-stabilité d'une méthode à pas multiples linéaire) *Si une méthode à pas multiples linéaire est A-stable, alors*

$$\mathrm{Re}\left(\frac{\rho(z)}{\sigma(z)}\right) > 0 \; si \; |z| > 1. \tag{8.117}$$

Si les racines de ρ et σ diffèrent, cette condition est aussi suffisante.

82. Plusieurs théories ont été développées pour combler ce manque et l'on renvoie le lecteur intéressé par ce sujet au dernier chapitre de [Lam91] pour une introduction.

DÉMONSTRATION. La méthode étant A-stable, toute racine ξ du polynôme $\rho - \nu\sigma$, avec $\nu \in \mathbb{C}$, satisfait $|\xi| \leq 1$ dès que $\mathrm{Re}(\nu) \leq 0$. La négation de cette application, à savoir que $|\xi| > 1$ implique que $\mathrm{Re}(\nu) > 0$, fournit alors la condition (8.117) puisque l'on a $\nu = \frac{\rho(\xi)}{\sigma(\xi)}$.
Supossons à présent que la condition (8.117) soit satisfaite et que les racines des polynôme ρ et σ diffèrent. Soit $\nu_0 \in \mathbb{C}$ tel que $\mathrm{Re}(\nu_0) \leq 0$ et soit ξ une racine du polynôme $\rho - \nu_0\sigma$. On a alors $\sigma(\xi) \neq 0$ et $\nu_0 = \frac{\rho(\xi)}{\sigma(\xi)}$, ce qui implique que $|\xi| \leq 1$. Pour que la méthode soit A-stable, il reste à prouver que ξ est une racine simple si son module est égal à un. Par un argument de continuité, il découle de (8.117) que les assertions $|\xi| = 1$ et $\mathrm{Re}(\nu_0) < 0$ sont contradictoires. A FINIR □

Cette notion de stabilité est extrêmement restrictive pour les méthodes à pas multiples linéaires, comme le montre le résultat suivant.

Théorème 8.47 (« seconde barrière de Dahlquist » [Dah63]) *Il n'existe pas de méthode à pas multiples linéaire explicite qui soit A-stable. De plus, l'ordre d'une méthode à pas multiples linéaire implicite et A-stable est au plus égal à deux.*

DÉMONSTRATION. A ECRIRE □

REPRENDRE De plus, si l'ordre est égal à deux, la constante d'erreur de la méthode satisfait $C \leq \frac{1}{12}$. La méthode de la règle du trapèze est le seule méthode A-stable d'ordre deux avec $C = -\frac{1}{12}$

Bien qu'il découle de ce théorème qu'une majorité de méthodes à pas multiples linéaires ne sont pas A-stables, ceci ne signifie pas pour autant qu'elles ne peuvent servir à la résolution numérique de systèmes raides. Des versions « affaiblies » de la condition de stabilité (8.116), également pertinentes, existent en effet. L'une d'entre elles mène à la notion de *A(α)-stabilité* suivante, adaptée aux systèmes raides dont les valeurs propres incriminées sont situées à proximité de l'axe réel négatif dans le plan complexe.

Définition 8.48 (A(α)-stabilité d'une méthode [Wil67]) *Une méthode numérique pour l'approximation de la solution de* (8.88) *est dite* ***A(α)-stable****, avec* $0 < \alpha < \frac{\pi}{2}$*, si*

$$\{z \in \mathbb{C} \,|\, |\arg(-z)| \leq \alpha,\ z \neq 0\} \subset \mathscr{S},$$

où $\mathscr{S}$ *désigne la région de stabilité absolue de la méthode. Elle est dite* ***A(0)-stable*** *si elle est A(α)-stable pour une valeur de* α *suffisamment petite.*

Les méthodes BDF à q pas zéro-stables, c'est-à-dire pour $1 \leq q \leq 6$, sont A(α)-stables pour les valeurs de l'angle α indiquées dans le tableau 8.5, ce qui les rend particulièrement attractives pour la résolution de systèmes raides.

q	α
1	90°
2	90°
3	$86,03^\circ$
4	$73,35^\circ$
5	$51,84^\circ$
6	$17,84^\circ$

TABLE 8.5 – Valeur maximale de l'angle α (mesuré en degrés) pour laquelle une méthode BDF à q pas est A(α)-stable.

Pour tout $\alpha < \frac{\pi}{2}$ donné et pour tout entier naturel q, il existe une méthode à q pas linéaires A(α)-stable d'ordre q (voir [JN82]). Il est en revanche illusoire de penser pouvoir s'affranchir aussi facilement de la seconde barrière de Dahlquist : pour des ordres élevés et des valeurs de α proches de $\frac{\pi}{2}$, la grandeur des constantes d'erreur principales de ces méthodes les rend en pratique inutiles.

MENTIONNER les autres concepts introduits : A_0-stabilité, L-stabilité, "stiff stability"...

Comme on peut l'observer sur les figures de la sous-section 8.4.5 et de la section 8.5, ces différentes propriétés de stabilité ne peuvent être satisfaites par une méthode méthode de Runge–Kutta explicite, une méthode à pas multiples linéaire explicite ou même une méthode de prédiction-correction de type $\mathrm{P(EC)}^\mu\mathrm{E}^{1-\tau}$ avec μ un entier fixé, ce qui rend l'emploi de méthodes implicites obligatoire. Les difficultés ne se trouvent cependant pas toutes éliminées, la résolution des systèmes d'équations non linéaires définissant ces méthodes par la méthode

des approximations successives imposant une forte restriction [83] sur la longueur du pas de discrétisation. Cette situation, quelque peu paradoxale puisque c'est ici le caractère implicite de la méthode (et non sa stabilité) qui pose problème, se règle par l'utilisation d'une méthode de Newton–Raphson modifiée (voir respectivement les relations de récurrence (8.50) et (8.61) pour les méthodes de Runge–Kutta et à pas multiples linéaires).

8.8 Application à la résolution numérique de problèmes aux limites **

8.8.1 Définition du problème

8.8.2 Méthodes de tir

PARLER ici des *méthodes de tir* (*shooting methods* en anglais), qui constituent dans ce cas une alternative aux méthodes de différences finies

8.9 Notes sur le chapitre *

En plus des références déjà conseillées pour l'ensemble du cours et traitant de la résolution numérique des équations différentielles ordinaires, nous renvoyons le lecteur intéressé à l'ouvrage particulièrement abordable de Lambert [Lam91]. Pour de nombreux autres développements, ainsi que des aspects techniques et historiques, on ne peut que recommander les livres de Hairer, Nørsett et Wanner [HNW93, HW96] et de Butcher [But08].

L'utilisation de la méthode d'Euler pour la résolution numérique des équations différentielles ordinaires fut décrite pour la première fois de manière détaillée dans le premier volume des *Intitutiones calculi integralis* (*sectio secunda, caput VII, De integratione aequationum differentialium per approximationem*), paru en 1768, pour les équations du premier ordre et dans le second volume (*sectio prima, caput XII, De aequationum differentio-differentialium integratione per approximationes*), paru en 1769, pour les équations du second ordre. D'un point de vue théorique, sa convergence fut prouvée par Cauchy dans les septième et huitième leçons de son cours traitant des équations différentielles ordinaire à l'École Polytechnique imprimé en 1824, conduisant au premier résultat (constructif) d'existence de solution pour une équation différentielle ordinaire « quelconque » dans le champ réel [84], sous l'hypothèse que l'application f définissant l'équation et ses dérivées sont continues et bornées. Ce résultat fut redémontré de manière indépendante et plus précise par Lipschitz en 1868 [Lip68], en supposant f continue et satisfaisant la condition (8.7), et constitue le théorème 8.10, dont la formulation « moderne » est l'œuvre de Picard [Pic93] et de Lindelöf [Lin94].

L'analyse de méthodes de Runge–Kutta, effectuée par Butcher dans une série d'articles, a donné lieu à une théorie algébrique d'une certaine classe de méthodes d'intégration numérique des équations différentielles ordinaires [But72], qui fait apparaître un groupe, le *groupe de Butcher*, représentable par une famille de fonctions à valeurs réelles définies sur l'algèbre de Hopf [85] des arbres enracinés [86] et dont l'intérêt dépasse largement celui de l'analyse numérique, puisqu'il intervient de manière fondamentale dans la formulation mathématique de la renormalisation en théorie quantique des champs [CK99, Bro00]. Le lecteur intéressé par l'histoire des méthodes de Runge-Kutta pourra consulter l'article [But96].

A FINIR Il existe une alternative à la théorie de Butcher, due à Albrecht [Alb87, Alb96], pour déterminer l'ordre des méthodes de Runge–Kutta, qui consiste à voir ces dernières comme des méthodes composites linéaires (A DEFINIR, lien avec méthodes d'intégration utilisée à chaque étape), avantages dans le cas des systèmes d'edo

Les méthodes d'Adams–Moulton et d'Adams–Bashforth sont l'œuvre du seul Adams. On en trouve un exposé complet dans un traité sur la théorie de la capillarité de Bashforth [BA83], dans lequel ce dernier s'intéresse à la

83. Rappelons que les conditions sur la longueur du pas assurant la convergence de la méthode des approximations successives sont les inégalités (8.44) (pour une méthode de Runge–Kutta implicite) et (8.59) (pour une méthode à pas multiples linéaire implicite) et que la constante de Lipschitz L, de l'ordre de $|Re(\overline{\lambda})|$ pour un système différentiel linéaire non homogène, peut être très grande pour un système raide.

84. Il démontre en effet que, sous certaines conditions de régularité sur la fonction f, les valeurs calculées par la méthode tendent, lorsque le pas de discrétisation tend vers zéro, vers celles d'une fonction qui est solution, au moins localement, du problème de Cauchy (8.1)-(8.5).

85. Heinz Hopf (19 novembre 1894 - 3 juin 1971) était un mathématicien allemand, connu pour ses travaux en topologie algébrique.

86. Un *arbre* est un graphe non orienté, acyclique et connexe. Il est possible de représenter cet objet dans un plan au moyen d'un plongement, ce qui permet d'orienter ses arêtes. Une *racine* d'un arbre plongé est la donnée d'une orientation d'une de ses arêtes. Un arbre plongé muni d'une racine est dit *enraciné*.

forme prise par une goutte de liquide reposant sur un plan [87]. Le nom de Moulton ne leur fut associé que lorsque ce dernier réalisa que les méthodes implicites d'Adams pouvaient être utilisées en conjonction avec leurs contreparties explicites pour former des paires de prédicteur-correcteur [Mou26].

Un résumé détaillé de l'histoire des méthodes à pas multiples est proposé dans l'article [Tou98].

Une référence historique pour le concept de système raide d'équations différentielles ordinaires est l'article de Curtiss et Hirschfelder [CH52], dans lequel les méthodes BDF furent introduites.

MENTIONNER AUSSI :

utilisation du jacobien de f dans une méthode à niveaux de type Runge–Kutta : méthodes de Rosenbrock [Ros63]

les méthodes de type Nordsieck [Nor62]

Des méthodes numériques, s'apparentant aux méthodes à pas multiples d'Adams et spécifiques aux équations différentielles ordinaires d'ordre deux de la forme particulière

$$x''(t) = f(t, x(t)),$$

c'est-à-dire dans lesquelles aucune dérivée n'apparaît au second membre, qui sont typiques de la modélisation de problèmes en mécanique céleste sans dissipation, furent introduites par des astronomes au dix-neuvième et vingtième siècles. On peut ainsi citer les travaux de Bond [88] [Bon49], de Størmer [89] [Stø07] pour expliquer le phénomène des aurores boréales, de Cowell [90] et Crommelin [91] pour la détermination trajectoire de la comète de Halley, dont les apports sont résumés en détail dans [Tou98]. ET Nÿstrom (?)

A REPRENDRE (lier avec le paragraphe précédent ?) Il existe des classes importantes d'équations différentielles ordinaires présentant des propriétés de structure spécifiques, souvent de nature géométrique, dont la préservation au niveau discret du schéma conduit à des méthodes de résolution performantes, notamment lorsque l'on résout numériquement sur de longs intervalles de temps. / Ces derniers systèmes sont des exemples typiques de systèmes différentiels pouvant s'écrire sous la forme d'*équations canoniques de Hamilton* [92]

$$\boldsymbol{p}'(t) = -\frac{\partial H}{\partial \boldsymbol{q}}(\boldsymbol{p}(t), \boldsymbol{q}(t), t),\ \boldsymbol{q}'(t) = \frac{\partial H}{\partial \boldsymbol{p}}(\boldsymbol{p}(t), \boldsymbol{q}(t), t),$$

dans lesquelles les quantités $\boldsymbol{p}(t)$ et $\boldsymbol{q}(t)$ à valeurs dans $\mathbb{R}^d$ représentent respectivement les positions et les moments du système au temps t et la fonction H, de $\mathbb{R}^{2d+1}$ à valeurs dans $\mathbb{R}^d$, est l'*hamiltonien* du système, représentant son énergie totale. On écrit encore

$$\boldsymbol{x}'(t) = J\,\boldsymbol{\nabla} H(\boldsymbol{x}(t), t),$$

en posant $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{pmatrix}$ et $J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}$, et l'on dit que le système est hamiltonien. (A VOIR)

Par exemple, les systèmes dit hamiltoniens, qui sont omniprésents en physique, ont la propriété géométrique de posséder un flot symplectique. / Les systèmes hamiltoniens occupent une place de toute première importance en mécanique classique, en physique statistique et en mécanique quantique en raison des propriétés remarquables dont ils jouissent. En particulier, le *flot* engendré par l'évolution dynamique d'un tel système à partir d'une condition initiale $\boldsymbol{x}_0$ au temps t_0, c'est-à-dire l'application ϕ_t telle que $\phi_t(\boldsymbol{x}_0) = \boldsymbol{x}(t)$, $t \geq t_0$, est *symplectique*. En d'autres

87. La mise en équation de la surface de la courbe méridienne de la surface conduit à une équation différentielle du second ordre. Bashforth en confia la résolution numérique à Adams, qui pour ce faire appliqua, après avoir remplacé l'équation en question par un système équivalent de deux équations différentielles du premier ordre, une méthode qu'il avait nouvellement imaginée. On notera qu'Adams faisait appel à la méthode de Newton–Raphson pour la résolution numérique de l'équation non linéaire (8.67) lors de l'utilisation de ses méthodes sous leur forme implicite.

88. George Phillips Bond (20 mai 1825 - 17 février 1865) était un astronome américain. Il découvrit avec son père, William Cranch Bond, le satellite de Saturne Hypérion en 1848 et fut l'un des premiers à faire usage de la photographie en astronomie, prenant les premiers clichés d'une étoile (Véga) en 1850 et d'une binaire visuelle (Mizar et Alcor) en 1857.

89. Fredrik Carl Mülertz Størmer (3 septembre 1874 - 13 août 1957) était un mathématicien et physicien norvégien. Il est connu à la fois pour ses travaux en théorie des nombres et pour ses études sur le mouvement des particules chargées dans la magnétosphère et la formation des aurores polaires.

90. Philip Herbert Cowell (7 août 1870 - 6 juin 1949) était un astronome britannique. COMPLETER

91. Andrew Claude de la Cherois Crommelin (6 février 1865 - 20 septembre 1939) était un astronome britannique. COMPLETER

92. Sir William Rowan Hamilton (4 août 1805 - 2 septembre 1865) était un mathématicien, physicien et astronome irlandais, dont les apports à la mécanique classique, à l'optique et à l'algèbre furent importants. On lui doit notamment une formulation alternative et fondamentale de la mécanique dite newtonienne, ainsi que la découverte des quaternions.

mots, ce flot vérifie une relation liée à la conservations des aires dans l'espace des phases, qui se traduit par le fait que la dérivée $D\phi_t$ satisfait

$$D\phi_t(\boldsymbol{x})^T J D\phi_t(\boldsymbol{x}) = J,\ \forall \boldsymbol{x},\ \forall t \text{ tels que } \phi_t(\boldsymbol{x}) \text{ existe.}$$

Cette relation est liée à la conservation des aires dans l'espace des phases.

Un intégrateur numérique est dit symplectique s'il conserve la structure hamiltonienne des équations, i.e. s'il préserve ces propriétés géométriques du flot exact... / En respectant la géométrie de l'espace des phases au niveau discret, on change de point de vue pour la méthode en se concentrant non pas sur l'approximation d'une solution (i.e. d'une trajectoire) mais en considérant plutôt la méthode comme un système dynamique discret approchant le *flot* d'une équation différentielle ordinaire, c'est-à-dire l'application φ_t associant à la valeur initiale x_0 à l'instant initial la valeur de la solution au temps t : $\varphi_t(x_0) = x(t)$.

Une méthode numérique à un pas est dite symplectique si le flot numérique ϕ_h est une application symplectique :

$$D\phi_h(t,x)^T J D\varphi_h(t,x) = J,$$

De telles méthodes existent. Par exemple, la méthode obtenue en appliquant la méthode d'Euler explicite aux variables de position $q \in \mathbb{R}^d$ et de moments $p \in \mathbb{R}^d$

$$\begin{aligned} p_{n+1} &= p_n - h\,\boldsymbol{\nabla} q^H(p_{n+1}, q_n) \\ q_{n+1} &= q_n - h\,\boldsymbol{\nabla} p^H(p_{n+1}, q_n) \end{aligned}$$

en est un exemple.

si la méthode symplectique, l'équation modifiée obtenue par analyse d'erreur inverse est elle aussi hamiltonienne, ce qui permet de prouver la quasi-conservation de l'énergie (en h^{-1}) sur des temps exponentiellement longs.

Une caractérisation des méthodes de Runge–Kutta symplectiques, obtenue de manière indépendante par Lasagni [Las88], Sanz-Serna [San88] et Suris en 1988, est que les coefficients d'une méthode symplectique à s niveaux doivent satisfaire

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0,\ i,j = 1,\dots,s,$$

ce qui implique que cette méthode est implicite. On peut montrer que les méthodes de Runge–Kutta implicites basées sur les formules de Gauss satisfont ces conditions.

exemple d'intégrateur symplectique utilisé de longue date est la *méthode de Størmer–Verlet* [93], proposée par Verlet pour la dynamique moléculaire [Ver67]. hamiltonien : $H(p,q) = \frac{1}{2} p^T M p + V(q)$ avec M une matrice symétrique définie positive, qui est la méthode explicite donnée par

$$\begin{aligned} p_{n+1/2} &= p_n - h/2\,\boldsymbol{\nabla} V(q_n) \\ q_{n+1} &= q_n + hM^{-1} p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - h/2\,\boldsymbol{\nabla} V(q_{n+1}) \end{aligned}$$

précédemment utilisée par Størmer en 1907

Note : autres méthodes/techniques d'intégration géométrique pour conserver d'autres propriétés, notamment la symétrie, la conservation d'intégrales premières, la structure de Poisson, etc. pour des classes particulières de systèmes d'edo

Références

[Alb87] P. ALBRECHT. A new theoretical approach to Runge–Kutta methods. *SIAM J. Numer. Anal.*, 24(2):391–406, 1987. DOI: 10.1137/0724030.

[Alb96] P. ALBRECHT. The Runge–Kutta theory in a nutshell. *SIAM J. Numer. Anal.*, 33(5):1712–1735, 1996. DOI: 10.1137/S0036142994260872.

[Ale77] R. ALEXANDER. Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s. *SIAM J. Numer. Anal.*, 14(6):1006–1021, 1977. DOI: 10.1137/0714068.

93. Loup Verlet (né le 24 mai 1931) est un physicien et philosophe français, pionnier de la simulation par ordinateur des modèles moléculaires dynamiques.

[Arz95] C. ARZELÀ. Sulle funzioni di linee. *Mem. Accad. Sci. Ist. Bologna Cl. Sci. Fis. Mat.*, 5(5):55–74, 1895.

[Asc84] G. ASCOLI. Le curve limiti di una varietà data di curve. *Atti R. Accad. Dei Lincei Memorie Cl. Sci. Fis. Mat. Nat.*, 18(3):521–586, 1884.

[BA83] F. BASHFORTH and J. C. ADAMS. *An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid, with an explanation of the method of integration employed in constucting the tables which give the theoretical forms of such drops*. Cambridge University Press, 1883.

[Bon49] G. P. BOND. On some applications of the method of mechanical quadratures. *Mem. Amer. Acad. Arts Sci.*, 4(1):189–208, 1849. DOI: 10.2307/25058159.

[Bro00] C. BROUDER. Runge–Kutta methods and renormalization. *European Phys. J. C*, 12(3):521–534, 2000. DOI: 10.1007/s100529900235.

[BS89] P. BOGACKI and L. F. SHAMPINE. A 3(2) pair of Runge-Kutta formulas. *Appl. Math. Lett.*, 2(4):321–325, 1989. DOI: 10.1016/0893-9659(89)90079-7.

[But08] J. C. BUTCHER. *Numerical methods for ordinary differential equations*. John Wiley & Sons Ltd, second edition, 2008. DOI: 10.1002/9780470753767.

[But64a] J. C. BUTCHER. Implicit Runge-Kutta processes. *Math. Comput.*, 18(85):50–64, 1964. DOI: 10.1090/S0025-5718-1964-0159424-9.

[But64b] J. C. BUTCHER. Integration processes based on Radau quadrature formulas. *Math. Comput.*, 18(86):233–234, 1964. DOI: 10.1090/S0025-5718-1964-0165693-1.

[But65] J. C. BUTCHER. On the attainable order of Runge-Kutta methods. *Math. Comput.*, 19(91):408–417, 1965. DOI: 10.1090/S0025-5718-1965-0179943-X.

[But72] J. C. BUTCHER. An algebraic theory of integration methods. *Math. Comput.*, 26(117):79–106, 1972. DOI: 10.1090/S0025-5718-1972-0305608-0.

[But76] J. C. BUTCHER. On the implementation of implicit Runge-Kutta methods. *BIT*, 16(3):237–240, 1976. DOI: 10.1007/BF01932265.

[But95] J. C. BUTCHER. On fifth order Runge–Kutta methods. *BIT*, 35(2):202–209, 1995. DOI: 10.1007/BF01737162.

[But96] J. C. BUTCHER. A history of Runge-Kutta methods. *Appl. Numer. Math.*, 20(3):247–260, 1996. DOI: 10.1016/0168-9274(95)00108-5.

[BWZ71] D. BARTON, I. M. WILLERS, and R. V. M. ZAHAR. The automatic solution of systems of ordinary differential equations by the method of Taylor series. *Comput. J.*, 14(3):243–248, 1971. DOI: 10.1093/comjnl/14.3.243.

[Ces61] F. CESCHINO. Modification de la longueur du pas dans l'intégration numérique par les méthodes à pas liés. *Chiffres*, 2 :101-106, 1961.

[CGH+96] R. M. CORLESS, G. H. GONNET, D. E. G. HARE, D. J. JEFFREY, and D. E. KNUTH. On the Lambert *W* function. *Adv. Comput. Math.*, 5(1):329–359, 1996. DOI: 10.1007/BF02124750.

[CH52] C. F. CURTISS and J. O. HIRSCHFELDER. Integration of stiff equations. *Proc. Nat. Acad. Sci. U.S.A.*, 38(3):235–243, 1952. DOI: 10.1073/pnas.38.3.235.

[Chi71] F. H. CHIPMAN. *A*-stable Runge–Kutta processes. *BIT*, 11(4):384–388, 1971. DOI: 10.1007/BF01939406.

[CK90] J. R. CASH and A. H KARP. A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides. *ACM Trans. Math. Software*, 16(3):201–222, 1990. DOI: 10.1145/79505.79507.

[CK99] A. CONNES and D. KREIMER. Lessons from quantum field theory: Hopf algebras and spacetime geometries. *Lett. Math. Phys.*, 48(1):85–96, 1999. DOI: 10.1023/A:1007523409317.

[CM75] D. M. CREEDON and J. J. H. MILLER. The stability properties of *q*-step backward difference schemes. *BIT*, 15(3):244–249, 1975. DOI: 10.1007/BF01933656.

[Cry72] C. W. CRYER. On the instability of high order backward-difference multistep methods. *BIT*, 12(1):17–25, 1972. DOI: 10.1007/BF01932670.

[Dah56] G. DAHLQUIST. Convergence and stability in the numerical integration of ordinary differential equations. *Math. Scand.*, 4:33–53, 1956.

[Dah63] G. G. DAHLQUIST. A special stability problem for linear multistep methods. *BIT*, 3(1):27–43, 1963. DOI: 10.1007/BF01963532.

[DP80] J. R. DORMAND and P. J. PRINCE. A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, 6(1):19–26, 1980. DOI: 10.1016/0771-050X(80)90013-3.

[Ehl69] B. L. EHLE. On Padé approximation to the exponential function and *A*-stable methods for the numerical solution of initial value problems. Technical report CS-RR 2010, Dept. AACS, University of Waterloo, 1969.

[Eng69] R. ENGLAND. Error estimates for Runge-Kutta type solutions to systems of ordinary differential equations. *Comput. J.*, 12(2):166–170, 1969. DOI: 10.1093/comjnl/12.2.166.

[Feh69] E. FEHLBERG. Low-order classical Runge-Kutta formulas with stepsize control and their application to some heat transfer problems. Technical report R-315, NASA, 1969.

[Goo67] R. M. GOODWIN. A growth cycle. In C. H. FEINSTEIN, editor, *Socialism, capitalism and economic growth. Essays presented to Maurice Dobb*, pages 54–58. Cambridge University Press, 1967.

[Heu00] K. HEUN. Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. Math. Phys.*, 45:23–38, 1900.

[HNW93] E. HAIRER, S. P. NØRSETT, and G. WANNER. *Solving ordinary differential equations. I Nonstiff problems*, volume 8 of *Springer series in computational mathematics*. Springer, second revised edition, 1993. DOI: 10.1007/978-3-540-78862-1.

[Hol59] C. S. HOLLING. Some characteristics of simple types of predation and parasitism. *Can. Entomol.*, 91(7):385–398, 1959. DOI: 10.4039/Ent9138-7.

[Huť56] A. HUŤA. Une amélioration de la méthode de Runge–Kutta–Nyström pour la résolution numérique des équations différentielles du premier ordre. *Acta Fac. Rerum Natur. Univ. Comenian. Math.*, 1(IV–VI) :201-224, 1956.

[HW83] E. HAIRER and G. WANNER. On the instability of the BDF formulas. *SIAM J. Numer. Anal.*, 20(6):1206–1209, 1983. DOI: 10.1137/0720090.

[HW96] E. HAIRER and G. WANNER. *Solving ordinary differential equations. II Stiff and differential-algebraic problems*, volume 14 of *Springer series in computational mathematics*. Springer, second revised edition, 1996. DOI: 10.1007/978-3-642-05221-7.

[JN82] R. JELTSCH and O. NEVANLINNA. Stability and accuracy of time discretizations for initial value problems. *Numer. Math.*, 40(2):245–296, 1982. DOI: 10.1007/BF01400542.

[Klo71] R. W. KLOPFENSTEIN. Numerical differentiation formulas for stiff systems of ordinary differential equations. *RCA Rev.*, 32:447–462, 1971.

[KM27] W. O. KERMACK and A. G. MCKENDRICK. A contribution to the mathematical theory of epidemics. *Proc. Roy. Soc. London Ser. A*, 115(772):700–721, 1927. DOI: 10.1098/rspa.1927.0118.

[Kro73] F. T. KROGH. Algorithms for changing the step size. *SIAM J. Numer. Anal.*, 10(5):949–965, 1973. DOI: 10.1137/0710081.

[Kun61] J. KUNTZMANN. Neuere Entwicklungen der Methode von Runge und Kutta. *Z. Angew. Math. Mech.*, 41(S1):T28–T31, 1961. DOI: 10.1002/zamm.19610411317.

[Kut01] W. KUTTA. Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. *Z. Math. Phys.*, 46:435–453, 1901.

[Lam91] J. D. LAMBERT. *Numerical methods for ordinary differential systems: the initial value problem*. John Wiley & Sons, 1991.

[Las88] F. M. LASAGNI. Canonical Runge-Kutta methods. *Z. Angew. Math. Phys.*, 39(6):952–953, 1988. DOI: 10.1007/BF00945133.

[Lin94] E. LINDELÖF. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *C. R. Acad. Sci. Paris*, 118 :454-457, 1894.

[Lip68] R. LIPSCHITZ. Disamina della possibilità d'integrare completamente un dato sistema di equazioni differenziali ordinarie. *Ann. Mat. Pura Appl. (2)*, 2(1):288–302, 1868. DOI: 10.1007/BF02419619.

[Lor63] E. N. LORENZ. Deterministic nonperiodic flow. *J. Atmospheric Sci.*, 20(2):130–141, 1963. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[Lot10] A. J. LOTKA. Contribution to the theory of periodic reaction. *J. Phys. Chem.*, 14(3):271–274, 1910. DOI: 10.1021/j150111a004.

[Lot20] A. J. LOTKA. Analytical note on certain rhythmic relations in organic systems. *Proc. Nat. Acad. Sci. U.S.A.*, 6(7):410–415, 1920. DOI: 10.1073/pnas.6.7.410.

[MC53] A. R. MITCHELL and J. W. CRAGGS. Stability of difference relations in the solution of ordinary differential equations. *Math. Comput.*, 7(42):127–129, 1953. DOI: 10.1090/S0025-5718-1953-0054350-0.

[Mer57] R. H. MERSON. An operational method for the study of integration processes. In *Proc. Symp. Data Processing*, pages 110–125, 1957.

[Mil26] W. E. MILNE. Numerical integration of ordinary differential equations. *Amer. Math. Monthly*, 33(9):455–460, 1926. DOI: 10.2307/2299609.

[Mol08] C. B. MOLER. *Numerical Computing with* MATLAB. SIAM, revised edition, 2008. DOI: 10.1137/1.9780898717952.

[Mou26] F. R. MOULTON. *New methods in exterior ballistics*. The University of Chicago Press, 1926.

[MTN08] P. J. MOHR, B. N. TAYLOR, and D. B. NEWELL. CODATA recommended values of the fundamental physical constants: 2006. *Rev. Mod. Phys.*, 80(2):633–730, 2008. DOI: 10.1103/RevModPhys.80.633.

[Mur02] J. D. MURRAY. *Mathematical biology. I. An introduction*, volume 17 of *Interdisciplinary applied mathematics*. Springer, third edition, 2002. DOI: 10.1007/b98868.

[Nor62] A. NORDSIECK. On numerical integration of ordinary differential equations. *Math. Comput.*, 16(77):22–49, 1962. DOI: 10.1090/S0025-5718-1962-0136519-5.

[Nør76] S. P. NØRSETT. Runge-Kutta methods with a multiple real eigenvalue only. *BIT*, 16(4):388–393, 1976. DOI: 10.1007/BF01932722.

[Nys25] E. J. NYSTRÖM. Über die numerische Integration von Differentialgleichungen. *Acta Soc. Sci. Fennicae*, 50(13):1–55, 1925.

[Oli75] J. OLIVER. A curiosity of low-order explicit Runge–Kutta methods. *Math. Comput.*, 29(132):1032–1036, 1975. DOI: 10.1090/S0025-5718-1975-0391514-5.

[Pea86] G. PEANO. Sull'integrabilità delle equazioni differenziali di primo ordine. *Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur.*, 21:677–685, 1886.

[Pea90] G. PEANO. Démonstration de l'intégrabilité des équations différentielles ordinaires. *Math. Ann.*, 37(2) :182-228, 1890. DOI : 10.1007/BF01200235.

[Phi58] A. W. PHILLIPS. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957. *Economica*, 25(100):283–299, 1958. DOI: 10.1111/j.1468-0335.1958.tb00003.x.

[Pic93] É. PICARD. Sur l'application des méthodes d'approximations successives à l'étude de certaines équations différentielles ordinaires. *J. Math. Pures Appl. (4)*, 9 :217-272, 1893.

[Ral62] A. RALSTON. Runge-Kutta methods with minimum error bounds. *Math. Comput.*, 16(80):431–437, 1962. DOI: 10.1090/S0025-5718-1962-0150954-0.

[RG27] L. F. RICHARDSON and J. A. GAUNT. The deferred approach to the limit. Part I. Single lattice. Part II. Interpenetrating lattices. *Philos. Trans. Roy. Soc. London Ser. A*, 226(636-646):299–361, 1927. DOI: 10.1098/rsta.1927.0008.

[Rob66] H. H. ROBERTSON. The solution of a set of reaction rate equations. In J. WALSH, editor, *Numerical analysis: an introduction*, pages 178–182. Academic Press, 1966.

[Ros63] H. H. ROSENBROCK. Some general implicit processes for the numerical solution of differential equations. *Comput. J.*, 5(4):329–330, 1963. DOI: 10.1093/comjnl/5.4.329.

[Run95] C. RUNGE. Über die numerische Auflösung von Differentialgleichungen. *Math. Ann.*, 46(2):167–178, 1895. DOI: 10.1007/BF01446807.

[San88] J. M. SANZ-SERNA. Runge-Kutta schemes for Hamiltonian systems. *BIT*, 28(4):877–883, 1988. DOI: 10.1007/BF01954907.

[SR97] L. F. SHAMPINE and M. W. REICHELT. The MATLAB ODE suite. *SIAM J. Sci. Comput.*, 18(1):1–22, 1997. DOI: 10.1137/S1064827594276424.

[Stø07] C. STØRMER. Sur les trajectoires des corpuscules électrisés dans l'espace sous l'action du magnétisme terrestre avec application aux aurores boréales. *Arch. Sci. Phys. Nat. Genève (4)*, 24 :5-18, 113-158, 221-247, 1907.

[Sun09] K. F. SUNDMAN. Nouvelles recherches sur le problème des trois corps. *Acta. Soc. Sci. Fennicae*, 35(9) :3-27, 1909.

[Tou98] D. TOURNÈS. L'origine des méthodes multipas pour l'intégration numérique des équations différentielles ordinaires. *Rev. Histoire Math.*, 4(1) :5-72, 1998.

[vdPol26] B. van der POL. On "relaxation-oscillations". *Philos. Mag.*, 2(11):978–992, 1926. DOI: 10.1080/14786442608564127.

[Ver38] P-F. VERHULST. Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Math. Phys.*, 10 :113-121, 1838.

[Ver67] L. VERLET. Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159(1):98–103, 1967. DOI: 10.1103/PhysRev.159.98.

[Vol26] V. VOLTERRA. Variazioni e fluttuazioni del numero d’individui in specie animali conviventi. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 2:31–113, 1926.

[Wan91] Q.-D. WANG. The global solution of the n-body problem. *Celestial Mech. Dynam. Astronom.*, 50(1):73–88, 1991. DOI: 10.1007/BF00048987.

[WHN78] G. WANNER, E. HAIRER, and S. P. NØRSETT. Order stars and stability theorems. *BIT*, 18(4):475–489, 1978. DOI: 10.1007/BF01932026.

[Wil67] O. B. WILDLUND. A note on unconditionally stable linear multistep methods. *BIT*, 7(1):65–70, 1967. DOI: 10.1007/BF01934126.

[Wri70] K. WRIGHT. Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties. *BIT*, 10(2):217–227, 1970. DOI: 10.1007/BF01936868.

[Zed90] H. ZEDAN. Avoiding the exactness of the Jacobian matrix in Rosenbrock formulae. *Comput. Math. Appl.*, 19(2):83–89, 1990. DOI: 10.1016/0898-1221(90)90011-8.

Chapitre 9

Résolution numérique des équations différentielles stochastiques

Bien qu'étant largement utilisés et étudiés, les modèles mathématiques dits *déterministes* basés sur des équations différentielles ordinaires ne rendent pas toujours compte de la réalité des phénomènes considérés de manière satisfaisante. En effet, dans de nombreux domaines d'application, les mesures expérimentales ne sont que rarement conformes aux solutions prédites, des effets fluctuants de l'environnement venant perturber l'évolution, quelque peu idéalisée par la représentation mathématique, du phénomène considéré. Dans de tels cas de figure, on peut toutefois essayer d'améliorer le modèle en le modifiant par l'introduction de processus aléatoires dans le système différentiel.

Un exemple historique de cette approche est celui de l'*équation de Langevin* [1] [Lan08], issue de l'écriture du principe fondamental de la dynamique pour la modélisation du mouvement d'une particule en suspension dans un fluide [2] en équilibre thermodynamique,

$$m\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -6\pi\mu r\,\boldsymbol{v} + \boldsymbol{\eta}, \tag{9.1}$$

dans laquelle le scalaire m désigne la masse de la particule considérée et le vecteur $\boldsymbol{v}$ est son champ de vitesse à un instant donné. Le terme $-6\pi\mu r\,\boldsymbol{v}$, avec μ la *viscosité dynamique* du fluide, dans le membre de droite de l'équation représente une force de frottements visqueux (en accord avec la *loi de Stokes* [3]), tandis que la *force complémentaire* [4] $\boldsymbol{\eta}$, résultant des chocs aléatoires incessants, causés par l'agitation thermique, des molécules constituant le fluide contre la particule, est à l'origine de ce qu'on appelle le *mouvement brownien*.

De manière plus générale, la prise en compte d'une composante aléatoire dans la modélisation d'un phénomène mène de manière courante à la résolution d'équations différentielles, qualifiées de *stochastiques*, de la forme

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t,\omega) = f(t,X(t,\omega)) + g(t,X(t,\omega))\,\eta(t,\omega), \tag{9.2}$$

dans lesquelles l'inconnue X et la quantité η, assimilée à un *« bruit »*, sont des *processus aléatoires*.

Dans le présent chapitre, nous nous intéressons à la résolution approchée d'équations différentielles stochastiques par des méthodes de discrétisation similaires à celles étudiées dans le chapitre 8 et utilisant des outils numériques permettant la simulation du hasard. Avant d'entrer dans le vif du sujet, nous allons dans un premier temps nous attacher à préciser la notion même de solution d'une équation comme (9.2) en lui donnant un sens mathématique et fournir quelques exemples concrets de problèmes dans lesquels des équations différentielles stochastiques interviennent.

1. Paul Langevin (23 janvier 1872 - 19 décembre 1946) était un physicien français, connu notamment pour sa théorie du magnétisme et l'organisation des Congrès Solvay.

2. On fait l'hypothèse que la taille de la particule, supposée sphérique de rayon r, est grande devant celle des molécules du fluide.

3. George Gabriel Stokes (13 août 1819 - 1er février 1903) était un mathématicien et physicien britannique. Il fit d'importantes contributions à la mécanique des fluides, à l'optique et à la physique mathématique.

4. Langevin écrit à propos de cette force *« qu'elle est indifféremment positive et négative, et sa grandeur est telle qu'elle maintient l'agitation de la particule que, sans elle, la résistance visqueuse finirait par arrêter »*.

9.1 Rappels de calcul stochastique

Cette section est consacrée au rappel de quelques notions de base de calcul stochastique qui permettront de rigoureusement introduire les équations différentielles stochastiques. Dans toute la suite, le triplet $(\Omega, \mathscr{A}, P)$ désigne un *espace de probabilité* [5].

9.1.1 Processus stochastiques en temps continu

Du point de vue de la modélisation, on peut assimiler une suite de *variables aléatoires réelles* [6] à des données fournies par une série d'observations effectuées au cours du temps. On concrétise mathématiquement cette notion avec la définition suivante.

Définition 9.1 (processus stochastique) *Étant donné un espace de probabilité $(\Omega, \mathscr{A}, P)$, un espace mesurable $(S, \mathscr{S})$ et un ensemble ordonné I, un* ***processus stochastique*** *(ou* ***aléatoire****) à valeurs dans S est une famille de variables aléatoires, définies de $(\Omega, \mathscr{A})$ dans $(S, \mathscr{B})$, indexée par I.*

L'espace S est appelé l'*espace des états* du processus. Un processus stochastique $X = \{X(t,\cdot),\ t \in I\}$ est dit *en temps discret* si l'ensemble I est dénombrable et *en temps continu* si c'est un intervalle de $\mathbb{R}$, la variable t étant généralement interprétée comme le temps (et prenant typiquement ses valeurs dans $[0, +\infty[$). Dans ce second cas, pour tout événement ω de Ω, on appelle *trajectoire* (*sample path* en anglais) du processus la fonction $t \mapsto X(t,\omega)$ définie sur I et à valeurs dans S. Dans toute la suite, sauf mention contraire, nous n'allons considérer que des processus stochastiques à temps continu, en posant $I = [0, +\infty[$, *réels*, c'est-à-dire pour lesquels l'espace des états est $\mathbb{R}$ (que l'on équipe de la tribu $\mathscr{B}(\mathbb{R})$). De plus, pour davantage de lisibilité, on notera le plus souvent $X(t)$ la variable aléatoire $X(t,\cdot)$ associée à tout élément t de I par un processus stochastique X.

Étant donné deux processus stochastiques définis sur un même espace, on peut entendre en différents sens la relation d'égalité entre ces processus. Ceci est l'objet des définitions suivantes.

Définition 9.2 (version d'un processus stochastique) *On dit qu'un processus stochastique Y défini sur $(\Omega, \mathscr{A}, P)$ est une* version *ou une* modification *d'un processus X défini sur le même espace si*

$$P(\{\omega \in \Omega \,|\, X(t,\omega) = Y(t,\omega)\}) = 1,\ \forall t \in I.$$

Définition 9.3 (égalité des lois fini-dimensionnelles de deux processus stochastiques) *On dit que deux processus stochastiques X et Y définis sur $(\Omega, \mathscr{A}, P)$* ***ont mêmes lois fini-dimensionnelles*** *si pour tout entier naturel non nul k et tout k-uplet $(t_1, t_2, \ldots, t_k)$ d'éléments de I, on a égalité des lois des vecteurs aléatoires $(X_{t_1}, \ldots, X_{t_k})$ et $(Y_{t_1}, \ldots, Y_{t_k})$.*

Définition 9.4 (processus stochastiques indistinguables) *Deux processus stochastiques X et Y définis sur $(\Omega, \mathscr{A}, P)$ sont dits* ***indistinguables*** *si $P(\{\omega \in \Omega \,|\, X(t,\omega) = Y(t,\omega),\ \forall t \in I\}) = 1$.*

Nous concluons cette sous-section en mentionnant quelques propriétés particulières que peut posséder un processus stochastique.

Définition 9.5 (processus stochastique stationnaire) *Un processus stochastique X est dit* ***stationnaire*** *si, pour tout entier naturel non nul k, tout k-uplet $(t_1, t_2, \ldots, t_k)$ de points de I et tout élément s de I, la loi du vecteur aléatoire $(X(t_1), \ldots, X(t_k))$ est celle du vecteur $(X(t_1+s), \ldots, X(t_k+s))$.*

Définition 9.6 (processus stochastique à accroissements indépendants) *Un processus stochastique X est dit* ***à accroissements indépendants*** *si, pour tout entier naturel non nul k et tout k-uplet $(t_1, t_2, \ldots, t_k)$ de points de I tels que $t_1 \leq t_2 \leq \cdots \leq t_k$, les variables aléatoires $X(t_1)$, $X(t_2) - X(t_1)$,..., $X(t_k) - X(t_{k-1})$ sont indépendantes.*

5. On rappelle que Ω désigne un ensemble non vide appelé *univers* (*sample space* en anglais), que $\mathscr{A}$ est une *tribu* ou σ-*algèbre* sur Ω (c'est-à-dire un ensemble non vide de parties de Ω, stable par passage au complémentaire et par union dénombrable) dont les éléments sont des *événements* et que P est une *mesure de probabilité* sur l'*espace mesurable* $(\Omega, \mathscr{A})$ (c'est-à-dire une application définie sur $\mathscr{A}$ à valeurs dans $[0,1]$, telle que la mesure de toute réunion dénombrable d'éléments de $\mathscr{A}$ deux à deux disjoints soit égale à la somme des mesures de ces éléments et telle que $P(\Omega) = 1$).

6. On rappelle qu'une fonction Z de Ω dans $\mathbb{R}$ est une variable aléatoire réelle si c'est une application *mesurable* de $(\Omega, \mathscr{A})$ dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, où $\mathscr{B}(\mathbb{R})$ est la *tribu borélienne de* $\mathbb{R}$. Cette condition de mesurabilité assure l'existence d'une mesure de probabilité P_Z sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, telle que

$$P_Z(B) = P(\{\omega \in \Omega \,|\, Z(\omega) \in B\}),\ \forall B \in \mathscr{B}(\mathbb{R}),$$

et qu'on appelle la *loi de probabilité* de la variable aléatoire. Enfin, la fonction F_Z de $\mathbb{R}$ dans $\mathbb{R}$, définie par $F_Z(z) = P_Z(]-\infty, z[) = P(\{\omega \in \Omega \,|\, Z(\omega) < z\})$, est la *fonction de répartition* de la variable aléatoire.

Définition 9.7 (processus gaussien) *Un processus stochastique réel X est dit **gaussien** si, pour tout entier naturel non nul k et tout k-uplet $(t_1, t_2, \ldots, t_k)$ de points de I, le vecteur aléatoire $(X(t_1), \ldots, X(t_k))$ est un vecteur gaussien*[7].

9.1.2 Filtrations et martingales *

Nous introduisons à présent une notion utilisée pour rendre compte de l'accroissement de la quantité d'information disponible au cours du temps, notamment celle fournie par l'observation d'un processus stochastique donné.

Définition 9.8 (filtration) *Une **filtration** sur un espace mesurable $(\Omega, \mathscr{A})$ est une famille croissante (au sens de l'inclusion) de sous-tribus de $\mathscr{A}$.*

On dit qu'une filtration $\{\mathscr{F}_t,\ t \geq 0\}$ est *continue à droite* (resp. *à gauche*) si A VERIFIER

$$\mathscr{F}_t = \bigcap_{\varepsilon > 0} \mathscr{F}_{t+\varepsilon},\ t \geq 0 \quad (\text{resp. } \mathscr{F}_t = \sigma\Big(\bigcup_{0 \leq s < t} \mathscr{F}_s\Big),\ t > 0).$$

Cette même filtration est dite *complète* par rapport à une mesure de probabilité P si $\mathscr{F}_0$ contient l'ensemble des parties de $\mathscr{A}$ *négligeables*, c'est-à-dire de mesure nulle, pour P.

On appelle *espace de probabilité filtré*, et l'on note $(\Omega, \mathscr{A}, \{\mathscr{F}_t,\ t \geq 0\}, P)$, l'espace de probabilité $(\Omega, \mathscr{A}, P)$ muni de la filtration compatible $\{\mathscr{F}_t,\ t \geq 0\}$.

Le concept de filtration permet de définir une notion de mesurabilité des processus stochastiques essentielle pour la construction de l'intégrale stochastique abordée dans la sous-section 9.1.4.

Définition 9.9 (processus stochastique adapté à une filtration) *On dit qu'un processus stochastique $\{X(t),\ t \geq 0\}$ est **adapté** à une filtration $\{\mathscr{F}_t,\ t \geq 0\}$ si, pour tout $t \geq 0$, la variable aléatoire $X(t)$ est mesurable par rapport à la tribu $\mathscr{F}_t$.*

Tout processus stochastique sur $(\Omega, \mathscr{A}, P)$ engendre une filtration, qui est la plus petite filtration rendant ce processus adapté et que l'on peut voir comme la quantité d'information apportée par la connaissance du processus à tout instant.

Définition 9.10 (filtration naturelle) *La **filtration naturelle** d'un processus stochastique $X = \{X(t),\ t \geq 0\}$, notée $\mathscr{F}^X$, est la famille croissante de tribus engendrées par $\{X(s),\ 0 \leq s \leq t\}$, $t \geq 0$, c'est-à-dire*

$$\mathscr{F}^X = \left\{\mathscr{F}_t^X = \sigma\left(\{X(s),\ 0 \leq s \leq t\}\right),\ t \geq 0\right\}.$$

REVOIR, UTILE ? L'augmentation/La complétion de la filtration naturelle permet d'affirmer que si deux variables aléatoires X et Y sont égales presque sûrement par rapport à la mesure P et que Y est mesurable par rapport à la tribu $\mathscr{F}_t$ alors X est également mesurable par rapport à $\mathscr{F}_t$.

A DEPLACER ? INTRODUIRE : espaces L^p, moments (espérance, etc.)

REPRENDRE Rappelons à présent la notion d'espérance conditionnelle, que l'on peut interpréter comme la meilleure prévision possible d'une variable aléatoire compte tenu de l'information à disposition à un moment donné.

Définition 9.11 (espérance conditionnelle) *Soit Z une variable aléatoire et $\mathscr{B}$ une sous-tribu de $\mathscr{A}$. On appelle **espérance conditionnelle de Z par rapport à**, ou **sachant**, $\mathscr{B}$ l'unique variable aléatoire, notée $E(Z|\mathscr{B})$, vérifiant*

$$E(E(Z|\mathscr{B})\,\mathbf{1}_B) = E(Z\,\mathbf{1}_B),\ \forall B \in \mathscr{B},$$

où $\mathbf{1}_B$ désigne la fonction caractéristique/indicatrice[8] *du sous-ensemble B.*

7. On rappelle qu'un vecteur aléatoire est *gaussien* si toute combinaison linéaire de ses composantes est une variable aléatoire suivant une loi normale, c'est-à-dire ayant une densité de probabilité égale à $\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, avec μ l'espérance et $\sigma > 0$ l'écart type de la loi.

8. On rappelle que $\mathbf{1}_B(\omega) = \begin{cases} 1 & \text{si } \omega \in B \\ 0 & \text{si } \omega \notin B \end{cases}$.

justification existence ? (projection hilbertienne pour le cas L^2, Radon-Nikodym pour L^1)
Nous pouvons à présent rappeler la notion de *martingale*.

Définitions 9.12 (martingale, sous-martingale, sur-martingale) *Soit $(\Omega, \mathscr{A}, \{\mathscr{F}_t,\ t \geq 0\}, P)$ un espace de probabilité filtré et $\{X(t),\ t \geq 0\}$ un processus stochastique adapté à la filtration $\{\mathscr{F}_t,\ t \geq 0\}$. On dit que X est une* ***martingale*** *(resp.* ***sous-martingale****, resp.* ***sur-martingale****) par rapport à $\{\mathscr{F}_t,\ t \geq 0\}$ si*

- $E(|X(t)|) < +\infty$*, pour tout* $t \geq 0$*,*
- $E(X(t)|\mathscr{F}_s) = X(s)$ *(resp.* $E(X(t)|\mathscr{F}_s) \geq X(s)$*, resp.* $E(X(t)|\mathscr{F}_s) \leq X(s)$*), pour tout* $0 \leq s \leq t$*.*

Il découle de sa définition qu'une martingale est un processus à espérance constante. Elle modélise ainsi est un *jeu équitable*, c'est-à-dire un jeu pour lequel le gain que l'on peut espérer faire en tout temps ultérieur est égal à la somme gagnée au moment présent. De la même façon, une sous-martingale est un processus à espérance croissante (un jeu favorable) et une sur-martingale un processus à espérance décroissante (un jeu défavorable).

9.1.3 Processus de Wiener et mouvement brownien *

Les *processus de Wiener*[9] forment une classe particulièrement importante de processus stochastiques en temps continu. Ils sont une réprésentation mathématique du phénomène physique de mouvement brownien et interviennent dans de nombreux modèles probabilistes utilisés, par exemple, en finance.

Définition 9.13 (processus de Wiener standard) *Un* ***processus de Wiener standard*** *est un processus stochastique réel en temps continu $\{W(t),\ t \geq 0\}$ issu*[10] *de 0, dont les accroissements sont indépendants et tel que, pour $0 \leq s \leq t$, la variable aléatoire $W(t) - W(s)$ suit une loi normale de moyenne nulle et de variance égale à $t - s$.*
AJOUTER continuité des trajectoires

REPRENDRE L'indépendance des accroissements d'un processus de Wiener standard se traduit encore par le fait que, pour $0 \leq s \leq t$, la variable aléatoire $W(t) - W(s)$ est indépendante de la tribu $\sigma(\{W(r),\ 0 \leq r \leq s\})$. La dernière condition de cette définition implique la stationnarité des accroissements du processus, au sens où, pour $0 \leq s \leq t$, la loi de la variable aléatoire $W(t) - W(s)$ est celle de la variable $W(t-s) - W(0)$. Il en résulte qu'un processus de Wiener standard est un processus gaussien centré ($E(W(t)) = 0$) tel que[11] $E(W(s)W(t)) = \min(s,t)$ (autocovariance).
REPRENDRE
Notons qu'il existe diverses façons de prouver rigoureusement l'existence d'un processus de Wiener. On peut tout d'abord choisir de prescrire des lois fini-dimensionnelles choisies de manière à imposer les propriétés d'indépendance et de normalité des accroissements et de stationnarité du processus EXPLICITER ?. Celles-ci s'avérant alors *consistantes*[12], le *théorème d'extension de Kolmogorov* garantit qu'il existe un processus stochastique les vérifiant. Si le processus ainsi obtenu n'est pas unique, il en existe une version dont les trajectoires sont presque sûrement continues en vertu du *critère*[13] *de Kolmogorov–Chentsov* [Che56]. Une deuxième approche possible est de se rappeler qu'un processus de Wiener est un processus gaussien et d'exploiter le fait que certains *espaces vectoriels*

9. Norbert Wiener (26 novembre 1894 - 18 mars 1964) était un mathématicien américain, connu comme le fondateur de la cybernétique. Il fut un pionnier dans l'étude des processus stochastiques et du bruit, contribuant ainsi par ses travaux à l'ingénierie électronique, aux télécommunications et aux systèmes de commande.

10. Dire qu'un processus $\{W(t),\ t \geq 0\}$ est *issu du point x* signifie que $P(\{W(0) = x\}) = 1$.

11. En prenant par exemple pour $0 \leq s < t$, il vient en effet

$$E(W(s)W(t)) = E(W(s)(W(s) + W(t) - W(s))) = E(W(s)^2) + E(W(s)(W(t) - W(s)))$$
$$= s + E(W(s))E(W(t) - W(s)) = s,$$

l'avant-dernière égalité découlant de l'indépendance des accroissements du processus.

12. On doit vérifier que : for all permutations π of $\{1, \ldots, k\}$ and measurable sets $F_i \subseteq \mathbb{R}$,

$$\nu_{t_{\pi(1)} \ldots t_{\pi(k)}}\left(F_{\pi(1)} \times \cdots \times F_{\pi(k)}\right) = \nu_{t_1 \ldots t_k}(F_1 \times \cdots \times F_k);$$

and for all measurable sets $F_i \subseteq \mathbb{R}^n, m \in \mathbb{N}$

$$\nu_{t_1 \ldots t_k}(F_1 \times \cdots \times F_k) = \nu_{t_1 \ldots t_k t_{k+1}, \ldots, t_{k+m}}(F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n).$$

13. REPRENDRE Ce résultat s'énonce de la manière suivante : *un processus stochastique $X = \{X(t)\}_{t \geq 0}$ tel que, pour tout $T > 0$, il existe des constantes α, β et C strictement positives telles que*

$$E(|X(t) - X(s)|^\alpha) \leq C\ |t - s|^{1+\beta},\ 0 \leq s, t \leq T,$$

gaussiens sont des espaces de Hilbert. (base hilbertienne de $L^2([0,1])$, *fonctions de Haar* [14] et approximation par fonctions de Schauder, qui sont les primitives des fonctions de Haar (convergence uniforme en temps p.s. vers une fonction continue). Cette construction est proche celle originelle de Wiener [Wie23] et due à Lévy et Ciesielski [Cie61]. On peut encore obtenir le processus de Wiener comme une limite, en un sens faible, de marches aléatoires normalisées sur tout intervalle borné. Ce résultat, qui porte le nom de *principe d'invariance de Donsker* [Don52], s'inspire de l'observation que, pour toute famille $\{\xi_i\}_{i\in\mathbb{N}}$ de variables aléatoires indépendantes, de même loi, centrées et de variance égale à $\sigma^2 > 0$, la suite des sommes partielles telle que

$$S_0 = 0,\ S_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} \xi_i,\ n \geq 1,$$

converge, d'après le *théorème de la limite centrale*, vers une variable aléatoire de loi normale centrée réduite. Pour le démontrer, on introduit la suite $(W^{(n)})_{n\in\mathbb{N}^*}$ de processus stochastiques à trajectoires continues définis par

$$W^{(n)}(t) = \frac{1}{\sigma\sqrt{n}} \left(\sum_{i=1}^{\lfloor nt \rfloor} \xi_i + (nt - \lfloor nt \rfloor)\, \xi_{\lfloor nt \rfloor + 1} \right),\ \forall t \in [0, +\infty[,\ \forall n \in \mathbb{N}^*,$$

dont les lois fini-dimensionnelles convergent, lorsque l'entier n tend vers l'infini, vers celles d'un processus possédant toutes les propriétés du processus de Wiener. Cette famille de processus étant par ailleurs tendue [15], on en déduit la convergence en loi de la suite vers le processus de Wiener.
4 - représentation par un développement en série analogue à la représentation en série de Fourier des fonctions (*développement de Karhunen* [16]*–Loève* [17]) ?, les coefficients du développement sont des variables aléatoires et les fonctions des fonctions trigonométriques :

$$W(t) = \sqrt{2} \sum_{k=0}^{+\infty} Z_k \frac{\sin\left(\left(k + \frac{1}{2}\right)\pi\, t\right)}{\left(k + \frac{1}{2}\right)\pi}$$

avec $\{Z_k\}_{k\in\mathbb{N}}$ un suite de variables aléatoires gaussiennes indépendantes centrées réduites.

Donnons quelques propriétés élémentaires d'un processus de Wiener.

Théorème 9.14 (propriété de martingale d'un processus de Wiener) *Un processus de Wiener est une martingale par rapport à la filtration naturelle.*

DÉMONSTRATION. VERIFIER et REPRENDRE Pour tout $t > s > 0$, on a

$$E(W(t)|\mathcal{F}_s) = E(W(t) - W(s)|\mathcal{F}_s) + E(W(s)|\mathcal{F}_s) = E(W(t-s)) + W(s) = W(s),$$

en vertu de la propriété différentielle et de la propriété des incréments indépendants. □

Il découle [18] de l'*inégalité de Jensen* [19] que les processus stochastiques $\{W(t)^2,\ t \geq 0\}$ et, pour tout réel σ strictement positif, $\{e^{\sigma W(t)},\ t \geq 0\}$ sont des sous-martingales par rapport à la filtration naturelle complétée de W. Le résultat suivant montre qu'en les modifiant de manière déterministe, on obtient des martingales.

---

possède une modification dont les trajectoires satisfont localement une condition de Hölder dont l'exposant est strictement compris entre 0 *et* $\frac{\beta}{\alpha}$. Pour un processus de Wiener, il est facile de montrer que le critère est vrai avec $\alpha = 4$, $\beta = 1$ et $C = 3$. En effet, ... PREUVE ? ? ? ? Par ailleurs, en introduisant une variable aléatoire N de loi normale centrée et d'écart type égal à 1, on a d'après la définition ref, pour tous $s, t \geq 0$ et $p \geq 1$,

$$E\left(|W(t) - W(s)|^p\right) = E\left(\left|\sqrt{|t-s|}\, N\right|^p\right) = |t-s|^{p/2}\, E\left(|N|^p\right).$$

En faisant alors tendre p vers l'infini, on obtient que l'exposant de la condition de Hölder a pour borne supérieure $1/2$.

14. Alfréd Haar (Haar Alfréd en hongrois, 11 octobre 1885 - 16 mars 1933) était un mathématicien hongrois. TRADUIRE His results are from the fields of mathematical analysis and topological groups, in particular he researched orthogonal systems of functions, singular integrals, analytic functions, partial differential equations, set theory, function approximation and calculus of variations.

15. Cette notion de tension est reliée à celle de compacité relative. Étant donné un espace métrique S, on dit qu'une famille de variables aléatoires dont les lois de probabilité sont définies sur $(S, \mathcal{B}(S))$ est *tendue* si, pour tout $\varepsilon > 0$, il existe un ensemble compact $K \subseteq S$ tel que $P(K) \geq 1 - \varepsilon$, pour toute mesure P associée à la famille.

16. Kari Onni Uolevi Karhunen (12 avril 1915 - 16 septembre 1992) était un mathématicien et statisticien finlandais. COMPLETER

17. Michel Loève (22 janvier 1907 - 17 février 1979) était un mathématicien et statisticien franco-américain. COMPLETER

18. On rappelle en effet que, si ϕ est une fonction convexe sur un intervalle $]a, b[$ et Z une variable aléatoire d'espérance finie à valeurs dans $]a, b[$, alors l'inégalité suivante, dite de Jensen,

$$\phi(E(Z)) \leq E(\phi(Z)),$$

est vraie.

19. Johan Ludwig William Valdemar Jensen (8 mai 1859 - 5 mars 1925) était un mathématicien et ingénieur danois. Il est surtout connu pour l'inégalité et la formule qui portent son nom.

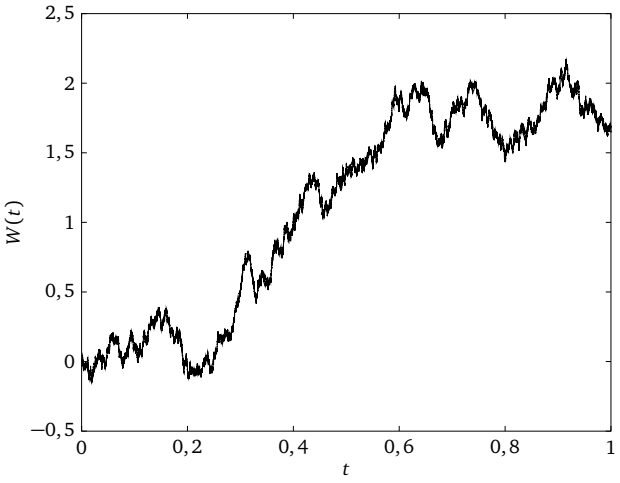


FIGURE 9.1 – Simulation numérique d'une réalisation d'un processus de Wiener standard sur l'intervalle $[0,1]$.

exemples de martingales construites à partir d'un processus de Wiener

Proposition 9.15 *Soit W un processus de Wiener et $\mathscr{F}^W$ sa filtration naturelle complétée. Alors, les processus $\{W(t)^2 - t,\ t \geq 0\}$ et, pour tout réel σ strictement positif, $\{e^{\sigma W(t) - \frac{\sigma^2}{2}t},\ t \geq 0\}$ sont des martingales par rapport à $\mathscr{F}^W$.*

DÉMONSTRATION. REPRENDRE

$$E(W(t)^2|\mathscr{F}_s) = E(W(s)^2 + 2W(s)(W(t)-W(s)) + (W(t)-W(s))^2|\mathscr{F}_s) = W(s)^2 + 0 + (t-s).$$

$$E(e^{\sigma W(t)}|\mathscr{F}_s) = e^{\sigma W(s)}\, E(e^{\sigma (W(t)-W(s))}|\mathscr{F}_s) = e^{\sigma W(s)}\, E(e^{\sigma W(t-s)}|\mathscr{F}_s),$$

et

$$E(e^{\sigma W(t-s)}) = \int_{-\infty}^{+\infty} e^{\sigma x} \frac{e^{-\frac{x^2}{2}(t-s)}}{\sqrt{2\pi(t-s)}}\,\mathrm{d}x = e^{\frac{\sigma^2}{2}(t-s)}$$

par complétion du carré. □

COMPLETER AVEC DEFINITIONS Les trajectoires d'un processus de Wiener ne sont presque sûrement nulle part lipschitziennes (voir [PWZ33]), la borne sup obtenue pour l'exposant de régularité holderienne est donc stricte, et donc non différentiables. Il en découle (?) qu'un processus de Wiener n'est pas à variation bornée (DEFINITION), mais seulement à variation quadratique bornée.

9.1.4 Calcul stochastique d'Itō **

Revenons à présent à l'équation (9.2), à laquelle on adjoint une condition initiale à la date $t = t_0$. On a coutume d'écrire cette équation sous la forme différentielle symbolique

$$\mathrm{d}X(t) = f(t,X(t))\,\mathrm{d}t + g(t,X(t))\,\eta(t)\,\mathrm{d}t,\ t \geq t_0, \tag{9.3}$$

ou encore, comme dans le cas des équations différentielles ordinaires, sous la forme intégrale suivante

$$X(t) = X(t_0) + \int_{t_0}^{t} f(s,X(s))\,\mathrm{d}s + \int_{t_0}^{t} g(s,X(s))\,\eta(s)\,\mathrm{d}s,\ t \geq t_0. \tag{9.4}$$

En considérant que le bruit qu'il modélise est à l'origine d'un mouvement brownien, on peut assimiler [20] le processus stochastique η à un *bruit blanc gaussien*, c'est-à-dire un processus stationnaire et à moyenne nulle, dont la *densité*

20. Les observations du mouvement brownien suggèrent en effet que la force complémentaire $\boldsymbol{\eta}$ apparaissant dans l'équation de Langevin (9.1) est nulle *en moyenne* et que son temps de corrélation est beaucoup plus court (de l'ordre du temps de collision de la particule avec les molécules du fluide) que le temps caractéristique de relaxation du champ de vitesse, donné par $\frac{m}{6\pi\mu r}$, ce qu'on idéalise en postulant que la corrélation est instantanée.

spectrale, c'est-à-dire la transformée de Fourier de son *autocovariance* $C_{\eta\eta}(t) = E(\eta(s)\,\eta(s+t))$, est constante [21], ce qui signifie encore que $C_{\eta\eta}(t) = \Gamma\,\delta(t)$, où Γ est une constante et δ est la *masse de Dirac* [22] (une *distribution* [23] telle que $\delta(t) = 0$ pour tout $t \neq 0$ et telle que

$$\int_{-\infty}^{+\infty} \psi(s)\,\delta(s)\,\mathrm{d}s = \psi(0),$$

pour toute fonction ψ continue en 0).

Il découle alors du fait qu'un processus de Wiener est une représentation mathématique du mouvement brownien qu'une relation entre η et W, obtenue empiriquement à partir de (9.3), est

$$\mathrm{d}W(t) = \eta(t)\,\mathrm{d}t.$$

Un bruit blanc gaussien apparaît par conséquent comme la « dérivée » d'un processus de Wiener. Les trajectoires de ce dernier n'étant cependant nulle part différentiables, on voit que le bruit blanc η n'existe pas en termes d'une dérivée au sens classique, mais *au sens des distributions*, du processus W. En réécrivant alors formellement (9.4) sous la forme

$$X(t) = X(t_0) + \int_{t_0}^{t} f(s,X(s))\,\mathrm{d}s + \int_{t_0}^{t} g(s,X(s))\,\mathrm{d}W(s),\ t \geq t_0,$$

on comprend que toute la difficulté pour définir la solution d'une équation différentielle stochastique se résume à la question délicate de donner un sens mathématique à la seconde des deux intégrales écrites ci-dessus. Pour cela, l'utilisation d'une formule d'intégration par parties n'est pas possible, l'application g n'étant généralement pas différentiable. Le recours à la notion d'*intégrale de Stieltjes* [24] n'est pas non plus envisageable puisque l'on a vu que les trajectoires d'un processus de Wiener ne sont pas à variation bornée, mais seulement à variation *quadratique* bornée. C'est néanmoins cette dernière propriété qui va permettre la construction de l'intégrale en question, définissant ainsi la base du calcul stochastique introduit par Itō [25].

Intégrale stochastique d'Itō

Étant donné un processus de Wiener W, ainsi qu'un processus stochastique X, nous allons chercher à définir l'intégrale stochastique

$$\int_0^t X(s)\,\mathrm{d}W(s),\ 0 \leq t < +\infty. \tag{9.5}$$

Pour cela, l'idée naturelle présidant à la définition de l'*intégrale stochastique d'Itō* [Itô44] est d'utiliser un procédé de construction similaire à celui de l'intégrale de Riemann [26] (voir la section B.4 de l'annexe B). Ceci suppose de la définir tout d'abord pour toute une classe de processus « élémentaires », jouant le rôle d'analogues aléatoires des fonctions en escalier, et d'en étendre la portée en approchant, en un sens convenable, tout intégrand par une suite de processus élémentaires. L'intégrale stochastique est ainsi obtenue par un passage à la limite, entendu la encore en un sens approprié (celui de la convergence en moyenne quadratique des suites de variables aléatoires).

Pour parvenir à une définition raisonnable, il va falloir restreindre la classe des intégrands considérés dans (9.5). Dans toute la suite de cette sous-section, nous faisons l'hypothèse que le processus stochastique X est

21. Cette propriété donne son appellation au processus en raison d'une analogie avec la *lumière blanche*, dans laquelle toutes les ondes électromagnétiques visibles à l'œil nu sont présentes avec la même intensité.

22. Paul Adrien Maurice Dirac (8 août 1902 - 20 octobre 1984) était un physicien et mathématicien anglais dont les contributions à la mécanique et l'électrodynamique quantiques furent fondamentales. Il a notamment formulé l'équation décrivant le comportement des fermions et a prévu l'existence de l'antimatière. Il fût par ailleurs colauréat avec Schrödinger du prix Nobel de physique de 1933 « pour la découverte de formes nouvelles et utiles de la théorie atomique ».

23. On rappelle qu'une distribution sur un ouvert borné Ω de $\mathbb{R}^d$, $d \geq 1$, est une application linéaire continue sur l'ensemble des fonctions à valeurs réelles indéfiniment différentiables et à support compact inclus dans Ω.

24. Thomas Joannes Stieltjes (29 décembre 1856 - 31 décembre 1894) était un mathématicien hollandais. Il travailla notamment sur les formules de quadrature de Gauss, les polynômes orthogonaux ou encore les fractions continues.

25. Kiyoshi Itō (伊藤 清 en japonais, 7 septembre 1915 - 10 novembre 2008) était un mathématicien japonais. Ses apports à la théorie des probabilités et des processus stochastiques, au nombre desquels figurent le calcul stochastique ou la théorie des excursions browniennes dits d'Itō, furent fondamentaux et ont aujourd'hui des applications dans des domaines aussi divers que la physique et l'économie.

26. Georg Friedrich Bernhard Riemann (17 septembre 1826 - 20 juillet 1866) était un mathématicien allemand. Ses contributions à l'analyse et la géométrie différentielle eurent une portée profonde, ouvrant notamment la voie aux géométries non euclidiennes et à la théorie de la relativité générale.

— mesurable par rapport à $\mathscr{B}([0,+\infty[) \times \mathscr{F}^W$, où $\mathscr{F}^W$ est la filtration naturelle du processus de Wiener W,
— adapté à la filtration $\mathscr{F}^W$,
— tel que

$$E\left(\int_0^t X(s)^2\,\mathrm{d}s\right) < +\infty,\ t \geq 0.$$

On dit encore que ce processus est une *fonctionnelle non-anticipative* du processus de Wiener, car la variable aléatoire $X(t)$, $t \geq 0$, ne dépend que de façon *causale* de la trajectoire de W.

REMARQUE sur la complétion de $\mathscr{F}^W$ si condition initiale aléatoire

Introduisons

Définition 9.16 (processus simple) *Un processus stochastique Θ, adapté à une filtration $\{\mathscr{F}_t,\ t \geq 0\}$, est dit **simple** s'il existe une suite réelle strictement croissante $(t_i)_{i\in\mathbb{N}}$, avec $t_0 = 0$ et $\lim_{i\to+\infty} t_i = +\infty$, et une suite $(\theta_i)_{i\in\mathbb{N}}$ de variables aléatoires bornées, pour laquelle la variable aléatoire θ_i est mesurable par rapport à $\mathscr{F}_{t_i}$, $i \in\in \mathbb{N}$, telles que*

$$\Theta(t,\omega) = \theta_0(\omega)\,\mathbf{1}_{\{0\}}(t) + \sum_{i=0}^{+\infty} \theta_i(\omega)\,\mathbf{1}_{]t_i,t_i+1]}(t),\ t \geq 0,\ \omega \in \Omega.$$

Pour tout processus simple Θ adapté à $\mathscr{F}^W$, on peut définir l'intégrale stochastique comme un processus donné par

$$\sum_{i=0}^{+\infty} \theta_i\,(W(\min(t_{i+1},t)) - W(\min(t_i,t))),\ t \geq 0,$$

soit encore, si $t \in]t_m, t_{m+1}]$,

$$\sum_{i=0}^{m} \theta_i\,(W(t_{i+1}) - W(t_i)) + \theta_m\,(W(t) - W(t_m)).$$

Définition 9.17 (intégrale stochastique d'un processus simple) *L'intégrale stochastique entre 0 et t, $t \leq T$, d'un processus simple $(\theta_t)_{0\leq t\leq T}$ est définie par*

$$\int_0^t \theta(s)\,\mathrm{d}W(s) =,$$

où l'entier m est tel que $t \in [t_m, t_{m+1}[$.

L'intégrale d'un processus simple vérifie des propriétés élémentaires suivante : linéarité, continuité par rapport à t p. s., processus adapté,

$$E\left(\int_0^t \theta(s)\,\mathrm{d}W(s)\right) = 0 \text{ si } \int_0^t E(|\theta(s)|)\,\mathrm{d}s < +\infty \tag{9.6}$$

et (*isométrie d'Itō*)

$$E\left(\left(\int_0^t \theta(s)\,\mathrm{d}W(s)\right)^2\right) = \int_0^t E(\theta(s)^2)\,\mathrm{d}s \text{ si } \int_0^t E(\theta(s)^2)\,\mathrm{d}s < +\infty,$$

propriété de martingale...

résultat (admis) de densité des processus simples de carré intégrables dans $\mathscr{L}^2_{\mathscr{F}}(\Omega)$ au sens de la convergence en norme quadratique

Il est donc naturel de définir cette intégrale comme une limite

$$\int_0^T \theta(s)\,\mathrm{d}W(s) = \lim_{\Pi_n\to 0} \sum_{i=0}^{n-1} \theta(t_i)(W(t_{i+1}) - W(t_i)),$$

avec convergence au sens des variables aléatoires dans $\mathscr{L}^2(\Omega)$, en imposant au processus θ d'être dans $\mathscr{L}^2(\Omega,[0,T])$ et d'être également $\mathscr{F}$-adapté afin que θ_{t_i} soit indépendant de $W_{t_{i+1}} - W_{t_i}$.

résultat d'existence de cette limite : intégrale stochastique d'Itō

outils : on utilise le résultat suivant, que l'on admettra.

Lemme 9.18 (« lemme de Borel [27]–Cantelli [28] » [Bor09, Can17]) *Si la somme des probabilités d'une suite d'événements $(A_n)_{n\geq 0}$ est finie, alors la probabilité qu'une infinité d'entre eux se réalisent simultanément est nulle, ce qu'on écrit encore*

$$\sum_{n\geq 0} P(A_n) < +\infty \Rightarrow P\left(\limsup_{n\geq 0} A_n\right) = 0,$$

où $\limsup_{n\geq 0} A_n = \bigcap_{n=0}^{+\infty}\left(\bigcup_{k=n}^{+\infty} A_k\right)$.

+ deux (?) lemmes préliminaires dont un de densité des processus simples dans les processus non anticipatifs

CONCLURE avec propriétés de l'intégrale : linéarité, additivité, $E(\int_0^t X(s)\,\mathrm{d}W(s)) = 0$, l'intégrale est mesurable par rapport à $\mathscr{F}^W(t)$ et c'est une martingale par rapport à $\mathscr{F}_t^W$

Exemple de calcul d'une intégrale stochastique d'Itō. Supposons que l'on cherche à calculer l'intégrale stochastique

$$\int_0^t W(s)\,\mathrm{d}W(s)$$

en appliquant la définition eqref. On a alors

$$\begin{aligned}\lim_{\Pi_n\to 0}\sum_{i=0}^{n-1} W(t_i)(W(t_{i+1})-W(t_i)) &= \frac{1}{2}\lim_{\Pi_n\to 0}\sum_{i=0}^{n-1}\left((W(t_{i+1})+W(t_i))(W(t_{i+1})-W(t_i))-(W(t_{i+1})-W(t_i))^2\right)\\ &= \frac{1}{2}W(t)^2-\frac{1}{2}\lim_{\Pi_n\to 0}\sum_{i=0}^{n-1}(W(t_{i+1})-W(t_i))^2,\end{aligned}$$

dont on déduit que

$$\int_0^t W(s)\,\mathrm{d}W(s) = \frac{1}{2}W(t)^2-\frac{1}{2}t. \tag{9.7}$$

Extensions de la classe d'intégrands : on peut admettre que $X(t)$ dépende de variables aléatoires supplémentaires, indépendantes de $W(t)$. Dans ce cas, il convient d'étendre la filtration $\mathscr{F}^W$ de manière adéquate, W devant rester une martingale par rapport à la filtration étendue.

On peut aussi affaiblir la dernière condition en $P(\int_0^t X(s)^2\,\mathrm{d}s < +\infty) = 1$, mais on n'a plus la propriété de martingale en général

Formule d'Itō

La *formule d'Itō* est un outil de base du calcul stochastique caractérisant l'effet d'un changement de variable sur l'évolution d'un type particulier de processus, appelé *processus d'Itō*.

Définition 9.19 (« processus d'Itō ») *Un **processus d'Itō** est un processus stochastique X de la forme*

$$X(t) = X_0 + \int_0^t \mu(s)\,\mathrm{d}s + \int_0^t \sigma(s)\,\mathrm{d}W(s),\ t\geq 0, \tag{9.8}$$

où X_0 est une variable aléatoire $\mathscr{F}_0$-mesurable, μ et σ sont des processus adaptés à $\mathscr{F}^W$ tels que

$$\forall t\geq 0,\ \int_0^t \left(|\mu(s)|+|\sigma(s)|^2\right)\mathrm{d}s < +\infty \text{ presque sûrement}.$$

premier terme : la valeur initiale, $X_0 = x_0$, pouvant être aléatoire, le deuxième : une composante continue évoluant lentement (dérive), le dernier une composante aléatoire continue à variations rapides (diffusion), c'est une intégrale stochastique d'Itō par rapport au processus de Wiener $W = \{W(t),\ t\geq 0\}$.

27. Félix Édouard Justin Émile Borel (7 janvier 1871 - 3 février 1956) était un mathématicien et homme politique français. Il figure parmi les pionniers de la théorie de la mesure et de son application à la théorie des probabilités.

28. Francesco Paolo Cantelli (décembre 1875 - 21 juillet 1966) était un mathématicien italien, surtout connu pour ses travaux en probabilités et en statistiques.

L'équation intégrale (9.8) est souvent écrite sous la forme différentielle

$$\mathrm{d}X(t) = \mu(t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W(t) \tag{9.9}$$

qui est appelée une *équation différentielle stochastique d'Itō*.

Proposition 9.20 (« formule d'Itō ») *REPRENDRE NOTATIONS ! Pour une fonction φ de deux variables t et x de classe $\mathscr{C}^2$ et un processus d'Itō défini par* (9.8)*, on a*

$$\varphi(t,X(t)) = \varphi(t_0,X(t_0)) + \int_{t_0}^{t} \left(\frac{\partial\varphi}{\partial s}(s,X(s)) + \mu(s)\,\frac{\partial\varphi}{\partial x}(s,X(s)) + \frac{1}{2}\,\sigma(s)^2\,\frac{\partial^2\varphi}{\partial x^2}(s,X(s)) \right) \mathrm{d}s + \int_{t_0}^{t} \sigma(s)\,\frac{\partial\varphi}{\partial x}(s,X(s))\,\mathrm{d}W(s), \tag{9.10}$$

ce qu'on écrit encore sous la forme différentielle suivante

$$\mathrm{d}\varphi(t,X(t)) = \left(\frac{\partial\varphi}{\partial t}(t,X(t)) + \mu(t)\,\frac{\partial\varphi}{\partial x}(t,X(t)) + \frac{1}{2}\,\sigma(t)^2\,\frac{\partial^2\varphi}{\partial x^2}(t,X(t)) \right) \mathrm{d}t + \sigma(t)\,\frac{\partial\varphi}{\partial x}(t,X(t))\,\mathrm{d}W(t). \tag{9.11}$$

DÉMONSTRATION. A ECRIRE □

On voit que la formule d'Itō constitue une généralisation stochastique de la formule de dérivation des fonctions composées. Son utilisation permet de calculer simplement certaines intégrales stochastiques sans revenir à la définition de ces dernières, à la manière d'une formule d'intégration par parties.

Exemples d'applications de la formule d'Itō. Retrouvons l'identité (9.7) en utilisant la formule d'Itō. En faisant le choix $\varphi(t,x) = \frac{1}{2}\,x^2$ et en posant $t_0 = 0$ dans (9.10), il vient

$$\frac{1}{2}\,W(t)^2 = \int_0^t \frac{1}{2}\,\mathrm{d}s + \int_0^t W(s)\,\mathrm{d}W(s),$$

d'où

$$\int_0^t W(s)\,\mathrm{d}W(s) = \frac{1}{2}\,W(t)^2 - \frac{1}{2}\,t.$$

Considérons à présent l'évaluation de l'intégrale stochastique

$$\int_0^t s\,\mathrm{d}W(s).$$

Pour cela, on pose $\varphi(t,x) = t\,x$. On trouve alors

$$t\,W(t) = \int_0^t W(s)\,\mathrm{d}s + \int_0^t s\,\mathrm{d}W(s),$$

d'où

$$\int_0^t s\,\mathrm{d}W(s) = t\,W(t) - \int_0^t W(s)\,\mathrm{d}s.$$

Intégrale stochastique de Stratonovich

lien intégrale d'Itō et somme de Riemann à gauche, non anticipatif, le choix du point milieu donne lieu à l'*intégrale stochastique de Stratonovich* [29] [Str66] : limite des sommes discrètes $\sum X\left(\frac{t_i+t_{i+1}}{2}\right)(W(t_{i+1}) - W(t_i))$.

si on intègre une variable déterministe, les deux intégrales fournissent le même résultat mais ce n'est pas forcément le cas lorsqu'elle est aléatoire

29. Ruslan Leontevich Stratonovich (Руслан Леонтьевич Стратонович en russe, 31 mai 1930 - 13 janvier 1997) était un physicien et mathématicien russe. Il est l'inventeur d'un calcul stochastique servant d'alternative à celui d'Itō et s'appliquant naturellement à la modélisation de phénomènes physiques.

Exemple de calcul d'une intégrale stochastique de Stratonovich. Supposons que l'on cherche à calculer l'intégrale stochastique

$$\int_0^t W(s) \circ \mathrm{d}W(s).$$

Il vient

$$\lim_{\Pi_n \to 0} \sum_{i=0}^{n-1} W\left(\frac{t_i + t_{i+1}}{2}\right)(W(t_{i+1}) - W(t_i)) = \frac{1}{2} W(t)^2 - \lim_{\Pi_n \to 0} \sum_{i=0}^{n-1} \left(W\left(\frac{t_i + t_{i+1}}{2}\right) - \frac{W(t_{i+1}) + W(t_i)}{2} \right) (W(t_{i+1}) - W(t_i)),$$

d'où

$$\int_0^t W(s) \circ \mathrm{d}W(s) = \frac{1}{2} W(t)^2,$$

que l'on ne manquera pas de comparer avec (9.7).

La différence notable avec l'intégrale d'Itō est que la variable aléatoire $X\left(\frac{t_i+t_{i+1}}{2}\right)$ n'est pas indépendante de la somme $W(t_{i+1}) - W(t_i)$ et l'on n'a alors généralement plus l'égalité (9.6), ce qui peut compliquer certains calculs. En revanche, aucune direction du temps n'est privilégiée par ce choix, ce qui fait que la prescription de Stratonovich est largement utilisée en physique statistique car les processus stochastiques qu'elle définit satisfont des équations différentielles stochastiques invariantes par renversement du temps.

MENTIONNER les mérites respectifs des deux intégrales

Il faut noter qu'il est possible de passer de l'une à l'autre des prescriptions en effectuant des changements de variables simples ce qui les rend équivalentes. Le choix du type d'intégrale stochastique reste donc avant tout une question de convenance.

9.1.5 Équations différentielles stochastiques *

Nous considérons à présent des équations différentielles stochastiques de la forme

$$\mathrm{d}X(t) = f(t, X(t))\,\mathrm{d}t + g(t, X(t))\,\mathrm{d}W(t), \tag{9.12}$$

avec f et g des fonctions déterministes mesurables, que nous munissons d'une condition initiale

$$X(0) = Z, \tag{9.13}$$

où Z est soit une constante, auquel cas la filtration $\mathscr{F}$ est uniquement celle engendrée par le processus de Wiener W, soit une variable aléatoire de carré intégrable et indépendante de W, $\mathscr{F}$ désignant alors la filtration engendrée par W et Z.

Définition 9.21 (solution forte d'une équation différentielle stochastique) *Un processus stochastique X est une* ***solution forte*** *de l'équation différentielle stochastique* (9.12) *sur l'intervalle* $[0, T]$*, satisfaisant la condition initiale* (9.13)*, s'il est adapté à la filtration $\mathscr{F}$ sur* $[0, T]$*, satisfait*

$$\int_0^t \left(|f(s, X(s))|\,\mathrm{d}s + |g(s, X(s))|^2 \right) \mathrm{d}s < +\infty \text{ presque sûrement}, \ \forall t \in [0, T],$$

et vérifie

$$X(t) = Z + \int_0^t f(s, X(s))\,\mathrm{d}s + \int_0^t g(s, X(s))\,\mathrm{d}W(s) \text{ presque sûrement } \forall t \in [0, T].$$

Certaines équations différentielles stochastiques peuvent ne pas posséder de solutions fortes, exemple de l'*équation de Tanaka*

$$X(t) = \mathrm{sign}(X(t))\,\mathrm{d}W(t),$$

avec $\mathrm{sign}(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{sinon} \end{cases}$ (REVOIR la définition de la fonction signe...)

EXPLICATION (brève) de la notion de solution faible (le processus de Wiener « porteur » n'est plus spécifié à l'avance, comme dans la notion de solution forte, mais fait partie intégrante de la solution)

unicité faible : les processus solutions ont tous la même loi/ unicité forte (par trajectoire) : l'espace de probabilité et le processus porteur étant fixés, deux solutions X_1 et X_2 de l'équation sont indistinguables ($P(\exists t \in [0,T] \,|\, X_1(t) \neq X_2(t)) = 0$)

Théorème 9.22 (existence et unicité d'une solution forte) *Soit W un processus de Wiener de filtration $\mathscr{F}$, $f : [0,T]\times\mathbb{R}\to\mathbb{R}$ et $g : [0,T]\times\mathbb{R}\to\mathbb{R}$ des fonctions mesurables pour les tribus produit de boréliens pour lesquelles il existe une constante $C>0$ telle que, $\forall t\in[0,T]$, $\forall (x,y)\in\mathbb{R}^2$, on a la condition de Lipschitz*

$$|f(t,x)-f(t,y)|+|g(t,x)-g(t,y)| \leq C\,|x-y|\,,$$

et la condition de restriction sur la croissance

$$|f(t,x)|+|g(t,x)| \leq C\,(1+|x|).$$

Alors, $\forall x\in\mathbb{R}$, $X_0 = x$, l'équation différentielle stochastique (9.12) *possède une unique solution forte, à trajectoires presque sûrement continues et vérifiant*

$$E\left(\int_0^T X(s)^2\,\mathrm{d}s\right) < +\infty.$$

DÉMONSTRATION. A ECRIRE □

explications + exemples d'équation explicitement intégrable ?

9.1.6 Développements d'Itō–Taylor *

Les *développements d'Itō–Taylor* [PW82] font partie des analogues stochastiques des développements de Taylor déterministes (voir par exemple le théorème B.116). Ils permettent, entre autres applications en calcul stochastique, de construire des méthodes numériques d'approximation des solutions d'équations différentielles stochastiques (voir la section 9.3).

Considérons la solution forte d'une équation différentielle stochastique,

$$X(t) = X(t_0) + \int_{t_0}^t f(s,X(s))\,\mathrm{d}s + \int_{t_0}^t g(s,X(s))\,\mathrm{d}W(s),$$

avec f et g des fonctions suffisamment régulières satisfaisant les hypothèses du théorème 9.22. En introduisant les opérateurs

$$L_0 = \frac{\partial}{\partial t} + f\,\frac{\partial}{\partial x} + \frac{1}{2}\,g^2\,\frac{\partial^2}{\partial x^2} \text{ et } L_1 = g\,\frac{\partial}{\partial x},$$

la formule d'Itō (9.10) donne alors, pour toute fonction φ de $[t_0,T]\times\mathbb{R}$ dans $\mathbb{R}$ deux fois continûment différentiable,

$$\varphi(t,X(t)) = \varphi(t_0,X(t_0)) + \int_{t_0}^t (L_0\varphi)(s,X(s))\,\mathrm{d}s + \int_{t_0}^t (L_1\varphi)(s,X(s))\,\mathrm{d}W(s).$$

Si la régularité de la fonction φ le permet, on peut de nouveau appliquer la formule d'Itō, en considérant cette fois les fonctions $L_0\varphi$ et $L_1\varphi$ en place de φ. On trouve ainsi

$$\varphi(t,X(t)) = \varphi(t_0,X(t_0)) + (L_0\varphi)(t_0,X(t_0))\int_{t_0}^t \mathrm{d}s + (L_1\varphi)(t_0,X(t_0))\int_{t_0}^t \mathrm{d}W(s) + R_1, \tag{9.14}$$

où le reste R_1 est égal à

$$R_1 = \int_{t_0}^t \left(\int_{t_0}^s (L_0L_0\varphi)(r,X(r))\,\mathrm{d}r + \int_{t_0}^s (L_1L_0\varphi)(r,X(r))\,\mathrm{d}W(r)\right)\mathrm{d}s$$
$$+ \int_{t_0}^t \left(\int_{t_0}^s (L_0L_1\varphi)(r,X(r))\,\mathrm{d}r + \int_{t_0}^s (L_1L_1\varphi)(r,X(r))\,\mathrm{d}W(r)\right)\mathrm{d}W(s).$$

L'égalité (9.14) est l'exemple le plus simple de développement d'Itō–Taylor non trivial. On peut poursuivre le développement du reste

$$R_1 = (L_0L_0\varphi)(t_0,X(t_0))\int_{t_0}^{t}\left(\int_{t_0}^{s}\mathrm{d}r\right)\mathrm{d}s + (L_1L_0\varphi)(t_0,X(t_0))\int_{t_0}^{t}\left(\int_{t_0}^{s}\mathrm{d}W(r)\right)\mathrm{d}s$$
$$+ (L_0L_1\varphi)(t_0,X(t_0))\int_{t_0}^{t}\left(\int_{t_0}^{s}\mathrm{d}r\right)\mathrm{d}W(s) + (L_1L_1\varphi)(t_0,X(t_0))\int_{t_0}^{t}\left(\int_{t_0}^{s}\mathrm{d}W(r)\right)\mathrm{d}W(s) + R_2,$$

pour obtenir le développement d'Itō–Taylor au second ordre suivant

$$\varphi(t,X(t)) = \varphi(t_0,X(t_0)) + (L_0\varphi)(t_0,X(t_0))\,I_{(0)}[1]_{t_0,t} + (L_1\varphi)(t_0,X(t_0))\,I_{(1)}[1]_{t_0,t}$$
$$+ (L_0L_0\varphi)(t_0,X(t_0))\,I_{(0,0)}[1]_{t_0,t} + (L_1L_0\varphi)(t_0,X(t_0))\,I_{(1,0)}[1]_{t_0,t}$$
$$+ (L_0L_1\varphi)(t_0,X(t_0))\,I_{(0,1)}[1]_{t_0,t} + (L_1L_1\varphi)(t_0,X(t_0))\,I_{(1,1)}[1]_{t_0,t} + R_2, \quad (9.15)$$

dans lequel COMPLETER

$$R_2 = \ldots$$

et l'intégrale multiple $I_\alpha[f(\cdot)]_{\rho,\tau}$, avec α un multi-indice de longueur $l(\alpha)$ supérieure ou égale à un et $0 \leq \rho(\omega) \leq \tau(\omega) \leq T$, est définie récursivement par

$$I_\alpha[f(\cdot)]_{\rho,\tau} = \begin{cases} f(\tau) & \text{si } l(\alpha) = 0 \\ \int_\rho^\tau I_{\alpha-}[f(\cdot)]_{\rho,\tau}\,\mathrm{d}s & \text{si } l(\alpha) \geq 1 \text{ et } \alpha_{l(\alpha)} = 0 \\ \int_\rho^\tau I_{\alpha-}[f(\cdot)]_{\rho,\tau}\,\mathrm{d}W(s) & \text{si } l(\alpha) \geq 1 \text{ et } \alpha_{l(\alpha)} = 1 \end{cases}$$

INTRODUIRE NOTATIONS pour multi-indices, etc.

parler de la généralisation à tout ordre

Notons enfin que l'on construit par un procédé identique des *développements de Stratonovich–Taylor*.

9.2 Exemples d'équations différentielles stochastiques

On a vu en début de chapitre, avec le cas de l'équation de Langevin et le phénomène du mouvement brownien, que les équations différentielles stochastiques permettent de modéliser des systèmes déterministes de grande dimension à un niveau microscopique par des systèmes stochastiques de moindre dimension à un niveau macroscopique. Ces équations servent évidemment à décrire des systèmes qui sont par essence aléatoires, comme en mécanique quantique, mais aussi dont la dynamique présente un comportement extrêmement complexe, de type *chaotique*, et qui sont par conséquent imprédictibles. Elles interviennent encore lorsque l'on ne connaît pas de façon précise le système déterministe étudié, pour combler le manque d'information sur les conditions initiales, les conditions aux limites ou les paramètres du modèle, comme en hydrogéologie par exemple.

9.2.1 Exemple issu de la physique ***

variantes stochastiques des exemples d'edo ?

Voir l'article de review de Chandrasekhar[30] [Cha43]

9.2.2 Modèle de Black–Scholes pour l'évaluation des options en finance

Le *modèle de Black–Scholes* est un modèle mathématique d'évolution des actifs financiers permettant de définir le prix des produits dérivés que sont les options et qui est aujourd'hui, avec ses diverses extensions, couramment utilisé sur les marchés.

30. Subrahmanyan Chandrasekhar (19 octobre 1910 - 21 août 1995) était un astrophysicien et mathématicien américain d'origine indienne, co-lauréat du prix Nobel de physique de 1983 pour ses études théoriques des processus physiques régissant la structure et l'évolution des étoiles.

Options

On rappelle qu'en finance, une *option d'achat européenne* (*european call option* en anglais) est un contrat entre un acheteur et un vendeur donnant le droit, mais pas l'obligation, à l'acheteur d'acquérir un *actif sous-jacent* [31] (*underlying asset* en anglais) à une date (future), dite *date d'échéance* ou *maturité* (*expiration date* ou *maturity date* en anglais), et à un *prix d'exercice* (*exercice price* ou *striking price* en anglais) tous deux fixés à l'avance. Ce contrat a lui-même un prix, appelé *prime* (*premium* en anglais).

Deux questions naturelles se posent au vendeur d'une option : quel doit être le prix de ce contrat (on parle d'*évaluation (du prix) de l'option*, *option pricing* en anglais) et, une fois un tel produit vendu, quelle attitude adopter pour se prémunir contre le risque endossé à la place de l'acheteur (c'est le problème de *couverture du risque*) ? Cette double problématique trouve sa réponse dans l'approche de Black [32], Scholes [33] [BS73] et Merton [34] [Mer73], qui consiste à mettre en œuvre une stratégie d'investissement dynamique supprimant tout risque possible dans n'importe quel scénario de marché.

Hypothèses sur le marché

L'incertitude sur le marché financier entre l'instant initial $t = 0$, correspondant à la vente de l'option, et le temps $t = T$, qui représente sa date d'échéance, est modélisée par un espace de probabilité filtré $(\Omega, \mathscr{A}, \{\mathscr{F}_t,\ 0 \leq t \leq T\}, P)$, où l'ensemble Ω contient les états du monde, la tribu $\mathscr{A}$ est l'ensemble de l'information disponible sur le marché, la filtration $\{\mathscr{F}_t,\ 0 \leq t \leq T\}$ décrit l'information accessible aux agents intervenant sur le marché au cours du temps et la mesure de probabilité P, dite *historique*, donne la probabilité *a priori* de tout événement considéré.

Sous sa forme la plus simple, le *modèle de Black-Scholes* ne considère que deux titres de base : un actif sans risque (typiquement une obligation émise par un état ne présentant pas de risque de défaut) et un actif risqué (une action sous-jacente à l'option par exemple), et fait un certain nombre d'hypothèses idéalisées sur le fonctionnement du marché, à savoir :

- on peut vendre à découvert sans restriction, ni pénalité,
- les actifs sont parfaitement divisibles,
- les échanges ont lieu sans coût de transaction,
- on peut emprunter et prêter de l'argent à un taux d'intérêt sans risque (*risk free rate* en anglais) constant r,
- les agents négocient en continu et l'on peut à tout moment trouver des acheteurs et des vendeurs pour les titres du marché.

L'évolution de la valeur de l'actif sans risque, dont le rendement est connu à l'avance, est gouvernée par l'équation différentielle ordinaire suivante

$$\mathrm{d}R(t) = r\,R(t)\,\mathrm{d}t,\ \ t \in [0, T], \tag{9.16}$$

et l'on a par conséquent $R(t) = R_0\, e^{r\,t},\ t \in [0, T]$.

On suppose également qu'aucun dividende sur l'actif sous-jacent n'est distribué durant la vie de l'option et que l'évolution du cours de cet actif est celle d'un processus de Wiener *géométrique*, c'est-à-dire qu'il satisfait l'équation différentielle stochastique suivante

$$\mathrm{d}S(t) = S(t)\,(\mu\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)),\ \ t \in [0, T] \tag{9.17}$$

où W est un processus de Wiener standard sous la probabilité P et les coefficients μ et σ, avec $\sigma > 0$, sont respectivement la *tendance* ou *dérive* (*drift* en anglais) et la *volatilité* (*volatility* en anglais) de l'actif [35], toutes deux

31. Les actifs sous-jacents sont généralement des actions, des obligations, des devises, des contrats à terme, des produits dérivés ou encore des matières premières.

32. Fischer Sheffey Black (11 janvier 1938 - 30 août 1995) était un économiste américain, connu pour avoir inventé, avec Myron Scholes, une formule d'évaluation du prix des actifs financiers.

33. Myron Samuel Scholes (né le 1er juillet 1941) est un économiste américain d'origine canadienne. Il reçut, avec Robert Merton, le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel en 1997 pour ses travaux sur la valorisation des produits dérivés, notamment les options.

34. Robert Carhart Merton (né le 31 juillet 1944) est un économiste américain, connu son application d'une approche mathématique des processus stochastiques en temps continu à l'économie, et plus particulièrement à l'étude des marchés financiers. Il a reçu, avec Myron Scholes, en 1997 le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel pour sa participation à la découverte du modèle de Black–Scholes de valorisation des options.

35. Ces deux quantités mesurent respectivement le rendement relatif espéré et l'ampleur des variations du cours de l'actif par unité de temps.

supposées constantes. Ce modèle est le plus simple que l'on puisse imaginer pour la dynamique du cours d'un actif tout en garantissant sa stricte positivité [36].

L'aléa provenant seulement du processus S dans ce cas, on a

$$\mathscr{F}_t = \sigma(\{S(s),\ 0 \le s \le t\}) = \sigma(\{W(s),\ 0 \le s \le t\}),\ t \in [0,T].$$

Le théorème (9.22) assure que l'équation (9.17), complétée par la donnée d'une condition initiale en $t = 0$, admet une unique solution forte, dont l'expression est

$$S(t) = S_0\, e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W(t)},\ t \in [0,T].$$

Stratégie de portefeuille autofinancée

Pour se prémunir contre le risque d'une possible évolution défavorable du cours de l'actif sous-jacent, le vendeur de l'option va, en investissant sur le marché financier, construire un *portefeuille de couverture* (*delta neutral portfolio* en anglais) répliquant (parfaitement) le comportement de l'option. Le prix de ce portefeuille, c'est-à-dire la somme que l'on doit y placer initialement pour réaliser la couverture, détermine alors le prix de l'option.

Pour le modèle considéré, la *stratégie financière* correspondante consiste en la donnée de deux processus stochastiques α et β, adaptés à la filtration $\{\mathscr{F}_t,\ 0 \le t \le T\}$, qui représentent les quantités respectives d'actifs sans risque et risqué détenues dans le portefeuille de couverture à chaque instant et sont déterminés sur la base des informations disponibles au cours du temps. La valeur du portefeuille à un temps t donné est alors

$$V(t) = \alpha(t)R(t) + \beta(t)S(t),\ t \in [0,T]. \tag{9.18}$$

La gestion dynamique après l'instant initial du portefeuille se faisant sans apport, ni retrait de fonds extérieurs (on parle de portefeuille *autofinançant*), la variation instantanée de V ne dépend que de la variation de cours de l'actif risqué et du rendement de la somme placée sur l'actif sans risque, c'est-à-dire que l'on a

$$\mathrm{d}V(t) = \alpha(t)\,\mathrm{d}R(t) + \beta(t)\,\mathrm{d}S(t),\ t \in [0,T], \tag{9.19}$$

soit encore, en tenant compte des équations (9.16), (9.17) et de la relation (9.18),

$$\mathrm{d}V(t) = (r\,V(t) + (\mu - r)\,\beta(t)S(t))\,\mathrm{d}t + \sigma\,\beta(t)S(t)\,\mathrm{d}W(t),\ t \in [0,T]. \tag{9.20}$$

On observera que, pour avoir un sens, la condition d'autofinancement (9.19) impose des restrictions sur les processus α et β. On suppose donc dans toute la suite que l'on a

$$\int_0^T \left(|\alpha(s)| + |\beta(s)|^2\right)\mathrm{d}s < +\infty \text{ presque sûrement.} \tag{9.21}$$

En pratique, on travaille généralement avec des *valeurs actualisées* (*discounted values* en anglais), par rapport à celle de l'actif sans risque, de l'actif et du portefeuille, c'est-à-dire les quantités

$$\widetilde{S}(t) = \frac{S(t)}{R(t)} \text{ et } \widetilde{V}(t) = \frac{V(t)}{R(t)},\ t \in [0,T].$$

On trouve, en utilisant la formule d'Itō (9.11), que ces valeurs évoluent respectivement selon les équations

$$\mathrm{d}\widetilde{S}(t) = \widetilde{S}(t)\,((\mu - r)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)),\ t \in [0,T],$$

et

$$\mathrm{d}\widetilde{V}(t) = \beta(t)\,\mathrm{d}\widetilde{S}(t),\ t \in [0,T],$$

montrant que la stratégie suivie est entièrement déterminée par la donnée de la somme initialement investie dans le portefeuille et la connaissance du processus adapté β.

36. La distribution de la variable $S(t)$ suit en effet une loi dite *log-normale*, c'est-à-dire que la variable aléatoire $\ln(S(t))$ suit une loi normale de moyenne $\left(\mu - \frac{\sigma^2}{2}\right)t$ et d'écart-type $\sigma^2 t$.

Principe d'arbitrage et mesure de probabilité risque-neutre

La classe des stratégies de portefeuille définies par la condition d'intégrabilité (9.21) reste trop large pour prévenir les *opportunités d'arbitrage*, c'est-à-dire les possibilités de faire, sans aucun investissement initial, une série de transactions conduisant de manière certaine à un profit [37]. Or, le *principe d'absence d'opportunité d'arbitrage*, nécessaire à la viabilité du marché, interdit de telles stratégies.

On peut montrer que l'existence d'une mesure de probabilité Q, équivalente [38] à la mesure P, sous laquelle le cours de l'actif risqué est une martingale, alliée au choix d'une stratégie *minorée* ou vérifiant une condition d'intégrabilité de la forme

$$E_Q\left(\int_0^T \left|\beta(s)\widetilde{S}(s)\right|^2 \mathrm{d}s\right) < +\infty, \tag{9.22}$$

implique l'absence d'opportunité d'arbitrage.

Ici, l'existence d'une telle mesure Q est une conséquence du *théorème de Girsanov* [39] [Gir60], sa densité de Radon [40]–Nikodym [41] par rapport à P sur $(\Omega, \mathscr{F}_T)$ étant donnée par

$$\left.\frac{\mathrm{d}Q}{\mathrm{d}P}\right|_{\mathscr{F}_T} = e^{-\frac{\lambda^2}{2}T - \lambda W(T)},$$

où $\lambda = \frac{\mu - r}{\sigma}$ est la *prime de risque* de l'actif, c'est-à-dire l'écart de rendement espéré en contrepartie de la prise de risque.

On observe que le prix actualisé de l'actif risqué est une martingale sous la mesure Q. Ceci découle en effet de la proposition 9.15, le processus $\widetilde{S}$ satisfaisant sous Q l'équation différentielle stochastique

$$\mathrm{d}\widetilde{S}(t) = \sigma\,\widetilde{S}(t)\,\mathrm{d}W^*(t),\ t \in [0,T],$$

où W^* est un processus de Wiener par rapport à Q tel que $W^*(t) = W(t) + \lambda\, t$. Le cours non actualisé de l'actif risqué évolue lui selon

$$\mathrm{d}S(t) = S(t)\,(r\,\mathrm{d}t + \sigma\,\mathrm{d}W^*(t)),\ t \in [0,T]. \tag{9.23}$$

Cette dernière équation fournit une interprétation de la mesure de probabilité Q, que l'on peut voir comme celle qui régirait le processus de prix de l'actif risqué si l'espérance du taux de rendement de celui-ci était le taux d'intérêt sans risque, lui donnant le nom de mesure de probabilité *risque-neutre* (puisque, sous elle, aucune prime n'est attribuée à la prise de risque).

Concernant la valeur actualisée du portefeuille, il vient sous la mesure Q

$$\mathrm{d}\widetilde{V}(t) = \sigma\beta(t)\widetilde{S}(t)\,\mathrm{d}W^*(t),\ t \in [0,T],$$

qui définit bien une martingale sous la condition d'intégrabilité (9.22), et l'on a alors, en vertu de la définition 9.12,

$$e^{-r\,t}\,V(t) = e^{-r\,T}\,E_Q\left(V(T)\,|\,\mathscr{F}_t\right),\ t \in [0,T],$$

avec

$$\mathrm{d}V(t) = r\,V(t)\,\mathrm{d}t + \beta(t)S(t)\,(r\,\mathrm{d}t + \sigma\,\mathrm{d}W^*(t)),\ t \in [0,T].$$

37. Mathématiquement, l'existence d'une opportunité d'arbitrage se traduit par celle d'une stratégie de portefeuille telle que l'on a

$$V(0) = 0,\ P(\{V(T) \geq 0\}) = 1 \text{ et } P(\{V(T) > 0\}) > 0.$$

38. Étant donné un ensemble Ω et une tribu $\mathscr{A}$ sur Ω, deux mesures de probabilités P et Q définies sur $(\Omega, \mathscr{A})$ sont *équivalentes* si, pour tout A appartenant à $\mathscr{A}$, $P(A) = 0$ si et seulement si $Q(A) = 0$

39. Igor Vladimirovich Girsanov (Игорь Влади́мирович Гирсанов en russe, 10 septembre 1934 - 16 mars 1967) était un mathématicien russe, connu pour ses contributions à la théorie des probabilités et ses applications.

40. Johann Karl August Radon (16 décembre 1887 - 25 mai 1956) était un mathématicien autrichien. Il œuvra en théorie de la mesure ainsi qu'en analyse fonctionnelle, et introduisit une transformée aujourd'hui couramment utilisée pour la reconstruction d'images en tomographie.

41. Otto Marcin Nikodým (13 août 1887 - 4 mai 1974) était un mathématicien polonais. Il a travaillé dans plusieurs domaines des mathématiques, comme la théorie de la mesure ou la théorie des opérateurs dans les espaces de Hilbert.

Réplication et évaluation de l'option

Considérons à présent une option d'achat européenne de maturité T et de prix d'exercice K sur un actif financier sous-jacent dont le prix à l'instant t, $0 \leq t \leq T$, est $S(t)$. L'acheteur, après avoir payé la prime C au vendeur à l'instant initial $t = 0$, reçoit à l'instant $t = T$ le gain (*pay-off* en anglais) $(S(T)-K)_+ = \max\{S(T)-K, 0\}$. Pour sa part, le vendeur investit l'intégralité de la prime reçue dans un portefeuille autofinançant de valeur $V(t)$ au temps t, $0 \leq t \leq T$, son objectif étant de couvrir le risque en faisant en sorte que la valeur finale $V(T)$ du portefeuille soit celle du gain $(S(T)-K)_+$. Si cela est faisable, l'option est dite *réplicable* et, en l'absence d'opportunité d'arbitrage, son prix à toute date est bien défini et égal à la valeur du portefeuille de couverture à cette date. Un marché dans lequel toutes les options, ou plus généralement tous les *actifs contingents*, sont réplicables est dit *complet*. Dans le cas présent, la complétude du marché découle du fait que $\sigma > 0$.

Compte tenu de ces considérations, et en vertu d'une *propriété de Markov* vérifiée par le processus S, le prix de l'option à l'instant t est donné par la quantité

$$V(t) = e^{-r(T-t)} E_Q\left((S(T)-K)_+ \,|\, \mathscr{F}_t\right) = e^{-r(T-t)} E_Q\left((S(T)-K)_+ \,|\, S(t)\right) = C(t, S(t)),\ t \in [0, T],$$

qui est une fonction de la variable $S(t)$. En particulier, le *juste prix* (*fair price* en anglais) de l'option est

$$C(0, S_0) = e^{-r\,T} E_Q\left((S(T)-K)_+\right) = e^{-r\,T} E\left((X(T)-K)_+\right),$$

avec

$$\mathrm{d}X(t) = X(t)\left(r\,\mathrm{d}t + \sigma\,\mathrm{d}W(t)\right),\ t \in [0, T],\ X(0) = S_0. \tag{9.24}$$

La fonction à valeurs réelles C ainsi introduite étant régulière, l'application de la formule d'Itō (9.11) et l'utilisation de (9.23) donnent

$$\mathrm{d}C(t, S(t)) = \left(\frac{\partial C}{\partial t}(t, S(t)) + \frac{1}{2}(\sigma S(t))^2 \frac{\partial^2 C}{\partial x^2}(t, S(t))\right)\mathrm{d}t + \frac{\partial C}{\partial x}(t, S(t))\,\mathrm{d}S(t),\ t \in [0, T],$$

ce qui permet, par identification avec la condition d'autofinancement (9.19), de déterminer la stratégie de couverture [42]

$$\alpha(t) = \frac{1}{r\,R(t)}\left(\frac{\partial C}{\partial t}(t, S(t)) + \frac{1}{2}(\sigma S(t))^2 \frac{\partial^2 C}{\partial x^2}(t, S(t))\right),\ \beta(t) = \frac{\partial C}{\partial x}(t, S(t)).$$

Formule de Black–Scholes

Les coefficients des équations différentielles régissant les cours des actifs étant constants, il est possible d'expliciter la fonction C, définie sur $[0, T] \times \mathbb{R}_+$ par

$$C(t, x) = e^{-r(T-t)} E\left((X(T)-K)_+\right),$$

où

$$\mathrm{d}X(s) = X(s)\left(r\,\mathrm{d}t + \sigma\,\mathrm{d}W(s)\right),\ s \in [t, T],\ X(t) = x,$$

en remarquant que la variable aléatoire X_s suit une loi log-normale. Il vient alors

$$C(t, x) = e^{-r(T-t)} \int_{\ln(K)}^{+\infty} (e^u - K)\, \frac{e^{-\frac{1}{2\sigma^2(T-t)}\left(u-\ln(x)-(T-t)\left(r-\frac{\sigma^2}{2}\right)\right)^2}}{\sigma\sqrt{2\pi(T-t)}}\,\mathrm{d}u,$$

dont on déduit, après quelques calculs, la fameuse *formule de Black–Scholes*

$$C(t, x) = x\,N(d_1(t, x)) - K\,e^{-r(T-t)} N(d_2(t, x)), \tag{9.25}$$

42. La quantité $\frac{\partial C}{\partial x}(t, S(t))$ est appelée le *delta* de l'option à l'instant t et représente la sensibilité du prix de cette option par rapport à la valeur de l'actif sous-jacent à cette date. Notons que l'on a encore

$$\alpha(t) = \frac{1}{R(t)}\left(C(t, S(t)) - S(t)\,\frac{\partial C}{\partial x}(t, S(t))\right)$$

en vertu de l'équation (11.1).

où N est la fonction de répartition de la loi centrée réduite

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} \, \mathrm{d}u,$$

$$d_1(t,x) = \frac{1}{\sigma\sqrt{T-t}} \left(\ln\left(\frac{x}{K}\right) + (T-t)\left(r + \frac{\sigma^2}{2}\right)\right),$$

et

$$d_2(t,x) = d_1(t,x) - \sigma\sqrt{T-t}.$$

On a ainsi établi que, dans le cadre du modèle de Black–Scholes, la prime d'une option de prix d'exercice K et de date de maturité T sur un actif sous-jacent, dont le cours évolue selon l'équation (9.17), est donnée par

$$C(0,x) = x\,N(d_1(0,x)) - K\,e^{-r\,T}\,N(d_2(0,x)),$$

où les quantités $d_1(0,x)$ et $d_2(0,x)$ ne dépendent que des prix x et K, de la date T, du taux d'intérêt sans risque r et de la volatilité σ, ce dernier paramètre étant le seul à être non directement observable.

Extensions et méthodes de Monte-Carlo

Le modèle de Black–Scholes permet également l'évaluation du prix d'une option de vente (*put option* en anglais). Pour le rendre plus réaliste, on peut l'étendre de façon à prendre en compte un nombre d, avec $d \geq 1$, d'actifs risqués (dont les tendances et les volatilités pourront être des fonctions déterministes du temps et/ou des cours des actifs ou même des processus stochastiques), le paiement de dividendes ou encore l'utilisation de *processus de Lévy*[43] dont les incréments ne suivent pas une loi normale et les trajectoires sont seulement des fonctions *continues à droite et admettant une limite à gauche* (*càdlàg* en abrégé) en tout point en place de processus de Wiener.

plus nécessairement de formule explicite à disposition

L'approche Monte-Carlo appliquée à l'évaluation du prix d'une option européenne [Boy77] consiste en

- la simulation d'un nombre M de réalisations indépendantes de la solution S jusqu'à la date de maturité T (avec taux rendement = taux sans risque),
- le calcul des gains correspondants,
- approximation de l'espérance par une moyenne

$$\frac{1}{M} \sum_{k=1}^{M} (X_T^{(k)} - K)_+,$$

où $X_T^{(k)}$ est (l'approximation d')une réalisation de la valeur à la date de maturité d'un processus satisfaisant (9.24).
- actualisation de la valeur en multipliant par $e^{-r\,T}$

estimation du delta ou des autres grecques par différences finies

+ remarque sur convergence lente, réduction de variance, variables antithétiques

9.2.3 Modèle de Vasicek d'évolution des taux d'intérêts en finance **

INTRO On considère ici, comme c'était le cas avec le modèle de Black–Scholes dans la section précédente, un modèle continu en temps.

A VOIR

43. Paul Pierre Lévy (15 septembre 1886 - 15 décembre 1971) était un mathématicien français, figurant parmi les fondateurs de la théorie moderne des probabilités. On lui doit d'importants travaux sur les lois stables et sur les fonctions aléatoires, ainsi que l'introduction du concept de martingale.

9.2.4 Quelques définitions

A zero-coupon bond (obligation zéro-coupon) price with maturity T is a security that pays 1 at time T and provides no other cash flows between time t and T. Suppose that for any T there exists a zero coupon with maturity T. Then, the price at time t of the zero coupon bond with maturity T is denoted $P(t,T)$. We have $P(T,T)=1$.

The yield to maturity at time t, denoted $Y(t,T)$, is defined by

$$P(t,T)=\exp\left(-(T-t)\,Y(t,T)\right).$$

the forward spot rate at time t with maturity T is

$$f(t,T)=-\left[\frac{\partial \ln(P)}{\partial \theta}(t,\theta)\right]_{\theta=T}.$$

we have

$$Y(t,T)=\frac{1}{T-t}\int_t^T f(t,u)\,\mathrm{d}u \text{ and } P(t,T)=\exp\left(-\int_t^T f(t,u)\,\mathrm{d}u\right).$$

the instantaneous spot rate is

$$R(t)=\lim_{T\to t} Y(t,T)=-\left[\frac{\partial \ln(P)}{\partial \theta}(t,\theta)\right]_{\theta=T}=f(t,t).$$

The yield curve (courbe des taux) is given by the function $\theta \mapsto Y(t,\theta)$.

taux court instantané (limite du taux moyen quand le temps restant à maturité tend vers zéro, c'est le taux à court terme)

la courbe des taux est la fonction qui donne les différents taux moyens de la date t en fonction de leur maturité restante $T-t$. on cherche à décrire la courbe des taux dans le futur en fonction de la courbe observée aujourd'hui.

Dans le *modèle de Vasicek* [44] [Vas77], on suppose que l'évolution du taux court (*spot rate* en anglais) $S(t)$ est, sous la probabilité historique P, gouvernée par l'équation différentielle stochastique

$$\mathrm{d}S(t)=\alpha(\nu-S(t))\,\mathrm{d}t+\sigma\,\mathrm{d}W(t) \tag{9.26}$$

où $\alpha>0$, ν, $\sigma>0$ constantes, où ν est la moyenne à long terme du taux, α est la vitesse de retour, ou d'ajustement, du taux court actuel vers sa moyenne à long terme, σ est la volatilité du taux

La moyenne instantanée est proportionnelle à la différence entre la valeur de ν et celle de $S(t)$. Une « force de rappel » tend à ramener $S(t)$ près de la valeur de ν.

une solution explicite de (9.26) est

$$S(t)=\nu+(S_0-\nu)e^{-\alpha t}+\mu+\sigma\int_0^t e^{-\alpha(t-u)}\,\mathrm{d}W_u.$$

appelée un *processus d'Ornstein* [45]*–Uhlenbeck* [46]

si $S_0\in\mathbb{R}$, alors S_y est une variable aléatoire gaussienne de moyenne $(S_0-\mu)\,e^{-\alpha t}+\nu$ et de variance $\frac{\sigma^2}{2\alpha}(1-e^{-2\alpha t})$. En particulier, cette variable n'est pas positive. Plus généralement, si S_0 est une variable gaussienne indépendante du processus de Wiener W, le processus S est une fonction aléatoire gaussienne d'espérance $E(S(t))=\mu(1-e^{-\alpha t})+e^{-\alpha t}\,E(S_0)$ et de variance ...

Ce modèle autorise donc les taux à devenir négatifs avec une probabilité non nulle, ce qui n'est pas satisfaisant en pratique. Ceci est corrigé par le *modèle de Cox–Ingersoll–Ross* [CIR85]

$$\mathrm{d}S(t)=\alpha(\nu-S(t))\,\mathrm{d}t+\sigma\sqrt{S(t)}\,\mathrm{d}W(t),\ S(0)=0,$$

qui présente le même effet de retour, mais reste positif.

44. Oldřich Alfons Vašíček (né en 1942) est un mathématicien tchèque. Ses travaux sur la courbe des taux d'intérêt ont conduit à la théorie « moderne » de ces derniers.

45. Leonard Salomon Ornstein (12 novembre 1880 - 20 mai 1941) était un physicien hollandais, principalement connu pour ses travaux en physique statistique.

46. George Eugene Uhlenbeck (6 décembre 1900 - 31 octobre 1988) était un physicien américain d'origine hollandaise. Il est connu pour avoir proposé, avec Samuel Goudsmit, l'hypothèse du spin de l'électron en 1925.

9.3 Méthodes numériques de résolution des équations différentielles stochastiques

Comme cela était le cas pour les équations différentielles ordinaires étudiées dans le précédent chapitre, on ne connaît que rarement une forme explicite de la solution d'une équation différentielle stochastique de la forme (9.9) et l'on fait donc appel à une méthode numérique pour approcher cette solution. La solution d'une équation différentielle stochastique étant un processus stochastique, il est important de noter que la méthode utilisée va calculer des trajectoires approchées, c'est-à -dire des approximations de *réalisations* du processus. Pour cette raison, les définitions de la consistance et la convergence d'une méthode diffèrent (mais coïncident en l'absence d'aléa) de celles données dans le cadre déterministe des équations différentielles ordinaires.

Note : intervalles de temps : $h_n = t_{n+1} - t_n$, $\forall n$, on ne fera aucune considération d'adaptation/de variation du pas et la grille de discrétisation est uniforme)
Dans toute la suite, $\{\mathscr{A}_t,\ t \geq 0\}$ désigne une filtration donnée, généralement associée au processus d'Itō que l'on approche ou du procesus de Wiener sous-jacent.

9.3.1 Simulation numérique d'un processus de Wiener *

La possibilité de résoudre numériquement une équation différentielle stochastique comme (9.9) repose de manière fondamentale sur le fait de disposer d'une représentation numérique de réalisations d'un processus de Wiener ou bien d'approximations de celles-ci. Nous allons pour celle raison nous intéresser à la simulation numérique d'un processus de Wiener en présentant deux méthodes, dont l'une s'inspire directement de la définition 9.13.

Ces techniques sont basées sur l'utilisation pratique d'une source de nombres susceptibles de représenter une réalisation d'une suite de variables aléatoires indépendantes et de loi de probabilité donnée. La question de la génération de tels nombres sur un calculateur étant absolument non triviale, nous commençons par l'aborder dans le détail.

Générateurs de nombres pseudo-aléatoires

Pour obtenir une suite de nombres aléatoires, une idée naturelle est d'avoir recours à l'observation de mécanismes « physiques » pouvant être considérés comme imprévisibles, tels le lancer de dés ou de pièces de monnaie, le jeu de roulette, le brassage de billes ou de cartes suivi de tirages au sort, ou encore la radioactivité, le *bruit de Johnson*[47]*–Nyquist*[48] généré par l'agitation thermique de porteurs de charge dans un conducteur, le *bruit de grenaille* dans un composant électronique ou d'*avalanche* dans un semi-conducteur, certains mécanismes de la physique quantique, etc. Ce type de procédé présente cependant plusieurs inconvénients. Tout d'abord, le phénomène en question et/ou l'appareil de mesure utilisé pour l'appréhender souffrent généralement d'asymétries ou de biais systématiques qui compromettent le caractère uniformément aléatoire des suites produites. Par ailleurs, la génération des nombres peut s'avérer trop lente pour certaines applications visées et le dispositif mis en jeu trop coûteux et pas toujours fiable. Enfin, les suites obtenues sont généralement non reproductibles.

Une alternative à cette première approche réside dans l'utilisation d'un *générateur de nombres pseudo-aléatoires* (*pseudorandom number generator* en anglais), qui est un algorithme fournissant une suite de nombres faite pour présenter, de manière approchée[49] (d'où la présence du préfixe *pseudo-*), certaines propriétés statistiques du hasard, telles que l'indépendance entre les termes de la suite et une distribution selon une loi de probabilité donnée. Dans la plupart des cas, la génération de tels nombres est accomplie en deux temps, avec tout d'abord la génération de valeurs jouant le rôle d'une réalisation d'une suite de variables aléatoires continues, indépendantes et identiquement distribuées suivant la loi uniforme sur l'intervalle $[0, 1]$, puis l'application d'une transformation à la suite produite de manière à finalement obtenir une suite de variables aléatoires simulées distribuées suivant la loi désirée. Compte tenu de cette observation, nous ne présenterons ici que des générateurs de nombres

47. John Bertrand Johnson (né Johan Erik Bertrand, 2 octobre 1887 - 27 novembre 1970) était un physicien et ingénieur américain d'origine suédoise. Il a été le premier à expliquer l'origine du bruit dû au passage de courant électrique dans un conducteur.

48. Harry Nyquist (né Harry Theodor Nyqvist, 7 février 1889 - 4 avril 1976) était un physicien et ingénieur américain d'origine suédoise. Il fut un important contributeur à la théorie de l'information.

49. Sur ce point, on peut reprendre la phrase célèbre de John von Neumann : *"Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin."* [vNeu51].

uniformément distribués, en mentionnant brièvement quelques techniques permettant la simulation de suites de variables aléatoires suivant d'autres lois de probabilité.

Les générateurs de nombres pseudo-aléatoires étant particulièrement simples à mettre en œuvre et souvent très rapides, ils sont aujourd'hui employés dans de nombreux domaines, majoritairement pour la simulation stochastique par des applications de la méthode de Monte-Carlo, mais aussi dans les machines automatiques de jeux de hasard ou encore en cryptographie (pour la fabrication de clés de cryptage). Ils se doivent par conséquent de satisfaire à une série de critères quantitatifs et qualitatifs, portant à la fois sur leur fonctionnement et sur les suites qu'ils produisent, au nombre desquels on peut signaler :

- l'*absence de corrélation ou de dépendance entre les termes des suites* et l'*adéquation des suites à la loi de distribution uniforme* (les suites de nombres pseudo-aléatoires obtenues doivent être en mesure de passer un certain nombre de tests statistiques visant à vérifier qu'elles ressemblent à des réalisations d'une suite de variables aléatoires indépendantes uniformément distribuées sur $[0,1]$; en particulier, des d-uplets de termes, consécutifs ou non, doivent recouvrir uniformément l'hypercube d-dimensionnel $[0,1]^d$ pour des valeurs « raisonnables » de l'entier d),
- la *longueur de la période des suites* (les générateurs conduisent à des suites périodiques, dont les périodes se doivent d'être les plus grandes possibles),
- la *reproductibilité* (on doit être en mesure de produire exactement la même suite de nombres pseudo-aléatoires à partir d'appels répétés d'un même générateur),
- la *portabilité* (on doit pouvoir générer exactement les mêmes suites de nombres pseudo-aléatoires sur des machines différentes),
- l'*efficacité* (on doit utiliser peu d'opérations arithmétiques et de ressources pour produire chaque nombre pseudo-aléatoire excessives et pouvoir exploiter les possibilités offertes par des processeurs vectoriels ou l'architecture parallèle d'un calculateur).

Le premier générateur de nombres pseudo-aléatoires, la *méthode du carré médian* (*middle-square method* en anglais), fut proposé par von Neumann [50] en 1949 [vNeu51]. Son principe est extrêmement simple : il consiste à prendre un nombre entier, appelé *graine* (*seed* en anglais), à t chiffres, avec t un entier naturel pair, à l'élever au carré pour obtenir un entier à $2t$ chiffres (des zéros non significatifs sont ajoutés si l'entier obtenu contient moins de $2t$ chiffres) dont on retient les t chiffres du milieu comme sortie (en divisant l'entier ainsi trouvé par 10^t, on obtient bien un nombre normalisé contenu dans l'intervalle $[0,1[$). Il suffit de réitérer le procédé pour construire une suite, le dernier nombre obtenu servant de nouvelle graine. Bien que rapide, ce générateur ne possède qu'un intérêt historique, car il présente plusieurs faiblesses rédhibitoires, telle qu'une période courte (celle-ci ne peut dépasser 8^t) et l'existence d'*états absorbants* [51].

Aujourd'hui largement répandue, la classe des *générateurs à congruence linéaire* (*linear congruential generators* en anglais) fut introduite par Lehmer en 1949 [Leh51]. Ces générateurs construisent une suite d'entiers naturels $(x_n)_{n\in\mathbb{N}}$ à partir d'une graine x_0 suivant une relation de récurrence de la forme

$$x_n = (a\,x_{n-1} + c) \mod m,\ n \geq 1, \tag{9.27}$$

avec a, le *multiplicateur*, et m, le *module*, deux entiers strictement positifs et c, l'*incrément*, un entier positif (le générateur étant dit *multiplicatif* lorsque $c = 0$). Cette suite prenant ses valeurs dans $\{0,\ldots,m-1\}$, on obtient effectivement une suite de nombres normalisés contenus dans l'intervalle $[0,1[$ en divisant chacun de ses termes par l'entier m. Des choix pratiques de valeurs des entiers a, c et m sont, par exemple, $a = 65539$, $c = 0$ et $m = 2^{31}$ pour le générateur RANDU de la bibliothèque de programmes scientifiques des machines IBM System/360 fabriquées dans les années 1960 et 1970, $a = 1103515245$, $c = 12345$ et $m = 2^{31}$ pour celui de la fonction `rand()` du langage ANSI C.

50. John von Neumann (Neumann János Lajos en hongrois, 28 décembre 1903 - 8 février 1957) était un mathématicien et physicien américano-hongrois. Il a apporté d'importantes contributions tant en théorie des ensembles, en analyse fonctionnelle, en théorie ergodique, en analyse numérique et en statistiques, ainsi qu'en mécanique quantique, en informatique, en sciences économiques et en théorie des jeux.

51. Si les chiffres du milieu du nombre élevé au carré sont tous égaux à 0 à une itération donnée, cela sera évidemment le cas pour tous les nombres suivants, rendant ainsi la suite constante à partir d'un certain rang. Pour $t = 4$, ce phénomène se produit également avec les nombres 100, 2500, 3792 et 7600. Par ailleurs, certaines graines peuvent conduire à des cycles courts se répétant indéfiniment (comme 540-2916-5030-3009 pour $t = 4$ par exemple). Mentionnons enfin que si la première moitié des chiffres d'un des nombre obtenu est uniquement composée de 0, les valeurs des nombres produits par l'algorithme décroissent vers 0. C'est le cas, pour $t = 4$, de la suite issue de l'entier 1926, qui « s'éteint » après seulement vingt-six itérations : 7094, 3248, 5495, 1950, 8025, 4006, 480, 2304, 3084, 5110, 1121, 2566, 5843, 1406, 9768, 4138, 1230, 5129, 3066, 4003, 240, 576, 24, 5, 0.

Une suite de nombres pseudo-aléatoires étant produite de façon déterministe une fois fixés ces trois paramètres et la graine, les propriétés, et donc la qualité, d'un générateur à congruence linéaire dépend crucialement des valeurs retenues, de mauvais choix conduisant à des générateurs ayant de très mauvaises propriétés statistiques. Un premier critère à prendre en compte pour la sélection des paramètres est celui de la longueur de la période du générateur. Lorsque l'incrément est non nul [52], des conditions nécessaires et suffisantes pour que cette longueur soit *maximale* (et donc égale à m) sont données par le résultat suivant.

Théorème 9.24 (longueur de période maximale pour un générateur congruentiel linéaire d'incrément non nul [HD62]) *La suite définie par la relation de récurrence* (9.27)*, avec $c \neq 0$, a une période de longueur égale à m si et seulement si*

— *les entiers c et m sont premiers entre eux,*
— *pour chaque nombre premier p divisant m, l'entier $a-1$ est un multiple de p (i.e., $a \equiv 1 \mod p$),*
— *si m est un multiple de 4, alors $a-1$ l'est également (i.e., $a \equiv 1 \mod 4$).*

Notons que des considérations d'ordre pratique peuvent s'ajouter à ces conditions, notamment pour le choix de la valeur du module. En effet, sur une machine fonctionnant avec un système de numération binaire, il est avantageux de prendre $m = 2^w$, où l'entier w représente le nombre de bits servant à coder la valeur absolue d'un entier signé (par exemple $w = 31$ pour un codage des entiers signés sur 32 bits), car la division euclidienne induite par la relation de congruence est alors ramenée à une simple troncature et la division des termes de la suite par le module à un déplacement du séparateur. Un inconvénient majeur de ce choix est que les bits dits de poids faible (c'est-à-dire ceux situés le plus à droite dans la notation positionnelle) des nombres de la suite produite possèdent une période significativement plus courte que celle de la suite elle-même, mettant ainsi facilement en évidence son caractère non aléatoire. Pour corriger ce problème, on choisit le module comme un *nombre premier de Mersenne* [53] (par exemple $m = M_{31} = 2^{31} - 1 = 2147483647$), la division euclidienne pouvant encore être évitée par une astuce [PRB69].

La faiblesse fondamentale des générateurs à congruence linéaire, identifiée par Marsaglia [54] [Mar64], ne provient cependant pas de la longueur de leur période, mais du fait qu'une suite de d-uplets de nombres normalisés produits par le générateur au cours d'une période complète ne recouvre pas uniformément le cube unité de dimension d avec une erreur de discrétisation de l'ordre attendu [55] et se répartit sur un nombre limité [56] d'hyperplans parallèles et équidistants. Cette structure particulière s'explique par la linéarité (à congruence près) de la relation de récurrence (9.27). La corrélation induite est particulièrement catastrophique dans le cas du générateur RANDU pour $d = 3$, puisque l'on peut montrer que les triplets de nombres successifs appartiennent à seulement quinze plans différents (voir la figure 9.2 pour une illustration), ce qui explique pourquoi il est aujourd'hui considéré comme l'un des plus mauvais jamais inventé.

Une généralisation des générateurs à congruence linéaire consiste à utiliser plus d'un état passé pour obtenir l'état courant [Gru73], ce qui conduit à une relation de récurrence de la forme

$$x_n = (a_1\, x_{n-1} + a_2\, x_{n-2} + \cdots + a_k\, x_{n-k} + c) \mod m,\ n \geq k,$$

52. Si l'incrément est nul, la longueur maximale de la période est $m-1$ (0 étant un état absorbant). Dans ce cas, le théorème s'énonce ainsi (et se déduit de résultats dus à Carmichael [Car10]) :

Théorème 9.23 (longueur de période maximale pour un générateur congruentiel linéaire multiplicatif) *La période de la suite produite par un générateur congruentiel linéaire multiplicatif est de longueur égale à $m-1$ si et seulement si le module m est un nombre premier et que le multiplicateur a est une racine primitive modulo m.*

Le concept de *racine primitive modulo un entier* est issu de l'arithmétique modulaire en théorie des nombres. Lorsque le module est un nombre premier impair, il est possible d'expliciter la seconde condition en utilisant une caractérisation disant que a est une racine primitive modulo m, si et seulement si $a^{(m-1)/p} - 1$ est un multiple de m pour tout facteur premier p de $m-1$. L'article [FM86] présente une recherche exhaustive de multiplicateurs vérifiant cette condition avec $m = M_{31}$. Pour cette valeur du module, le choix $a = 7^5 = 16807$ est recommandé dans [PM88], donnant lieu au générateur portant le nom de *minimal standard* (MINSTD).

53. Marin Mersenne (8 septembre 1588 - 1er septembre 1648) est un religieux, érudit, mathématicien et philosophe français. On lui doit les premières lois de l'acoustique, qui portèrent longtemps son nom.

54. George Marsaglia (12 mars 1924 - 15 février 2011) était un mathématicien et informaticien américain. On lui doit le développement de méthodes parmi les plus courantes pour la générations de nombres pseudo-aléatoires et leur utilisation pour la production d'échantillons de distributions diverses, ainsi que l'élaboration de séries de tests visant à mesurer la qualité d'un générateur en déterminant si les suites de nombres qu'il fournit possèdent certaines propriétés statistiques.

55. Les termes de la suite prenant leurs valeurs dans un sous-ensemble *discret* de l'intervalle $[0,1]$ de la forme $\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}, 1\}$, une suite de d-uplets peut au mieux recouvrir $[0,1]^d$ avec un réseau régulier de points dont l'espacement est de l'ordre de $\frac{1}{m}$.

56. Il a été établi que ce nombre d'hyperplans est majoré par $(d!m)^{\frac{1}{d}}$.

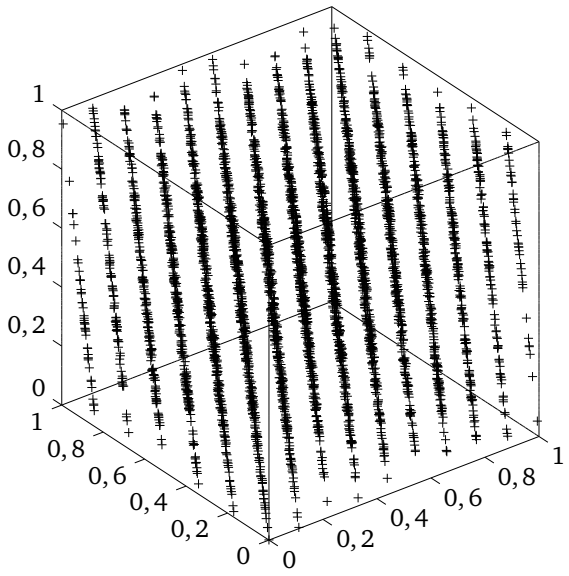


FIGURE 9.2 – Représentation de 4000 triplets de nombres pseudo-aléatoires consécutifs issus d'une suite de 12000 termes construite par le générateur RANDU.

l'entier k étant l'*ordre* du générateur. Si $c = 0$, le générateur est de type multiplicatif (on parle en anglais de *multiple recursive multiplicative congruential generator*) et la longueur maximale de sa période est égale à $m^k - 1$ lorsque le module m est un nombre premier et que le polynôme $z^k - (a_1 z^{k-1} + a_2 z^{k-2} + \cdots + a_k)$ satisfait certaines conditions. Les suites qu'il produit ne possèdent généralement pas de bien meilleures propriétés vis-à-vis de la corrélation que celles issues des générateurs simples. Il est néanmoins possible, en combinant plusieurs générateurs de ce type et moyennant un choix adapté de leurs paramètres, d'obtenir à un coût comparable des suites ayant un comportement satisfaisant une quantité raisonnable de tests statistiques [LEc96].

On remarquera qu'en annulant tous les coefficients, sauf deux que l'on prend égaux à 1, dans la relation de récurrence ci-dessus, on obtient des suites récurrentes rappelant la suite de Fibonacci, par exemple

$$x_n = x_{n-j} + x_{n-k} \mod m,\ 0 < j < k,\ n \geq k.$$

Diverses généralisations de cette relation de récurrence mènent à des familles de générateurs, portant en anglais le nom de *lagged Fibonacci generators*, définies par

$$x_n = x_{n-j} \star x_{n-k} \mod m,\ 0 < j < k,\ n \geq k, \tag{9.28}$$

où $\star$ désigne une loi de composition interne parmi l'addition, la multiplication ou encore l'opération bit à bit de ou exclusif (XOR) si le système de numération est binaire. Dans ce dernier cas, on dit que le générateur est basé sur un *registre à décalage à rétroaction* (*feedback shift register* en anglais) et fait partie de la classe de générateurs introduite par Tausworthe [Tau65] puis généralisée par Lewis et Payne [LP73]. Indiquons que le générateur *Mersenne twister* [MN98], particulièrement réputé pour ses qualités, est basé sur une modification de ce dernier type de générateur (*twisted generalized feedback shift register generator* en anglais). Tirant son nom de la longueur de sa période, qui est égale au nombre $2^{19937} - 1$, ce générateur produit des suites uniformément distribuées sur un très grand nombre de dimensions (623) tout en étant généralement plus rapide que la plupart des autres générateurs.

Indiquons pour conclure cette énumération que les générateurs présentés peuvent s'avérer suffisants pour la plupart des applications à l'exception de celles relatives à la cryptographie. Dans ce domaine, les générateurs doivent en effet pouvoir résister à des attaques en plus de satisfaire les tests statistiques classiques. Dans ce contexte particulier, la rapidité avec laquelle le générateur produit des nombres n'est pas primordiale et c'est l'imprévisibilité

des sorties qui prime. Parmi les générateurs que l'on peut qualifier de *cryptographiques*, c'est-à-dire particulièrement sûrs, on peut citer *Blum Blum Shub* [BBS86], dont la sécurité est assurée par la complexité théorique du *problème de la résidualité quadratique*[57].

Parlons à présent des tests statistiques empiriques complétant l'analyse mathématique théorique portant sur la longueur des périodes et l'uniformité de la distribution des termes des suites produites par un générateur de nombres pseudo-aléatoires. Ceux-ci vont en effet permettre d'assurer que les suites obtenues sont de bonnes imitations de réalisations d'une suite de variables aléatoires. Pour cela, on formule une hypothèse « nulle », qui postule que toute suite de nombres pseudo-aléatoires produite par un générateur est effectivement une réalisation d'une suite de variables aléatoires indépendantes et identiquement distribuées, de loi uniforme sur l'intervalle $[0,1]$. On sait que cette hypothèse est formellement fausse et un test statistique a pour but de le détecter à partir d'un nombre fini de terme de cette suite. S'il est bien sûr formellement impossible qu'un générateur passe tous les tests imaginables, un compromis heuristique est de se satisfaire d'un générateur qui réussit à un certain nombre de tests jugés « raisonnables ». Ainsi, on dira qu'un générateur est « mauvais » s'il échoue aux tests statistiques les plus simples, alors qu'un « bon » générateur les passera avec succès et ne sera mis en échec que par des tests élaborés. En pratique, différents tests statistiques visent à mettre en évidence différents types de défauts et des batteries de tests prédéfinis ont été réalisées, parmi lesquelles on peut citer DIEHARD de Marsaglia et TestU01 de L'Écuyer et Simard [LS07] (on pourra consulter cette dernière référence pour une description détaillée de toute une variété de tests statistiques).

On obtient des distributions autres qu'uniformes par des transformations :
variables aléatoires discrètes (A VOIR) : découpage d'intervalles, *méthode de rejet* (*rejection sampling* ou *acceptance-rejection method* en anglais), dont le principe date du problème de l'aiguille de Buffon, amélioration de la complexité : méthode de Walker [Wal77]

variables aléatoires continues : *méthode de la transformée inverse* (*inverse transform sampling* en anglais) utilisée lorsque l'on sait évaluer le réciproque de la fonction de répartition, ce qui n'est pas sans difficulté en pratique, méthode de rejet
méthode pour des lois spécifiques, reposent sur les propriétés des lois en question : Pour une distribution gaussienne centrée, il existe des méthodes algorithmiquement plus efficaces : rectangle-wedge-tail de Marsaglia [Mar61], la *méthode de Box*[58]*–Muller* [BM58] (méthode de rejet qui génère une paire de nombres aléatoires à distribution normale centrée réduite à partir d'une paire de nombres aléatoires à distribution uniforme), la *méthode polaire de Marsaglia* [MB64] (variante de la précédente méthode qui évite l'évaluation des fonctions trigonométriques cosinus et sinus) ou encore la *méthode ziggourat* [MT00] (dont le nom provient du fait qu'elle revient à recouvrir l'aire sous la courbe de la densité de la loi avec des rectangles empilés par ordre de taille décroissante, produisant une figure ressemblant à une *ziggourat*[59]).

Approximation d'un processus de Wiener

première méthode : Les variables aléatoires ΔW_n sont indépendantes et suivent une loi normale de moyenne nulle et de variance égale à $h_n = t_{n+1} - t_n$, i.e. $E(\Delta W_n) = 0$ et $E\left((\Delta W_n)^2\right) = h_n$, $\forall n = 0, \dots, N-1$.

inconvénient si l'on souhaite raffiner la subdivision de l'intervalle de temps (tous les calculs doivent être refaits)

seconde méthode ne possédant pas ce défaut : renormalisation de marches aléatoires, basé sur Donsker (A VOIR)

9.3.2 Méthode d'Euler–Maruyama

La méthode la plus simple est celle dite *d'Euler–Maruyama*[60] [Mar55].

57. Ce problème est celui de la distinction, à l'aide de calculs, des *résidus quadratiques modulo n* (c'est-à-dire des nombres possédant une racine carrée en arithmétique modulaire de module n), avec n un *nombre composé* (c'est-à-dire un entier naturel non nul possédant un diviseur strictement positif autre que 1 ou lui-même) fixé. Il est communément admis que ce problème est « difficile », au sens où sa résolution nécessite un nombre d'opérations arithmétiques grand par rapport à la taille de l'entier n.

58. George Edward Pelham Box (né le 18 octobre 1919) est un statisticien anglais. Il a apporté d'importantes contributions aux domaines de l'analyse des séries temporelles, de l'inférence bayésienne et du contrôle qualité.

59. Une ziggourat est un édifice religieux mésopotamien à degrés, constitué de plusieurs terrasses.

60. Gishirō Maruyama (丸山 儀四郎 en japonais, 4 avril 1916 - 5 juillet 1986) était un mathématicien japonais. Il est connu pour ses contributions à l'étude des processus stochastiques.

explications : comme pour les edo, consiste en l'approximation de la solution de l'équation différentielle stochastique aux instants discrets t_n (que l'on interpole si des valeurs à des temps intermédiaires sont requises)

$$X_{n+1} = X_n + f(t_n, X_n)(t_{n+1} - t_n) + g(t_n, X_n)(W(t_{n+1}) - W(t_n)),\ n \geq 0, \tag{9.29}$$

Obtenue en fixant les valeurs des intégrands sur chaque intervalle de discrétisation en temps à celle qu'ils prennent au début de celui-ci/à partir d'un développement d'Itō–Taylor au premier ordre (UTILISER (9.14))

$\hat{X}(t)$ valeur au temps t de l'interpolée P_1 par morceaux des valeurs approchées produites par la méthode :

$$\hat{X}(t) = X_i + \frac{t - t_i}{t_{i+1} - t_i}(X_{i+1} - X_i),\ \forall t \in [t_i, t_{i+1}[.$$

9.3.3 Différentes notions de convergence et de consistance

A VOIR erreur globale

$$E\left(\max_{i=0,\ldots,N} |X(t_i) - X_i|\right)$$

ou encore $E\left(\sup_{t\in[0,T]} \left|X(t) - \hat{X}(t)\right|\right)$.

Définition 9.25 (convergence forte) *on a convergence forte avec un ordre $p \in]0, +\infty]$ s'il existe une constante $K > 0$ et $\delta_0 > 0$ tels que*

$$E\left(\sup_{i=0,\ldots,N} |X(t_i) - X_i|\right) \leq K h^p,\ h \in]0, \delta_0[$$

NOTER que l'ordre p peut être fractionnaire car $\sqrt{E((W(t_{n+1}) - W(t_n)^2)}$ est d'ordre $\sqrt{h_n}$.

Dans le cas déterministe ($g \equiv 0$, cette notion de convergence coïncide avec celle introduite pour les approximations de solutions d'équations différentielles ordinaires (voir la définition 8.33). On a la notion de *consistante forte* associée suivante.

Définition 9.26 (consistance forte) $\exists c(h) \geq 0$ *telle que* $c(h) \to 0$ *quand* $h \to 0$*,*

$$E\left(\left|E\left(\frac{X_{n+1} - X_n}{h_n} \mid \mathscr{A}_{t_n}\right) - f(t_n, X_n)\right|^2\right) \leq c(h) \tag{9.30}$$

et

$$E\left(\frac{1}{h}\left|X_{n+1} - X_n - E\left(X_{n+1} - X_n \mid \mathscr{A}_{t_n}\right) - g(t_n, X_n)(W(t_{n+1}) - W(t_n))\right|^2\right) \leq c(h) \tag{9.31}$$

La condition (9.30) exprime que la moyenne des incréments de l'approximation converge vers celle du processus d'Itō lorsque la longueur du pas de discrétisation tend vers 0. En l'absence d'aléa, ceci correspond à la condition de consistance d'une méthode à un pas déterministe (ESSAYER DE FAIRE LE LIEN avec $\frac{1}{h_n}\tau_{n+1}$).

La condition (9.31) dit que la variance de la différence entre la partie aléatoire de l'approximation et celle du processus d'Itō tend vers 0 avec h.

Cette notion de consistance traduit donc la proximité des trajectoires des approximations de celles du processus d'Itō.

Lemme 9.27 *La méthode d'Euler–Maruyama est fortement consistante.*

DÉMONSTRATION. A ECRIRE □

Dans de nombreuses applications cependant, il n'est pas nécessaire d'avoir une approximation fidèle des trajectoires du processus d'Itō. On n'est souvent intéressé que par la *valeur d'une certaine fonction* du processus à l'instant final, comme celle des moments $E(X(T))$, $E(|X(T)|^2)$ ou, plus généralement, $E(\varphi(X(T)))$ pour une fonction φ donnée dans une classe particulière. Dans ce cas, il suffit de seulement bien approcher la distribution de probabilité de la variable aléatoire X_T et la convergence requise pour l'approximation est alors entendue dans un sens plus faible que celui vu plus haut.

Définition 9.28 (convergence faible) *On dit qu'on a convergence faible avec un ordre $p \in]0,+\infty[$ quand h tend vers 0 si, pour toute fonction ϕ appartenant à l'espace $\mathscr{C}_P^{2(p+1)}(\mathbb{R}^d,\mathbb{R})$ des fonctions $2(p+1)$ fois continûment différentiables qui ont, avec leurs dérivées jusqu'à l'ordre $2(p+1)$, une croissance polynomiale, il existe des constantes $K>0$ et $\delta_0>0$ telles que*

$$|E(\phi(X(T)))-E(\phi(X_N))| \le K\,h^p,\ \forall h \in]0,\delta_0[.$$

Définition 9.29 (consistance faible) $\exists c(h)\ge 0$ *telle que* $c(h)\to 0$ *quand* $h\to 0$, *que* (9.30) *est vérifiée et*

$$E\left(\left|E\left(\frac{1}{h_n}(X_{n+1}-X_n)^2 \mid \mathscr{A}_{t_n}\right)-(g(t_n,X_n))^2\right|^2\right) \le c(h) \tag{9.32}$$

La condition (9.32) traduit le fait que la variance de l'approximation doit être proche de celle du processus d'Itō.

??? Une manière naturelle de classer les méthodes numériques pour la résolution des équations différentielles stochastiques est de les comparer avec des approximations fortes et faibles obtenues en tronquant des formules d'Itō–Taylor.

un schéma pouvant converger en deux sens, il peut donc avoir deux ordres de convergence distincts. L'ordre de convergence forte est cependant parfois moindre dans le cas stochastique (par rapport au cas déterministe), essentiellement parce que les quantités $\sqrt{E((\Delta W(t_n))^)}$ sont d'ordre $h^{1/2}$ (voir plus bas). Nous allons ainsi montrer que la méthode d'Euler–Maruyama converge fortement à l'ordre $\frac{1}{2}$ et faiblement à l'ordre 1

cas autonome

preuve de convergence forte sous l'hypothèse de l'existence d'une unique solution forte du problème continu considéré

Théorème 9.30 (convergence forte de la méthode d'Euler–Maruyama) *sous hypothèses :*

$E(|X(t_0)|^2)<+\infty$, $\sqrt{E(|X(t_0)-X_0|^2)}\le C_1\sqrt{h}$, *condition de Lipschitz*

$$|f(t,x)-f(t,y)|+|g(t,x)-g(t,y)| \le C_2\,|x-y|,$$

croissance linéaire

$$|f(t,x)|+|g(t,x)| \le C_3\,(1+|x|),$$

$$|f(t,x)-f(s,x)|+|g(t,x)-g(s,x)| \le C_4\,(1+|x|)\,\sqrt{|t-s|},$$

$\forall t,s\in[t_0,t_0+T]$, $\forall x,y\in\mathbb{R}^d$, *où les constantes* $C_1,\dots,C_4$ *ne dépendant pas de h. Alors, on a l'estimation*

$$E\left(\sup_{i=0,\dots,N}|X(t_i)-X_i|\right) \le C_5\sqrt{h},$$

avec C_5 *indépendante de h, et la méthode d'Euler–Maruyama converge donc fortement à l'ordre* $\frac{1}{2}$.

DÉMONSTRATION. A ECRIRE □

Note : lorsque le bruit est additif, i.e. $g(t,x)=g(t)$, on a, en faisant des hypothèse de régularité sur f et g un résultat de convergence forte à l'ordre 1

Concernant la convergence faible de la méthode d'Euler–Maruyama, on a le résultat suivant dans le cas autonome.

Théorème 9.31 (convergence faible de la méthode d'Euler–Maruyama) *sous hypothèses ..., la méthode d'Euler–Maruyama converge faiblement à l'ordre un.*

DÉMONSTRATION. A ECRIRE □

A VOIR : Talay-Tubaro [TT90] (article sur la mise en œuvre de l'extrapolation à la Richardson via une expression pour le coefficient d'erreur principal d'E-M) ?

L'écart entre l'ordre de convergence forte de la méthode d'Euler–Maruyama et l'ordre de convergence de son analogue déterministe s'explique en observant qu'on utilise le calcul différentiel « habituel », et non le calcul d'Itō, dans l'établissement du schéma (9.29) à partir de l'équation eqref. Ce faisant, on ne tient pas compte du fait que

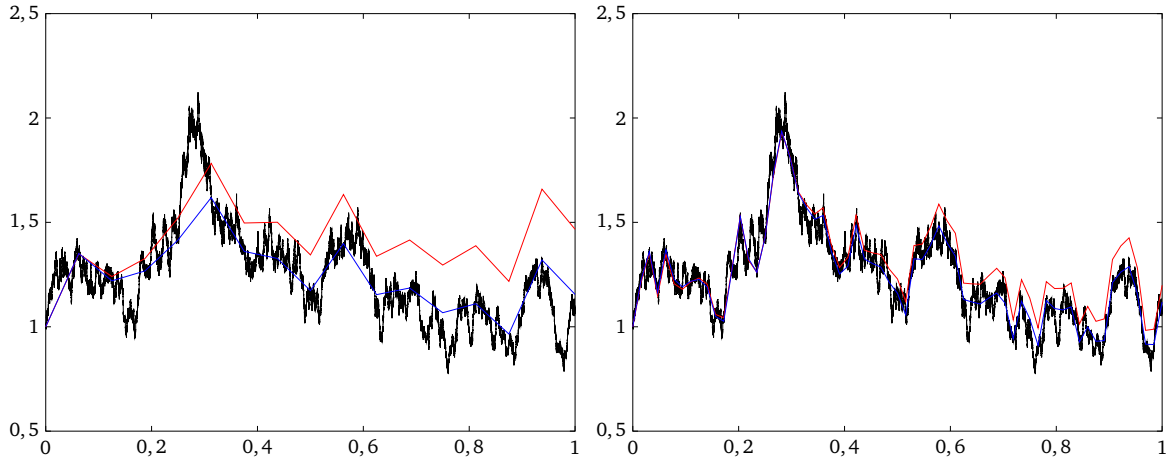


FIGURE 9.3 – Simulation d'une réalisation de la trajectoire du processus d'Itō solution du problème eqref, avec $a = 1,5$, $b = 1$ et $X_0 = 1$, sur l'intervalle de temps $[0,1]$ (en noir) et approximations numériques par les méthodes d'Euler–Maruyama (en rouge) et de Milstein (en bleu) pour $h = 2^{-4}$ (à gauche) et $h = 2^{-6}$ (à droite).

le mouvement brownien n'est pas à variation quadratique bornée, de qui conduit à négliger un terme d'ordre h, rendant ainsi impossible l'obtention d'un schéma fortement convergent à l'ordre un. Le prise en compte de ce terme conduit à l'obtention d'une méthode convergeant à un ordre supérieur (voir la section 9.3.5).

ILLUSTRER ORDRE DE CONVERGENCE, par méthode de Monte-Carlo : on effectue M calculs de trajectoires approchées pour différentes simulations/réalisation du mouvement brownien et on approche l'erreur globale par la moyenne empirique

$$\frac{1}{M}\sum_{k=1}^{M}\left(\sup_{i=0,\dots,N}\left|X(t_i)^{(k)} - X_i^{(k)}\right|\right).$$

on doit en théorie tenir compte de l'erreur statistique (décroît en $\frac{1}{\sqrt{M}}$), des erreurs inhérentes au générateur de nombres pseudo-aléatoires (problème d'indépendance des tirages quand h diminue), des erreurs d'arrondi propres à tout calcul numérique réalisé en arithmétique en précision finie

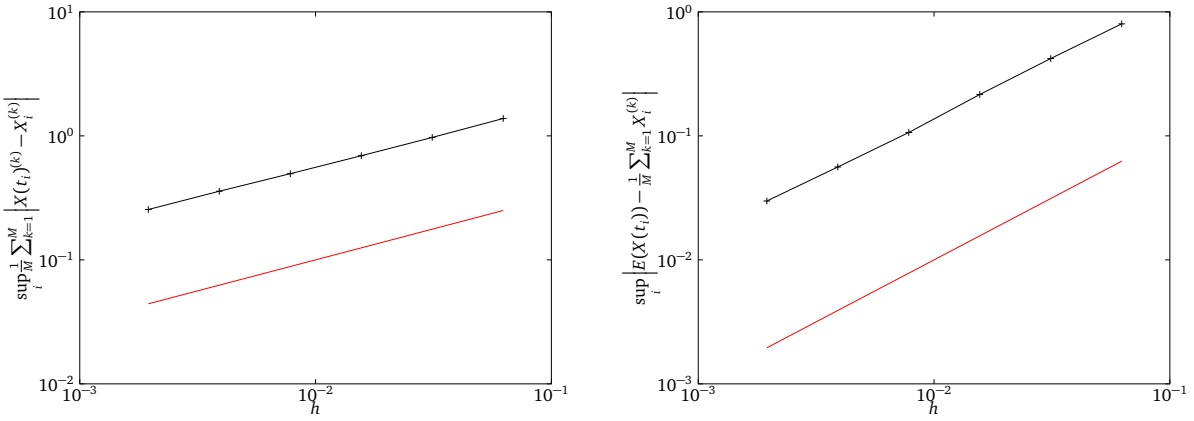


FIGURE 9.4 – illustration de la convergence forte et faible de la méthode d'Euler–Maruyama.

9.3.4 Stabilité

A ECRIRE

notion(s) de stabilité absolue pour ces méthodes

théorie *linéaire* comme pour les edo : on considère l'équation différentielle stochastique $\mathrm{d}X(t) = \lambda X(t)\,\mathrm{d}t + \mathrm{d}W(t)$, avec $\mathrm{Re}(\lambda) < 0$

9.3.5 Méthodes d'ordre plus élevé

INTRO A ECRIRE, méthodes de Taylor

Méthode de Milstein

ce schéma est basé sur un développement d'Itō–Taylor à l'ordre un : faire le calcul en conservant tous les termes d'ordre un et en déduire la *méthode de Milstein* [Mil75], faire aussi le lien avec les méthodes de Taylor pour les edo (voir la sous-section 8.3.4)

$$X_{n+1} = X_n + f(t_n,X_n)(t_{n+1}-t_n) + g(t_n,X_n)(W(t_{n+1})-W(t_n)) + \frac{1}{2}\,g(t_n,X_n)\,\frac{\partial g}{\partial x}(t_n,X_n)\big((W(t_{n+1})-W(t_n))^2-(t_{n+1}-t_n)\big) \quad (9.33)$$

Résultat de convergence à l'ordre 1 fort et faible (à démontrer)

Théorème 9.32 (convergence forte de la méthode de Milstein) *sous hypothèses, la méthode de Milstein converge fortement à l'ordre un.*

DÉMONSTRATION. A ECRIRE □

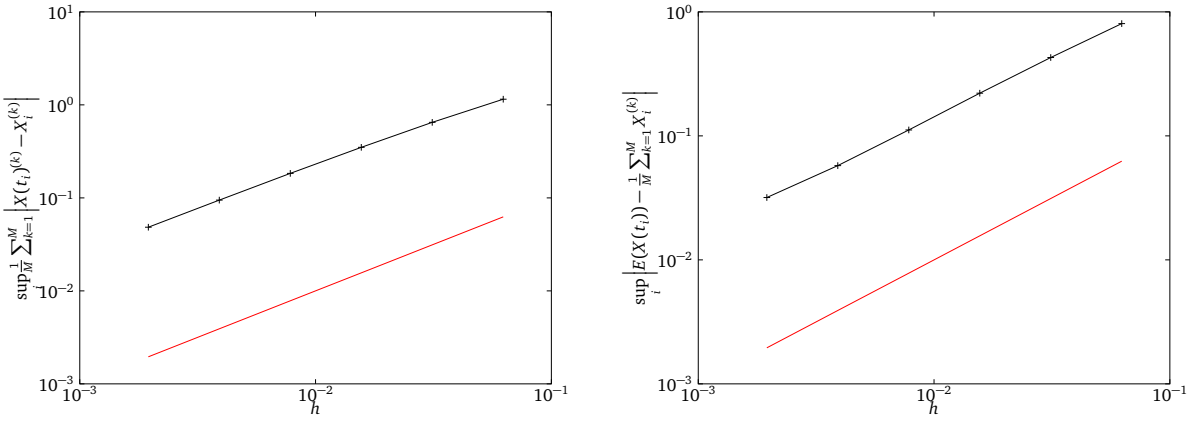


FIGURE 9.5 – illustration de la convergence forte et faible de la méthode de Milstein ($M = 100000$).

Théorème 9.33 (convergence faible de la méthode de Milstein) *sous hypothèses, la méthode de Milstein converge faiblement à l'ordre un.*

DÉMONSTRATION. A ECRIRE □

Méthodes de Runge–Kutta stochastiques

il ne suffit pas d'étendre formellement des méthodes de Runge–Kutta définies pour des équations différentielles ordinaires pour obtenir des méthodes convergeant vers la solution attendue... Application formelle peut conduire à des méthodes convergeant vers des limites qui ne sont pas solution au sens d'Itō de l'équation (voir [Wri74, Rüm82])

Méthodes multipas stochastiques

dérivées de manière heuristique et d'intérêt limité (a voir)...

A DEPLACER Remarques sur les schémas implicites : on n'implicite pas la partie brownienne méthode d'Euler–Maruyama implicite

$$X_{n+1} = X_n + f(t_{n+1}, X_{n+1})(t_{n+1} - t_n) + g(t_n, X_n)(W(t_{n+1}) - W(t_n)),\ n \geq 0,$$

9.4 Notes sur le chapitre

L'ouvrage de référence sur ce chapitre est sans nul doute le livre de Kloeden et Platen [KP99]. Cependant, on recommande également l'article de Higham [Hig01], pour une introduction rapide et pédagogique à la résolution numérique des équations différentielles stochastiques, et celui de Burrage, Burrage et Tian [BBT04], pour un tour d'horizon relativement complet et récent des développements de méthodes à un pas dans ce domaine.

Brown [61] fut l'un des premiers à observer le mouvement brownien, lors de l'étude de grains de pollen en suspension dans l'eau, en 1827 [Bro28]. En 1900, Bachelier [62], ayant perçu le caractère aléatoire des fluctuations des cours de la bourse, proposa dans sa thèse [Bac00] la première théorie mathématique de ce mouvement. Voulant tester la théorie cinétique moléculaire de la chaleur dans les liquides, Einstein [63], dans une série de trois articles publiés en 1905 et 1906 [Ein05, Ein06a, Ein06b], donna une théorie du mouvement brownien et montra comment ses mesures pouvaient conduire à une détermination précises des dimensions moléculaires et du *nombre* [64] *d'Avogadro* [65] ; ce programme, réalisé expérimentalement par Perrin [66] en 1908 [Per09], permit d'établir définitivement l'existence des atomes et des molécules. De manière indépendante, von Smoluchowski [67] mis à profit sa conception du mouvement brownien en termes d'une théorie cinétique pour le décrire comme une limite de promenades aléatoires [vSmo06]. Ce n'est que près d'une vingtaine d'années plus tard que Wiener construisit de manière rigoureuse, en s'appuyant sur la théorie de la mesure et l'analyse harmonique, un objet mathématique décrivant le phénomène [Wie23].

Le nom de méthode de Monte-Carlo, qui fait allusion aux jeux de hasard pratiqués dans le célèbre casino d'un des quartiers de la cité-État de la principauté de Monaco, a été inventé en 1947 par Metropolis [68] et publié pour la première fois en 1949 dans un article écrit avec Ulam [69] [MU49].

Une référence incontournable sur les générateurs de nombres pseudo-aléatoires est le second volume de la monographie *The art of computer programming* de Knuth [70] [Knu97].

61. Robert Brown (21 décembre 1773 - 10 juin 1858) était un botaniste écossais. Il fit de nombreuses contributions à la taxinomie des plantes. Son usage pionnier du microscope le conduisit à découvrir le noyau des cellules et la cyclose, ainsi qu'à faire une des premières observations du mouvement portant aujourd'hui son nom.

62. Louis Jean-Baptiste Alphonse Bachelier (11 mars 1870 - 26 avril 1946) était un mathématicien français. Il est aujourd'hui considéré comme le fondateur des mathématiques financières, ayant introduit dans sa thèse l'utilisation du mouvement brownien en finance.

63. Albert Einstein (14 mars 1879 - 18 avril 1955) était un physicien théoricien ayant eu diverses nationalités. Il contribua de façon considérable au développement de la mécanique quantique et de la cosmologie par l'introduction de sa théorie de la relativité restreinte en 1905, qu'il étendit en une théorie de la gravitation en 1915. Il reçut le prix Nobel de physique en 1921, notamment pour son explication de l'effet photoélectrique.

64. Ce nombre est le nombre d'entités élémentaires (des atomes, des molécules, des ions par exemple) dans une *mole* de matière et correspond au nombre d'atomes contenus dans douze grammes de carbone 12 (un isotope stable du carbone de masse atomique égale à 12 u). Dans le système international d'unités, sa valeur recommandée est $6,02214179(30)\ 10^{23}\,\mathrm{mol}^{-1}$ [MTN08].

65. Lorenzo Romano Amedeo Carlo Avogadro, comte de Quaregna et de Cerreto, (9 août 1776 - 9 juillet 1856) était un physicien et chimiste italien. Il est connu pour ses contributions à la théorie atomique et moléculaire de la matière.

66. Jean Baptiste Perrin (30 septembre 1870 - 17 avril 1942) était un physicien, chimiste et homme politique français. Il reçut le prix Nobel de physique en 1926 pour ses travaux sur la discontinuité de la matière et sa découverte de l'équilibre de sédimentation.

67. Marian von Smoluchowski (28 mai 1872 - 5 septembre 1917) était un physicien polonais, pionnier de la physique statistique. On lui doit d'importants travaux sur la théorie cinétique des gaz, au nombre desquels figurent notamment une description du mouvement brownien et une explication du phénomène d'opalescence critique.

68. Nicholas Constantine Metropolis (11 juin 1915 - 17 octobre 1999) était un physicien américain. Il est connu pour son développement des méthodes de Monte-Carlo.

69. Stanisław Marcin Ulam (13 avril 1909 - 13 mai 1984) était un mathématicien américain d'origine polonaise, à l'origine de l'architecture des bombes thermonucléaires. Il fit des contributions à la théorie des ensembles, à la théorie ergodique, à la topologie algébrique.

70. Donald Ervin Knuth (né le 10 janvier 1938) est un informaticien américain. Il est un des pionniers de l'algorithmique et on lui doit de nombreuses contributions dans plusieurs branches de l'informatique théorique. Il est aussi l'auteur de l'interpréteur et langage TEX et du langage METAFONT, qui permettent la composition de documents, notamment scientifiques.

Références

[Bac00] L. BACHELIER. Théorie de la spéculation. *Ann. Sci. École Norm. Sup. (3)*, 17 :21-86, 1900. DOI : 10.24033/asens.476.

[BBS86] L. BLUM, M. BLUM, and M. SHUB. A simple unpredictable pseudo-random number generator. *SIAM J. Comput.*, 15(2):364–383, 1986. DOI: 10.1137/0215025.

[BBT04] K. BURRAGE, P. M. BURRAGE, and T. TIAN. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proc. Roy. Soc. London Ser. A*, 460(2041):373–402, 2004. DOI: 10.1098/rspa.2003.1247.

[BM58] G. E. P. BOX and M. E. MULLER. A note on the generation of random normal deviates. *Ann. Math. Statist.*, 29(2):610–611, 1958. DOI: 10.1214/aoms/1177706645.

[Bor09] É. BOREL. Les probabilités dénombrables et leurs applications arithmétiques. *Rend. Circ. Mat. Palermo*, 27(1) :247-271, 1909. DOI : 10.1007/BF03019651.

[Boy77] P. P. BOYLE. Options: a Monte Carlo approach. *J. Finan. Econ.*, 4(3):323–338, 1977. DOI: 10.1016/0304-405X(77)90005-8.

[Bro28] R. BROWN. A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants ; and on the general existence of active molecules in organic and inorganic bodies. *Edinburgh New Philos. J.*, 5:358–371, 1828. DOI: 10.1080/14786442808674769.

[BS73] F. BLACK and M. SCHOLES. The pricing of options and corporate liabilities. *J. Polit. Economy*, 81(3):637–654, 1973. DOI: 10.1086/260062.

[Can17] F. P. CANTELLI. Sulla probabilità come limite della frequenza. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 26(1):39–45, 1917.

[Car10] R. D. CARMICHAEL. Note on a new number theory function. *Bull. Amer. Math. Soc.*, 16(5):232–238, 1910. DOI: 10.1090/S0002-9904-1910-01892-9.

[Cha43] S. CHANDRASEKHAR. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15(1):1–89, 1943. DOI: 10.1103/RevModPhys.15.1.

[Che56] N. N. CHENTSOV. Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov–Smirnov tests. *Theory Probab. Appl.*, 1(1):140–144, 1956. DOI: 10.1137/1101013.

[Cie61] Z. CIESIELSKI. Hölder conditions for realizations of gaussian processes. *Trans. Amer. Math. Soc.*, 99(3):403–413, 1961. DOI: 10.1090/S0002-9947-1961-0132591-2.

[CIR85] J. C. COX, J. E. INGERSOLL, JR., and S. A. ROSS. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985. DOI: 10.2307/1911242.

[Don52] M. D. DONSKER. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.*, 23(2):277–281, 1952. DOI: 10.1214/aoms/1177729445.

[Ein05] A. EINSTEIN. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Physik*, 322(8):549–560, 1905. DOI: 10.1002/andp.19053220806.

[Ein06a] A. EINSTEIN. Eine neue Bestimmung der Moleküldimensionen. *Ann. Physik*, 324(2):289–306, 1906. DOI: 10.1002/andp.19063240204.

[Ein06b] A. EINSTEIN. Zur Theorie der Brownschen Bewegung. *Ann. Physik*, 324(2):371–381, 1906. DOI: 10.1002/andp.19063240208.

[FM86] G. S. FISHMAN and L. R. MOORE, III. An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31} - 1$. *SIAM J. Sci. Statist. Comput.*, 7(1):24–45, 1986. DOI: 10.1137/0907002.

[Gir60] I. V. GIRSANOV. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory Probab. Appl.*, 5(3):285–301, 1960. DOI: 10.1137/1105027.

[Gru73] A. GRUBE. Mehrfach rekursiv-erzeugte Pseudo-Zufallszahlen. *Z. Angew. Math. Mech.*, 53(12):T223–T225, 1973. DOI: 10.1002/zamm.197305312116.

[HD62] T. T. HULL and A. R. DOBELL. Random number generators. *SIAM Rev.*, 4(3):230–254, 1962. DOI: 10.1137/1004061.

[Hig01] D. J. HIGHAM. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001. DOI: 10.1137/S0036144500378302.

[Itô44] K. ITÔ. Stochastic integral. *Proc. Imp. Acad.*, 20(8):519–524, 1944. DOI: 10.3792/pia/1195572786.

[Knu97] D. E KNUTH. *Seminumerical Algorithms*, volume 2 of *The art of computer programming*. Addison-Wesley, third edition, 1997.

[KP99] P. E. KLOEDEN and E. PLATEN. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of mathematics*. Springer, corrected third printing edition, 1999. DOI: 10.1007/978-3-662-12616-5.

[Lan08] P. LANGEVIN. Sur la théorie du mouvement brownien. *C. R. Acad. Sci. Paris*, 146 :530-532, 1908.

[LEc96] P. L'ECUYER. Combined multiple recursive random number generators. *Operations Res.*, 44(5):816–822, 1996. DOI: 10.1287/opre.44.5.816.

[Leh51] D. H. LEHMER. Mathematical methods in large-scale computing units. In *Proceedings of a second symposium on large-scale digital calculating machinery*, pages 141–146. Harvard University Press, 1951. volume 26 of the annals of the computation laboratory of Harvard University.

[LP73] T. G. LEWIS and W. H. PAYNE. Generalized feedback shift register pseudorandom number algorithm. *J. ACM*, 20(3):456–468, 1973. DOI: 10.1145/321765.321777.

[LS07] P. L'ECUYER and R. SIMARD. TestU01: a C library for empirical testing of random number generators. *ACM Trans. Math. Software*, 33(4), 2007. DOI: 10.1145/1268776.1268777.

[Mar55] G. MARUYAMA. Continuous Markov processes and stochastic equations. *Rend. Circ. Mat. Palermo*, 4(1):48–90, 1955. DOI: 10.1007/BF02846028.

[Mar61] G. MARSAGLIA. Expressing a random variable in terms of uniform random variables. *Ann. Math. Statist.*, 32(3):894–898, 1961. DOI: 10.1214/aoms/1177704983.

[Mar64] G. MARSAGLIA. Random numbers fall mainly in the planes. *Proc. Nat. Acad. Sci. U.S.A.*, 61(1):25–28, 1964. DOI: 10.1073/pnas.61.1.25.

[MB64] G. MARSAGLIA and T. A. BRAY. A convenient method for generating normal variables. *SIAM Rev.*, 6(3):260–264, 1964. DOI: 10.1137/1006063.

[Mer73] R. C. MERTON. Theory of rational option pricing. *Bell J. Econ. Manage. Sci.*, 4(1):141–183, 1973. DOI: 10.2307/3003143.

[Mil75] G. N. MIL'SHTEJN. Approximate integration of stochastic differential equations. *Theory Probab. Appl.*, 19(3):557–562, 1975. DOI: 10.1137/1119062.

[MN98] M. MATSUMOTO and T. NISHIMURA. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, 1998. DOI: 10.1145/272991.272995.

[MT00] G. MARSAGLIA and W. W. TSANG. The ziggurat method for generating random variables. *J. Statist. Software*, 5(8):1–7, 2000. DOI: 10.18637/jss.v005.i08.

[MTN08] P. J. MOHR, B. N. TAYLOR, and D. B. NEWELL. CODATA recommended values of the fundamental physical constants: 2006. *Rev. Mod. Phys.*, 80(2):633–730, 2008. DOI: 10.1103/RevModPhys.80.633.

[MU49] N. METROPOLIS and S. ULAM. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44(247):335–341, 1949. DOI: 10.1080/01621459.1949.10483310.

[Per09] J. PERRIN. Mouvement brownien et réalité moléculaire. *Ann. Chim. Phys. (8)*, 18 :5-114, 1909.

[PM88] S. K. PARK and K. W. MILLER. Random number generators: good ones are hard to find. *Comm. ACM*, 31(10):1192–1201, 1988. DOI: 10.1145/63039.63042.

[PRB69] W. H. PAYNE, J. R. RABUNG, and T. P. BOGYO. Coding the Lehmer pseudo-random number generator. *Comm. ACM*, 12(2):85–86, 1969. DOI: 10.1145/362848.362860.

[PW82] E. PLATEN and W. WAGNER. On a Taylor formula for a class of Itô processes. *Probab. Math. Statist.*, 3(1):37–51, 1982.

[PWZ33] R. E. A. C. PALEY, N. WIENER, and A. ZYGMUND. Notes on random functions. *Math. Z.*, 37(1):647–668, 1933. DOI: 10.1007/BF01474606.

[Rüm82] W. RÜMELIN. Numerical treatment of stachastic differential equations. *SIAM J. Numer. Anal.*, 19(3):604–613, 1982. DOI: 10.1137/0719041.

[Str66] R. L. STRATONOVICH. A new representation for stochastic integrals and equations. *SIAM J. Control*, 4(2):362–371, 1966. DOI: 10.1137/0304028.

[Tau65] R. C. TAUSWORTHE. Random numbers generated by linear recurrence modulo two. *Math. Comput.*, 19(90):201–209, 1965. DOI: 10.1090/S0025-5718-1965-0184406-1.

[TT90] D. TALAY and L. TUBARO. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509, 1990. DOI: 10.1080/07362999008809220.

[Vas77] O. VASICEK. An equilibrium characterisation of the term structure. *J. Finan. Econ.*, 5(2):177–188, 1977. DOI: 10.1016/0304-405X(77)90016-2.

[vNeu51] J. von NEUMANN. Various techniques used in connection with random digits. In A. S. HOUSEHOLDER, G. E. FORSYTHE, and H. H. GERMOND, editors, *Monte Carlo Method*. Volume 12, Applied mathematics series, pages 36–38. National Bureau of Standards, 1951.

[vSmo06] M. von SMOLUCHOWSKI. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Ann. Physik*, 326(14):756–780, 1906. DOI: 10.1002/andp.19063261405.

[Wal77] A. J. WALKER. An efficient method for generating discrete random variables with general distributions. *ACM Trans. Math. Software*, 3(3):253–256, 1977. DOI: 10.1145/355744.355749.

[Wie23] N. WIENER. Differential space. *MIT J. Math. Phys.*, 2:131–174, 1923. DOI: 10.1002/sapm192321131.

[Wri74] D. J. WRIGHT. The digital simulation of stochastic differential equations. *IEEE Trans. Automat. Control*, 19(1):75–76, 1974. DOI: 10.1109/TAC.1974.1100468.

Chapitre 10

Méthodes de résolution des systèmes hyperboliques de lois de conservation

COMPLETER INTRO

on s'intéresse à la résolution de problèmes de Cauchy composés d'un système d'équations aux dérivées partielles de la forme

$$\frac{\partial \boldsymbol{u}}{\partial t}(t,\boldsymbol{x}) + \sum_{j=1}^{d} \left(A_j(\boldsymbol{u}) \frac{\partial \boldsymbol{u}}{\partial x_j} \right)(t,\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{x} \in \mathbb{R}^d,\ t > 0, \tag{10.1}$$

d'inconnue $\boldsymbol{u}$ une fonction à valeurs dans $\mathbb{R}^p$, avec A_j, $1 \leq j \leq d$, fonction régulière d'un sous-ensemble de $\mathbb{R}^p$ à valeurs dans $M_p(\mathbb{R})$, complété par la donnée d'une condition initiale

$$\boldsymbol{u}(0,\boldsymbol{x}) = \boldsymbol{u_0}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^d.$$

bien que systèmes d'EDP, proches des systèmes d'EDO

10.1 Généralités sur les systèmes hyperboliques

Définition 10.1 (système hyperbolique) *Le système d'équations* (10.1) *est dit* ***hyperbolique*** *dans un ensemble $\mathcal{U}$ de $\mathbb{R}^p$ si et seulement si la matrice $A(\boldsymbol{u},\boldsymbol{\alpha}) = \sum_{j=1}^{d} \alpha_j A_j(\boldsymbol{u})$ ne possède que des valeurs propres réelles et est diagonalisable pour tout vecteur $\boldsymbol{u}$ de $\mathcal{U}$ et tout $\boldsymbol{\alpha}$ de $\mathbb{R}^d$. Il est dit* ***strictement hyperbolique*** *si toutes les valeurs propres de $A(\boldsymbol{u},\boldsymbol{\alpha})$ sont de plus distinctes.*

Le sens du qualificatif « hyperbolique » n'apparaît pas clairement dans cette définition, qui concerne des systèmes d'équations aux dérivées partielles d'ordre un. Il provient en effet d'une classification particulière des équations aux dérivées partielles linéaires d'ordre *deux*, que l'on rappelle dans le préambule de cette partie. Nous aurons l'occasion de donner d'illustrer le lien entre ces deux types d'équations avec l'exemple de l'*équation des ondes* dans la sous-section 10.2.7.

IMPORTANCE de l'hyperbolicité pour le caractère bien posé [1] d'un problème de Cauchy

EXEMPLE à deux équations linéaires ($p = 2$ et $A(\boldsymbol{u}) \equiv A$) en une dimension d'espace ($d = 1$). : Considérons le cas où la matrice A n'est pas diagonalisable dans $\mathbb{R}$ mais dans $\mathbb{C}$: $A = P \left(\begin{smallmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{smallmatrix}\right)$. On peut supposer sans perte de généralité que $\mathrm{Im}(\lambda) < 0$ et on note $\overline{\boldsymbol{q}}$ un vecteur propre associée à la valeur propre $\overline{\lambda}$. En prenant $\boldsymbol{u}_0(x) = e^{\mathrm{i}kx}\overline{\boldsymbol{q}}$, on obtient que la solution du problème est donnée par $\boldsymbol{u}(t,x) = e^{\mathrm{Im}(\lambda)kx} e^{\mathrm{i}(\mathrm{Re}(\lambda)kt - kx)}\overline{\boldsymbol{q}}$. Il est clair que l'amplitude de la solution croît avec le temps t alors que la donnée $\boldsymbol{u}_0$ est bornée dans $L^2(\mathbb{R})$: on n'a pas dépendance continue par rapport à la donnée.

définition système hyperbolique linéaire/non-linéaire

De très nombreux exemples de systèmes hyperboliques résultent de l'écriture d'une *loi de conservation* (*conservation law* en anglais). Pour le voir, considérons un domaine arbitraire Ω de $\mathbb{R}^d$, de frontière $\partial\Omega$ suffisamment

1. renvoyer à la section 1.4.2 du premier chapitre

régulière pour que le vecteur normal $\boldsymbol{n}$, unitaire et orienté vers l'extérieur de Ω, existe en tout point du bord. Soit alors l'intégrale

$$\int_\Omega \boldsymbol{u}(t,\boldsymbol{x})\,\mathrm{d}x,$$

avec $\boldsymbol{u}$ une fonction de $[0,+\infty[\times\mathbb{R}^d$ à valeurs dans $\mathbb{R}^p$, représentant la quantité d'un champ (masse, quantité de mouvement, énergie...) contenue dans Ω à un instant donné $t \geq 0$. À tout moment, la variation de cette quantité est égale au flux de $\boldsymbol{u}$ à travers la frontière, ce que l'on résume dans l'équation de bilan

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \boldsymbol{u}(t,\boldsymbol{x})\,\mathrm{d}x + \sum_{j=1}^{d}\int_{\partial\Omega} \boldsymbol{f}_j(\boldsymbol{u})(t,\boldsymbol{x})n_j\,\mathrm{d}S = 0,\ t>0,$$

avec $\boldsymbol{f}_j \in \mathscr{C}^1(\mathbb{R}^p,\mathbb{R}^p)$. Le *théorème de la divergence* (ou *théorème d'Ostrogradski*[2]) permet de réecrire l'intégrale sur la frontière sous la forme d'une intégrale sur le domaine, conduisant à

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \boldsymbol{u}(t,\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \sum_{j=1}^{d}\int_\Omega \left(\frac{\partial \boldsymbol{f}_j(\boldsymbol{u})}{\partial x_j}\right)(t,\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0,\ t>0. \tag{10.2}$$

En supposant la fonction $\boldsymbol{u}$ est suffisamment régulière, on a alors en tout point

$$\frac{\partial \boldsymbol{u}}{\partial t}(t,\boldsymbol{x}) + \sum_{j=1}^{d}\left(\frac{\partial \boldsymbol{f}_j(\boldsymbol{u})}{\partial x_j}\right)(t,\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{x}\in\Omega,\ t>0, \tag{10.3}$$

ce que l'on peut encore écrire

$$\frac{\partial \boldsymbol{u}}{\partial t}(t,\boldsymbol{x}) + \sum_{j=1}^{d}\left(\boldsymbol{\nabla}\boldsymbol{f}_j(\boldsymbol{u})\frac{\partial \boldsymbol{u}}{\partial x_j}\right)(t,\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{x}\in\Omega,\ t>0. \tag{10.4}$$

On retrouve alors un système hyperbolique de la forme (10.1) en posant $A_j(\boldsymbol{u}) = \boldsymbol{\nabla}\boldsymbol{f}_j(\boldsymbol{u})$.

La loi de conservation décrite par l'équation (10.3) est dite sous forme *conservative*, par opposition à celle, dite forme *non conservative*, de l'équation (10.4). Dans la suite de ce chapitre, l'accent sera mis sur l'étude et la résolution numérique approchée de lois de conservation scalaires en une dimension d'espace, c'est-à-dire pour $p = d = 1$.

AJOUTER : systèmes symétriques, mentionner propriétés

10.2 Exemples de systèmes d'équations hyperboliques et de lois de conservation *

ondes de diverses natures, particules, etc.

De nombreux modèles de la physique ou de la mécanique des milieux continus font intervenir des équations hyperbolique non-linéaires sous forme conservative. Celles-ci traduisent en effet de manière fondamentale la conservation d'une quantité macroscopique (masse, quantité de mouvement, énergie totale... d'un fluide ou d'un solide) lorsque l'on néglige des phénomènes ayant lieu à une petite échelle microscopique (dûs à la viscosité, la capillarité ou encore la conduction thermique). COMPLETER les non-linéarités entraînent l'apparition de singularités (ondes de choc) en temps fini...

2. Mikhail Vasilyevich Ostrogradski (Михаил Васильевич Остроградский en russe, 24 septembre 1801 - 1er janvier 1862) était un mathématicien et physicien russe. Ses travaux portèrent notamment sur le calcul intégral, l'algèbre, la physique mathématique et la mécanique classique.

10.2.1 Équation d'advection linéaire **

Dans sa version linéaire en une dimension d'espace, l'équation de transport est l'une des équations de type hyperbolique les plus simples

$$\frac{\partial u}{\partial t}(t,x) + a\,\frac{\partial u}{\partial x}(t,x) = 0,\ x \in \mathbb{R},\ t > 0,$$

loi de conservation avec $f(u) = a\,u$, a : vitesse de propagation

A VOIR : intervient, seule ou couplée à d'autres équations, dans de nombreux modèles en physique : modèle de trafic routier, équations cinétiques, démographie ou renouvellement cellulaire...

10.2.2 Modèle de trafic routier *

REPRENDRE modèle macroscopique : On considère la partie d'une section d'autoroute correspondant à l'un des sens de parcours, et plus particulièrement une section sans bretelle d'accès ou de sortie. Le domaine d'étude étant supposé grand devant la taille des véhicules, on assimile le trafic au mouvement d'un milieu continu monodimensionnel. On désigne par $\rho(t,x)$ la densité de véhicules au temps t et au point d'abscisse x, et par $v(\rho)(t,x)$ la vitesse moyenne des véhicules en ce point.

Un modèle de trafic routier consiste à relier via une loi la densité de voitures et leur vitesse moyenne. On suppose que les conducteurs adaptent leur vitesse aux conditions de circulation de sorte que la vitesse est une fonction décroissante de la densité, et qu'il existe une valeur de saturation $\rho_{\max}$ pour laquelle v s'annule : l'espace des états est l'intervalle $[0, \rho_{\max}]$. La densité ρ satisfait une loi de conservation dont le flux est $f(\rho) = \rho\, v(\rho)$ qui est une fonction concave.

DONNER une forme typique pour le flux en donnant le modèle de Lighthill[3]–Whitham–Richards [LW55, Ric56], correspondant au choix

$$v(\rho) = v_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right)$$

A FAIRE : analyse qualitative

This model does not respect the fact that a car — in contrast to gas particles – is only stimulated by the traffic flow situation in front of it, whereas a gas particle is also influenced by the situation behind it. MODELE PLUS REALISTE [AR00]

10.2.3 Équation de Boltzmann en mécanique statistique **

En théorie cinérique des gaz, l'*équation de Boltzmann*[4] [Bol72] est une équation intégro-differentielle linéaire de type hyperbolique, décrivant l'évolution statistique d'un fluide (gaz) peu dense hors d'équilibre thermodynamique. Elle s'écrit

$$\frac{\partial f}{\partial t}(t,x,v) + v\,\frac{\partial f}{\partial x}(t,x,v) + \sigma(x,v) f(t,x,v) - (K * f)(t,x,v) = 0,$$

où f est la fonction de distribution à une particule du gaz, t est la variable de temps, x la variable d'espace, v la vitesse

terme de collisions, etc.

10.2.4 Équation de Burgers pour la turbulence

L'*équation de Burgers*[5] non visqueuse,

$$\frac{\partial u}{\partial t}(t,x) + u(t,x)\,\frac{\partial u}{\partial x}(t,x) = 0,$$

3. Sir Michael James Lighthill (23 janvier 1924 - 17 juillet 1998) était un mathématicien anglais, connu pour ses travaux de recherche novateurs sur les ondes en dynamique des fluides, et plus particulièrement dans le domaine de l'aéroacoustique.

4. Ludwig Eduard Boltzmann (20 février 1844 - 5 septembre 1906) était un physicien autrichien. Il est l'un des initiateurs de la mécanique statistique et fut un fervent défenseur de la théorie atomique de la matière.

5. Johannes Martinus Burgers (13 janvier 1895 - 7 juin 1981) était un physicien néerlandais. On lui attribue notamment l'invention d'une équation aux dérivées partielles en mécanique des fluides et celle d'un vecteur caractérisant la déformation d'un cristal engendrée par une dislocation.

ici écrite en une dimension d'espace et dans laquelle le champ u désigne la vitesse d'un fluide, constitue un prototype d'équation hyperbolique scalaire non linéaire. Elle correspond à un cas particulier de l'équation

$$\frac{\partial u}{\partial t}(t,x) + u(t,x)\,\frac{\partial u}{\partial x}(t,x) = \nu\,\frac{\partial^2 u}{\partial x^2}(t,x),$$

étudiée par Burgers dans le cadre de la modélisation de la turbulence [Bur48], pour lequel la *viscosité cinématique* du fluide ν a été négligée.

10.2.5 Système des équations de la dynamique des gaz en description eulérienne

Le système d'équations aux dérivées partielles gouvernant l'évolution d'un écoulement compressible de fluide non visqueux et sans conductivité thermique s'écrit, en description eulérienne et sous forme conservative,

$$\begin{aligned}
&\frac{\partial \rho}{\partial t}(t,\boldsymbol{x}) + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(\rho\, u_j)(t,\boldsymbol{x}) = 0\\
&\frac{\partial}{\partial t}(\rho\, u_i)(t,\boldsymbol{x}) + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}(\rho\, u_i u_j + p\,\delta_{ij})(t,\boldsymbol{x}) = 0,\ i = 1,2,3,\\
&\frac{\partial E}{\partial t}(t,\boldsymbol{x}) + \sum_{j=1}^{3} \frac{\partial}{\partial x_j}\left((E+p)u_i\right)(t,\boldsymbol{x}) = 0,
\end{aligned}$$

les champs ρ, $\boldsymbol{u}$, p et $E = \rho\left(\frac{\|\boldsymbol{u}\|^2}{2} + e\right)$ désignant respectivement la masse volumique, la vitesse, la pression et l'énergie totale par unité de volume du fluide, avec e l'énergie interne par unité de masse du fluide. Ces équations, établies par Euler en 1755 et publiées en 1757 [Eul57], expriment la conservation de la masse, de la quantité de mouvement et de l'énergie totale du fluide au sein de l'écoulement. Pour fermer ce système (on a en effet seulement cinq équations pour six inconnues), il faut prescrire une relation constitutive, une *équation d'état*, entre les variables décrivant les états d'équilibre du fluide vu comme un système thermodynamique. Dans le cas d'un *gaz parfait*, on utilise la relation

$$p = \rho e(\gamma - 1),$$

dérivant de la *loi des gaz parfaits*, la constante $\gamma > 1$ étant le rapport des *chaleurs spécifiques à pression constante C_P et à volume constant C_V* du gaz. Le système alors obtenu est hyperbolique.

10.2.6 Équations de Barré de Saint-Venant **

Barré de Saint-Venant[6] 1871

version unidimensionnelle des équations des écoulements en eau peu profonde (de)*shallow water equations* en anglais)

10.2.7 Équation des ondes linéaire *

INTRO

en dimension 1 :

$$\frac{\partial^2 u}{\partial t^2}(t,x) - c^2\,\frac{\partial^2 u}{\partial x^2}(t,x) = 0,\ x \in \mathbb{R},\ t > 0, \tag{10.5}$$

+ $u(0,x) =$ et $\frac{\partial u}{\partial t}(0,x) = ...$

Cette équation constitue le premier modèle de corde vibrante, énoncé par D'Alembert[7] en 1747 [DAl49].

...

6. Adhémar Jean Claude Barré de Saint-Venant (23 août 1797 - 6 janvier 1886) était un ingénieur, mécanicien et mathématicien français. Ses recherches portèrent notamment sur la résistance et l'élasticité des solides, ainsi que sur les écoulements dans les cours d'eaux.

7. Jean le Rond D'Alembert (16 novembre 1717 - 29 octobre 1783) était un mathématicien, physicien, philosophe, encyclopédiste et théoricien de la musique français. Il est célèbre pour avoir dirigé avec Denis Diderot l'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*.

On peut se ramener à un système du premier ordre en posant $v = \frac{\partial u}{\partial t}$ et $w = c\,\frac{\partial u}{\partial x}$, car on a alors

$$\begin{cases} \dfrac{\partial v}{\partial t} - c\,\dfrac{\partial w}{\partial x} = 0 \\ \dfrac{\partial w}{\partial t} - c\,\dfrac{\partial v}{\partial x} = 0 \end{cases}$$

la première équation étant une simple réécriture de l'équation (10.5) dans les nouvelles variables, la seconde étant la traduction dans ces mêmes variables du théorème de Schwarz pour la solution u supposée deux fois dérivable

donc $A = \begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix}$, premier exemple de système hyperbolique linéaire

équation des télégraphistes : modèle de ligne électrique développé par Heaviside vers 1880)

inconnues : tension U et intensité de courant I dans la ligne

$$\begin{cases} L\,\dfrac{\partial I}{\partial t} + \dfrac{\partial U}{\partial x} + R\,I = 0 \\ C\,\dfrac{\partial U}{\partial t}\,\dfrac{\partial I}{\partial x} + G\,U = 0 \end{cases}$$

avec R résistance linéique du conducteur, L inductance linéique, C capacité linéique, G conductance linéique + conditions initiales

10.2.8 Système des équations de Maxwell en électromagnétisme *

REPRENDRE ET COMPLETER Les *équations de Maxwell* [8] sont un système d'équations aux dérivées partielles traduisant les lois de base de l'électromagnétisme qui régissent l'électrodynamique et l'optique classiques. Elles décrivent les interactions entre l'induction magnétique $\boldsymbol{B}$, le champ électrique $\boldsymbol{E}$, le déplacement électrique $\boldsymbol{D}$ et le champ magnétique $\boldsymbol{H}$ via

loi de Faraday [9]

$$\frac{\partial \boldsymbol{B}}{\partial t}(t,\boldsymbol{x}) + \mathbf{rot}\,(\boldsymbol{E}(t,\boldsymbol{x})) = \mathbf{0} \tag{10.6}$$

loi d'Ampère [10]

$$\frac{\partial \boldsymbol{D}}{\partial t}(t,\boldsymbol{x}) - \mathbf{rot}\,(\boldsymbol{H}(t,\boldsymbol{x})) = \mathbf{0} \tag{10.7}$$

définition de l'opérateur rotationnel

on ferme le système avec des relations constitutives. Par exemple, pour un conducteur parfait, on a

$$\boldsymbol{D} = \varepsilon\,\boldsymbol{E} \text{ et } \boldsymbol{B} = \mu\,\boldsymbol{B},$$

avec ε la *permittivité diélectrique* et μ la *perméabilité magnétique* du milieu.

10.3 Problème de Cauchy pour une loi de conservation scalaire

Cette section est consacrée à la construction et à l'étude de solutions du problème de Cauchy suivant, basée sur une loi de conservation scalaire en une dimension d'espace [11],

$$\frac{\partial u}{\partial t}(t,x) + \frac{\partial f(u)}{\partial x}(t,x) = 0,\ t > 0,\ x \in \mathbb{R}, \tag{10.8}$$

$$u(0,x) = u_0(x),\ x \in \mathbb{R}, \tag{10.9}$$

8. James Clerk Maxwell (13 juin 1831 - 5 novembre 1879) était un physicien et mathématicien écossais. Il est principalement connu pour avoir unifié les équations de l'électricité, du magnétisme et de l'optique au sein d'une théorie consistante de l'électromagnétisme et pour avoir développé une méthode statistique de description en théorie cinétique des gaz.

9. Michael Faraday (22 septembre 1791 - 25 août 1867) était un physicien et chimiste britannique, connu pour ses travaux fondamentaux dans le domaine de l'électromagnétisme et de l'électrochimie.

10. André-Marie Ampère (20 janvier 1775 - 10 juin 1836) était un mathématicien et physicien français. Il inventa le premier télégraphe électrique et est généralement considéré comme le principal initiateur de la théorie de l'électromagnétisme avec ses travaux sur l'électrodynamique.

11. L'ensemble des résultats énoncés reste valable pour des problèmes posés en plusieurs dimensions d'espace.

les fonctions réelles f et u_0 étant données, pour lequel la théorie mathématique est aujourd'hui quasiment complète, à la différence de celle pour les systèmes de lois de conservation que nous ne ferons qu'évoquer dans ces pages. Indiquons que ce modèle peut être rendu plus général en considérant une dépendance explicite par rapport aux variables t et x du flux f ou encore un *terme source* de la forme $g(u(t,x),t,x)$ dans l'équation (10.8) (on parle dans ce dernier cas de *balance law* en anglais).

Afin de simplifier la présentation tout en facilitant l'introduction des principales notions, nous traitons dans un premier temps le cas d'une équation *linéaire*, avant de développer la théorie dans le cadre général.

10.3.1 Le cas linéaire

Pour un problème linéaire, la fonction f est de la forme $f(u) = a\,u$, avec a un réel non nul, et l'on est ramené à l'étude du problème de Cauchy suivant :

$$\frac{\partial u}{\partial t}(t,x) + a\,\frac{\partial u}{\partial x}(t,x) = 0,\ t>0,\ x\in\mathbb{R}, \tag{10.10}$$

$$u(0,x) = u_0(x),\ x\in\mathbb{R}. \tag{10.11}$$

Supposons la donnée initiale u_0 de classe $\mathscr{C}^1$ sur $\mathbb{R}$. La fonction u donnée par

$$u(t,x) = u_0(x - a\,t) \tag{10.12}$$

est alors de classe $\mathscr{C}^1$ sur $[0,+\infty[\times\mathbb{R}$ et satisfait à la fois l'équation (10.10), puisque l'on a

$$\frac{\partial u}{\partial t}(t,x) = -a\,u_0'(x-a\,t) \text{ et } \frac{\partial u}{\partial x}(t,x) = u_0'(x-a\,t),$$

et la condition initiale (10.11). Nous verrons dans la sous-section 10.3.2 que cette solution est l'unique *solution classique* du problème. Dans l'immédiat, nous allons simplement nous contenter de remarquer qu'elle vérifie un certain nombre de propriétés.

Tout d'abord, cette solution est *globale*, c'est-à-dire qu'elle est définie pour toute valeur positive de la variable t, et *constante* sur chacune des droites (parallèles et disjointes) du plan (x,t), d'équations respectives $x = a\,t + x_0$, $x_0\in\mathbb{R}$. Elle préserve de plus la régularité de la donnée initiale et, le cas échéant, sa norme L^q, $1\le q\le+\infty$, au sens où

$$\|u(t,\cdot)\|_{L^q(\mathbb{R})} = \|u_0\|_{L^q(\mathbb{R})},\ \forall t>0,$$

cette dernière propriété découlant de l'invariance par translation de l'intégrale. On peut s'attendre à ce qu'il en soit de même pour une donnée initiale moins régulière, la question étant alors de donner une signification appropriée à la notion de solution, l'équation (10.10) ne pouvant plus être entendue dans le sens habituel des fonctions continues (voir la sous-section 10.3.3). Enfin, l'information (liée à la solution) *se propage à vitesse finie*. La formule (10.12) montre en effet que, en tout point (t,x), la valeur $u(t,x)$ de la solution du problème dépend uniquement de la valeur de la donnée initiale au point $x - a\,t$.

REPRENDRE Le cas d'un système de p équations linéaires,

$$\frac{\partial \boldsymbol{u}}{\partial t}(t,x) + A\,\frac{\partial \boldsymbol{u}}{\partial x}(t,x) = \mathbf{0},\ t>0,\ x\in\mathbb{R}, \tag{10.13}$$

$$\boldsymbol{u}(0,x) = \boldsymbol{u}_0(x),\ x\in\mathbb{R}, \tag{10.14}$$

avec $\boldsymbol{u}$ une fonction à valeurs dans $\mathbb{R}^p$ et A une matrice d'ordre p, se traite de manière analogue en supposant que le système satisfait la propriété d'hyperbolicité introduite dans la définition 10.1, c'est-à-dire que la matrice A est diagonalisable et ne possède que des valeurs propres réelles. Il existe alors une matrice P inversible telle que $A = P\Lambda P^{-1}$, où Λ est une matrice diagonale contenant les valeurs propres λ_i, $i=1,\dots,p$, de A. En faisant le changement d'inconnue $\boldsymbol{v} = P^{-1}\boldsymbol{u}$, on obtient le problème

$$\frac{\partial \boldsymbol{v}}{\partial t}(t,x) + \Lambda\,\frac{\partial \boldsymbol{v}}{\partial x}(t,x) = \mathbf{0},\ t>0,\ x\in\mathbb{R},$$

$$\boldsymbol{v}(0,x) = P^{-1}\boldsymbol{u}_0(x),\ x\in\mathbb{R},$$

qui est un ensemble de p équations différentielles linéaires découplées, dont les solutions respectives sont donc explicitement connues. On a ainsi obtenu que :

$$\boldsymbol{u}(t,x) = P \begin{pmatrix} v_1(0, x-\lambda_1 t) \\ \vdots \\ v_p(0, x-\lambda_p t) \end{pmatrix}.$$

Lorsque le système est strictement hyperbolique, les valeurs propres de A sont distinctes et la valeur $\boldsymbol{u}(t,x)$ de la solution dépend alors des valeurs de la condition initiale en p points eux aussi distincts. On donne à cet ensemble le nom de *domaine de dépendance* de la solution du système hyperbolique linéaire et on le note

$$\mathscr{D}(t,x) = \{z \in \mathbb{R} \,|\, z = x - \lambda_i\, t,\ i = 1\ldots, p\}\,. \tag{10.15}$$

DEFINITION dans un cas plus général (non linéaire) : ensemble des points d'espace en lesquels un changement dans la donnée initiale est susceptible d'affecter la valeur de $u(t,x)$. Ce domaine est borné (délimité par des caractéristiques) pour les équations hyperbolique car l'information se propage à vitesse finie

A VOIR : possibilité d'imposer des condition aux limites de type Dirichlet

10.3.2 Solutions classiques

Revenons à présent à l'étude du problème de Cauchy pour une loi de conservation scalaire générale. Nous supposons à partir de maintenant la fonction f régulière (au moins de classe $\mathscr{C}^2$) et la donnée initiale u_0 *bornée*, c'est-à-dire qu'elle appartient à l'espace $L^\infty(\mathbb{R})$.

Nous commençons par donner une signification précise à la notion de solution telle qu'elle a été sous-entendue jusqu'à présent.

Définition 10.2 (solution classique) *On dit que u est une **solution classique** de l'équation* (10.8) *dans un ouvert $\mathscr{O}$ de $[0,+\infty[\times\mathbb{R}$ si c'est une fonction de classe $\mathscr{C}^1$ satisfaisant* (10.8) *point par point dans $\mathscr{O}$.*

Évidemment, on ne peut avoir de solution classique du problème de Cauchy (10.8)-(10.9) que si la donnée initiale est au moins de classe $\mathscr{C}^1$. Cependant, même lorsque c'est cas, il n'y a pas nécessairement existence globale de ce type de solution lorsque l'équation présente une non-linéarité. Pour le voir, nous allons généraliser la technique de construction de solution précédemment proposée, en montrant que celle-ci reste valable, au moins sur un intervalle de temps borné. Nous introduisons tout d'abord la notion de *caractéristique*.

Définition 10.3 (caractéristique) *Soit u une solution classique de l'équation* (10.8). *On appelle **caractéristique** associée à l'équation* (10.8) *toute courbe du plan (x,t) définie par une courbe intégrale de l'équation différentielle ordinaire*

$$x'(t) = f'(u(t,x(t))),\ t > 0. \tag{10.16}$$

*Le point $(x_0,0)$ en lequel une telle courbe coupe l'axe des abscisses du plan (x,t) est appelé le **pied** de la caractéristique.*

On notera que l'on a existence et unicité, au moins localement, de solutions des problèmes définissant les caractéristiques dès que les hypothèses du théorème de Cauchy–Lipschitz (voir le théorème 8.10) sont vérifiées. On voit là un lien fort entre la théorie des équations différentielles ordinaires, rappelée dans le chapitre 8, et la résolution des lois de conservation.

Le fait que les caractéristiques dépendent de la solution du problème que l'on cherche à résoudre semble *a priori* constituer une obstruction à leur détermination. La propriété suivante, qui repose sur une observation déjà faite dans le cas linéaire, montre qu'il n'en est rien.

Proposition 10.4 *Une solution classique de l'équation* (10.8) *est constante le long de toute caractéristique définie par l'équation différentielle* (10.16).

DÉMONSTRATION. Soit u une solution classique de l'équation (10.8). Le long d'une caractéristique, il vient

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(u(t,x(t))) = \frac{\partial u}{\partial t}(t,x(t)) + x'(t)\,\frac{\partial u}{\partial x}(t,x(t)) &= \frac{\partial u}{\partial t}(t,x(t)) + f'(u(t,x(t)))\,\frac{\partial u}{\partial x}(t,x(t)) \\ &= \frac{\partial u}{\partial t}(t,x(t)) + \frac{\partial f(u)}{\partial x}(t,x(t)),\ t>0,\end{aligned}$$

qui est nulle en vertu de (10.8). □

On déduit de ce résultat que les caractéristiques associées à l'équation (10.8) sont des droites, dont les pentes dépendent des valeurs initiales de la solution et, par conséquent, de la donnée initiale u_0. Le long d'une caractéristique, on a en effet

$$u(t, x(t)) = u(0, x(0)),\ t \geq 0,$$

d'où

$$x'(t) = f'(u(0, x(0))) = f'(u_0(x(0)),\ t > 0,$$

la caractéristique issue de $(x_0, 0)$ ayant alors pour équation

$$x(t) = x_0 + f'(u_0(x_0))\, t,\ t \geq 0.$$

Pour résoudre le problème (10.8)-(10.9), il suffit donc de déterminer les caractéristiques qui lui sont associées, la valeur d'une solution classique u en un point (t, x) étant celle de la donnée initiale au pied de la caractéristique passant par ce point .

Ce procédé de construction explicite de solutions classiques porte le nom de *méthode des caractéristiques*. On remarque que pour que solution soit définie sans ambiguïté, il faut qu'une – et une seule – caractéristique passe en chaque point du plan. Or, pour une fonction f quelconque, ces caractéristiques ne sont pas forcément parallèles [12]. En effet, supposons qu'il existe deux réels x_1 et x_2 tels que $x_1 < x_2$ et

$$\frac{1}{f'(u_0(x_1))} < \frac{1}{f'(u_0(x_2))}.$$

On sait d'après la précédente proposition que la solution est constante sur toute caractéristique et donc, en particulier, sur les droites d'équations respectives

$$x(t) = x_1 + f'(u_0(x_1))\, t \text{ et } x(t) = x_2 + f'(u_0(x_2))\, t,\ t \geq 0.$$

Or, ces deux droites se croisent (voir la figure 10.1), ce qui implique que la solution doit prendre à la fois la valeur $u_0(x_1)$ et la valeur $u_0(x_2)$ en leur point d'intersection. Elle ne peut donc y être continue et cesse alors d'être une solution classique.

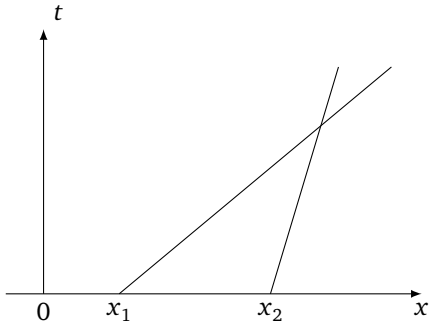


FIGURE 10.1 – Exemple de caractéristiques se croisant pour un problème de Cauchy (10.8)-(10.9) de flux non linéaire.

Insistons sur le fait que ce phénomène d'apparition de discontinuités en temps fini est de nature purement non linéaire et n'est pas lié à la régularité de la donnée initiale. On peut résumer ces constatations dans la proposition suivante.

Proposition 10.5 (existence d'une solution classique globale) *Supposons que la fonction f soit de classe $\mathscr{C}^2$ et que la fonction u_0 soit de classe $\mathscr{C}^1$, bornée ainsi que sa dérivée première. Il existe un temps maximal $T^* > 0$ tel que le problème* (10.8)-(10.9) *admet une unique solution classique sur $[0, T^*[\times\mathbb{R}$, donnée par la méthode des caractéristiques. De plus, on a*

$$T^* = +\infty$$

12. On se rappelle qu'il en était ainsi pour les caractéristiques associées à l'équation linéaire (10.10).

si la fonction $f''(u_0)u_0{}'$ est positive sur $\mathbb{R}$,

$$T^* = \inf_{x\in\mathbb{R}} \left\{ -\frac{1}{f''(u_0(x))\,u_0'(x)} \right\}$$

sinon.

DÉMONSTRATION. A REPRENDRE Les caractéristiques associées au problème existent, les hypothèses du théorème de Cauchy–Lipschitz étant vérifées. On a donc $x(t) = x_0 + f'(u_0(x_0))\,t$ pour toute caractéristique issue de x_0 et par conséquent $u(t,x) = u_0(x_0)$ avec x_0 solution de $x = x_0 + f'(u_0(x_0))\,t$. Considérons alors, pour tout $t \in \mathbb{R}_+$, la fonction $\phi_t : y \mapsto y + f'(u_0(y))\,t$. Celle-ci est dérivable sur $\mathbb{R}$ et $\phi_t'(y) = 1 + f''(u_0(y))u_0'(y)$. Si $f''(u_0(y))u_0'(y) \geq 0$ pour tout $y \in \mathbb{R}$, ϕ_t est strictement croissante sur $\mathbb{R}$ et, u_0 étant bornée et f' continue, telle que $\lim_{y\to\pm\infty} \phi_t(y) = \pm\infty$, $\forall t$. Elle définit donc une bijection sur $\mathbb{R}$ et les caractéristiques ne peuvent se croiser. Sinon, il existe au moins une valeur de y telle que $t = -\frac{1}{f''(u_0(y))u_0'(y)}$ annulant ϕ_t' et le temps maximal T^* est alors donné par eqref

UNICITE

Pour tout $t \geq 0$, l'application $\xi \to \xi + f'(u_0(\xi))\,t$ définit une bijection $\mathbb{R}$ dans $\mathbb{R}$. En effet, sa dérivée $1 + f''(u_0(\xi))\,u_0'(\xi)$ est, par hypothèse, strictement positive $\forall\xi$ et elle a pour limite $\pm\infty$ en $\pm\infty$. Par conséquent, $\forall(t,x)$, il existe un unique $\xi(t,x)$ tel que

$$x = \xi(t,x) + a(u_0(\xi(t,x)))\,t. \tag{10.17}$$

Si la solution classique existe, elle est donnée par $u(t,x) = u_0(\xi(t,x))$ et l'on a

$$\frac{\partial u}{\partial t}(t,x) + a(u(t,x))\,\frac{\partial u}{\partial x}(t,x) = u_0(\xi(t,x))\left(\frac{\partial \xi}{\partial t}(t,x) + f'(u_0(\xi(t,x))\,\frac{\partial u}{\partial x}(t,x)\right).$$

En dérivant (10.17), il vient

$$\frac{\partial \xi}{\partial t}(t,x)\big(1 + a'(u_0(\xi(t,x)))u_0'(\xi(t,x))\big) = -a(u_0(\xi(t,x))),$$

$$\frac{\partial \xi}{\partial x}(t,x)\big(1 + a'(u_0(\xi(t,x)))u_0'(\xi(t,x))\big) = 1,$$

et $a'(u_0(\xi(t,x)))u_0'(\xi(t,x)) \geq 0$ par hypothèse. L'équation (10.8) est donc bien vérifiée le long d'une caractéristique pour tout $t \geq 0$. □

Dans le cas où la fonction f est supposée convexe, la condition d'existence globale énoncée ci-dessus devient simplement la croissance de la fonction u_0.

10.3.3 Solutions faibles

Nous venons de voir qu'il peut ne pas exister de solution classique de l'équation (10.8) pour tout temps. Pour prolonger l'existence d'une solution du problème au delà de l'instant d'apparition d'une discontinuité, il s'avère donc nécessaire d'étendre le concept de solution. Pour cela, nous allons faire appel au formalisme des distributions, afin de donner un sens plus « faible » à cette notion. Ce faisant, nous rendrons aussi possible la présence de discontinuités dans la donnée initiale.

Pour justifier ce procédé, il faut voir l'équation (10.8) comme l'expression d'une loi de conservation, qui est elle-même la conséquence d'une relation intégrale de la forme (10.2) et qui reste valable pour une fonction discontinue. L'idée est par conséquent de définir une solution du problème en faisant porter les dérivées partielles dans l'équation (10.8) non pas sur la fonction u, mais sur un ensemble de fonctions tests régulières. Dans toute la suite, on désigne par L^∞_{loc} l'ensemble des fonctions à valeurs réelles mesurables *localement bornées* [13].

Définition 10.6 (solution faible) *Soit u_0 une fonction de $L^\infty_{loc}(\mathbb{R})$. Une fonction u de $L^\infty_{loc}([0,+\infty[\times\mathbb{R})$ est appelée une **solution faible** du problème* (10.8)-(10.9) *si elle satisfait*

$$\int_0^{+\infty}\int_{\mathbb{R}} \left(u(t,x)\,\frac{\partial \varphi}{\partial t}(t,x) + f(u)(t,x)\frac{\partial \varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\varphi(0,x)\,\mathrm{d}x = 0, \tag{10.18}$$

pour toute fonction test φ de classe $\mathscr{C}^1$ à support [14] *compact dans $[0,+\infty[\times\mathbb{R}$.*

13. Pour tout ouvert Ω de $\mathbb{R}^d$, on définit cet ensemble par

$$L^\infty_{\text{loc}}(\Omega) = \left\{ v : \Omega \to \mathbb{R} \text{ mesurable} \,|\, v_{|K} \in L^\infty(K),\ \forall K \subset \Omega,\ K \text{ compact} \right\}.$$

14. On rappelle que le *support* d'une fonction est l'adhérence de l'ensemble des points en lesquels cette fonction est non nulle.

On voit immédiatement que la notion de solution faible permet d'étendre directement l'utilisation de la méthode des caractéristiques au cas d'une donnée initiale discontinue, la solution résultante satisfaisant le problème au sens de la relation (10.18).

De plus, on remarque, en choisissant la fonction test dans l'ensemble $\mathscr{C}_0^\infty([0,+\infty[\times\mathbb{R})$ des fonctions indéfiniment dérivables à support compact dans $[0,+\infty[\times\mathbb{R}$, que toute solution faible du problème (10.8)-(10.9) satisfait l'équation (10.8) au sens des distributions. Cette observation permet d'établir le résultat suivant, qui montre que la notion de solution faible étend celle de solution classique.

Lemme 10.7 (lien entre les notions de solution classique et de solution faible) *Une solution classique du problème* (10.8)-(10.9) *est aussi une solution faible de ce problème. Réciproquement, une solution faible du problème appartenant à* $\mathscr{C}^1(]0,+\infty[\times\mathbb{R})\cap\mathscr{C}^0([0,+\infty[\times\mathbb{R})$ *est une solution classique.*

DÉMONSTRATION. Soit u une solution classique du problème (10.8)-(10.9) et φ une fonction de classe $\mathscr{C}^1$ à support compact dans $[0,+\infty[\times\mathbb{R}$. Multiplions l'équation (10.8) par φ et intégrons par parties sur $[0,+\infty[\times\mathbb{R}$. Il vient alors, en utilisant (10.9),

$$\begin{aligned}0=\int_0^{+\infty}\int_{\mathbb{R}}\left(\frac{\partial u}{\partial t}(t,x)+\frac{\partial f(u)}{\partial x}(t,x)\right)\varphi(t,x)\,\mathrm{d}x\,\mathrm{d}t&\\ =-\int_0^{+\infty}\int_{\mathbb{R}}\left(u(t,x)\,\frac{\partial\varphi}{\partial t}(t,x)+f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right)\mathrm{d}x\,\mathrm{d}t-\int_{\mathbb{R}}u_0(x)\varphi(0,x)\,\mathrm{d}x,&\end{aligned}$$

de sorte que u est une solution faible du problème.
Réciproquement, si u est une solution faible du problème appartenant à $\mathscr{C}^1(]0,+\infty[\times\mathbb{R})\cap\mathscr{C}^0([0,+\infty[\times\mathbb{R})$, on a, pour toute fonction test φ de $\mathscr{C}_0^\infty(]0,+\infty[\times\mathbb{R})$,

$$\int_0^{+\infty}\int_{\mathbb{R}}\left(u(t,x)\,\frac{\partial\varphi}{\partial t}(t,x)+f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right)\mathrm{d}x\,\mathrm{d}t=0,$$

ce qui signifie encore que u satisfait l'équation (10.8) au sens des distributions dans $]0,+\infty[\times\mathbb{R}$, et donc point par point puisqu'elle est de classe $\mathscr{C}^1$ sur ce domaine. Soit alors une fonction φ de $\mathscr{C}_c^1([0,+\infty[\times\mathbb{R})$. En réalisant une intégration par parties dans (10.18) et se servant de la précédente conclusion, on obtient

$$\int_{\mathbb{R}}(u_0(x)-u(0,x))\varphi(0,x)\,\mathrm{d}x=0.$$

Le choix de la fonction φ étant arbitraire et la fonction u étant continue, on en déduit que la condition initiale (10.9) est vérifiée point par point. □

Nous allons maintenant nous intéresser à une classe particulière de solutions faibles.

Définition 10.8 (fonction de classe $\mathscr{C}^1$ par morceaux) *Une fonction u définie sur $[0,+\infty[\times\mathbb{R}$ est dite* ***de classe $\mathscr{C}^1$ par morceaux*** *si, sur tout ouvert borné $\mathscr{O}$ de $[0,+\infty[\times\mathbb{R}$, il existe un nombre fini de courbes $\Sigma_1,\ldots,\Sigma_p$, de paramétrisations respectives $(t,\xi_i(t))$, $t\in]t_i^+,t_i^-[$, $i=1,\ldots,p$, avec ξ_i une fonction de classe $\mathscr{C}^1$, telles que u est de classe $\mathscr{C}^1$ dans chaque composante connexe de $\mathscr{O}\setminus(\Sigma_1\cup\cdots\cup\Sigma_p)$.*

Voyons comment cet ensemble de fonctions permet de construire des solutions faibles du problème (10.8)-(10.9) en utilisant la méthode des caractéristiques. Pour cela, nous supposons que la donnée initiale u_0 est à présent de classe $\mathscr{C}^1$ par morceaux sur $\mathbb{R}$.

Théorème 10.9 (conditions nécessaires et suffisantes de caractérisation d'une solution faible de classe $\mathscr{C}^1$ par morceaux) *Une fonction u de classe $\mathscr{C}^1$ par morceaux dans $\mathbb{R}\times[0,+\infty[$ est une solution faible du problème* (10.8)-(10.9) *si et seulement si*

- *c'est une solution classique de* (10.8)-(10.9) *dans tout domaine où elle est de classe $\mathscr{C}^1$,*
- *elle satisfait la condition*

$$\xi'(u^+-u^-)=f(u^+)-f(u^-) \tag{10.19}$$

le long de toute courbe de discontinuité Σ de la fonction u paramétrée par la fonction ξ, la valeur u^- (resp. u^+) désignant la limite de u à gauche (resp. à droite) de Σ.

DÉMONSTRATION. Considérons un ouvert D de $[0,+\infty[\times\mathbb{R}$ que la courbe de discontinuité Σ partage en deux parties (voir la figure 10.2), respectivement notées D^- et D^-, avec

$$D^- = \{(t,x)\in D \,|\, x<\xi(t)\} \text{ et } D^+ = \{(t,x)\in D \,|\, x>\xi(t)\}.$$

On note $\boldsymbol{n}$ le vecteur normal unitaire à Σ, orienté dans la direction des valeurs croissantes de la variable x. Compte tenu de la paramétrisation de la courbe Σ, on a

$$\boldsymbol{n} = \begin{pmatrix} n_t \\ n_x \end{pmatrix} = \frac{1}{\sqrt{1+(\xi')^2}} \begin{pmatrix} -\xi' \\ 1 \end{pmatrix}.$$

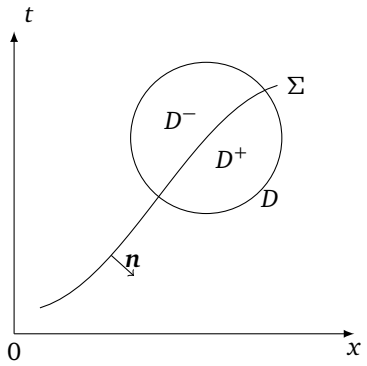


FIGURE 10.2 – Représentation dans le plan (x,t) de la courbe de discontinuité Σ, des ouverts D, D^- et D^+, ainsi que du vecteur normal unitaire $\boldsymbol{n}$ introduits dans la preuve du théorème 10.9.

Soit φ une fonction de $\mathscr{C}_0^\infty(D)$. Si la fonction u est une solution faible du problème, on a

$$\int_D \left(u(t,x)\frac{\partial\varphi}{\partial t}(t,x) + f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t = 0.$$

Décomposons cette intégrale en la somme de deux intégrales, l'une sur D^-, l'autre sur D^+. La fonction u étant régulière sur ces sous-ensembles, on a

$$\begin{aligned}\int_{D^-} \left(u(t,x)\frac{\partial\varphi}{\partial t}(t,x) + f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t = &-\int_{D^-} \left(\frac{\partial u}{\partial t}(t,x) + \frac{\partial f(u)}{\partial x}(t,x)\right)\varphi(t,x)\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{\Sigma\cap D} \big(u^-(t,x)n_t(t,x) + f(u^-)(t,x)n_x(t,x)\big)\varphi(t,x)\,\mathrm{d}\sigma,\end{aligned}$$

et

$$\begin{aligned}\int_{D^+} \left(u(t,x)\frac{\partial\varphi}{\partial t}(t,x) + f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t = &-\int_{D^+} \left(\frac{\partial u}{\partial t}(t,x) + \frac{\partial f(u)}{\partial x}(t,x)\right)\varphi(t,x)\,\mathrm{d}x\,\mathrm{d}t \\ &- \int_{\Sigma\cap D} \big(u^+(t,x)n_t(t,x) + f(u^+)(t,x)n_x(t,x)\big)\varphi(t,x)\,\mathrm{d}\sigma.\end{aligned}$$

On déduit alors que

$$\begin{aligned}\int_D \left(u(t,x)\frac{\partial\varphi}{\partial t}(t,x) + f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t = &-\int_{D^-\cup D^+} \left(\frac{\partial u}{\partial t}(t,x) + \frac{\partial f(u)}{\partial x}(t,x)\right)\varphi(t,x)\,\mathrm{d}x\,\mathrm{d}t \\ &- \int_{\Sigma\cap D} \big((u^+-u^-)(t,x)n_t(t,x) + (f(u^+)-f(u^-))(t,x)n_x(t,x)\big)\varphi(t,x)\,\mathrm{d}\sigma, \quad (10.20)\end{aligned}$$

d'où

$$\int_{\Sigma\cap D} \big((u^+-u^-)(t,x)n_t(t,x) + (f(u^+)-f(u^-))(t,x)n_x(t,x)\big)\varphi(t,x)\,\mathrm{d}\sigma = 0,$$

puisque u vérifie l'équation (10.8) au sens classique dans D^- et D^+. Cette égalité étant vraie quelle que soit la fonction test φ, on en déduit, en utilisant les expressions des composantes du vecteur normal $\boldsymbol{n}$, que

$$\xi'(t)(u^+-u^-)(t,\xi(t)) = (f(u^+)-f(u^-))(t,\xi(t)).$$

Réciproquement, si la fonction u de classe $\mathscr{C}^1$ par morceaux satisfait les deux conditions du théorème, on vérifie qu'elle est solution du problème au sens des distributions. □

Ce résultat montre que, pour être admissibles, les discontinuités d'une solution faible doivent satisfaire l'équation (10.19). Cette dernière est appelée *condition de Rankine* [15]*–Hugoniot* [16], par analogie avec une condition connue de longue date en dynamique des gaz [Ran70, Hug87]. On la résume souvent, en notant respectivement $[u]_{|_\Sigma} = u^+ - u^-$ et $[f(u)]_{|_\Sigma} = f(u^+) - f(u^-)$ les *sauts* des fonctions u et $f(u)$ au travers de la courbe Σ et en posant $\sigma = \xi'$, par la relation

$$\sigma\,[u]_{|_\Sigma} = [f(u)]_{|_\Sigma}\,, \tag{10.21}$$

vérifiée en tout point de Σ. On interprète alors la fonction σ comme la *vitesse de propagation* de la discontinuité.

Exemple de construction d'une solution faible de l'équation de Burgers non visqueuse. Considérons le problème de Cauchy composé de l'équation de Burgers non visqueuse

$$\frac{\partial u}{\partial t}(t,x) + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)(t,x) = 0,\ t>0,\ x\in\mathbb{R}, \tag{10.22}$$

et d'une condition initiale dont la donnée est une fonction continue

$$u(0,x) = u_0(x) = \begin{cases} 1 & \text{si } x<0,\\ 1-\dfrac{x}{\alpha} & \text{si } 0\le x\le\alpha,\\ 0 & \text{si } x>\alpha, \end{cases} \tag{10.23}$$

avec α un réel strictement positif. Nous allons construire une solution faible de ce problème.
Commençons par déterminer le temps maximal d'existence d'une solution continue. La caractéristique issue du pied $(x_0,0)$ a pour équation

$$\begin{cases} x(t) = t + x_0 & \text{si } x_0\le 0,\\ x(t) = \left(1-\dfrac{x_0}{\alpha}\right)t + x_0 & \text{si } 0\le x_0\le\alpha,\\ x(t) = x_0 & \text{si } x_0\ge\alpha, \end{cases}$$

et l'on a un croisement de caractéristiques au temps $T^* = \alpha$ (voir la figure 10.3). Nous pouvons alors obtenir la solution (classique) du problème pour $0\le t<T^*$ en utilisant la méthode des caractéristiques. Si $x\le t$, on a $x_0 = x-t\le 0$ et alors $u(t,x) = u_0(x_0) = 1$. Si $t\le x\le\alpha$, il vient $x_0 = \frac{x-t}{1-\frac{t}{\alpha}}$. On a donc $0\le x_0\le\alpha$, d'où $u(t,x) = u_0(x_0) = \frac{\alpha-x}{\alpha-t}$. Enfin, si $x\ge\alpha$, on a $x_0\ge\alpha$, d'où $u(t,x) = u_0(x_0) = 0$.
Il nous reste à prolonger cette solution classique au delà de l'instant T^*. Pour $t = T^* = \alpha$, la solution est discontinue et vaut $u_- = 1$ si $x<\alpha$ et $u_+ = 0$ si $x>\alpha$. En utilisant la relation de Rankine–Hugoniot (10.21), on trouve que la courbe de discontinuité Σ issue du point (α,α) est une demi-droite ayant pour pente $\sigma = \frac{1}{2}$ et donc pour équation $\xi(t) = \frac{\alpha}{2} + \frac{t}{2}$, $t\ge\alpha$ (voir la figure 10.3).
On a donc trouvé

$$u(t,x) = \begin{cases} 1 & \text{si } x\le t,\\ \dfrac{\alpha-x}{\alpha-t} & \text{si } t\le x\le\alpha,\\ 0 & \text{si } x\ge\alpha, \end{cases} \text{ pour } t<\alpha \text{ et } u(t,x) = \begin{cases} 1 & \text{si } x<\dfrac{\alpha}{2}+\dfrac{t}{2},\\ 0 & \text{si } x>\dfrac{\alpha}{2}+\dfrac{t}{2}, \end{cases} \text{ pour } t>\alpha.$$

Indiquons que la classe des fonctions de classe $\mathscr{C}^1$ par morceaux n'est pas assez grande pour décrire l'ensembles des solutions faibles des systèmes de lois de conservation généraux, un cadre plus approprié étant celui des *fonctions à variation bornée* [17] en espace. Ce dernier dépassant les objectifs d'un cours introductif, nous renvoyons

15. William John Macquorn Rankine (5 juillet 1820 - 24 décembre 1872) était un ingénieur et physicien écossais. Pionnier de la thermodynamique, il élabora une théorie complète de la machine à vapeur et plus généralement des moteurs thermiques.

16. Pierre-Henri Hugoniot (5 juin 1851 - 1887) était un physicien et mathématicien français. On lui doit une théorie, basée sur la conservation de la masse, de la quantité de mouvement et de l'énergie, qui permit l'amélioration des études des écoulements de fluides.

17. Pour tout ouvert Ω de $\mathbb{R}^d$, l'espace des fonctions à variation bornée sur Ω est

$$\mathrm{BV}(\Omega) = \left\{ v\in L^1_{\mathrm{loc}}(\Omega)\,|\,TV(v,\Omega)<+\infty \right\},$$

la *variation totale* de la fonction v sur Ω étant définie par

$$TV(v,\Omega) = \sup\left\{ \int_\Omega v(\boldsymbol{x})\,\mathrm{div}(\boldsymbol{\phi})(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\ \boldsymbol{\phi}\in\mathscr{C}^1_c(\Omega,\mathbb{R}^d),\ \|\boldsymbol{\phi}\|_{L^\infty(\Omega)}\le 1 \right\},$$

où $\mathscr{C}^1_c(\Omega,\mathbb{R}^d)$ est l'ensemble des fonctions continûment différentiables à support compact contenu dans Ω et à valeurs dans $\mathbb{R}^d$.

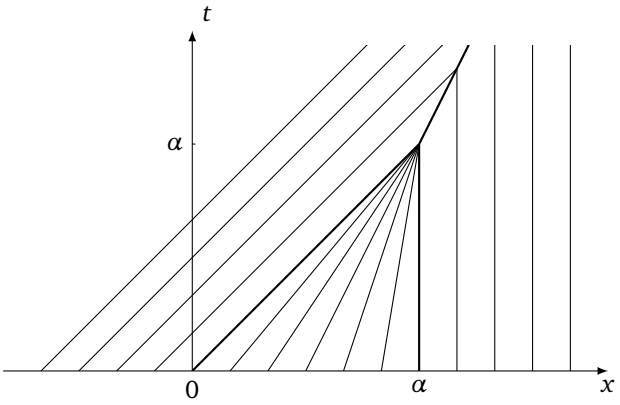


FIGURE 10.3 – Caractéristiques et droite de discontinuité Σ de la solution faible du problème de Cauchy (10.22)-(10.23).

le lecteur intéressé vers la référence [GR96]. Pour de nombreux systèmes issus de la physique cependant, les solutions pertinentes sont effectivement régulières par morceaux.

Si l'existence d'une solution faible du problème (10.8)-(10.9) est toujours assurée, on observera que celle-ci n'est pas nécessairement unique, comme le montre l'exemple suivant.

Contre-exemple à l'unicité des solutions faibles de l'équation de Burgers non visqueuse. Considérons le problème de Cauchy composé de l'équation de Burgers non visqueuse

$$\frac{\partial u}{\partial t}(t,x)+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)(t,x)=0,\ t>0,\ x\in\mathbb{R},$$

et d'une condition initiale de donnée triviale

$$u(0,x)=u_0(x)=0,\ x\in\mathbb{R}.$$

Ce problème possède une solution classique, triviale, $u\equiv 0$. Une solution faible non triviale (en fait une famille infinie de solutions faibles dépendant d'un paramètre réel) est donnée par la fonction

$$u(t,x)=\begin{cases}0 & \text{si } x<-\alpha\, t,\\ -2\alpha & \text{si } -\alpha\, t<x<0,\\ 2\alpha & \text{si } 0<x<\alpha\, t,\\ 0 & \text{si } x>\alpha\, t,\end{cases}$$

avec $\alpha>0$, qui satisfait bien la condition de Rankine–Hugoniot le long des lignes de discontinuité de la solution, d'équations respectives $x=0$ et $x=\pm\alpha\, t$.

10.3.4 Solutions entropiques

Pour parer au problème de non unicité des solution faibles du problème que nous venons d'évoquer, il nous faut trouver un critère discriminant. Or, il est apparu avec certains des exemples de la section 10.2, que les lois de conservation correspondaient à un reflet quelque peu idéalisé de la réalité physique [18], négligeant les mécanismes de dissipation ou de diffusion. Afin de sélectionner parmi l'ensemble des solutions faibles la solution possédant une pertinence « physique », on va requérir qu'elle satisfasse une condition supplémentaire, basée sur concept d'*entropie* [19] *mathématique*, qui la rend admissible en un sens que l'on précisera.

18. C'est le cas lorsque les systèmes de lois de conservation sont utilisés pour décrire des transformations irréversibles, comme en dynamique des gaz par exemple. On voit en effet que l'équation (10.8) ne contient aucune trace d'un tel phénomène, puisque l'application $(t,x)\mapsto u(-t,x)$ est solution si et seulement $(t,x)\mapsto u(t,-x)$ l'est.

19. La terminologie est empruntée à la dynamique des gaz, dans laquelle (l'opposé de) la densité d'entropie thermodynamique joue le rôle d'une entropie mathématique (VOIR la sous-section 10.2.5 ?).

Définition 10.10 (paire d'entropie-flux d'entropie [20]) *On appelle* ***paire d'entropie-flux d'entropie*** *pour l'équation* (10.8) *tout couple* (S,F) *de fonctions de* $\mathbb{R}$ *dans* $\mathbb{R}$ *telles que*

- *la fonction* S *est continue et convexe,*
- *on a* $F' = S'f'$*, éventuellement au sens des distributions.*

Exemple de paire d'entropie-flux d'entropie. Un choix possible de famille de paires d'entropie-flux d'entropie dépendant d'un paramètre est celui, considéré par Vol'pert [Vol67] et Kruzkov [Kru70], des fonctions

$$S(u) = |u-k| \text{ et } F(u) = \operatorname{sign}(u-k)(f(u)-f(k)),\ k \in \mathbb{R}. \tag{10.24}$$

où sign est la fonction signe définie par

$$\operatorname{sign}(v) = \begin{cases} -1 & \text{si } v < 0, \\ 0 & \text{si } v = 0, \\ 1 & \text{si } v > 0. \end{cases}$$

Ces paires jouent un rôle important dans la théorie des solutions entropiques dans le cas scalaire, lié au fait que le cône convexe fermé engendré par l'ensemble des fonctions affines et des fonctions S données par (10.24) est l'ensemble des fonctions convexes.

A VOIR : cas des systèmes symétriques ($A_j(u) = f_j(u)$ symétriques), alors $S(u) = \frac{1}{2}\sum_{i=1}^{p} {u_i}^2$

Il est facile de voir qu'une solution classique u de l'équation (10.8) satisfait aussi la loi de conservation

$$\frac{\partial S(u)}{\partial t}(t,x) + \frac{\partial F(u)}{\partial x}(t,x) = 0,\ t > 0,\ x \in \mathbb{R},$$

pour toute paire d'entropie-flux d'entropie (S,F) régulière, mais ce n'est généralement pas le cas pour une solution faible, en raison de la possible présence de discontinuités. On peut en revanche montrer qu'une solution faible particulière du problème s'obtient par passage à la limite (au sens des distributions) sur la solution [21] du problème perturbé

$$\frac{\partial u^\varepsilon}{\partial t}(t,x) + \frac{\partial f(u^\varepsilon)}{\partial x}(t,x) = \varepsilon \frac{\partial^2 u^\varepsilon}{\partial x^2}(t,x),\ t > 0,\ x \in \mathbb{R}, \tag{10.25}$$

$$u^\varepsilon(0,x) = u_0(x),\ x \in \mathbb{R}, \tag{10.26}$$

lorsque le paramètre ε tend vers 0, jouant le rôle d'une viscosité artificielle évanescente comme suggéré dans [Gel59]. C'est de cette solution limite que sera tiré le critère de sélection recherché.

Théorème 10.11 *Soit une paire d'entropie-flux d'entropie* (S,F) *pour l'équation* (10.8) *et une suite* $(u^\varepsilon)_{\varepsilon>0}$ *de solutions régulières du problème* (10.25)-(10.26)*, telle que*

- $\|u^\varepsilon\|_{L^\infty([0,+\infty[\times\mathbb{R})} \le C,\ \forall \varepsilon > 0,$
- u^ε *tend vers une limite* u *lorsque* ε *tend vers* 0 *presque partout dans* $[0,+\infty[\times\mathbb{R}$*,*

avec $C > 0$ *une constante indépendante de* ε*. Alors, la fonction* u *est une solution faible du problème* (10.8)-(10.9)*, satisfaisant l'****inégalité d'entropie***

$$\frac{\partial S(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \le 0 \tag{10.27}$$

au sens des distributions dans $[0,+\infty[\times\mathbb{R}$*, c'est-à-dire*

$$\int_0^{+\infty}\int_{\mathbb{R}} \left(S(u)(t,x)\frac{\partial \varphi}{\partial t}(t,x) + F(u)(t,x)\frac{\partial \varphi}{\partial x}(t,x)\right) \mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} S(u_0)(x)\varphi(0,x)\,\mathrm{d}x \ge 0,$$

pour toute fonction φ *de* $\mathscr{C}_0^\infty([0,+\infty[\times\mathbb{R})$ *à valeurs positives.*

20. La définition donnée ici est spécifique au cas d'une loi de conservation scalaire en une dimension d'espace. Pour un système de lois de conservation, c'est-à-dire pour $p > 1$, en dimension d'espace quelconque $d \ge 1$, une fonction convexe S d'un ensemble convexe de $\mathbb{R}^p$ à valeurs dans $\mathbb{R}$ est une entropie pour le système s'il existe des fonctions flux d'entropie F_j, $j = 1,\dots,d$, de $\mathbb{R}^p$ à valeurs dans $\mathbb{R}$, telles que l'on a $\boldsymbol{\nabla} F_j(\boldsymbol{u})^T = \boldsymbol{\nabla} S(\boldsymbol{u})^T \boldsymbol{\nabla} \boldsymbol{f}_j(\boldsymbol{u})$. Trouver une telle paire revient donc à déterminer $d+1$ fonctions satisfaisant un système de dp équations aux dérivées partielles, ce qui peut être compliqué, voire impossible, lorsque $p > 1$, ce système étant en général surdéterminé. On n'est donc pas toujours assuré de pouvoir d'exhiber des paires d'entropie-flux d'entropie (et encore moins l'ensemble de ces dernières) pour un système quelconque, alors que, pour une loi de conservation *scalaire*, toute fonction S convexe définit une entropie. L'existence d'une entropie pour un système de lois de conservation est donc une propriété importante et il est remarquable que la plupart des systèmes de lois de conservation issus de la physique ou de la mécanique possède une paire d'entropie-flux d'entropie associée, qui a de plus une signification physique.

21. REPRENDRE Le problème (10.25)-(10.26) possède une unique solution appartenant à $L^\infty([0,+\infty[\times\mathbb{R})$, qui est de plus très régulière sur $]0,+\infty[\times\mathbb{R}$ (REF !).

DÉMONSTRATION. Soit φ une fonction de $\mathscr{C}_0^\infty([0,+\infty[\times\mathbb{R})$; en multipliant (10.25) par φ et en intégrant par partie sur $[0,+\infty[\times\mathbb{R}$, on obtient

$$\int_0^{+\infty}\int_{\mathbb{R}} u^\varepsilon(t,x)\frac{\partial\varphi}{\partial t}(t,x)\,\mathrm{d}t\,\mathrm{d}x+\int_{\mathbb{R}} u_0(x)\varphi(0,x)\,\mathrm{d}x+\int_0^{+\infty}\int_{\mathbb{R}} f(u^\varepsilon)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\,\mathrm{d}t\,\mathrm{d}x + \int_0^{+\infty}\int_{\mathbb{R}} u^\varepsilon(t,x)\frac{\partial^2\varphi}{\partial x^2}(t,x)\,\mathrm{d}t\,\mathrm{d}x=0.$$

Il découle alors des hypothèses sur la suite $(u^\varepsilon)_{\varepsilon>0}$ et du théorème de convergence dominée de Lebesgue (REF !) que

$$\int_0^{+\infty}\int_{\mathbb{R}} u^\varepsilon(t,x)\frac{\partial\varphi}{\partial t}(t,x)\,\mathrm{d}t\,\mathrm{d}x\to\int_0^{+\infty}\int_{\mathbb{R}} u(t,x)\frac{\partial\varphi}{\partial t}(t,x)\,\mathrm{d}t\,\mathrm{d}x,$$

$$\int_0^{+\infty}\int_{\mathbb{R}} f(u^\varepsilon)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\,\mathrm{d}t\,\mathrm{d}x\to\int_0^{+\infty}\int_{\mathbb{R}} f(u)(t,x)\frac{\partial\varphi}{\partial x}(t,x)\,\mathrm{d}t\,\mathrm{d}x$$

et

$$\int_0^{+\infty}\int_{\mathbb{R}} u^\varepsilon(t,x)\frac{\partial^2\varphi}{\partial x^2}(t,x)\,\mathrm{d}t\,\mathrm{d}x\to 0$$

quand $\varepsilon\to 0$. On en conclut que u est une solution faible du problème (10.8)-(10.9) par densité de $\mathscr{C}_0^\infty([0,+\infty[\times\mathbb{R})$ dans $\mathscr{C}_0^1([0,+\infty[\times\mathbb{R})$ (REF !).
Considérons à présent une entropie S de classe $\mathscr{C}^2$ et multiplions l'équation (10.25) par $S'(u^\varepsilon)$. Par propriété des paires d'entropie-flux d'entropie, il vient

$$\frac{\partial S(u^\varepsilon)}{\partial t}(t,x)+\frac{\partial F(u^\varepsilon)}{\partial x}(t,x)=\varepsilon\left(\frac{\partial^2 S(u^\varepsilon)}{\partial x^2}(t,x)-S''(u^\varepsilon)(t,x)\left(\frac{\partial u^\varepsilon}{\partial x}\right)^2(t,x)\right),\ t>0,\ x\in\mathbb{R},$$

d'où

$$\frac{\partial S(u^\varepsilon)}{\partial t}(t,x)+\frac{\partial F(u^\varepsilon)}{\partial x}(t,x)\le\varepsilon\,\frac{\partial^2 S(u^\varepsilon)}{\partial x^2}(t,x),\ t>0,\ x\in\mathbb{R}.$$

Multiplions maintenant cette inégalité par une fonction test φ de $\mathscr{C}_0^\infty([0,+\infty[\times\mathbb{R})$ à valeurs positives et intégrons par parties sur $[0,+\infty[\times\mathbb{R}$. Nous arrivons à

$$\int_0^{+\infty}\int_{\mathbb{R}}\left(S(u^\varepsilon)(t,x)\left(\frac{\partial\varphi}{\partial t}(t,x)+\varepsilon\,\frac{\partial^2\varphi}{\partial x^2}(t,x)\right)+F(u^\varepsilon)\frac{\partial\varphi}{\partial x}(t,x)\right)\mathrm{d}t\,\mathrm{d}x+\int_{\mathbb{R}} S(u_0)(x)\varphi(0,x)\,\mathrm{d}x\le 0,$$

et, par passage à la limite sur ε,

$$\int_0^{+\infty}\int_{\mathbb{R}}\left(S(u)(t,x)\frac{\partial\varphi}{\partial t}(t,x)+F(u)\frac{\partial\varphi}{\partial x}(t,x)\right)\mathrm{d}t\,\mathrm{d}x+\int_{\mathbb{R}} S(u_0)(x)\varphi(0,x)\,\mathrm{d}x\le 0,\tag{10.28}$$

qui n'est autre que l'écriture (10.27) au sens des distributions.

Il reste à passer d'une entropie de classe $\mathscr{C}^2$ à une entropie générale. Pour cela, on introduit, pour toute entropie S, une suite $(S_n)_{n\in\mathbb{N}}$ de fonctions définies par le produit de convolution $S_n=S*(n\rho(n\cdot))$, avec ρ une fonction de $\mathscr{C}_0^\infty(\mathbb{R})$. La suite des flux d'entropie associés est alors donnée par

$$F_n(v)=\int_0^v f'(y)S_n'(y)\,\mathrm{d}y,\ n\in\mathbb{N}.$$

On a

$$F_n(v)=f'(v)S_n(v)-f'(0)S_n(0)-\int_0^v f''(y)S_n(y)\,\mathrm{d}y,\ \forall n\in\mathbb{N},$$

dont on déduit que la suite $(F_n)_{n\in\mathbb{N}}$ converge uniformément vers la fonction continue

$$F(v)=f'(v)S(v)-f'(0)S(0)-\int_0^v f''(y)S(y)\,\mathrm{d}y,$$

par convergence uniforme de la suite $(S_n)_{n\in\mathbb{N}}$ vers S. L'inégalité (10.28) étant vraie pour les paires (S_n,F_n), on montre qu'elle le reste pour le couple (S,F) en faisant tendre n vers l'infini. □

Ce résultat conduit à l'introduction de la notion de *solution entropique*.

Définition 10.12 (solution entropique) *Une fonction u de $L^\infty([0,+\infty[\times\mathbb{R})$ est une **solution entropique** du problème* (10.8)-(10.9) *si elle satisfait l'inégalité d'entropie* (10.27)*, au sens des distributions dans $[0,+\infty[\times\mathbb{R}$, pour toute paire d'entropie-flux d'entropie pour l'équation* (10.8).

Le résultat suivant est immédiat.

Proposition 10.13 *Une solution entropique du problème* (10.8)-(10.9) *est une solution faible de ce problème.*

DÉMONSTRATION. La preuve est directe, en faisant successivement les choix $S(u)=\pm u$ et $F(u)=\pm f(u)$ dans l'inégalité (10.27). □

La notion de solution entropique vise à donner à une solution faible du problème un caractère irréversible, en lui demandant de satisfaire une condition additionnelle. Si la fonction u est une solution faible de (10.8)-(10.9), on voit en effet que la fonction $(t,x)\mapsto u(s-t,-x)$ est une solution faible dans la bande $]0,s[\times\mathbb{R}$ pour la donnée initiale $u(s,-x)$, mais qu'elle n'est une solution entropique que si l'inégalité d'entropie (10.27) est une égalité. On note par ailleurs qu'il y a équivalence entre les notions de solution faible et de solution entropique dans le cas linéaire.

Nous allons maintenant donner une première caractérisation des solutions entropiques de classe $\mathscr{C}^1$ par morceaux.

Théorème 10.14 *Une solution faible u du problème* (10.8)-(10.9) *de classe $\mathscr{C}^1$ par morceaux est une solution entropique si et seulement si elle satisfait, pour tout couple d'entropie-flux d'entropie associé à* (10.8)*, l'inégalité*

$$\sigma\,[S(u)]_{|\Sigma} \geq [F(u)]_{|\Sigma} \tag{10.29}$$

le long de toute courbe de discontinuité Σ.

DÉMONSTRATION. La preuve étant en tout point similaire à celle du théorème 10.9, elle est laissée en exercice. □

Il existe d'autres formes, plus exploitables en pratique, de la condition (10.29), dues à Oleinik [22].

Lemme 10.15 (« conditions d'entropie d'Oleinik » [Ole59]) *Une solution faible u du problème* (10.8)-(10.9) *de classe $\mathscr{C}^1$ par morceaux vérifie l'inégalité* (10.29) *pour toute paire d'entropie-flux d'entropie (S,F) si et seulement si l'une des trois conditions suivantes est vérifiée,*

$$\sigma \geq \frac{f(u_+)-f(k)}{u_+-k} \text{ pour tout réel } k \text{ compris entre } u_- \text{ et } u_+, \tag{10.30}$$

$$\sigma \leq \frac{f(u_-)-f(k)}{u_--k} \text{ pour tout réel } k \text{ compris entre } u_- \text{ et } u_+, \tag{10.31}$$

$$\begin{cases} f(\alpha u_- + (1-\alpha)u_+) \geq \alpha f(u_-) + (1-\alpha) f(u_+) & \text{si } u_+ > u_- \\ f(\alpha u_- + (1-\alpha)u_+) \leq \alpha f(u_-) + (1-\alpha) f(u_+) & \text{si } u_+ < u_- \end{cases}, \ 0\leq\alpha\leq 1, \tag{10.32}$$

en tout point d'une courbe de discontinuité Σ, la fonction σ désignant la vitesse de propagation de la discontinuité considérée.

DÉMONSTRATION. Il suffit de prouver le résultat pour les paires de Kruzkov (10.24) (EXPLIQUER [23]). Dans ce cas, l'inégalité (10.29) se réécrit

$$\sigma\,(|u_+-k|-|u_--k|) \geq \operatorname{sign}(u_+-k)(f(u_+)-f(k)) - \operatorname{sign}(u_--k)(f(u_-)-f(k)),$$

cette inégalité devant être vérifiée pour tout nombre réel k.
Suppose que $u_+ > u_-$, le cas $u_+ < u_-$ se traitant de manière similaire. On obtient alors successivement

$$\sigma\,(u_+-u_-) \leq f(u_+)-f(u_-)$$

22. Olga Arsenievna Oleinik (Óльга Арсéньевна Олéйник en russe, 2 juillet 1925 - 13 octobre 2001) était une mathématicienne russe. Elle fit de remarquables contributions à la géométrie algébrique, à l'étude des équations aux dérivées partielles et à la théorie mathématique des milieux élastiques inhomogènes et des couches limites.

23. DEBUT a préciser : On a vu quel sens il fallait donner à l'inégalité d'entropie pour des entropies continues dans la démonstration du théorème 10.11. Il suffit ensuite de remarquer que pour toute fonction S continue et convexe, il existe une suite (S_α) définie par $S_\alpha(s)=b_0+b_1 s+\sum_j a_j\left|s-k_j\right|$, $a_j>0$, qui converge uniformément vers S (il en va de même pour le flux d'entropie associé).

pour $k \geq u_+$ et

$$\sigma\,(u_+ - u_-) \geq f(u_+) - f(u_-)$$

pour $k \leq u_-$, qui, prises ensemble, expriment simplement le fait que la solution satisfait la condition de Rankine–Hugoniot (10.21). Il reste à considérer le cas $u_- < k < u_+$, c'est-à -dire $k = \alpha u_- + (1-\alpha)u_+$, $\alpha \in [0,1]$. On a

$$\sigma\,(u_+ + u_- - 2k) \geq f(u_+) + f(u_-) - 2f(k). \tag{10.33}$$

En additionnant (10.33) à (10.21), il vient

$$2\sigma\,(u_+ - k) \geq 2\,(f(u_+) - f(k)),$$

dont on déduit (10.30), alors qu'on trouve (10.31) en soustrayant (10.33) à (10.21). Enfin, on a, en utilisant de nouveau (10.21),

$$\sigma\,(u_+ + u_- - 2k) = \sigma\,(2\alpha - 1)(u_+ - u_-) = (2\alpha - 1)\,(f(u_+) - f(u_-)),$$

et on obtient alors, en substituant dans (10.33),

$$(2\alpha - 1)\,(f(u_+) - f(u_-)) \geq f(u_+) + f(u_-) - 2f(\alpha u_- + (1-\alpha)u_+),$$

qui n'est autre que la première inégalité de (10.32). On a donc montré que les conditions de l'énoncé sont nécessaires. Leur suffisance est aisément établie et laissée en exercice. □

Une interprétation géométrique de la condition (10.32) est la suivante : une discontinuité admissible est telle que le graphe de l'application f sur le segment $[u_-, u_+]$ (resp. $[u_+, u_-]$) est situé au-dessus (resp. en dessous) de la corde joignant les points $(u_-, f(u_-))$ et $(u_+, f(u_+))$ (resp. $(u_+, f(u_+))$ et $(u_-, f(u_-))$) lorsque $u_- < u_+$ (resp. $u_+ < u_-$). Ainsi, dans le cas d'une fonction f strictement convexe, cette condition est satisfaite si et seulement si

$$u_+ < u_-, \tag{10.34}$$

ou, de manière équivalente (il suffit d'utiliser la condition de Rankine–Hugoniot (10.19), le théorème des accroissements finis et le théorème des valeurs intermédiaires), si et seulement si

$$f'(u_+) < \sigma < f'(u_-), \tag{10.35}$$

en tout point d'une courbe de discontinuité Σ. Cette condition, qui dans le cas général d'un système [24] de lois de conservation porte le nom de *condition d'entropie de Lax* [25] [Lax57], traduit le fait les caractéristiques convergent vers une courbe de discontinuité de la solution (voir par exemple la figure 10.3). Dans le cas de la dynamique des gaz, cette condition traduit en termes mathématiques une conséquence du second principe de la thermodynamique exprimant le fait que l'entropie thermodynamique des particules croît à la traversée d'une discontinuité (appelée dans ce contexte un *choc*).

REMARQUE sur les résultats d'existence de solutions entropiques : l'approximation parabolique, solution de la famille de problèmes (10.25)-(10.26), fournit une possibilité.

Nous terminons cette section par un résultat important de Kruzkov [Kru70].

Théorème 10.16 *Soit u et v deux solutions entropiques du problème* (10.8)-(10.9) *respectivement associées aux données initiales u_0 et v_0 de $L^\infty(\mathbb{R})$. On pose $M = \sup\{|f'(s)|\ |s \in [\inf(u_0, v_0), \sup(u_0, v_0)]\}$. Pour tout $R > 0$ et (presque) tout $t > 0$, on a*

$$\int_{|x|\leq R} |u(t,x) - v(t,x)|\ \mathrm{d}x \leq \int_{|x|\leq R+M\,t} |u_0(x) - v_0(x)|\ \mathrm{d}x.$$

DÉMONSTRATION. A ECRIRE □

Une conséquence immédiate de cette estimation *a priori* est l'unicité d'une solution entropique d'une loi de conservation scalaire [26], mais elle permet encore d'établir un certain nombre d'autres assertions.

24. Une condition, plus stricte que celle de Lax, également employée pour les systèmes de lois de conservation est le critère introduit par Liu [Liu76].

25. Peter David Lax (Lax Péter Dávid en hongrois, né le 1er mai 1926) est un mathématicien américain d'origine hongroise. Ses nombreuses contributions recouvrent plusieurs domaines d'étude des mathématiques et de la physique, parmi lesquels on peut citer les équations aux dérivées partielles, les systèmes hyperboliques de lois de conservation, les ondes de choc en mécanique des fluides, les systèmes intégrables, la théorie des solitons, la théorie de la diffusion, l'analyse numérique et le calcul scientifique.

26. Pour une autre preuve, basée sur un argument de dualité, de ce résultat dans le cas d'une fonction flux f strictement convexe, on pourra consulter l'article [Ole57].

Corollaire 10.17 *Soit une solution entropique u du problème* (10.8)-(10.9) *associée à une donnée initiale u_0 de $L^\infty(\mathbb{R})$. On a les assertions suivantes.*

1. *La fonction u est unique.*
2. *Si u_0 appartient à $L^1(\mathbb{R})$, alors $u(t,\cdot)$ appartient à $L^1(\mathbb{R})$ et $\|u(t,\cdot)\|_{L^1(\mathbb{R})} \leq \|u_0\|_{L^1(\mathbb{R})}$.*
3. *Si u_0 appartient à $L^1(\mathbb{R})$ et que la fonction v est une solution entropique du problème* (10.8)-(10.9) *associée à une donnée initiale v_0 de $L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$ telle que $u_0 \leq v_0$ presque partout sur $\mathbb{R}$, alors, pour presque tout $t > 0$, $u(t,\cdot) \leq v(t,\cdot)$ presque partout sur $\mathbb{R}$.*

DÉMONSTRATION. A ECRIRE □

Indiquons que ce résultat reste valable pour une loi de conservation scalaire en plusieurs dimensions d'espace, dont le flux dépend explicitement du temps et de l'espace et en présence d'un terme source (moyennant des hypothèses de régularité sur ces derniers). Le cas des systèmes ($p \geq 2$) ne possède pas de théorie aussi avancée et ce résultat y reste une conjecture.

10.3.5 Le problème de Riemann

Un exemple particulièrement éclairant sur la nature des solutions discontinues d'une loi de conservation est, malgré sa simplicité, celui du problème de Cauchy en une dimension d'espace suivant, appelé *problème de Riemann* par analogie avec un problème étudié en dynamique des gaz[27] [Rie60],

$$\frac{\partial u}{\partial t}(t,x) + \frac{\partial f(u)}{\partial x}(t,x) = 0,\ t > 0,\ x \in \mathbb{R}, \tag{10.36}$$

$$u(0,x) = u_0(x) = \begin{cases} u_g & \text{si } x < 0, \\ u_d & \text{si } x > 0, \end{cases} \tag{10.37}$$

où f est une fonction réelle de classe $\mathscr{C}^2$ et u_g et u_d sont deux constantes données. Dans ce cas, l'existence et unicité d'une solution entropique du problème est assurée par le corollaire 10.17. Nous cherchons à construire explicitement cette solution.

Pour cela, nous allons tout d'abord établir une propriété satisfaite par la solution du problème (10.36)-(10.37), qui découle du fait que ce dernier est invariant par tout changement de variables homothétique $(t,x) \mapsto (\lambda t, \lambda x)$, avec $\lambda > 0$.

Lemme 10.18 *La solution du problème* (10.36)-(10.37) *est* ***auto-semblable****, c'est-à-dire qu'elle est de la forme*

$$u(t,x) = v\left(\frac{x}{t}\right).$$

DÉMONSTRATION. Soit u l'unique solution entropique du problème (10.36)-(10.37) et λ un réel strictement positif ; la fonction $u(\lambda\cdot, \lambda\cdot)$ est alors une solution de l'équation (10.36), satisfaisant la condition d'entropie et une condition initiale ayant pour donnée la fonction $u_0(\lambda\cdot)$. La fonction u_0 définie par (10.37) étant telle que

$$u_0(\lambda\cdot) = u_0,$$

il vient, par unicité de la solution entropique, que

$$u(\lambda\cdot, \lambda\cdot) = u,\ \forall \lambda > 0,$$

ce qui signifie exactement que u est auto-semblable. □

Sur tout domaine où la solution est de classe $\mathscr{C}^1$, on a

$$\frac{\partial u}{\partial t}(t,x) = -\frac{x}{t^2} v'\left(\frac{x}{t}\right) \text{ et } \frac{\partial f(u)}{\partial x}(t,x) = \frac{1}{t} f'\left(v\left(\frac{x}{t}\right)\right) v'\left(\frac{x}{t}\right),$$

27. Dans le problème en question, on considère un gaz au repos, contenu dans un tube cylindrique long et fin que l'on suppose divisé en deux parties par une membrane, le fluide possédant une densité et une pression plus élevées dans une partie que dans l'autre. À l'instant initial, la membrane est déchirée et l'on s'intéresse à l'écoulement qui s'ensuit.

l'équation (10.36) est donc satisfaite dans un tel domaine si et seulement si

$$v'\left(\frac{x}{t}\right)\left[f'\left(v\left(\frac{x}{t}\right)\right)-\frac{x}{t}\right]=0. \tag{10.38}$$

En excluant les état constants, correspondant au cas

$$v'\left(\frac{x}{t}\right)=0,$$

la fonction v est obtenue en résolvant

$$f'\left(v\left(\frac{x}{t}\right)\right)=\frac{x}{t}, \tag{10.39}$$

ce qui est possible si f' est monotone, ce qui revient à ce que f soit convexe ou concave sur le domaine considéré. On dit alors que deux états u_g et u_d sont liés par une *onde de raréfaction* ou *de détente* (*rarefaction wave* en anglais) si $f'(u_g)<f'(u_d)$ et qu'il existe une fonction v de $[f'(u_g),f'(u_d)]$ à valeurs dans $\mathbb{R}$ vérifiant l'équation (10.39), la solution u, continue pour tout temps strictement positif, correspondante étant donnée par

$$u(t,x)=\begin{cases}u_g & \text{si } x\le f'(u_g)\,t,\\ (f')^{-1}\left(\dfrac{x}{t}\right) & \text{si } f'(u_g)\,t\le x\le f'(u_d)\,t\\ u_d & \text{si } x\ge f'(u_d)\,t.\end{cases}$$

Une autre solution de l'équation (10.36) liant deux états u_g et u_d est fournie par la fonction discontinue, portant le nom d'*onde de choc* (*shock wave* en anglais),

$$u(t,x)=\begin{cases}u_g & \text{si } x<\sigma\,t,\\ u_d & \text{si } x>\sigma\,t,\end{cases}$$

où

$$\sigma=\frac{f(u_d)-f(u_g)}{u_d-u_g}$$

d'après la condition de Rankine–Hugoniot (10.21), satisfaisant de plus la condition d'entropie.

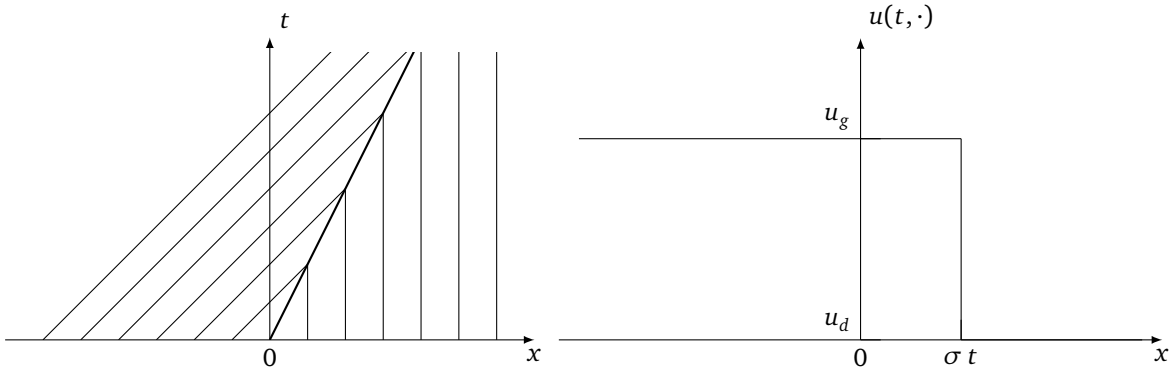


FIGURE 10.4 – Tracé de caractéristiques (à gauche) et représentation de l'onde de choc entropique solution (à droite) du problème de Riemann à deux états pour l'équation de Burgers avec $u_g=1$ et $u_d=0$.

Il reste à déterminer la solution entropique du problème de Riemann. Supposons dans un premier temps que la fonction f soit strictement convexe, ce qui couvre un bon nombre de situations rencontrées en pratique [28]. Trois cas se présentent suivant la monotonie de la donnée initiale u_0.

- Si $u_g=u_d$, l'état constant $u(t,x)=u_g=u_d$, $t\ge 0$, $x\in\mathbb{R}$, est l'unique solution entropique du problème.
- Si $u_g>u_d$, la discontinuité de la donnée initiale est admissible (car elle satisfait la condition (10.34)) et la solution entropique est une onde de choc entropique (voir la figure 10.4 pour un exemple).

28. On traite de manière symétrique le cas d'un problème pour lequel la fonction de flux est strictement concave.

- Enfin, si $u_g < u_d$, la discontinuité de la donnée initiale n'est pas admissible et donne lieu à une onde de raréfaction. Il faut résoudre l'équation (10.39), qui admet une unique solution, puisque l'on a supposé la fonction f strictement convexe. Cette solution est une fonction continue et croissante en espace (voir la figure (10.5) pour un exemple).

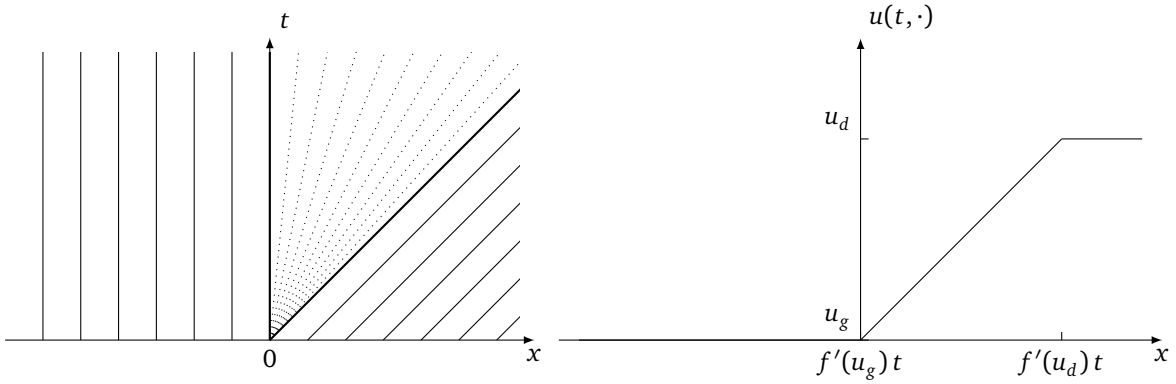


FIGURE 10.5 – Tracé de caractéristiques (à gauche) et représentation de l'onde de raréfaction solution du problème de Riemann à deux états pour l'équation de Burgers avec $u_g = 0$ et $u_d = 1$.

Pour une fonction f générale mais possédant un nombre fini de points d'inflexion, la structure de la solution du problème de Riemann demeure la même, consistant en deux états constants séparés par une combinaison d'ondes de raréfaction et/ou de choc de manière à satisfaire la condition d'entropie.

A FAIRE Exemple du modèle de Buckley–Leverett : [BL42] modèle scalaire simple d'écoulement d'un fluide contenant deux phases dans un milieu poreux. flux

$$f(u) = \frac{u^2}{u^2 + a(1-u)^2}$$

avec a une constante.

10.4 Méthodes de discrétisation par différences finies **

Nous allons à présent nous intéresser à des méthodes de résolution numérique du problème de Cauchy scalaire (10.8)-(10.9). Pour cela, nous ne considérons que des méthodes de discrétisation basées sur une approximation des opérateurs différentiels par la méthode des différences finies. Cette approche fut historiquement la première utilisée pour la résolution des systèmes hyperboliques de lois de conservation et reste communément employée aujourd'hui, après avoir connu de nombreux raffinements et améliorations. D'autres techniques de discrétisation, toutes aussi adaptées aux discontinuités que peuvent présenter les solutions, existent par ailleurs. Elles seront mentionnées dans la section 10.5.

Dans l'ensemble de cette section, nous supposons que la fonction de flux f est de classe $\mathscr{C}^2$ et que la donnée initiale u_0 appartient à $L^\infty(\mathbb{R})$.

PARLER de la généralisation des schémas à plusieurs dimensions d'espace, aux systèmes et de l'extension des résultats ? ? ?

10.4.1 Principe

Commençons par revenir sur la description de la méthode des différences finies pour l'approximation de solutions d'équations aux dérivées partielles, déjà esquissée dans la sous-section 2.1.2 du chapitre 2.

Cette méthode consiste en premier lieu en une discrétisation du plan $\{(x,t) \,|\, x \in \mathbb{R},\ t \in [0, +\infty[\}$, domaine sur lequel est posé le problème, par une grille régulière (voir la figure 10.6), obtenue par le choix de la longueur Δt d'un *pas de temps*, celle Δx d'un *pas d'espace* et la définition de *points de grille* par la donnée des couples (x_j, t_n), $j \in \mathbb{Z}$, $n \in \mathbb{N}$, tels que

$$t_n = n\,\Delta t \text{ et } x_j = j\,\Delta x.$$

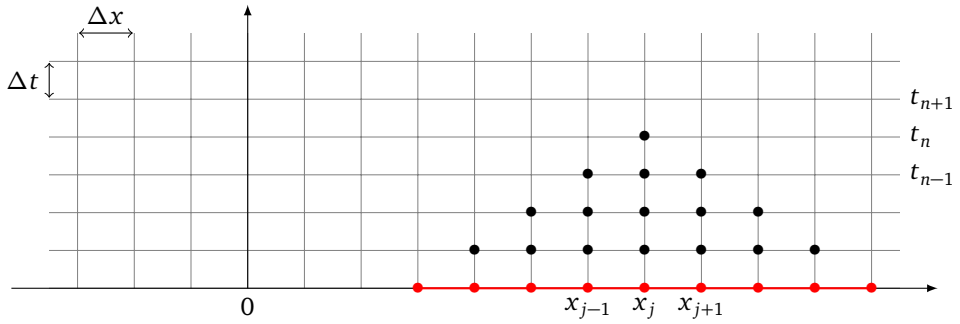


FIGURE 10.6 – Grille de discrétisation régulière du plan $\mathbb{R} \times [0, +\infty[$. On a représenté en rouge le domaine de dépendance numérique $\mathscr{D}_{\Delta t}(t_n, x_j)$ pour un schéma aux différences finies explicite à trois points.

En pratique, la grille n'est pas nécessairement régulière, mais, même lorsqu'elles sont variables, on a coutume de supposer que les longueurs des pas en temps et en espace sont liées entre elles, c'est-à-dire qu'il existe une fonction régulière Λ telle que $\Delta x = \Lambda(\Delta t)$, avec $\Lambda(0) = 0$. Pour les problèmes hyperboliques par exemple, on suppose en général que le rapport $\lambda = \frac{\Delta t}{\Delta x}$ reste constant lorsque les longueurs Δt et Δx tendent vers zéro. Nous reviendrons plus en détail sur ce point.

Ensuite, l'équation aux dérivées partielles du problème que l'on cherche à résoudre est remplacée chacun des points de la grille par une équation algébrique, encore appelée schéma, obtenue en substituant aux valeurs des opérateurs différentiels en ces points des quotients, ou différences finies, les approchant. La solution du système d'équations ainsi obtenu doit alors fournir une approximation des valeurs de la solution du problème aux points de la grille.

Pour cela, l'approximation par différences finies de la dérivée d'une fonction en un point de grille repose sur l'utilisation de développements de Taylor de cette fonction en d'autres points de grille bien choisis. Soit en effet une fonction v d'une variable réelle de classe $\mathscr{C}^2$ sur $\mathbb{R}$. Pour tout réel x, il existe, en vertu de la formule de Taylor–Lagrange (voir le théorème B.116), un réel θ_+, strictement compris entre 0 et 1, tel que

$$v(x + \Delta x) = v(x) + \Delta x\, v'(x) + \frac{(\Delta x)^2}{2}\, v''(x + \theta_+ \Delta x),\ \Delta x > 0,\ \theta_+ \in]0, 1[$$

dont on déduit l'approximation dite *décentrée à droite*

$$v'(x) \simeq \frac{v(x + \Delta x) - v(x)}{\Delta x}.$$

Observons qu'en utilisant le développement

$$v(x - \Delta x) = v(x) - \Delta x\, v'(x) + \frac{(\Delta x)^2}{2}\, v''(x - \theta_- \Delta x),\ \Delta x > 0,\ \theta_- \in]0, 1[,$$

on peut obtenir une approximation dite *décentrée à gauche*,

$$v'(x) \simeq \frac{v(x) - v(x - \Delta x)}{\Delta x},$$

ou encore une approximation *centrée*

$$v'(x) \simeq \frac{v(x + \Delta x) - v(x - \Delta x)}{2\,\Delta x},$$

toutes deux aussi légitimes. On peut donc approcher la dérivée d'une fonction en un point de différentes façons. Toutefois, le choix effectué pour la résolution d'une équation aux dérivées partielles par rapport au temps et à l'espace n'est absolument pas anodin, que ce soit en termes de précision de l'approximation obtenue ou de stabilité de la méthode.

Dans la suite, nous n'allons essentiellement considérer que des schémas explicites à un pas de temps, au sens où l'approximation de la solution au temps t_{n+1} sera obtenue de manière explicite à partir de l'approximation à l'instant t_n.
A VOIR : remarques sur les schémas implicites à pas multiples en temps (exemples en fin de chapitre), l'implicitation, etc.

Les schémas implicites sont rarement utilisés pour les équations hyperboliques, mais s'avèrent importants pour la classe des équations paraboliques abordées dans le prochain chapitre.

L'application de cette technique à la résolution de l'équation (10.8) conduit, pour une méthode explicite à un pas en temps et $2k$ pas (ou $2k+1$ points) en espace, $k \in \mathbb{N}^*$, à des schémas de la forme

$$u_j^{n+1} = H(u_{j-k}^n, \dots, u_{j+k}^n),\ n \in \mathbb{N},\ j \in \mathbb{Z}, \tag{10.40}$$

où H est une fonction continue de $\mathbb{R}^{2k+1}$ dans $\mathbb{R}$ et où la quantité u_j^n désigne une approximation de la solution u du problème au point de grille (t_n, x_j), $n \in \mathbb{N}$, $j \in \mathbb{Z}$. L'ensemble $\{u_{j-k}^n, \dots, u_{j+k}^n\}$ est appelé le *support* (*stencil* en anglais) du schéma au point (t_n, x_j). Étant donné une approximation u_j^0 de la donnée initiale, ce schéma permet de calculer une approximation de la solution pour tout temps t_n, $n \geq 0$.

Un premier exemple de schéma pour une loi de conservation scalaire. De manière relativement naturelle, on peut chercher à approcher la solution de l'équation (10.8) en remplaçant la dérivée en temps par une différence finie décentrée progressive et la dérivée en espace par une différence finie centrée. On a

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f(u_{j+1}^n) - f(u_{j-1}^n)}{2\Delta x} = 0,\ j \in \mathbb{Z},\ n \in \mathbb{N},$$

soit encore

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right),\ j \in \mathbb{Z},\ n \in \mathbb{N}, \tag{10.41}$$

dont on déduit que $H(u_{j-1}^n, u_j^n, u_{j+1}^n) = u_j^n - \frac{\Delta t}{2\Delta x}(f(u_{j+1}^n) - f(u_{j-1}^n))$. Le schéma explicite ainsi obtenu est à deux pas (ou trois points) en espace. En anglais, on utilise souvent l'acronyme *FTCS* (*forward in time, centered in space*) pour le désigner, par analogie avec un schéma utilisée pour l'équation de la chaleur (voir le schéma (11.5)).

Pour la suite, il est utile d'introduire l'opérateur $S_{\Delta t}$, associant à toute suite $\boldsymbol{v} = (v_j)_{j\in\mathbb{Z}}$ la suite $S_{\Delta t}(\boldsymbol{v})$ définie par

$$(S_{\Delta t}(\boldsymbol{v}))_j = H(v_{j-k}, \dots, v_{j+k}),\ j \in \mathbb{Z}, \tag{10.42}$$

afin d'écrire le schéma (10.40) de manière compacte suivante

$$\boldsymbol{u}^{n+1} = S_{\Delta t}(\boldsymbol{u}^n),\ n \in \mathbb{N}, \tag{10.43}$$

en justifiant du fait que la notation retenue fait seulement intervenir la quantité Δt parce que le rapport $\lambda = \frac{\Delta t}{\Delta x}$ est supposé rester constant lorsque les longueurs de ces pas de discrétisation tendent vers zéro.

À l'instar des lois de conservation qu'ils visent à approcher, les schémas de certaines méthodes peuvent être écrits sous une forme conservative, via l'introduction d'un *flux numérique*.

Définition 10.19 (forme conservative d'un schéma aux différences finies) *On dit que le schéma aux différences finies* (10.40) *pour la résolution numérique de l'équation* (10.8) *peut être mis sous* ***forme conservative*** *s'il existe une fonction continue h de $\mathbb{R}^{2k}$ dans $\mathbb{R}$, appelée* ***flux numérique*** *(**numerical flux** en anglais), telle que*

$$H(v_{j-k}, \dots, v_{j+k}) = v_j - \frac{\Delta t}{\Delta x}\left(h(v_{j-k+1}, \dots, v_{j+k}) - h(v_{j-k}, \dots, v_{j+k-1})\right),\ j \in \mathbb{Z}. \tag{10.44}$$

Il est clair que le flux numérique d'un schéma est défini à une constante additive près. On a coutume d'introduire la notation

$$h_{j+\frac{1}{2}}^n = h(u_{j-k+1}^n, \dots, u_{j+k}^n) \text{ et } h_{j-\frac{1}{2}}^n = h(u_{j-k}^n, \dots, u_{j+k-1}^n),\ n \in \mathbb{N},\ j \in \mathbb{Z},$$

la forme conservative du schéma (10.40) devenant alors

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(h_{j+\frac{1}{2}}^n - h_{j-\frac{1}{2}}^n\right),\ n \in \mathbb{N},\ j \in \mathbb{Z}. \tag{10.45}$$

Flux numérique associé au schéma FTCS. Considérons le premier exemple de schéma aux différences finies introduit plus haut pour approcher la loi de conservation (10.8). En utilisant la relation (10.41) et la précédente définition, on obtient

$$h(v_j, v_{j+1}) = \frac{1}{2}\left(f(v_j) + f(v_{j+1})\right),\ j \in \mathbb{Z}. \tag{10.46}$$

On notera que tous les schémas ne peuvent être mis sous forme conservative. On a la caractérisation suivante des schémas pour lesquels c'est le cas.

Proposition 10.20 *Le schéma aux différences finies* (10.40) *pour la résolution numérique de l'équation* (10.8) *peut être écrit sous forme conservative si et seulement si l'on a, pour toute suite* $\boldsymbol{v}$ *de* $\ell^1_{\Delta x}(\mathbb{Z})$ *telle que la suite* $S_{\Delta t}(\boldsymbol{v})$*, définie par* (10.42)*, appartient à* $\ell^1_{\Delta x}(\mathbb{Z})$*,*

$$\sum_{j\in\mathbb{Z}} H(v_{j-k}, \ldots, v_{j+k}) = \sum\nolimits_{-} j \in \mathbb{Z} v_j. \tag{10.47}$$

DÉMONSTRATION. On observe tout d'abord qu'on obtient l'égalité (10.47) par sommation de la relation (10.44) en remarquant que la seconde somme dans le membre de droite est télescopique. La condition est donc nécessaire. Pour montrer qu'elle aussi suffisante, on pose

$$G(v_{j-k}, \ldots, v_{j+k}) = \frac{\Delta x}{\Delta t}\left(v_j - H(v_{j-k}, \ldots, v_{j+k})\right).$$

En vertu de (10.47), on a alors

$$\sum_{j\in\mathbb{Z}} G(v_{j-k}, \ldots, v_{j+k}) = 0.$$

On va maintenant construire une fonction continue h telle que

$$G(v_{j-k}, \ldots, v_{j+k}) = h(v_{j-k+1}, \ldots, v_{j+k}) - h(v_{j-k}, \ldots, v_{j+k-1}) = h_{j+\frac{1}{2}} - h_{j-\frac{1}{2}},\ j \in \mathbb{Z}.$$

Pour cela, il suffit de déterminer $h_{j+\frac{1}{2}}$ et $h_{j-\frac{1}{2}}$ pour $j = 0$. L'égalité (10.47) étant satisfaite pour toute suite sommable, on choisit tout d'abord une suite $\boldsymbol{v}$ telle que $v_j = 0$ si $j \le -k$ ou $j \ge k+1$. On trouve ainsi

$$\begin{aligned} 0 &= \sum_{j\in\mathbb{Z}} G(v_{j-k}, \ldots, v_{j+k}) \\ &= G(0, \ldots, 0, v_{-k+1}) + G(0, \ldots, v_{-k+1}, v_{-k+2}) + \cdots + G(0, v_{-k+1}, \ldots, v_k) \\ &\quad + G(v_{-k+1}, \ldots, v_k, 0) + G(v_{-k+2}, \ldots, v_k, 0, 0) + \cdots + G(v_k, 0, \ldots, 0) \\ &= h_{\frac{1}{2}} + B_k \end{aligned}$$

où $h_{\frac{1}{2}} = G(0, \ldots, 0, v_{-k+1}) + G(0, \ldots, v_{-k+1}, v_{-k+2}) + \cdots + G(0, v_{-k+1}, \ldots, v_k)$ et $B_k = G(v_{-k+1}, \ldots, v_k, 0) + G(v_{-k+2}, \ldots, v_k, 0, 0) + \cdots + G(v_k, 0, \ldots, 0)$. De la même manière, en choisissant une suite $\boldsymbol{v}$ telle que $v_j = 0$ si $j \le -k-1$ ou $j \ge k+1$, il vient

$$\begin{aligned} 0 &= \sum_{j\in\mathbb{Z}} G(v_{j-k}, \ldots, v_{j+k}) \\ &= G(0, \ldots, 0, v_{-k}) + \cdots + G(0, v_{-k}, \ldots, v_{k-1}) + G(v_{-k}, \ldots, v_k) \\ &\quad + G(v_{-k+1}, \ldots, v_k, 0) + G(v_{-k+2}, \ldots, v_k, 0, 0) + \cdots + G(v_k, 0, \ldots, 0) \\ &= h_{-\frac{1}{2}} + G(v_{-k}, \ldots, v_k) + B_k \end{aligned}$$

où $h_{-\frac{1}{2}} = G(0, \ldots, 0, v_{-k}) + \cdots + G(0, v_{-k}, \ldots, v_{k-1})$. Le résultat voulu se déduit alors des deux dernières identités. □

On voit avec ce résultat que les solutions approchées obtenues avec des méthodes dont les schémas peuvent être écrits sous forme conservative possèdent une propriété analogue à la propriété ref des solutions d'une loi de conservation. On voit en effet que l'opérateur $S_{\Delta t}$ préserve l'intégrale discrète ref... TERMINER

Nous verrons plus loin, notamment avec le théorème 10.28, l'importance des méthodes dont le schéma peut être écrit sous forme conservative.

NOTER enfin qu'en pratique, on ne peut effectuer les calculs que sur un nombre *fini* de points alors que le problème est posé sur un domaine non borné ($\mathbb{R}$) en espace (noter que le domaine en temps ne présente pas de difficulté puisque que l'on résout le problème de Cauchy sur un intervalle de temps borné). Pour contourner cette impossibilité, on peut se ramener à un domaine borné $[a, b]$ en imposant à la solution u des conditions de périodicité de la forme

$$u(t, a) = u(t, b),\ t \ge 0.$$

Ceci revient à considérer un problème de Cauchy dont la donnée de la condition initiale est périodique : la solution du problème étant périodique, on peut se restreindre à ne la calculer que sur une période en utilisant la périodicité dans la méthode des différences finies.
AJOUTER que supposer de telles les conditions de périodicité est aussi utile pour l'analyse des schémas (cf. analyse de von Neumann)

10.4.2 Analyse des schémas

Si l'obtention d'un schéma d'approximation pour une loi de conservation est une chose facile (c'est là l'une des principales caractéristiques de la méthode des différences finies), nous allons voir que la dérivation d'un schéma efficace et précis demande un certain travail...

Dans la mesure du possible, les définitions données ici seront générales, au sens où elles peuvent s'appliquer à des schémas pour d'autres équations aux dérivées partielles que celles considérées dans ce chapitre.

On cherche à mesurer comment les quantités u_j^n approchent la solution u du problème de Cauchy lorsque Δt et Δx tendent vers 0. Pour cela, on va définir une erreur globale entre la solution et l'approximation numérique à partir de l'erreur ponctuelle au point de grille (t_n, x_j), définie par

$$E_\Delta(t_n, x_j) = u(t_n, x_j) - u_j^n$$

pour une solution classique, continue en tout point, et par

$$E_\Delta(t_n, x_j) = \left(\frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t_n, x)\,\mathrm{d}x \right) - u_j^n$$

pour une solution faible présentant des discontinuités. Cette définition de l'erreur ponctuelle peut être étendue à tout point (t, x) en construisant une fonction constante par morceaux E_Δ à partir des valeurs $E_\Delta(t_n, x_j)$ en posant

$$E_\Delta(t, x) = E_\Delta(t_n, x_j) \text{ pour } (t, x) \in [t_n, t_{n+1}[\times[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[.$$

On dit alors que la méthode converge lorsque l'erreur globale au temps t, $E_\Delta(t, \cdot)$, tend vers 0 lorsque Δt et Δx tendent vers 0 et que u^0 tend vers la donnée initiale (EXPLIQUER), pour toute valeur de t et toute donnée initiale u_0 dans une certaine classe, pour un certain choix de norme.

Idéalement, on souhaiterait obtenir la convergence en norme L^∞, c'est-à-dire pour la norme discrète

$$\|\boldsymbol{v}\|_{\infty,\Delta x} = \Delta x \max_{j\in\mathbb{Z}} \left|v_j\right|$$

pour toute suite $\boldsymbol{v} = (v_j)_{j\in\mathbb{Z}}$ de valeurs définies aux nœuds x_j, $j \in \mathbb{Z}$, mais ceci est irréaliste lorsque la solution est discontinue car l'erreur ponctuelle ne tend généralement pas vers zéro uniformément au voisinage des discontinuités quand Δt tend vers 0 alors que les résultats numériques sont par ailleurs satisfaisants.

Pour une loi de conservation, un choix naturel de norme est celui de la norme L^1, de norme discrète associée

$$\|\boldsymbol{v}\|_{1,\Delta x} = \Delta x \sum_{j\in\mathbb{Z}} \left|v_j\right|,$$

car celle-ci requiert seulement d'avoir une solution intégrable, la loi de conservation permettant en principe de donner un sens à l'intégrale (A VOIR, relier aux propriétés des solutions dans la sous-section 10.3.1 par exemple).

Pour les problèmes linéaires, une norme particulièrement intéressante est la norme L^2

$$\|\boldsymbol{v}\|_{2,\Delta x} = \left(\Delta x \sum_{j\in\mathbb{Z}} {v_j}^2 \right)^{\frac{1}{2}},$$

car elle permet de faire appel à l'analyse de Fourier, simplifiant ainsi considérablement l'étude comme nous allons le voir. AJOUTER que cette norme est aussi reliée à une énergie associée à la solution

L'analyse directe de la convergence d'une méthode n'est généralement pas chose aisée. Fort heureusement, la notion de convergence est liée à deux autres notions bien plus simples à vérifier qui sont celles de consistance et de stabilité.

Comme cela était le cas pour les méthodes de résolution numérique des équations différentielles ordinaires du chapitre 8, l'analyse d'un schéma de discrétisation de l'équation (10.8) comporte deux étapes fondamentales, qui sont d'une part l'étude de sa consistance, visant à mesurer l'erreur commise en substituant aux opérateurs différentiels des opérateurs aux différences finies, et d'autre part l'étude de sa stabilité, assurant que l'opérateur discret mis en jeu est bien inversible et que la norme de son inverse est bornée indépendamment des pas de discrétisation choisis.

Consistance

La consistance d'un schéma aux différences finies pour la résolution de l'équation (10.8) repose naturellement sur la notion d'erreur de troncature locale. Celle-ci quantifie comment le schéma approche l'équation localement. Pour l'obtenir, on insère la solution de l'équation, que l'on suppose régulière, dans le schéma, qu'elle ne vérifie que de manière approchée.

Définition 10.21 (erreur de troncature locale d'un schéma aux différences finies) *Pour tout entier n de $\mathbb{N}$ et tout entier j de $\mathbb{Z}$, l'**erreur de troncature locale** au point (t_{n+1}, x_j) du schéma aux différences finies* (10.40) *pour la résolution de l'équation* (10.8) *est définie par*

$$\varepsilon_j^{n+1} = u(t_{n+1}, x_j) - H\big(u(t_n, x_{j-k}), \dots, u(t_n, x_{j+k})\big), \tag{10.48}$$

où u est une solution classique de (10.8).

erreur de troncature = résidu obtenu lorsque l'on introduit une solution régulière de l'équation dans le schéma
Le schéma est consistant si

$$\varepsilon_j^{n+1}/\Delta t \to 0 \text{ quand } \Delta t, \Delta x \to 0.$$

ORDRE DU SCHEMA :
le schéma est d'ordre p en temps et q en espace, $p, q \in \mathbb{N}^*$, si

$$\varepsilon_j^{n+1}/\Delta t = O(\Delta t^p) + O(\Delta x^q),$$

quand $\Delta t, \Delta x \to 0$.

On obtient l'ordre en effectuant des développements de Taylor de la fonction u et de $H(u, \dots, u)$. Un schéma est consistant s'il est au moins d'ordre un en temps et en espace.

exemple FTCS : ce schéma est d'ordre un en temps et d'ordre deux en espace, il est donc consistant

Cette définition de l'ordre du schéma n'est pas entièrement satisfaisante. En effet, il se peut dans certains cas que des termes de la forme $O(\Delta x^p \Delta t^{-q})$ apparaissent et on ne peut alors appliquer la définition (voir l'exemple du schéma de Lax–Friedrichs plus loin).

On introduit dans ce cas la définition, de portée plus générale, suivante.

Définition 10.22 (consistance et ordre d'un schéma aux différences finies) *Le schéma aux différences finies* (10.40) *pour la résolution de l'équation* (10.8) *est dit **consistant** si l'erreur de troncature définie par* (10.48) *est un $O(\Delta t^2)$ lorsque le pas de discrétisation en temps tend Δt tend vers zéro, le rapport $\lambda = \frac{\Delta t}{\Delta x}$ étant supposé constant. On dit que ce même schéma est **d'ordre** p, avec p un entier naturel, si, pour toute solution suffisamment régulière, les longueurs Δt et Δx étant liées, on a*

$$u(t + \Delta t, x) - H\left(u(t, x - k\,\Delta x), \dots, u(t, x + k\,\Delta x)\right) = O(\Delta t^{p+1}),\ t \geq 0,\ x \in \mathbb{R},$$

quand le pas Δt tend vers 0.

Pour un schéma pouvant s'écrire sous forme conservative, on a le résultat suivant.

Proposition 10.23 (condition nécessaire et suffisante de consistance d'un schéma écrit sous forme conservative) *Le schéma aux différences finies* (10.45) *est consistant avec l'équation* (10.8) *si l'on a*

$$\forall v \in \mathbb{R},\ h(v, \dots, v) = f(v),$$

à une constante additive près.

DÉMONSTRATION. A ECRIRE □

Consistance du schéma FTCS. Le résultat précédent permet de conclure directement à la consistance du schéma FTCS puisque l'on a dans ce cas

$$\forall v \in \mathbb{R},\ h(v,v) = \frac{1}{2}\left(f(v)+f(v)\right) = f(v).$$

A VOIR : un schéma consistant dont le flux numérique est une fonction de classe $\mathscr{C}^1$ est au moins d'ordre un, expression pour l'erreur de troncature d'un schéma consistant ?

Stabilité

On introduit l'espace $\ell^p_{\Delta x}(\mathbb{Z})$ l'ensemble des suites de norme finie, $p = 1, 2, \ldots, +\infty$,

$$\ell^p_{\Delta x}(\mathbb{Z}) = \left\{ \boldsymbol{v} = (v_j)_{j\in\mathbb{Z}} \,|\, \|\boldsymbol{v}\|_{p,\Delta x} < +\infty \right\}.$$

Définition 10.24 (stabilité en norme L^p) *Le schéma aux différences finies est dit* ***stable en norme*** ℓ^p *s'il existe des constantes positives K et β, indépendantes de Δt et Δx éventuellement choisis suffisamment petits, telles que, pour toute donnée initiale $\boldsymbol{u}^0$ appartenant à $\ell^p_{\Delta x}(\mathbb{Z})$, on a*

$$\|\boldsymbol{u}^n\|_{p,\Delta x} \leq K\, e^{\beta T} \|\boldsymbol{u}^0\|_{p,\Delta x},\ \forall n \in \mathbb{N},\ n\Delta t \leq T. \tag{10.49}$$

En utilisant l'opérateur $S_{\Delta t}$ introduit dans la sous-section 10.4.1, on peut reformuler la condition de stabilité (10.49), au moyen de la norme d'opérateur associée à la norme choisie, de la manière suivante : pour T donné, il existe existe des constantes positives K et β telles que

$$\|{S_{\Delta t}}^n\|_{\mathscr{L}(\ell^p_{\Delta x}(\mathbb{Z}))} \leq C e^{\beta T},$$

pour tous Δt et Δx éventuellement choisis suffisamment petits et tout entier naturel n tels que $0 \leq n\Delta t \leq T$.

Proposition 10.25 (condition nécessaire et suffisante de stabilité en norme L^p) *le schéma est stable en norme ℓ^p s'il existe une constante positive C telle que, pour tous Δt et Δx éventuellement choisis suffisamment petits,*

$$\|S_{\Delta t}\|_{\mathscr{L}(\ell^p_{\Delta x}(\mathbb{Z}))} \leq 1 + C\,\Delta t.$$

DÉMONSTRATION. A ECRIRE □

On voit qu'on a donc besoin de connaître une expression de $\|S_{\Delta t}\|_{\mathscr{L}(\ell^p_{\Delta x}(\mathbb{Z}))}$.

A VOIR : a priori seulement utilisée pour des problèmes dans lesquels un mécanisme de croissance de la solution existe. Condition suffisante : que la norme soit inférieure ou égale à 1.

stabilité en norme L^∞ et lien avec un principe du maximum discret/ Certains schémas ne vérifient pas le principe du maximum discret mais sont néanmoins de « bons » schéma d'approximation.

A REVOIR Pour une équation non linéaire, il n'existe pas de technique générale d'analyse de stabilité. Des notions de stabilité heuristiques ont cependant été introduites, notamment en considérant une linéarisation de l'équation autour d'une solution (?). On vérifie alors la stabilité en norme L^2 dans le cas *linéaire* : ne garantit pas la stabilité dans le cas général mais vise a éliminer les schémas linéairement instables. Cette dernière se prête bien à l'étude de stabilité en domaine infini ou lorsque les conditions aux limites sont périodiques, via l'analyse de Fourier, ou de inégalité d'énergie pour d'autres conditions aux limites...

On peut peut parfois obtenir un résultat de stabilité en norme ℓ^2 dans le cas d'un problème linéaire à coefficients constants du type (10.8)-(10.9), le schéma numérique (10.40) étant alors de la forme

$$v_i^{n+1} = \sum_{j=-k}^{k} c_j\, v_{i+j}^n,$$

au moyen de l'analyse de Fourier, par le biais d'une technique étant due à von Neumann [CFvN50]. Pour cela, introduisons la *transformée de Fourier* $\hat{v}$ d'une suite $(v_j)_{j\in\mathbb{Z}}$ de $\ell^2_{\Delta x}(\mathbb{Z})$, définie par

$$\hat{v}(\xi) = \frac{\Delta x}{\sqrt{2\pi}} \sum_{j\in\mathbb{Z}} e^{-\mathrm{i}\xi j\Delta x}\, v_j,\ \xi \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]. \tag{10.50}$$

Cette fonction appartient à $L^2\left(\left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]\right)$, l'espace des classes de fonctions mesurables (au sens de Lebesgue) sur [29] l'intervalle $\left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]$ dont la norme

$$\|\hat{v}\|_2 = \left(\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} |\hat{v}(\xi)|^2 \, \mathrm{d}\xi\right)^{\frac{1}{2}}$$

est finie et sa transformée inverse est donnée par

$$v_j = \frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{\mathrm{i}\xi j\Delta x} \, \hat{v}(\xi) \, \mathrm{d}\xi, \ j \in \mathbb{Z}.$$

De manière classique [30], les normes $\|\hat{v}\|_2$ et $\|\boldsymbol{v}\|_{2,\Delta x}$ sont liées par l'*égalité de Parseval* [31] suivante

$$\|\hat{v}\|_2 = \|\boldsymbol{v}\|_{2,\Delta x}. \tag{10.51}$$

Appliquons la transformation de Fourier à la relation (10.43). Compte tenu des propriétés de la transformée de Fourier et après quelques manipulations, on trouve

$$\hat{u}^{n+1}(\xi) = g(\xi)\hat{u}^n(\xi), \ \xi \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right],$$

et l'on a

$$\|S_{\Delta t}\|_{\mathscr{L}(\ell^2_{\Delta x}(\mathbb{Z}))} = \max_{\xi \in \left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]} |g(\xi)|,$$

$g(\xi)$ étant appelée le *facteur d'amplification* (*amplification factor* en anglais) de la composante de Fourier de *nombre d'onde* (*wave number* en anglais) ξ de la suite $\boldsymbol{v}$.

une procédure équivalente pour dériver $g(\xi)$ est de remplacer dans le schéma u_j^n par $u_\Delta^n(\xi)\, e^{\mathrm{i}\xi x_j}$ (onde plane) et de chercher une relation de récurrence pour la suite $(u_\Delta^n(\xi))_n$.

Analyse de von Neumann du schéma FTCS. On trouve

$$g(\xi) = 1 - a\frac{\Delta t}{2\Delta x}\left(e^{\mathrm{i}\xi\Delta x} - e^{-\mathrm{i}\xi\Delta x}\right) = 1 - \mathrm{i}a\frac{\Delta t}{\Delta x}\sin(\xi\Delta x)$$

il est inconditionnellement instable si le rapport $\lambda = \frac{\Delta t}{\Delta x}$ est constant (voir figure 10.7).

Note : stabilité si $\frac{\Delta t}{(\Delta x)^2}$ constant, pas intéressant en pratique

Remarques : analyse réservée aux équations linéaires à coef. constants, possibilité de prendre en compte des conditions aux limites périodiques (car équivalence en général avec problème posé sur la droite avec une condition initiale de donnée périodique). Théorie de stabilité (linéaire) avec des conditions plus générales est souvent difficiles (relation subtile entre les discrétisations de l'edp et des conditions aux limites)

Proposition 10.26 *CNS de stabilité en norme L^2 pour les schémas consistants à trois points que l'on peut écrire sous forme conservative ? ? ?*

Note : cette proposition va faire apparaître le condition CFL

A VOIR : notion de dissipation et de dispersion numériques d'un schéma

A VOIR : schéma dissipatif au sens de Kreiss [Kre64] (surtout important pour les systèmes)

29. *A priori*, la fonction donnée par (10.50) est définie pour toute valeur du nombre réel ξ. On observe cependant qu'elle est $\frac{2\pi}{\Delta x}$-périodique (on parle de *repliement de spectre* (*aliasing* en anglais)) et l'on se restreint donc à l'intervalle $\left[-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right]$ pour la définition de la transformée.

30. On a en effet

$$\|\hat{v}\|_2{}^2 = \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} |\hat{v}(\xi)|^2 \, \mathrm{d}\xi = \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \overline{\hat{v}(\xi)} \frac{\Delta x}{\sqrt{2\pi}} \sum_{j\in\mathbb{Z}} e^{-\mathrm{i}\xi j\Delta x} \, v_j \, \mathrm{d}\xi = \Delta x \sum_{j\in\mathbb{Z}} \left(\frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{-\mathrm{i}\xi j\Delta x} \overline{\hat{v}(\xi)} \, \mathrm{d}\xi\right) v_j$$

$$= \Delta x \sum_{j\in\mathbb{Z}} \overline{\frac{1}{\sqrt{2\pi}} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{-\mathrm{i}\xi j\Delta x} \hat{v}(\xi) \, \mathrm{d}\xi} \, v_j = \Delta x \sum_{j\in\mathbb{Z}} \overline{v_j} v_j = \|\boldsymbol{v}\|_{2,\Delta x}{}^2,$$

EN JUSTIFIANT l'échange des signes somme et intégrale.

31. Marc-Antoine Parseval des Chênes (27 avril 1755 - 16 août 1836) était un mathématicien français, célèbre pour sa découverte d'une formule fondamentale de la théorie des séries de Fourier.

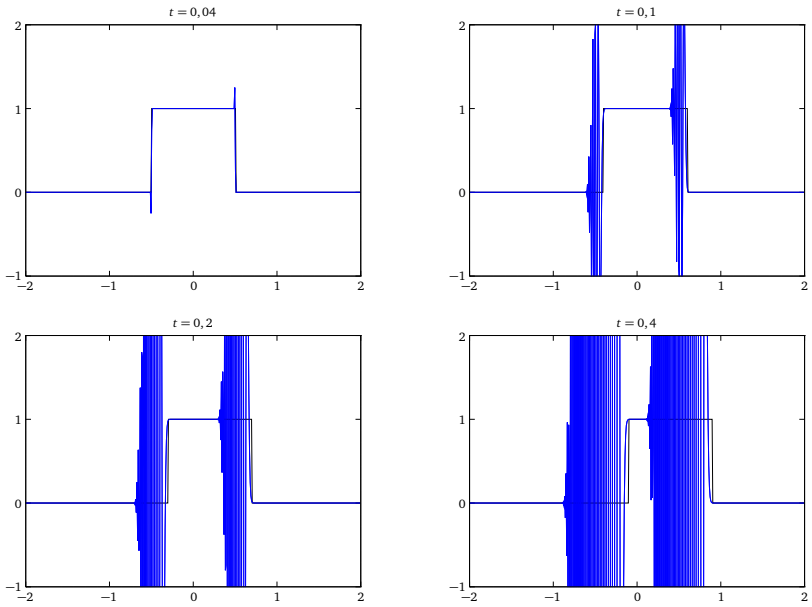


FIGURE 10.7 – simulation avec le schéma FTCS, equation de transport $a = 1$, donnée initiale de type créneau, condition aux limites périodiques, cfl $= 0,5$. On observe que l'instabilité du schéma se traduit par l'apparition et la croissance rapide d'oscillations à haute fréquence au voisinage des discontinuités.

Convergence

La résolution approchée du problème (10.8)-(10.9) par une méthode numérique passe tout d'abord par l'approximation de la condition initiale (10.9). On suppose dans la suite que l'on procède comme suit pour construire la donnée initiale $\boldsymbol{u}^0 = (u_j^0)_{j\in\mathbb{Z}}$ pour le schéma

$$u_j^0 = u_0(x_j),\ j \in \mathbb{Z},$$

si la fonction u_0 est continue,

$$u_j^0 = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\,\mathrm{d}x,\ j \in \mathbb{Z},$$

sinon.

résultat fondamental pour les équations linéaires

Théorème 10.27 (« théorème de Lax–Richtmyer[32] » ou « théorème d'équivalence de Lax » [LR56]) *Une condition nécessaire et suffisante à la convergence d'une approximation de la solution d'un problème linéaire et bien posé par une méthode aux différences finies consistante est que la méthode soit stable.*

DÉMONSTRATION. A ECRIRE
On montre d'abord que la stabilité de la méthode implique sa convergence. On montre ensuite qu'un schéma instable ne peut converger, ce qui complète la preuve. □

On retrouve ici un résultat du même type que le théorème d'équivalence de Dahlquist pour les méthodes à pas multiples (voir le théorème 8.39).

L'analyse de convergence pour des problèmes non linéaires est bien plus difficile. Pour des solutions régulières, la stabilité d'une *linéarisation* d'un schéma consistant suffit pour prouver sa convergence (voir [Str64]).

32. Robert Davis Richtmyer (10 octobre 1910 - 24 septembre 2003) était un physicien et mathématicien américain. Il s'intéressa notamment à la résolution numérique de problèmes d'hydrodynamique et réalisa quelques-unes des premières applications à grande échelle de la méthode de Monte-Carlo sur le calculateur SSEC d'IBM.

Résultat important relatif à la convergence vers une solution faible (mais pas forcément entropique) d'un schéma sous forme conservative convergent

Théorème 10.28 (« théorème de Lax–Wendroff[33] » [LW60]) *Si un schéma sous forme conservative et consistant converge dans L^1_{loc}, c'est vers une solution faible (pouvant éventuellement violer la condition d'entropie).*

DÉMONSTRATION. A ECRIRE □

REMARQUE sur la convergence de schémas qui ne sont pas sous forme conservative

La condition de Courant–Friedrichs–Lewy

introduction de la *condition de Courant–Friedrichs*[34]*–Lewy*[35] [CFL28] (une traduction anglaise de cet article est disponible [CFL67]) qui est une condition nécessaire de convergence pour un schéma

EXPLICATIONS : le domaine de dépendance de la solution de l'équation au point (x,t), avec $t>0$, est l'ensemble des points en espace en lesquels la donnée initiale à $t=0$ affecte la valeur $u(t,x)$ de la solution. Pour une équation hyperbolique, ce domaine est borné pour tout (x,t). Par exemple, pour l'équation d'advection linéaire cet ensemble est le singleton $\{x-a\,t\}$, mais il peut être plus généralement contenu entre des caractéristiques.

énoncé heuristique : L'approximation numérique de la solution possède aussi un domaine de dépendance en tout point (x_j,t_n). Avec un schéma explicite, l'approximation à l'instant t_n dépend d'un nombre fini de valeurs obtenues aux instants précédents. Dans ce cas, le domaine de dépendance numérique est constitué de l'ensemble des nœuds en espace appartenant à la base d'un triangle issu du point (x_j,t_n) et contenant tous les points de grille en lesquels la valeur de la solution intervient dans le calcul de u_j^n (note : ce triangle est isocèle pour certaines formules). Pour tout Δt, cet ensemble est discret, mais ce qui importe est l'ensemble limite lorsque $\Delta t\to 0$. Cet ensemble limite est un sous-ensemble fermé de l'espace, c'est-à-dire un intervalle. Discussion sur la taille de l'intervalle en fonction de la valeur de $\lambda=\frac{\Delta t}{\Delta x}$ pour un schéma explicite à trois points

FAIRE UN DESSIN

(note pour un schéma implicite : domaine non borné)

condition de CFL : un schéma ne peut converger s'il ne prend en compte toutes les données nécessaires : à la limite le domaine de dépendance numérique doit contenir le domaine de dépendance de la solution

Exemple pour une équation linéaire et un schéma à trois points :

$$|a|\,\frac{\Delta t}{\Delta x}\le 1 \tag{10.52}$$

Un schéma ne satisfaisant pas cette condition ne peut converger et donc, en vertu du théorème 10.27, être stable. Il n'est donc nullement étonnant de retrouver en (10.52) la condition de stabilité obtenue dans la proposition 10.26.

VOIR Aussi : Strang (1962) : lien entre CFL et stabilité de von Neumann

A VOIR : Consistance avec une condition d'entropie

DEFINITION

Sous les conditions du théorème de Lax–Wendroff, un schéma consistant avec toute condition d'entropie (schéma dit entropique ? ? ?) converge vers l'unique solution entropique du problème

Monotonie lien avec un principe du maximum satisfait par la solution que l'on peut retrouver au niveau discret

Définition 10.29 (schéma monotone) *Un schéma est dit **monotone** si,* $\forall n\in\mathbb{N}$,

$$\{u_j^n\ge v_j^n,\ \forall j\in\mathbb{Z}\}\Rightarrow\{u_j^{n+1}\ge v_j^{n+1},\ \forall j\in\mathbb{Z}\}$$

33. Burton Wendroff (né le 10 mars 1930) est un mathématicien américain, connu pour ses contributions au développement de schémas numériques de résolution des équations aux dérivées partielles hyperboliques.

34. Kurt Otto Friedrichs (28 septembre 1901 - 31 décembre 1982) était un mathématicien germano-américain. L'essentiel de ses travaux fut consacré à l'étude théorique des équations aux dérivées partielles, à leur résolution numérique et leurs application en physique quantique, en mécanique des fluides et en élasticité.

35. Hans Lewy (20 octobre 1904 - 23 août 1988) était un mathématicien américain, connu pour ses travaux sur les équations aux dérivées partielles et sur la théorie des fonctions de plusieurs variables complexes.

Proposition 10.30 (condition nécessaire et suffisante de monotonie d'un schéma) *Un schéma de la forme* (10.40) *est monotone si et seulement si la fonction H est une fonction croissante de chacun de ses arguments.*

DÉMONSTRATION. A ECRIRE □

condition réalisée en pratique sous condition type CFL

Lemme 10.31 (propriétés d'un schéma monotone) *A ECRIRE*

DÉMONSTRATION. A ECRIRE □

A VOIR : lien entre schémas entropiques et monotones : un schéma sous forme conservative, consistant et monotone est entropique

aussi : dans le cas scalaire, un schéma monotone convergent converge vers la solution entropique [HHL$^+$76]

Le résultat suivant concerne l'ordre d'un schéma monotone.

Théorème 10.32 (« théorème de Godunov[36] » [God59]) *Un schéma sous forme conservative, consistant avec l'équation* (10.8) *et monotone est exactement d'ordre un.*

DÉMONSTRATION. A ECRIRE □

10.4.3 Méthodes pour les équations hyperboliques linéaires **

INTRO On va ici traiter majoritairement de la résolution numérique de l'équation d'advection. On présentera aussi quelques schémas pour l'équation des ondes (EDP hyperbolique d'ordre **deux**)

INTRODUIRE quelques exemples de schémas et les analyser en supposant que $\lambda = \frac{\Delta t}{\Delta x}$ est fixé

classe de schémas à trois points

Méthode de Lax–Friedrichs

Le problème d'instabilité de la précédente méthode peut être aisément résolu. La *méthode de Lax–Friedrichs* [Lax54] fait en effet le choix de remplacer la quantité u_j^n dans le schéma (10.41) par la moyenne de u_{j+1}^n et u_{j-1}^n, ce qui conduit à

$$u_j^{n+1} = \frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right) - a\frac{\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right). \tag{10.53}$$

nous verrons que cette modification correspond à l'ajout d'un terme de viscosité artificielle au schéma FTCS, assurant sa stabilité sous CFL.

résultat via analyse von Neumann

$$g(\xi) = \frac{1}{2}\left(e^{\mathrm{i}\xi\Delta x} + e^{-\mathrm{i}\xi\Delta x}\right) - a\frac{\Delta t}{2\Delta x}\left(e^{\mathrm{i}\xi\Delta x} - e^{-\mathrm{i}\xi\Delta x}\right) = \cos(\xi\Delta x) - \mathrm{i}a\frac{\Delta t}{\Delta x}\sin(\xi\Delta x)$$

+ stabilité en norme L^p

au final : conservatif, consistant, linéairement stable sous condition CFL, monotone

hautement dissipatif (voir équation équivalente)

Schéma décentré amont

schéma décentré amont ("upwind" en anglais) [CIR52], particulier au cas linéaire (ou pour f monotone dans le cas non-linéaire)

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\,\frac{u_{j+1}^n - u_{j-1}^n}{2\,\Delta x} - |a|\,\frac{u_{j+1}^n - 2\,u_j^n + u_{j-1}^n}{2\,\Delta x} = 0 \tag{10.54}$$

soit encore

36. Sergei Konstantinovich Godunov (Сергéй Константúнович Годунóв en russe, né le 17 juillet 1929) est un mathématicient russe. Il est connu pour ses apports fondamentaux aux méthodes d'approximation utilisées en mécanique des fluides numérique.

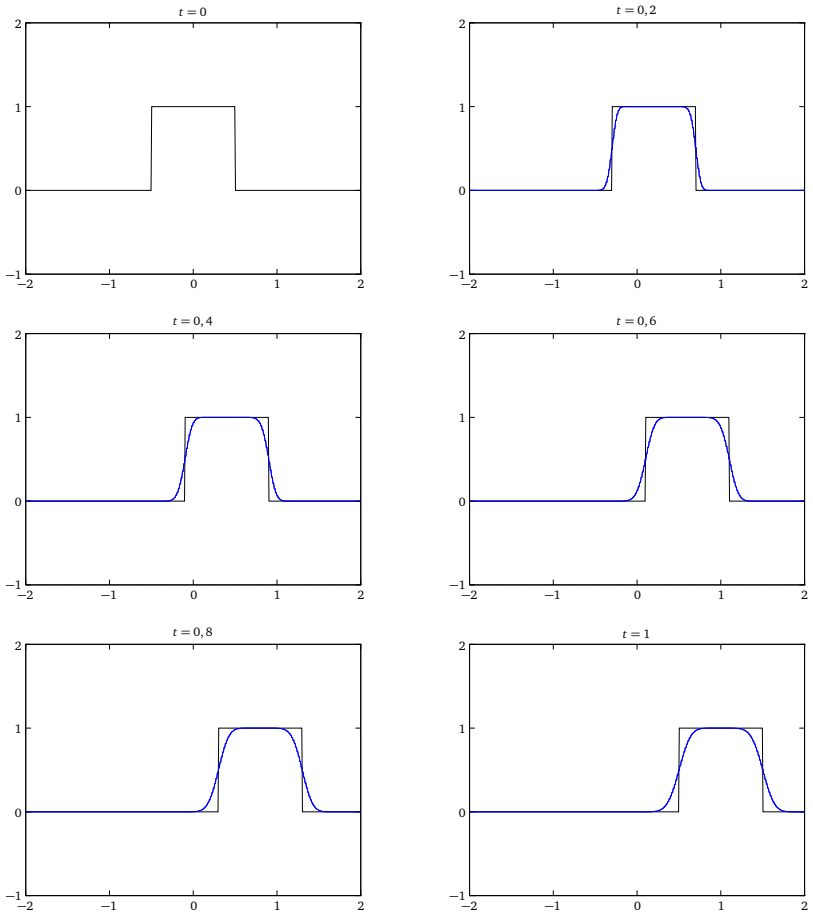


FIGURE 10.8 – simulation avec la méthode de Lax–Friedrichs, equation de transport $c = 1$, donnée initiale de type créneau, condition aux limites périodiques, cfl $= 0,5$

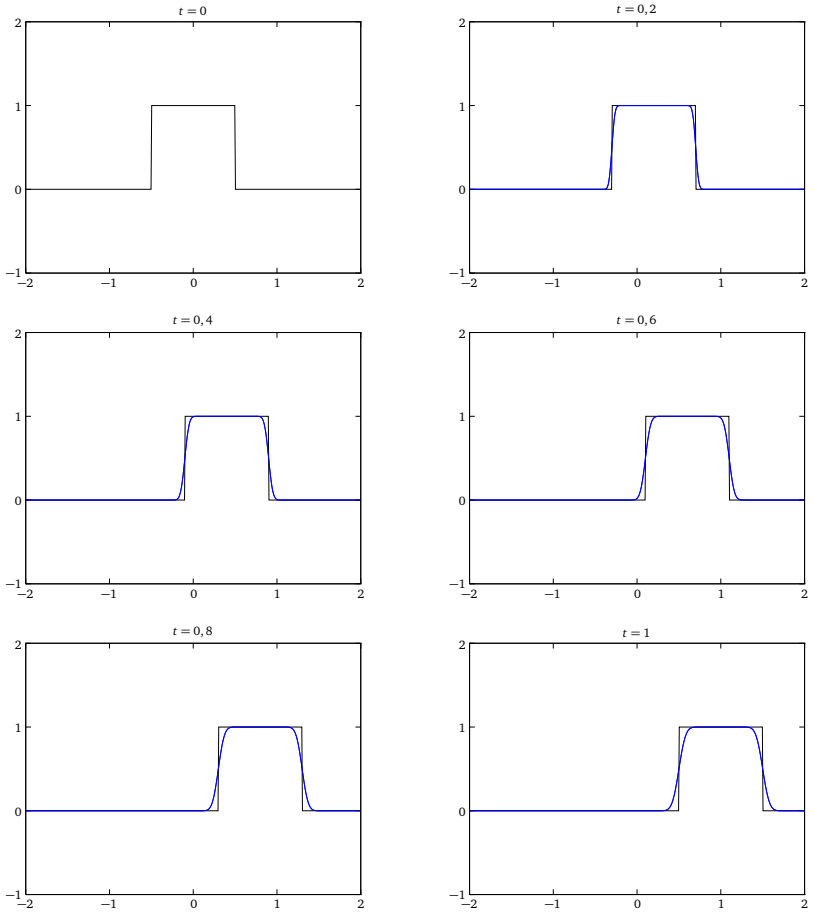


FIGURE 10.9 – simulation avec le schéma décentré amont, équation de transport $c = 1$, donnée initiale de type créneau, condition aux limites périodiques, cfl $= 0,5$

$$\begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + a \dfrac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \text{ si } a > 0 \\ \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + a \dfrac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \text{ si } a < 0 \end{cases}$$

facteur d'amplification pour $a > 0$

$$g(\xi) = 1 - a\frac{\Delta t}{\Delta x} + a\frac{\Delta t}{\Delta x}\, e^{-\mathrm{i}\xi\Delta x}$$

d'où stabilité si $a\frac{\Delta t}{\Delta x} \leq 1$

Méthode de Lax–Wendroff

Le schéma de *la méthode de Lax–Wendroff* [LW60] peut se construire en utilisant un développement de Taylor en $t = t_n$ tronqué au second ordre, etc.

on trouve

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{2\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right) + a^2\frac{(\Delta t)^2}{2(\Delta x)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) \tag{10.55}$$

A VOIR : autre dérivation du schéma par résolution d'un problème de Riemann avec flux numérique

conservatif, consistant, linéairement stable, non entropique (voir simulation)
stabilité sous CFL :

$$\begin{aligned} g(\xi) &= 1 - a\frac{\Delta t}{2\Delta x}\left(e^{\mathrm{i}\xi\Delta x} - e^{-\mathrm{i}\xi\Delta x}\right) + a^2\frac{(\Delta t)^2}{2(\Delta x)^2}\left(e^{\mathrm{i}\xi\Delta x} - 2 + e^{-\mathrm{i}\xi\Delta x}\right) \\ &= 1 - \mathrm{i}a\frac{\Delta t}{\Delta x}\sin(\xi\Delta x) + a^2\frac{(\Delta t)^2}{(\Delta x)^2}\left(\cos(\xi\Delta x) - 1\right) \end{aligned}$$

on utilise alors que $\cos(\xi\Delta x) - 1 = -2\sin^2\left(\frac{\xi\Delta x}{2}\right)$ pour trouver

$$|g(\xi)|^2 = 1 - 4a^2\frac{(\Delta t)^2}{(\Delta x)^2}\sin^2\left(\frac{\xi\Delta x}{2}\right) + 4a^4\frac{(\Delta t)^4}{(\Delta x)^4}\sin^4\left(\frac{\xi\Delta x}{2}\right) + a^2\frac{(\Delta t)^2}{(\Delta x)^2}\sin^2(\xi\Delta x)$$

en appliquant l'identité $\sin^2(\theta) = 4\sin^2\left(\frac{\theta}{2}\right)\cos^2\left(\frac{\theta}{2}\right) = 4\left(\sin^2\left(\frac{\theta}{2}\right) - \sin^4\left(\frac{\theta}{2}\right)\right)$ au dernier terme, on trouve

$$|g(\xi)|^2 = 1 + 4a^2\frac{(\Delta t)^2}{(\Delta x)^2}\left(a^2\frac{(\Delta t)^2}{(\Delta x)^2} - 1\right)\sin^4\left(\frac{\xi\Delta x}{2}\right),$$

et le schéma est donc stable si $4a^2\frac{(\Delta t)^2}{(\Delta x)^2}\left(a^2\frac{(\Delta t)^2}{(\Delta x)^2} - 1\right)\sin^4\left(\frac{\xi\Delta x}{2}\right)$ prend une valeur négative pour toute valeur de ξ, ce qui est le cas si $|a|\frac{\Delta t}{\Delta x} \leq 1$.

ce schéma approche l'équation (équivalente) dispersive

$$\frac{\partial u}{\partial t}(t,x) + a\frac{\partial u}{\partial x}(t,x) = \varepsilon\frac{\partial^3 u}{\partial x^3}(t,x)$$

avec $\varepsilon = \frac{\Delta x^2}{6}a\left(a^2\frac{\Delta t}{\Delta x}^2 - 1\right)$

oscillations au voisinage des discontinuités (voir la figure 10.10) : *phénomène de Gibbs* (voir la référence [HH79]), illustration que ce schéma ne vérifie pas le principe du maximum discret

Autres schémas

schéma de Beam–Warming [WB76] : schéma décentré vers l'amont d'ordre deux (dérivation similaire à celle de L–W)

$$\begin{aligned} &\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{3u_j^n - 4u_{j-1}^n + u_{j-2}^n}{2\Delta x} - a^2\frac{\Delta t}{2}\frac{u_j^n - 2u_{j-1}^n + u_{j-2}^n}{(\Delta x)^2} = 0 \text{ si } a > 0 \\ &\frac{u_j^{n+1} - u_j^n}{\Delta t} - a\frac{3u_j^n - 4u_{j+1}^n + u_{j+2}^n}{2\Delta x} - a^2\frac{\Delta t}{2}\frac{u_j^n - 2u_{j+1}^n + u_{j+2}^n}{(\Delta x)^2} = 0 \text{ si } a < 0 \end{aligned}$$

stabilité pour $a > 0$:

$$\begin{aligned} g(\xi) &= \frac{1}{2}\left(2 - a\frac{\Delta t}{\Delta x}\right)\left(1 - a\frac{\Delta t}{\Delta x}\right) + a\frac{\Delta t}{\Delta x}\left(2 - a\frac{\Delta t}{\Delta x}\right)e^{-\mathrm{i}\xi\Delta x} - \frac{1}{2}a\frac{\Delta t}{\Delta x}\left(1 - a\frac{\Delta t}{\Delta x}\right)e^{-2\mathrm{i}\xi\Delta x} \\ &= \left(-a\frac{\Delta t}{\Delta x}\left(1 - a\frac{\Delta t}{\Delta x}\right)\cos(\xi\Delta x) + \left(1 - a\frac{\Delta t}{\Delta x}\right)e^{\mathrm{i}\xi\Delta x} + a\frac{\Delta t}{\Delta x}\left(2 - a\frac{\Delta t}{\Delta x}\right)\right)e^{-\mathrm{i}\xi\Delta x} \end{aligned}$$

d'où

$$\begin{aligned} |g(\xi)|^2 &= \left(\left(1 - a\frac{\Delta t}{\Delta x}\right)^2\cos(\xi\Delta x) + a\frac{\Delta t}{\Delta x}\left(2 - a\frac{\Delta t}{\Delta x}\right)\right)^2 + \left(\left(1 - a\frac{\Delta t}{\Delta x}\right)\sin(\xi\Delta x)\right)^2 \\ &= 1 - a\frac{\Delta t}{\Delta x}\left(1 - a\frac{\Delta t}{\Delta x}\right)^2\left(2 - a\frac{\Delta t}{\Delta x}\right)(1 - \cos(\xi\Delta x))^2 \end{aligned}$$

en remarquant que $\left(a\frac{\Delta t}{\Delta x}\right)^2\left(2 - a\frac{\Delta t}{\Delta x}\right)^2 + \left(1 - a\frac{\Delta t}{\Delta x}\right)^2 = 1 - \left(1 - a\frac{\Delta t}{\Delta x}\left(1 - a\frac{\Delta t}{\Delta x}\right)^2\left(2 - a\frac{\Delta t}{\Delta x}\right)\right)$.

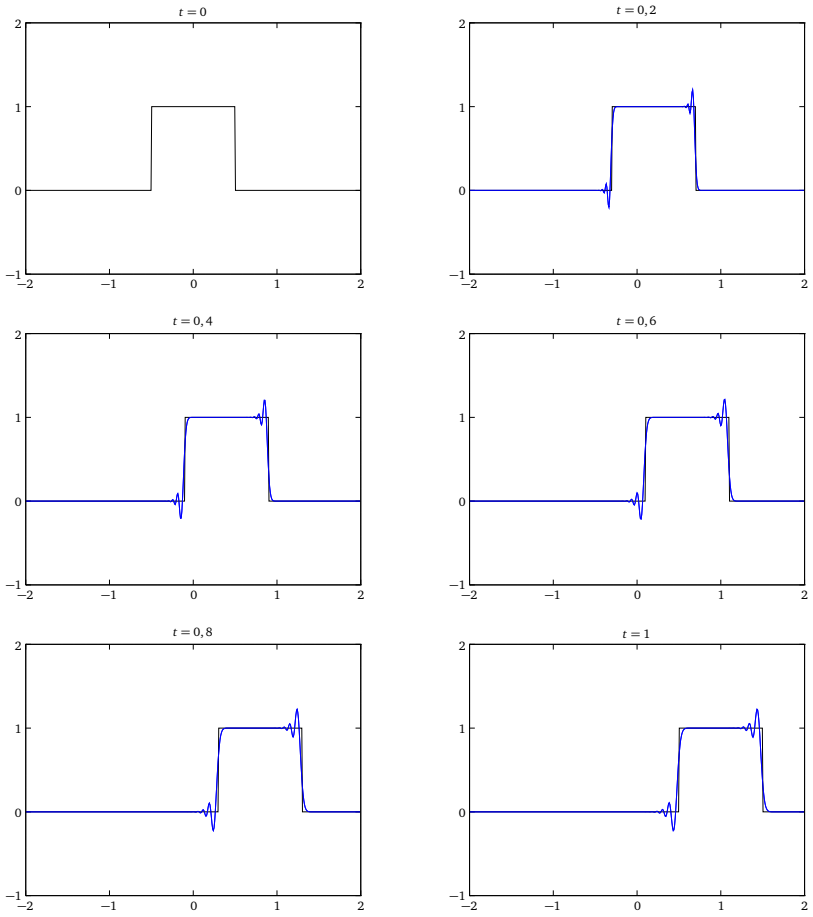


FIGURE 10.10 – simulation avec la méthode de Lax–Wendroff, équation de transport $c = 1$, donnée initiale de type créneau, condition aux limites périodiques, cfl $= 0,5$

Le coefficient $a\frac{\Delta t}{\Delta x}\left(1-a\frac{\Delta t}{\Delta x}\right)^2\left(2-a\frac{\Delta t}{\Delta x}\right)$ est positif sous la condition

$$a\frac{\Delta t}{\Delta x} \leq 2,$$

qui traduit le bénéfice apporté par le plus grand support de ce schéma par rapport à ceux présentés jusqu'à présent.

schéma centré implicite

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{4\,\Delta x}\left(u_{j+1}^{n+1} - u_{j-1}^{n+1}\right) \tag{10.56}$$

ou

$$u_j^{n+1} = u_j^n - a\frac{\Delta t}{4\,\Delta x}\left((u_{j+1}^{n+1} + u_{j+1}^n) - (u_{j-1}^{n+1} + u_{j-1}^n)\right)$$

schéma à plusieurs pas de temps : schéma dit *« saute-mouton »* (*leapfrog* en anglais)

$$u_j^{n+1} = u_j^{n-1} - a\frac{\Delta t}{\Delta x}\left(u_{j+1}^n - u_{j-1}^n\right)$$

requiert une procédure de démarrage particulière (deux données)

schéma de Carlson [Car59], dit *« diamant »*

$$\frac{u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n}{\Delta t} + a\,\frac{u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}}}{\Delta x} = 0, \tag{10.57}$$

$$u_{j+\frac{1}{2}}^{n+1} + u_{j+\frac{1}{2}}^n = u_{j+1}^{n+\frac{1}{2}} - u_j^{n+\frac{1}{2}} = 2\,u_{j+\frac{1}{2}}^{n+\frac{1}{2}}. \tag{10.58}$$

(utilisé pour la simulation du transport de neutrons par une équation linéaire)

Cas de l'équation des ondes ***

Pour la résolution de l'équation du second ordre (10.5)

schéma saute-mouton

$$u_j^{n+1} - 2\,u_j^n + u_j^{n-1} = \left(c\frac{\Delta t}{\Delta x}\right)^2\left(u_{j+1}^n - 2\,u_j^n + u_{j-1}^n\right)$$

méthode de Newmark [New59]

$$\begin{aligned} u_j^{n+1} - u_j^n &= (\Delta t)v_j^n + \left(c\tfrac{\Delta t}{\Delta x}\right)^2\left(\beta\, w_j^{n+1} + \left(\tfrac{1}{2}-\beta\right)w_j^n\right) \\ v_j^{n+1} - v_j^n &= \tfrac{1}{\Delta t}\left(c\tfrac{\Delta t}{\Delta x}\right)^2\left(\theta\, w_j^{n+1} + (1-\theta)\,w_j^n\right) \end{aligned}$$

avec $w_j = u_{j+1} - 2\,u_j + u_{j-1}$ et où les paramètres β et θ satisfont $0 \leq \beta \leq \frac{1}{2}$, $0 \leq \theta \leq 1$.

10.4.4 Méthodes pour les lois de conservation non linéaires

INTRO cas passablement plus compliqué que pour les équations linéaires, notamment en raison des notions de solutions entropiques. On va exiger que le schéma satisfasse davantage de propriétés, par exemple pour l'approximation de lois de conservation.

Extensions des schémas précédemment introduits

Le schéma FTCS pour la résolution d'une équation non-linéaire est

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\,\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right)$$

La généralisation de la méthode de Lax–Friedrichs à la résolution d'une équation non-linéaire prend la forme

$$u_j^{n+1} = \frac{1}{2}\left(u_{j+1}^n + u_{j-1}^n\right) - \frac{\Delta t}{2\,\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right). \tag{10.59}$$

Ce schéma peut être écrit sous forme conservative en posant

$$h(u_j,u_{j+1})=\frac{1}{2}\left(f(u_j)+f(u_{j+1})\right)+\frac{\Delta x}{2\Delta t}\left(u_j-u_{j+1}\right)$$

méthode de Courant–Isaacson–Rees [CIR52], dont découle le schéma décentré amont (10.54) dans le cas linéaire

$$u_j^{n+1}=u_j^n-\frac{\Delta t}{\Delta x}f'(u_j^n)\left(u_j^n-u_{j-1}^n\right)\text{ si }f'(u_j^n)>0$$

permet approximation naturelle d'une loi de conservation sous forme non conservative

$$\frac{\partial u}{\partial t}+f'(u)\frac{\partial u}{\partial x}=0$$

mais pas adapté aux problèmes pour lesquels la solution présente des discontinuités (car non écrit sous forme conservative sauf si f est monotone)

nous présenterons dans la suite plusieurs schémas pour lesquels on retrouve le schéma décentré dans le cas linéaire/qui étendent de manière conservative ce schéma au cas non linéaire.

généralisations de la méthode de Lax–Wendroff pour le non-linéaire : on peut procéder de plusieurs manières :

extension conservative du schéma :

$$u_j^{n+1}=u_j^n-\frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n)-f(u_{j-1}^n)\right)+\frac{(\Delta t)^2}{2(\Delta x)^2}\left(f'\left(\frac{u_{j+1}^n+u_j^n}{2}\right)\left(f(u_{j+1}^n)-f(u_j^n)\right)\right.$$
$$\left.-f'\left(\frac{u_j^n+u_{j-1}^n}{2}\right)\left(f(u_j^n)-f(u_{j-1}^n)\right)\right)$$

deux approches de type prédicteur-correcteur qui évitent d'utiliser f' :

- méthode à deux étapes de Richtmyer (*Richtmyer two-step Lax–Wendroff method* en anglais) [Ric62]
  prédiction :
  $$u_{j+\frac{1}{2}}^{n+\frac{1}{2}}=\frac{1}{2}\left(u_j^n+u_{j+1}^n\right)+\frac{\Delta t}{2\Delta x}\left(f(u_j^n)-f(u_{j+1}^n)\right)$$
  correction :
  $$u_j^{n+1}=u_j^n-\frac{\Delta t}{\Delta x}\left(f(u_{j+\frac{1}{2}}^{n+\frac{1}{2}})-f(u_{j-\frac{1}{2}}^{n+\frac{1}{2}})\right)$$
- méthode de MacCormack [Mac69], plus simple à mettre en œuvre (?)
  prédiction (forward differencing) :
  $$u_j^{\overline{n+1}}=u_j^n-\frac{\Delta t}{\Delta x}\left(f(u_{j+1}^n)-f(u_j^n)\right)$$
  correction (backward differencing) :
  $$u_j^{n+1}=\frac{1}{2}\left(u_j^{\overline{n+1}}+u_j^n\right)-\frac{\Delta t}{2\Delta x}\left(f(u_j^{\overline{n+1}})-f(u_{j-1}^{\overline{n+1}})\right)$$

  On peut aussi renverser l'ordre (backward puis forward differencing) dans les étapes.

Les deux derniers schémas sont sous forme conservative (à vérifier).

Pour chacune de ces trois méthodes, on retrouve le schéma (10.55) dans le cas $f(u)=a\,u$. Elles sont d'ordre deux.

Méthode de Godunov

schéma reposant sur la résolution exacte de problèmes de Riemann locaux [God59]

la solution au centre d'une cellule en espace $u(t,x_{j+\frac{1}{2}})$ est approchée par $u_{j+\frac{1}{2}}^n$ pour $t\in]t_n,t_{n+1}[$ où $u_{j+\frac{1}{2}}^n$ est la valeur au point $x_{j+\frac{1}{2}}$ de la solution (exacte) du problème de Riemann local suivant

$$\frac{\partial w}{\partial t}+\frac{\partial f(w)}{\partial x}=0,\ t\in]t_n,t_{n+1}[$$
$$w(t_n,x)=\begin{cases}u_j^n & x<x_{j+\frac{1}{2}}\\ u_{j+1}^n & x>x_{j+\frac{1}{2}}\end{cases}$$

COMPLETER
très diffusif, amélioration par montée en ordre, voir les notes de fin de chapitre

méthode de Murman–Roe

Murman–Roe [Mur74, Roe81] (conservatif, consistant, linéairement stable dans certains cas, non entropique)
flux-difference splitting leading to an approximate solution of a Riemann problem (see Steger and Warming)

Méthode d'Engquist–Osher

Engquist–Osher [EO80] (conservatif, consistant, linéairement stable, monotone)
schéma basé sur la résolution exacte de problèmes de Riemann approchés

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right) + \frac{\Delta t}{2\Delta x}\left(\int_{u_j^n}^{u_{j+1}^n} |a(v)|\,\mathrm{d}v - \int_{u_{j-1}^n}^{u_j^n} |a(v)|\,\mathrm{d}v\right)$$

flux numérique

$$g(v_j, v_{j+1}) = \frac{1}{2}\left(f(v_j) + f(v_{j+1}) - \int_{v_j}^{v_{j+1}} |a(v)|\,\mathrm{d}v\right)$$

dans les domaines où le signe de f' est constant, on retrouve le schéma décentré amont

Autres schémas

REPRENDRE autres schémas d'ordre deux de construction analogue : schéma de Beam–Warming [WB76] (VERIFIER)

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n) + \frac{1-\eta}{2}\left(f(u_j^n) - 2f(u_{j-1}^n) + f(u_{j-2}^n)\right)\right)$$

erreur plus faible ($\varepsilon = \frac{\Delta x^2}{6} a\left(2 - 3a\frac{\Delta t}{\Delta x} + a^2\frac{\Delta t}{\Delta x}\right)$), oscillations en aval des discontinuités, l'avantage de ce schéma est qu'il permet de prendre des pas de temps deux fois plus grands que les schémas décentré amont, de Lax–Friedrichs ou de Lax–Wendroff, tout en restant explicite.

et schéma de Fromm [Fro68] (utilisation d'une formule décentrée vers l'amont à trois pas en espace)

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n) + \frac{1-\eta}{4}\left(f(u_{j+1}^n) - f(u_j^n) - f(u_{j-1}^n) + f(u_{j-2}^n)\right)\right)$$

(le flux numérique de ce schéma peut être vu comme la demi-somme des flux de Lax–Wendroff et de Beam–Warming)

autres approches prédicteur-correcteur : Lerat–Peyret

10.4.5 Analyse par des techniques variationnelles **

en domaine borné, avec condition aux limites.
reprendre notamment l'exemple du schéma de Carlson

10.4.6 Remarques sur l'implémentation

pour la résolution d'un problème périodique, il est utile de stocker les valeurs aux points x_0 et x_M, même si elles sont identiques compte tenu de la condition de périodicité.

10.5 Notes sur le chapitre **

ouvrages de référence sur ce chapitre : [LeV92, GR96, Str04], pour les aspects non numériques : [Ser96]

A DEPLACER ? : ouverture sur des modèles faisant intervenir des équations cinétiques : équation de Vlasov[37]–Poisson, équation aux dérivées partielles non linéaire limite de « champ moyen » (i.e. $N \to +\infty$) du problème à N corps (voir la sous-section 8.2.1) l'évolution de systèmes stellaires sur de grandes échelles de temps.

Le problème de Riemann et sa résolution sont une brique de base de la *méthode de Glimm* [Gli65], qui est une preuve constructive d'existence de solutions de système de lois de conservation non linéaires dans le cadre des fonctions à variation totale bornée. Le schéma introduit par cette méthode possède un aspect aléatoire (au sens probabiliste du terme) et fut par la suite développé en une méthode numérique à part entière pour résoudre des problèmes de dynamique des gaz [Cho76].

problème des schémas d'ordre élevé à préserver la positivité des quantités approchées du fait des oscillations, le théorème de Godunov montrant la limitation des schémas linéaires monotones en terme d'ordre.
Pour corriger ce défaut, il faut avoir recours à des schémas pour lesquels l'inconnue à l'instant t_{n+1} n'est pas obtenue comme une combinaison linéaire (A VOIR si l'EDP est non linéaire) de ses valeurs aux temps précédents. Diverses approches existent. La famille des *schémas MUSCL* (acronyme anglais de *Monotone Upstream-centered Schemes for Conservation Laws*), introduite par van Leer [vLee79], repose sur le choix d'un *limiteur de pente* (*slope limiter* en anglais) pour la construction du flux numérique au moyen d'une combinaison non linéaire de deux approximations distinctes de la fonction flux.

EXPLICATIONS

choix des limiteurs : van Leer, minmod, superbee...

AUSSI : schémas ENO [HEO+87] et WENO [LOC94] (acronymes anglais respectifs de *Essentially Non-Oscillatory* et *Weighted Essentially Non-Oscillatory*)

Autres méthodes de résolution numérique : À la différence de la méthode des différences finies, qui discrétise la loi de conservation écrite sous une forme locale, la *méthode des volumes finis*, introduite dans les articles [TS62, TS63], considère le problème sous forme intégrale. COMPLETER

lien entre les lois de conservation scalaires et les équations d'Hamilton–Jacobi

De nombreux schémas numériques pour ces équations sont donc issus de ceux développés pour les lois de conservation (voir par exemple [CL84])

Références

[AR00] A. Aw and M. Rascle. Resurrection of "second order" models of traffic flow. *SIAM J. Appl. Math.*, 60(3):916–938, 2000. DOI: 10.1137/S0036139997332099.

[BL42] S. E. Buckley and M. C. Leverett. Mechanism of fluid displacement in sands. *Trans. AIME*, 146(1):107–116, 1942. DOI: 10.2118/942107-G.

[Bol72] L. Boltzmann. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe*, 66(1):275–370, 1872.

[Bur48] J. M. Burgers. *A mathematical model illustrating the theory of turbulence*. In *Advances in applied mechanics*. R. von Mises and T. von Kármán, editors. Volume 1. Academic Press Inc., 1948, pages 171–199. DOI: 10.1016/S0065-2156(08)70100-5.

[Car59] B. Carlson. Numerical solution of transient and steady-state neutron transport problems. Technical report LA-2260, Los Alamos scientific laboratory, 1959. DOI: 10.2172/4198642.

[CFL28] R. Courant, K. Friedrichs und H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100(1):32–74, 1928. DOI: 10.1007/BF01448839.

[CFL67] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM J.*, 11(2):215–234, 1967. DOI: 10.1147/rd.112.0215.

[CFvN50] J. G. Charney, R. Fjörtoft, and J. von Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950. DOI: 10.1111/j.2153-3490.1950.tb00336.x.

37. Anatoly Alexandrovich Vlasov (Анато́лий Алекса́ндрович Вла́сов en russe, 20 août 1908 - 22 décembre 1975) était un physicien théoricien russe dont les avancées dans les domaines de la mécanique statistique, de la physique des cristaux et de la physique des plasmas furent particulièrement marquantes.

[Cho76] A. J. CHORIN. Random choice solution of hyperbolic systems. *J. Comput. Phys.*, 22(4):517–533, 1976. DOI: 10.1016/0021-9991(76)90047-4.

[CIR52] R. COURANT, E. ISAACSON, and M. REES. On the solution of nonlinear hyperbolic differential equations by finite differences. *Comm. Pure Appl. Math.*, 5(3):243–255, 1952. DOI: 10.1002/cpa.3160050303.

[CL84] M. G. CRANDALL and P.-L. LIONS. Two approximations of solutions of Hamilton-Jacobi equations. *Math. Comput.*, 43(167):1–19, 1984. DOI: 10.1090/S0025-5718-1984-0744921-8.

[DAl49] J. D'ALEMBERT. *Recherches sur la courbe que forme une corde tenduë mise en vibration*. In tome 3 (année 1747). Histoire de l'Académie royale des sciences et belles lettres. Haude et Spener, Berlin, 1749, pages 214-219.

[EO80] B. ENGQUIST and S. OSHER. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comput.*, 34(149):45–75, 1980. DOI: 10.1090/S0025-5718-1980-0551290-1.

[Eul57] L. EULER. Principes généraux du mouvement des fluides. *Hist. Acad. Roy. Sci. Belles-Lettres Berlin*, 11 :274-315, 1757.

[Fro68] J. E. FROMM. A method for reducing dispersion in convective difference schemes. *J. Comput. Phys.*, 3(2):176–189, 1968. DOI: 10.1016/0021-9991(68)90015-6.

[Gel59] I. M. GEL'FAND. Some problems in the theory of quasi-linear equations (russian). *Uspekhi Mat. Nauk*, 14(2(86)):87–158, 1959.

[Gli65] J. GLIMM. Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.*, 18(4):697–715, 1965. DOI: 10.1002/cpa.3160180408.

[God59] S. K. GODUNOV. A difference scheme for numerical solution of discontinuous solution of fluid dynamics (russian). *Math. Sb.*, 47(89):271–306, 1959.

[GR96] E. GODLEWSKI and P.-A. RAVIART. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied mathematical sciences*. Springer, 1996.

[HEO+87] A. HARTEN, B. ENGQUIST, S. OSHER, and S. R. CHAKRAVARTHY. Uniformly high order accurate essentially non-oscillatory schemes, III. *J. Comput. Phys.*, 71(2):231–303, 1987. DOI: 10.1016/0021-9991(87)90031-3.

[HH79] E. HEWITT and R. E. HEWITT. The Gibbs–Wilbraham phenomenon: an episode in Fourier analysis. *Arch. History Exact Sci.*, 21(2):129–160, 1979. DOI: 10.1007/BF00330404.

[HHL+76] A. HARTEN, J. M. HYMAN, P. D. LAX, and B. KEYFITZ. On finite-difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.*, 29(3):297–322, 1976. DOI: 10.1002/cpa.3160290305.

[Hug87] H. HUGONIOT. Sur la propagation du mouvement dans les corps et spécialement dans les gaz parfaits. *J. École Polytechnique*, 57 :3-98, 1887.

[Kre64] H.-O. KREISS. On difference approximations of the dissipative type for hyperbolic differential equations. *Comm. Pure Appl. Math.*, 17(3):335–353, 1964. DOI: 10.1002/cpa.3160170306.

[Kru70] S. N. KRUŽKOV. First order quasilinear equations in several independent variables. *Math. USSR-Sb.*, 10(2):217–243, 1970. DOI: 10.1070/SM1970v010n02ABEH002156.

[Lax54] P. D. LAX. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, 7(1):159–193, 1954. DOI: 10.1002/cpa.3160070112.

[Lax57] P. D. LAX. Hyperbolic systems of conservation laws II. *Comm. Pure Appl. Math.*, 10(4):537–566, 1957. DOI: 10.1002/cpa.3160100406.

[LeV92] R. J. LEVEQUE. *Numerical methods for conservation laws, Lectures in Mathematics. ETH Zürich*. Birkhäuser, second edition, 1992. DOI: 10.1007/978-3-0348-8629-1.

[Liu76] T.-P. LIU. The entropy condition and the admissibility of shocks. *J. Math. Anal. Appl.*, 53(1):78–88, 1976. DOI: 10.1016/0022-247X(76)90146-3.

[LOC94] X.-D. LIU, S. OSHER, and T. CHAN. Weighted essentially non-oscillatory schemes. *J. Comput. Phys.*, 115(1):200–212, 1994. DOI: 10.1006/jcph.1994.1187.

[LR56] P. D. LAX and R. D. RICHTMYER. Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9(2):267–293, 1956. DOI: 10.1002/cpa.3160090206.

[LW55] M. J. LIGHTHILL and G. B. WHITHAM. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. London Ser. A*, 229(1178):317–345, 1955. DOI: 10.1098/rspa.1955.0089.

[LW60] P. D. LAX and B. WENDROFF. Systems of conservation laws. *Comm. Pure Appl. Math.*, 13(2):217–237, 1960. DOI: 10.1002/cpa.3160130205.

[Mac69] R. W. MACCORMACK. The effect of viscosity in hypervelocity impact cratering. In *AIAA hypervelocity impact conference*, Cincinnati, Ohio, 1969. AIAA paper 69-354.

[Mur74] E. M. MURMAN. Analysis of embedded shock waves calculated by relaxation methods. *AIAA J.*, 12(5):626–633, 1974. DOI: 10.2514/3.49309.

[New59] N. M. NEWMARK. A method of computation for structural dynamics. *J. Engrg. Mech. Div.*, 85(7):67–94, 1959. DOI: 10.1061/JMCEA3.0000098.

[Ole57] O. A. OLEINIK. Discontinuous solutions of non-linear differential equations (russian). *Uspekhi Mat. Nauk*, 12(3(75)):3–73, 1957.

[Ole59] O. A. OLEINIK. Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation (russian). *Uspekhi Mat. Nauk*, 14(2(86)):165–170, 1959.

[Ran70] W. J. M. RANKINE. On the thermodynamic theory of waves of finite longitudinal disturbances. *Philos. Trans. Roy. Soc. London*, 160:277–288, 1870. DOI: 10.1098/rstl.1870.0015.

[Ric56] P. I. RICHARDS. Shock waves on the highway. *Operations Res.*, 4(1):42–51, 1956. DOI: 10.1287/opre.4.1.42.

[Ric62] R. D. RICHTMYER. A survey of difference methods for non-steady fluid dynamics. Technical report NCAR technical note 63-2, National Center for Atmospheric Research, Boulder, Colorado, 1962. DOI: 10.5065/D67P8WCQ.

[Rie60] B. RIEMANN. *Ueber die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite*. Verlag der Dieterichschen Buchhandlung, 1860.

[Roe81] P. L. ROE. Approximate riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43(2):357–372, 1981. DOI: 10.1016/0021-9991(81)90128-5.

[Ser96] D. SERRE. *Systèmes de lois de conservation I. Hyperbolicité, entropies, ondes de choc, Fondations*. Diderot éditeur, arts et sciences, 1996.

[Str04] J. C. STRIKWERDA. *Finite difference schemes and partial differential equations*. SIAM, second edition, 2004. DOI: 10.1137/1.9780898717938.

[Str64] G. L. STRANG. Accurate partial difference methods ii. non-linear problems. *Numer. Math.*, 6(1):37–46, 1964. DOI: 10.1007/BF01386051.

[TS62] A. N. TIKHONOV and A. A. SAMARSKII. Homogeneous difference schemes. *USSR Comput. Math. Math. Phys.*, 1(1):5–67, 1962. DOI: 10.1016/0041-5553(62)90005-8.

[TS63] A. N. TIKHONOV and A. A. SAMARSKII. Homogeneous difference schemes on non-uniform nets. *USSR Comput. Math. Math. Phys.*, 2(5):927–953, 1963. DOI: 10.1016/0041-5553(63)90505-6.

[vLee79] B. van LEER. Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *J. Comput. Phys.*, 32(1):101–136, 1979. DOI: 10.1016/0021-9991(79)90145-1.

[Vol67] A. I. VOL'PERT. The spaces *BV* and quasilinear equations. *Math. USSR-Sb.*, 2(2):225–267, 1967. DOI: 10.1070/SM1967v002n02ABEH002340.

[WB76] R. F. WARMING and R. M. BEAM. Upwind second-order difference schemes and applications in aerodynamic flows. *AIAA J.*, 14(9):1241–1249, 1976. DOI: 10.2514/3.61457.

Chapitre 11

Résolution numérique des équations paraboliques

Nous intéressons à présent à des équations aux dérivées partielles rendant compte mathématiquement de phénomènes de *diffusion*. Dans le cas linéaire [1] et pour une inconnue scalaire en l'absence de source, celles-ci prennent la forme générale

$$\frac{\partial u}{\partial t}(t,\boldsymbol{x})+(Lu)(t,\boldsymbol{x})=0$$

l'opérateur L étant défini par

$$Lu=-\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left(\sum_{j=1}^{d}a_{ij}\frac{\partial u}{\partial x_j}\right)+\sum_{i=1}^{d}b_i\frac{\partial u}{\partial x_i}+c\,u,$$

avec $A=(a_{ij})$, $\boldsymbol{b}=(b_i)$...

COMPLETER

11.1 Quelques exemples d'équations paraboliques *

REPRENDRE Les équations paraboliques sont dans de nombreux cas issus d'une description macroscopique de phénomènes dont la description microscopique est essentiellement de nature aléatoire, comme le transfert de chaleur dans un solide lié à la cession de l'énergie cinétique, provenant de l'agitation thermique, d'un atome ou d'une molécule à ses voisins, les forces visqueuses dans un fluide liées aux mouvements moléculaires, etc.

11.1.1 Un modèle de conduction thermique *

L'*équation de la chaleur* est une équation aux dérivées partielles parabolique, initialement introduite par Fourier [Fou22] pour décrire le phénomène physique de distribution de la chaleur dans un milieu continu.

pour un matériau homogène, elle prend la forme

$$\frac{\partial u}{\partial t}(t,x)-\alpha\,\frac{\partial^2 u}{\partial x^2}(t,x)=0$$

u : température, α : *diffusivité thermique* [2] (en $\mathrm{m^2\,s^{-1}}$).

Note : l'analyse de Fourier fut introduite pour la résolution de cette équation

1. Quelques exemples d'équations paraboliques seront donnés dans la section 11.1.
2. Cette grandeur physique dépend des capacités du matériau à conduire et accumuler la chaleur. On a la formule

$$\alpha=\frac{\lambda}{\rho c_p},$$

dans laquelle λ désigne la *conductivité thermique* (en $\mathrm{W\,m^{-1}\,K^{-1}}$), ρ est la *masse volumique* (en $\mathrm{kg\,m^{-3}}$) et c_p est la *capacité thermique massique* (en $\mathrm{J\,kg^{-1}\,K^{-1}}$).

11.1.2 Systèmes d'advection-réaction-diffusion **

équations de Fokker[3]*–Planck*[4] [Fok14] (encore appelée équation "forward" de Kolmogorov dans la communauté probabiliste en raison de l'article [Kol31])

en particulier :

Retour sur le modèle de Black–Scholes *

Dans cette sous-section, une approche déterministe de résolution du problème de couverture d'une option d'achat européenne dans le cadre du modèle de Black–Scholes, déjà présenté dans la sous-section 9.2.2, est proposée. Celle-ci repose sur une équation aux dérivées partielles décrivant l'évolution du prix de l'option.

On rappelle que le prix d'une option à un instant t est égal à la valeur V_t de son portefeuille de couverture, qui est une fonction du temps et de la valeur de l'actif risqué sous-jacent, modélisé par le processus S. En notant C cette fonction, que l'on suppose de classe $\mathscr{C}^1$ par rapport à sa première variable et de classe $\mathscr{C}^2$, il vient par utilisation de la formule d'Itō (voir la proposition 9.20)

$$\mathrm{d}C(t,S_t)=\left(\frac{\partial C}{\partial t}(t,S_t)+\mu S_t\,\frac{\partial C}{\partial x}(t,S_t)+\frac{1}{2}\sigma^2{S_t}^2\,\frac{\partial^2 C}{\partial x^2}(t,S_t)\right)\mathrm{d}t+\sigma S_t\,\frac{\partial C}{\partial x}(t,S_t)\,\mathrm{d}W_t.$$

En identifiant alors avec l'équation (9.20), on obtient

$$\frac{\partial C}{\partial t}(t,S_t)+\mu S_t\,\frac{\partial C}{\partial x}(t,S_t)+\frac{1}{2}\sigma^2{S_t}^2\,\frac{\partial^2 C}{\partial x^2}(t,S_t)=r\,C(t,S_t)+(\mu-r)\beta_t S_t,$$

et

$$\sigma S_t\,\frac{\partial C}{\partial x}(t,S_t)=\sigma\,\beta_t S_t,$$

dont on déduit l'*équation de Black–Scholes*

$$\frac{\partial C}{\partial t}(t,S_t)+\frac{1}{2}\sigma^2 S_t^2\,\frac{\partial^2 C}{\partial x^2}(t,S_t)+rS_t\,\frac{\partial C}{\partial x}(t,S_t)-r\,C(t,S_t)=0,\ 0<t<T, \tag{11.1}$$

que l'on munit de la condition *terminale*

$$C(T,S_T)=(S_T-K)_+,$$

ces deux relations étant vérifiées presque sûrement par rapport à la mesure de probabilité historique P. Le support de la loi de probabilité de la variable S_t étant, pour tout $t\geq 0$, $[0,+\infty[$, l'équation reste satisfaite lorsque l'on remplace S_t par la variable x, avec $x>0$. On en déduit alors que la fonction C est solution du problème

$$\frac{\partial C}{\partial t}(t,x)+\frac{1}{2}\sigma^2x^2\,\frac{\partial^2 C}{\partial x^2}(t,x)+rx\,\frac{\partial C}{\partial x}(t,x)-r\,C(t,x)=0,\ 0<t<T,\ x\in]0,+\infty[, \tag{11.2}$$

$$C(T,x)=(x-K)_+,\ x\in]0,+\infty[. \tag{11.3}$$

EXPLICATIONS pour les CONDITIONS AUX LIMITES

$$C(t,0)=0,\ C(t,x)\to x \text{ quand } x\to\infty,\ \forall t.$$

- si $S_t=0\ \forall t$, le bénéfice à terme est nul et il n'y a aucun intérêt à exercer l'option, d'où $C(t,0)=0\ \forall t$.

- si le prix augmente considérablement au cours du temps ($S_t\to+\infty$), l'option sera exercée et le prix d'exercice de l'option sera négligeable, d'où $C(t,x)\sim x$, $x\to+\infty$.

3. Adriaan Daniël Fokker (17 août 1887 - 24 septembre 1972) était un physicien et musicien néerlandais. Il a a apporté plusieurs contributions à la relativité restreinte, en particulier pour la précession géodétique. Il a également conçu et construit divers claviers permettant de jouer de la musique microtonale.

4. Max Karl Ernst Ludwig Planck (23 avril 1858 - 4 octobre 1947) était un physicien allemand, souvent considéré comme le fondateur de la mécanique quantique. Il est d'ailleurs lauréat du prix Nobel de physique de 1918 pour ses travaux en théorie des quanta.

Il découle d'une extension de la *formule de Feynman* [5]*–Kac* [6] [Kac49] que la solution de ce problème peut être écrite sous la forme d'une espérance conditionnelle,

$$C(t,x) = e^{-r(T-t)}\, E\left((X_T - K)_+ | X_t = x\right),$$

où le processus X est solution de l'équation différentielle stochastique

$$\mathrm{d}X_s = rX_s\,(\mathrm{d}s + \sigma\,\mathrm{d}W_s),\ s \in [t,T].$$

On retrouve alors le résultat, obtenu dans la sous-section 9.2.2, conduisant à la formule de Black–Scholes (9.25). On peut cependant dériver cette formule sans faire appel au calcul stochastique. En effet, en introduisant le changement de variables

$$\tau = T - t,\ y = \ln\left(\frac{x}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)\tau,\ u(\tau, y) = C(t,x)\, e^{r\,\tau},$$

l'équation de Black–Scholes s'écrit alors

$$\frac{\partial u}{\partial \tau}(\tau, y) - \frac{\sigma^2}{2}\,\frac{\partial^2 u}{\partial y^2}(\tau, y) = 0,\ \ 0 < \tau < T,\ y \in \mathbb{R},$$

la condition terminale devenant une condition initiale

$$u(0,y) = K\,(e^{\max\{0,y\}} - 1).$$

METHODES DE RESOLUTION :
Par une méthode standard (convolution noyau de Green) de résolution, il vient

$$u(\tau, y) = \frac{1}{\sigma\sqrt{2\pi\tau}} \int_{-\infty}^{\infty} K\,(e^{\max\{z,0\}} - 1) \exp\left(-\frac{(y-z)^2}{2\sigma^2\tau}\right)\mathrm{d}z,$$

d'où, après quelques manipulations,

$$u(\tau, y) = K\, e^{y + \frac{\sigma^2}{2}\tau} N(d_1) - K\, N(d_2)$$

avec

$$d_1 = \frac{(y + \frac{\sigma^2}{2}\,\tau) + \frac{\sigma^2}{2}\,\tau}{\sigma\sqrt{\tau}},\ d_2 = d_1 - \sigma\sqrt{\tau}.$$

Le retour aux variables originelles dans l'expression de cette solution conduit alors à la formule de Black–Scholes.
extension pour options européennes payant des dividendes, résolution numérique, etc.

11.1.3 Systèmes de réaction-diffusion semi-linéaires **

modèles mathématiques décrivant l'évolution des concentrations d'une ou plusieurs substances spatialement distribuées et soumises à deux processus : un processus de réactions chimiques locales, dans lequel les différentes substances se transforment, et un processus de diffusion qui provoque une répartition de ces substances dans l'espace. Ils sont utilisés en chimie, chaque composante de l'inconnue $\boldsymbol{u}$ étant alors la concentration d'une substance en un point donné à un instant donné, mais ils décrivent aussi des phénomènes de nature différente ayant lieu dans des systèmes biologiques, écologiques ou sociaux.
systèmes non linéaires :

5. Richard Phillips Feynman (11 mai 1918 - 15 février 1988) était un physicien américain, comptant parmi les scientifiques les plus influents de la seconde moitié du vingtième siècle. Il est l'auteur de travaux sur la reformulation de la mécanique quantique à l'aide d'intégrales de chemin, l'électrodynamique quantique relativiste, la physique des particules ou encore la superfluidité de l'hélium liquide.

6. Mark Kac (Marek Kac en polonais, 3 août 1914 - 26 octobre 1984) était un mathématicien américain d'origine polonaise, spécialiste de la théorie des probabilités. Sa question, devenue célèbre et à laquelle la réponse est en général négative, « *Peut-on entendre la forme d'un tambour ?* » donna lieu à d'importants développements dans le domaine de la géométrie spectrale.

une composante (et une dimension d'espace),

$$\frac{\partial u}{\partial t} = D\,\frac{\partial^2 u}{\partial x^2} + R(u)$$

$R(u) = u(1-u)$: équation de Fisher [7]–Kolmogorov–Petrovskii–Piscounov [Fis37] ref?(originellement utilisée pour simuler la propagation d'un gène dans une population)
$R(u) = u(1-u^2)$: équation de Newell–Whitehead–Segel [NW69, Seg69] (décrit la convection de Rayleigh–Bénard [8])
$R(u) = u(1-u)(u-\alpha)$ avec $0 < \alpha < 1$: équation de Zeldovich [9] (apparaît en théorie de la combustion comme modèle de propagation de flamme), dont un cas dégénéré est $R(u) = u^2 - u^3$ (à comparer avec (8.113))

équation d'Allen-Cahn (modélise le processus de séparation de phase dans les alliages ferreux)

A VOIR : équation de Ginzburg [10]–Landau [11] complexe

- deux composantes, plusieurs (deux?) dimensions d'espace citer (avec explications) l'article précurseur de Turing [Tur52] sur la fondation chimique du phénomène de morphogenèse, poursuivre sur le *modèle de Schnakenberg* [Sch79]

$$\begin{aligned}\frac{\partial u}{\partial t} &= \Delta u - \gamma\left(a - u + u^2 v\right)\\ \frac{\partial v}{\partial t} &= d\,\Delta v + \gamma\left(b - u^2 v\right)\end{aligned}$$

le *modèle de Gray–Scott* [GS83]

$$\begin{aligned}\frac{\partial u}{\partial t} &= D_u\,\Delta u - u\,v^2 + F\,(1-u)\\ \frac{\partial v}{\partial t} &= D_v\,\Delta v + u\,v^2 - (F+k)\,v\end{aligned}$$

avec $D_u > 0$ et $D_v > 0$ des paramètres de diffusion et F et k peuvent être vus comme des paramètres de bifurcation

simulations numériques à partir de ce derniers modèle conduisant à la formation de motifs observés dans la nature [Pea93]

11.2 Existence et unicité d'une solution, propriétés **

INTRODUIRE le problème à résoudre en domaine borné $]0,+\infty[\times]0,L[$ avec condition initiale

$$\frac{\partial u}{\partial t}(t,x) - \alpha\,\frac{\partial^2 u}{\partial x^2}(t,x) = 0, t > 0,\ 0 < x < L, \tag{11.4}$$

$$u(0,x) = u_0(x)\ 0 < x < L.$$

7. Ronald Aylmer Fisher (17 février 1890 - 29 juillet 1962) était un statisticien et biologiste britannique. Il introduisit dans le domaine des statistiques de nombreux concepts clés parmi lesquels on peut citer le maximum de vraisemblance, l'information portant son nom et l'analyse de la variance. Il est également l'un des fondateurs de la génétique moderne, son approche statistique de la génétique des populations contribuant à la formalisation mathématique du principe de sélection naturelle.

8. Henri Bénard (25 octobre 1874 - 29 mars 1939) était un physicien français. Il est connu pour ses travaux de recherche sur les phénomènes de convection dans les liquides.

9. Yakov Borisovich Zeldovich (Я́ков Бори́сович Зельдо́вич en russe, 8 mars 1914 - 2 décembre 1987) était un physicien russe. Il joua un rôle important dans le développement des armes nucléaire et thermonucléaire soviétiques et fit d'importantes contributions dans les domaines de l'adsorption et de la catalyse, des ondes de chocs, de la physique nucléaire, de la physique des particules, de l'astrophysique, de la cosmologie et de la relativité générale.

10. Vitaly Lazarevich Ginzburg (Вита́лий Ла́заревич Ги́нзбург en russe, 4 octobre 1916 - 8 novembre 2009) était un physicien et astrophysicien russe, considéré comme l'un des pères de la bombe à hydrogène soviétique.

11. Lev Davidovich Landau (Ле́в Дави́дович Ланда́у en russe, 22 janvier 1908 - 1er avril 1968) était un physicien théoricien russe. COMPLETER

et conditions aux limites (homogènes). Plusieurs choix sont possibles : *Dirichlet* [12], *Neumann* [13], *Robin* [14] (voir [GA98] à propos de l'appellation de cette condition), périodiques ou toute combinaison compatible de celles-ci

Il existe plusieurs façons de démontrer l'existence d'une solution du problème eqref. Certaines sont abstraites, comme celles fondées sur la théorie des semi-groupes ou utilisant une approche variationnelle.

technique via base hilbertienne

Résultats (conditions de Dirichlet homogènes, pas de source, mais on peut adapter la théorie) :

Théorème 11.1 *Si $u_0 \in L^2(]0,L[)$, il existe une unique solution du problème u telle que*

$$u \in \mathscr{C}\left([0,+\infty[;L^2(]0,L[)\right) \cap \mathscr{C}\left(]0,+\infty[;H^2(]0,L[\cap H_0^1(]0,L[)\right),\ u \in \mathscr{C}^1\left(]0,+\infty;L^2(]0,L[)\right).$$

De plus, $\forall \varepsilon > 0$, $u \in \mathscr{C}^\infty([\varepsilon,T] \times [0,L])$.

DÉMONSTRATION. A ECRIRE □

COMMENTAIRES : effet fortement régularisant de l'équation sur la donnée initiale : même si u_0 est discontinue, la solution est de classe $\mathscr{C}^\infty$ dès que $t > 0$)

cet effet a pour conséquence la non réversibilité de l'équation : on ne peut généralement pas résoudre le problème avec condition terminale (il est impossible de retrouver la condition initiale à partir de la connaissance de la solution à un instant donné $t > 0$).

Théorème 11.2 *Si $u_0 \in L^\infty(]0,L[)$, $f = 0$, alors $u \in L^\infty(]0,T[\times]0,L[)$ et $\|u\|_\infty \leq \|u_0\|_\infty$ (principe du maximum/stabilité L^∞, note : aussi vrai pour équations de transport)*

Corollaire 11.3 *Si $u_0 \geq 0$, $f \geq 0$ (régularité ?) alors $u(t,\cdot) \geq 0$, $\forall t \geq 0$ (principe de positivité).*

enfin, si $u_0 \neq 0$, $u_0 \geq 0$ presque partout et $f \equiv 0$ alors $u(t,x) > 0$ $\forall x \in]0,L[$, $\forall t > 0$ (propagation à vitesse infinie, limitation du modèle)

Ces propriétés sont tout à fait différentes de celles des solutions des équations hyperboliques du chapitre précédent

11.3 Résolution approchée par la méthode des différences finies

INTRODUCTION

11.3.1 Analyse des méthodes **

consistance

stabilité, faire lien avec la stabilité pour les edo

domaine en espace étant borné, on peut réaliser l'analyse de stabilité dans le cas des conditions de Dirichlet via un calcul de rayon spectral

Remarque sur le principe du maximum discret

11.3.2 Présentation de quelques schémas **

méthodes à un pas en temps

FTCS (forward in time, centered in space) : Euler explicite en temps + différences centrées en espace

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} - \alpha \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = 0 \tag{11.5}$$

12. Johann Peter Gustav Lejeune Dirichlet (13 février 1805 - 5 mai 1859) était un mathématicien allemand. On lui doit des contributions profondes à la théorie analytique des nombres et la théorie des séries de Fourier, ainsi que divers travaux en analyse ; l'introduction du concept moderne de fonction lui est notamment attribué.

13. Carl Gottfried Neumann (7 mai 1832 - 27 mars 1925) était un mathématicien allemand. Il travailla sur le principe de Dirichlet et fut l'un des pionniers de la théorie des équations intégrales.

14. Victor Gustave Robin (17 mai 1855 - 1897) était un mathématicien et physicien français, connu pour ses contributions à la théorie du potentiel et à la thermodynamique.

ordre un en temps et deux en espace, stable si $\frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$

BTCS (backward in time, centered in space) : Euler implicite en temps + différences centrées en espace

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} - \alpha \frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{(\Delta x)^2} = 0$$

ordre un en temps et deux en espace, inconditionnellement stable

méthode de Crank [15]–Nicolson [16] [CN47]

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} - \frac{1}{2}\alpha\left(\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{(\Delta x)^2} + \frac{u_{j+1}^{n}-2u_j^{n}+u_{j-1}^{n}}{(\Delta x)^2}\right) = 0$$

inconditionnellement stable

Plus généralement : classe des θ-schémas

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} - \alpha\theta\,\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{(\Delta x)^2} - \alpha(1-\theta)\,\frac{u_{j+1}^{n}-2u_j^{n}+u_{j-1}^{n}}{(\Delta x)^2} = 0,\ \theta\in[0,1].$$

$\theta = 0$: FTCS, $\theta = 1$: BTCS, $\theta = \frac{1}{2}$: Crank–Nicolson

analyse de von Neumann : on trouve

$$g(\xi) = \frac{1-4(1-\theta)\frac{\Delta t}{(\Delta x)^2}\alpha\,\sin^2\left(\frac{\xi\Delta x}{2}\right)}{1+4\theta\frac{\Delta t}{(\Delta x)^2}\alpha\,\sin^2\left(\frac{\xi\Delta x}{2}\right)}$$

On doit distinguer deux cas : la méthode est stable sous la condition $\frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2(1-2\theta)}$ si $0 \leq \theta < \frac{1}{2}$, elle est inconditionnellement stable sinon.

Lien avec les méthodes pour la résolution des EDO : après semi-discrétisation en espace, on est conduit à résoudre le système différentiel linéaire

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}(t) = A\,\boldsymbol{u}(t)$$

que l'on résout par l'une des méthodes introduites au chapitre 8. On voit alors que la méthode Euler explicite correspond alors à (8.26), Euler implicite à (8.28) et le schéma de Crank–Nicolson au choix de la méthode de la règle du trapèze (8.29).

méthodes à deux pas de temps (mentionner le problème pratique de l'initialisation de la relation de récurrence : deux valeurs étant nécessaires et la condition initiale du problème n'en fournissant qu'une, on doit avoir recours à une méthode à un pas pour obtenir la valeur manquante)

méthode de Richardson [Ric11] (explicite, second ordre)

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\,\Delta t} - \alpha\frac{u_{j+1}^{n}-2u_j^{n}+u_{j-1}^{n}}{(\Delta x)^2},\ n\in\mathbb{N}^*,\ j\in\mathbb{Z}. \tag{11.6}$$

analyse de von Neumann :

$$\hat{u}^{n+1}(\xi) + 8\,\frac{\Delta t}{(\Delta x)^2}\alpha\,\sin^2\left(\frac{\xi\,\Delta x}{2}\right)\hat{u}^n(\xi) - \hat{u}^{n-1}(\xi) = 0,\ n\in\mathbb{N}^*.$$

L'analyse de stabilité doit être conduite en utilisant les résultat de la sous-section 8.4.1. Les racines du polynôme caractéristique associé sont réelles et distinctes (discriminant égal à $\frac{64(\Delta t)^2}{(\Delta x)^4}\,\sin^4\left(\frac{\xi\,\Delta x}{2}\right) + 4 > 0$) et de produit égal

15. John Crank (6 février 1916 - 3 octobre 2006) était un mathématicien anglais, connu pour ses travaux sur la résolution numérique des équations aux dérivées partielles pour des problèmes de conduction de la chaleur.

16. Phyllis Nicolson (21 septembre 1917 - 6 octobre 1968) était une physicienne britannique. Elle est, avec John Crank, à l'origine d'une méthode de résolution numérique de l'équation de la chaleur.

à -1. L'une d'entre elles est donc de valeur absolue strictement plus grande que l'unité et la méthode est donc inconditionnellement instable.

On peut remédier à cet inconvénient de taille en remplaçant dans (11.6) la quantité u_j^n par la moyenne $\frac{u_j^{n+1}+u_j^{n-1}}{2}$. On obtient ainsi la *méthode de Du Fort–Frankel* [DF53], dont le schéma s'écrit

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\,\Delta t}-\alpha\frac{u_{j+1}^n-u_j^{n+1}-u_j^{n-1}+u_{j-1}^n}{(\Delta x)^2}=0,\ n\in\mathbb{N}^*,\ j\in\mathbb{Z}. \tag{11.7}$$

La méthode reste explicite puisque l'on a

$$u_j^{n+1}=\frac{2\alpha\,\frac{\Delta t}{(\Delta x)^2}}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}}\left(u_{j+1}^n+u_{j-1}^n\right)+\frac{1-2\alpha\,\frac{\Delta t}{(\Delta x)^2}}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}}\,u_j^{n-1},\ n\in\mathbb{N}^*,\ j\in\mathbb{Z}.$$

et elle est inconditionnellement stable. En effet, l'analyse de von Neumann conduit à

$$\left(1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}\right)\hat{u}^{n+1}(\xi)-2\,\frac{\Delta t}{(\Delta x)^2}\,\cos(\xi\,\Delta x)\,\hat{u}^n(\xi)-\left(1-2\,\frac{\Delta t}{(\Delta x)^2}\right)\hat{u}^{n-1}(\xi)=0,\ n\in\mathbb{N}^*.$$

Le produit des racines du polynôme caractéristique associé à cette équation aux différences linéaire vaut

$$-\frac{1-2\alpha\,\frac{\Delta t}{(\Delta x)^2}}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}},$$

et le discriminant vaut

$$4\left(1-4\alpha^2\,\frac{(\Delta t)^2}{(\Delta x)^4}\,\sin^2(\xi\,\Delta x)\right).$$

Si ce dernier est strictement négatif, les racines sont complexes conjuguées, ce qui implique alors qu'elles sont de module strictement inférieur à un (il en va de même si le discriminant est nul et qu'il n'y a qu'une unique racine réelle de multiplicité double). Si le discriminant est strictement positif, les racines sont réelles distinctes et il suffit alors d'observer que l'on a dans ce cas

$$0<1-4\alpha^2\,\frac{(\Delta t)^2}{(\Delta x)^4}\,\sin^2(\xi\,\Delta x)<1,$$

d'où

$$\frac{-2\alpha\,\frac{\Delta t}{(\Delta x)^2}-1}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}}<\frac{2\alpha\,\frac{\Delta t}{(\Delta x)^2}\,\cos(\xi\,\Delta x)\pm\sqrt{1-4\alpha^2\,\frac{(\Delta t)^2}{(\Delta x)^4}\,\sin^2(\xi\,\Delta x)}}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}}<\frac{2\alpha\,\frac{\Delta t}{(\Delta x)^2}+1}{1+2\alpha\,\frac{\Delta t}{(\Delta x)^2}}.$$

Les racines sont donc strictement comprises entre -1 et 1.

En revanche, elle n'est que conditionnellement consistante. Pour le montrer, il suffit de voir le schéma (11.7) comme une perturbation du schéma de la méthode de Richardson, c'est-à-dire

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\,\Delta t}-\alpha\frac{u_{j+1}^n-2\,u_j^n+u_{j-1}^n}{(\Delta x)^2}=-\alpha\frac{(\Delta t)^2}{(\Delta x)^2}\,\frac{u_j^{n+1}-2\,u_j^n+u_j^{n-1}}{(\Delta t)^2},\ n\in\mathbb{N}^*,\ j\in\mathbb{Z}.$$

L'erreur de troncature du schéma est donc celle de la méthode de Richardson perturbée par le terme

$$\frac{(\Delta t)^3}{(\Delta x)^2}\,\frac{\partial^2 u}{\partial t^2}(t_n,x_j)+O\left(\frac{(\Delta t)^5}{(\Delta x)^2}\right).$$

La méthode n'est donc consistante que que si le rapport $\frac{\Delta t}{\Delta x}$ tend vers zéro avec Δt et Δx. La méthode est dans ce cas d'ordre deux en temps et en espace. Bien qu'à première vue surprenant, ce résultat était prévisible. La méthode étant explicite, elle ne propage l'information qu'à une vitesse finie, égale en l'occurrence à $\frac{\Delta x}{\Delta t}$. Pour que la solution approchée qu'elle fournit puisse converger vers la solution de l'équation, il faut donc que cette vitesse tende vers l'infini lorsque les longueurs des pas de discrétisation tendent vers zéro.

11.3.3 Remarques sur l'implémentation de conditions aux limites **

PLACER ici des résultats numériques

Références

[CN47] J. CRANK and P. NICOLSON. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Math. Proc. Cambridge Philos. Soc.*, 43(1):50–67, 1947. DOI: 10.1017/S0305004100023197.

[DF53] E. C. DU FORT and S. P. FRANKEL. Stability conditions in the numerical treatment of parabolic differential equations. *Math. Tables Aids Comp.*, 7(43):135–152, 1953. DOI: 10.1090/S0025-5718-1953-0059077-7.

[Fis37] R. A. FISHER. The wave of advance of advantageous genes. *Ann. Eugenics*, 7(4):335–369, 1937. DOI: 10.1111/j.1469-1809.1937.tb02153.x.

[Fok14] A. D. FOKKER. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Ann. Physik*, 348(5):810–820, 1914. DOI: 10.1002/andp.19143480507.

[Fou22] J. FOURIER. *Théorie analytique de la chaleur*. Firmin Didot, père et fils, 1822.

[GA98] K. GUSTAFSON and T. ABE. The third boundary condition – Was it Robin's? *Math. Intelligencer*, 20(1):63–71, 1998. DOI: 10.1007/BF03024402.

[GS83] P. GRAY and S. K. SCOTT. Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Isolas and other forms of multistability. *Chem. Engrg. Sci.*, 38(1):29–43, 1983. DOI: 10.1016/0009-2509(83)80132-8.

[Kac49] M. KAC. On distributions of certain Wiener functionals. *Trans. Amer. Math. Soc.*, 65(1):1–13, 1949. DOI: 10.1090/S0002-9947-1949-0027960-X.

[Kol31] A. KOLMOGOROFF. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, 104(1):415–458, 1931. DOI: 10.1007/BF01457949.

[NW69] A. C. NEWELL and J. A. WHITEHEAD. Finite bandwidth, finite amplitude convection. *J. Fluid Mech.*, 38(2):279–303, 1969. DOI: 10.1017/S0022112069000176.

[Pea93] J. E. PEARSON. Complex patterns in a simple system. *Science*, 261(5118):189–192, 1993. DOI: 10.1126/science.261.5118.189.

[Ric11] L. F. RICHARDSON. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philos. Trans. Roy. Soc. London Ser. A*, 210(459-470):307–357, 1911. DOI: 10.1098/rsta.1911.0009.

[Sch79] J. SCHNAKENBERG. Simple chemical reaction systems with limit cycle behaviour. *J. Theor. Biol.*, 81(3):389–400, 1979. DOI: 10.1016/0022-5193(79)90042-0.

[Seg69] L. A. SEGEL. Distant side-walls cause slow amplitude modulation of cellular convection. *J. Fluid Mech.*, 38(1):203–224, 1969. DOI: 10.1017/S0022112069000127.

[Tur52] A. M. TURING. The chemical basis of morphogenesis. *Philos. Trans. Roy. Soc. London Ser. B*, 237(641):37–72, 1952. DOI: 10.1098/rstb.1952.0012.

Quatrième partie

Annexes

Annexe A

Rappels et compléments d'algèbre linéaire

On rappelle dans cette annexe un certain nombre de définitions et de résultats relatifs à l'algèbre linéaire en dimension finie et à l'analyse matricielle. La plupart des notions abordées sont supposées déjà connues du lecteur, à l'exception peut-être des normes matricielles. Comme dans le reste du document, on désigne par $\mathbb{N}$ l'ensemble des nombres entiers naturels, par $\mathbb{Z}$ l'ensemble des nombres entiers relatifs et par $\mathbb{Q}$ l'ensemble des nombres rationnels, $\mathbb{Q} = \{\frac{p}{q} \,|\, p \in \mathbb{Z},\ q \in \mathbb{Z}\backslash\{0\}\}$, par $\mathbb{R}$ l'ensemble des nombres réels et par $\mathbb{C}$ l'ensemble des nombres complexes.

A.1 Ensembles et applications

Nous commençons par rappeler, de manière intuitive, des notions relatives aux ensembles et aux applications en adoptant le point de vue de la *théorie naïve des ensembles*.

A.1.1 Généralités sur les ensembles

En mathématiques, on étudie des objets de différents types : des nombres, des points ou encore des vecteurs par exemple. Ces *éléments* forment, en vertu de certaines propriétés, des collections appelées *ensembles* (*sets* en anglais). Dans la suite, on désignera généralement un élément par une lettre minuscule (l'élément x par exemple) et un ensemble par une lettre majuscule (l'ensemble E par exemple). L'*appartenance* d'un élément à un ensemble est par ailleurs notée avec le symbole $\in$ (on a ainsi $x \in E$) et la *non-appartenance* par $\notin$.

Un ensemble peut être *fini* ou *infini*, selon que le nombre d'éléments qui le constituent est fini ou infini (voir la sous-section A.1.4). S'il est fini, il peut être donné en *extension*, c'est-à-dire par la liste (non ordonnée) de ses éléments, *a priori* supposés distincts. S'il est infini (ou même fini), l'ensemble peut être donné en *compréhension*, c'est-à-dire par une ou des propriétés caractérisant ses éléments.

Exemple d'ensemble fini. Un cas particulier d'ensemble fini est le *singleton*, qui est formé d'un unique élément. Si cet élément est noté x, on désigne l'ensemble par $\{x\}$.

Une première notion essentielle est celle d'*égalité entre ensembles*.

Définition A.1 (égalité entre ensembles) *On dit qu'un ensemble E est **égal à** un ensemble F, et l'on note $E = F$, si tout élément de E est un élément de F et si tout élément de F est un élément de E. Lorsque les ensembles E et F ne sont pas égaux, ils sont dits **distincts** et l'on note $E \neq F$.*

Une autre notion importante, la *relation d'inclusion*, se définit de la manière suivante.

Définition A.2 (inclusion entre ensembles) *On dit qu'un ensemble E est **inclus** dans un ensemble F, ce que l'on note $E \subset F$, si et seulement si tout élément de E appartient à F,*

$$E \subset F \iff (\forall x \in E,\ x \in F).$$

L'inclusion d'un ensemble E dans un ensemble F peut encore se noter $F \supset E$ tandis que la négation de cette relation se note $E \not\subset F$. Lorsque $E \subset F$ et qu'il existe au moins un élément de F qui n'appartient pas à E, on dit

que E est un *sous-ensemble propre de* F, ce qui est noté $E \subsetneq F$. Pour tout ensemble E, on a $E \subset E$, et si E, F et G trois ensembles tels que $E \subset F$ et $F \subset G$, alors $E \subset G$. On dit que l'inclusion est une *relation transitive* (voir la sous-section A.1.2).

Le résultat suivant est immédiat. Il permet de démontrer l'égalité entre deux ensembles par un principe de double inclusion.

Proposition A.3 *Étant donné deux ensembles E et F, on a $E = F$ si et seulement si l'on a simultanément $E \subset F$ et $F \subset E$.*

Définition A.4 (partie d'un ensemble) *Soit E un ensemble. On appelle **partie** (ou **sous-ensemble**) de E tout ensemble A vérifiant $A \subset E$.*

Exemple de partie d'un ensemble. On nomme *ensemble vide*, et l'on note $\emptyset$, l'ensemble n'ayant aucun élément. C'est une partie de tout ensemble E. En effet, si cela n'était pas le cas, il existerait au moins un élément appartenant à $\emptyset$ qui n'appartiendrait pas à E. Or, ceci est impossible, puisque l'ensemble vide n'a pas d'élément. L'assertion $\emptyset \subset E$ est donc vraie.

Toutes les parties d'un ensemble E constituent un nouvel ensemble, noté $\mathscr{P}(E)$, que l'on nomme *ensemble des parties de* E. Pour tout ensemble E, E et $\emptyset$ appartiennent à $\mathscr{P}(E)$.

Définition A.5 (partition d'un ensemble) *Soit E un ensemble et $\mathscr{P}$ une partie de $\mathscr{P}(E)$. On dit que $\mathscr{P}$ est une **partition de** E si et seulement si*

- $\forall A \in \mathscr{P},\ A \neq \emptyset$,
- $\forall A \in \mathscr{P},\ \forall B \in \mathscr{P},\ (A \neq B \iff A \cap B = \emptyset)$,
- $\forall x \in E,\ \exists A \in \mathscr{P},\ x \in A$.

Nous introduisons à présent des opérations sur les parties d'un ensemble, en commençant par définir deux *lois de composition internes* dans l'ensemble de ses parties.

Définition A.6 (intersection d'ensembles) *Soit E un ensemble et A et B deux parties de E. On appelle **intersection des ensembles** A **et** B l'ensemble des éléments qui appartiennent à la fois à A et à B. On le note noté $A \cap B$. Lorsque $A \cap B = \emptyset$ (c'est-à-dire lorsque A et B n'ont aucun élément commun), on dit que A et B sont **disjoints**.*

Définition A.7 (réunion d'ensembles) *Soit E un ensemble et A et B deux parties de E. On appelle **réunion des ensembles** A **et** B l'ensemble des éléments qui appartiennent à A ou à B. Cet ensemble est noté $A \cup B$.*

Soit A, B et C trois sous-ensembles d'un ensemble E. L'intersection et la réunion d'ensembles sont des lois sont *commutatives*,

$$A \cap B = B \cap A,\ A \cup B = B \cup A,$$

et *associatives*,

$$A \cup (B \cup C) = (A \cup B) \cup C,\ A \cap (B \cap C) = (A \cap B) \cap C.$$

Elles sont également *distributives* l'une pour l'autre,

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

Pour prouver ces propriétés, on fait appel aux *tableaux de vérité* et aux *synonymies* utilisés en logique. Si l'on définit la proposition $(P(x))$ (resp. $(Q(x))$, resp. $(R(x))$) comme étant vraie si et seulement si $x \in A$ (resp. $x \in B$, resp. $x \in C$), la proposition $x \in B \cap C$ est alors équivalente à $(P(x)$ et $Q(x))$ et $x \in B \cup C$ à $(P(x)$ ou $Q(x))$. Ainsi, $x \in A \cup (B \cap C)$ est équivalente à $(P(x)$ ou $(Q(x)$ et $R(x)))$, ce qui est équivalent à $((P(x)$ ou $Q(x))$ et $(P(x)$ ou $R(x)))$, ou encore à $x \in (A \cup B) \cap (A \cup C)$. On procède de manière identique pour démontrer les autres assertions.

Nous continuons avec les notions de *différence d'ensembles* et de *complémentaire d'une partie d'un ensemble*.

Définition A.8 (différence de deux ensembles) *Soit A et B deux parties d'un ensemble E. On appelle **différence de** A **et de** B, et on note $A \backslash B$, l'ensemble des éléments de E appartenant à A mais pas à B.*

Définition A.9 (complémentaire d'une partie) *Soit A une partie d'un ensemble E. On appelle **complémentaire de** A **dans** E, et l'on note $C_E(A)$, l'ensemble des éléments de E qui n'appartiennent pas à A.*

Les démonstrations des propriétés suivantes sont laissées en exercice au lecteur. Soit A et B deux parties d'un ensemble E. On a

- $A\cap\varnothing=\varnothing$, $A\cap A=A$, $A\cap E=A$ (on dit que E est l'*élément neutre* pour $\cap$), $A\cap B=B \iff A\subset B$,
- $A\cup\varnothing=A$ (on dit que $\varnothing$ est l'*élément neutre* pour $\cup$), $A\cup A=A$, $A\cup E=E$, $A\cup B=B \iff A\subset B$.
- $A\cap(A\cup B)=A\cup(A\cap B)=A$,
- $C_E(\varnothing)=E$, $C_E(E)=\varnothing$, $C_E(C_E(A))=A$, $A\cap C_E(A)=\varnothing$, $A\cup C_E(A)=E$ (on déduit de ces deux dernières égalités que $\{A, C_E(A)\}$ est une partition de E),
- $C_E(A\cap B)=C_E(A)\cup C_E(B)$, $C_E(A\cup B)=C_E(A)\cap C_E(B)$ (ce sont les *lois de De Morgan* [1]),
- $E\backslash A=C_E(A)$, $A\backslash B=\varnothing \iff A\subset B$, $A\backslash B=A\cap C_E(B)=A\backslash(A\cap B)$.

Étant donné deux ensembles E et F, on peut associer à tous éléments $x\in E$ et $y\in F$ le nouvel objet (x,y) appelé *couple ordonné*. Ce couple est un élément d'un nouvel ensemble, que l'on nomme *ensemble produit de E par F*.

Définition A.10 (ensemble produit) *Soit E et F deux ensembles. On appelle* ***ensemble produit de*** *E* ***par*** *F, et l'on note* $E\times F$*, l'ensemble défini par*

$$E\times F=\{(x,y)\mid x\in E \text{ et } y\in F\}.$$

L'ensemble produit de deux ensembles est encore appelé *produit cartésien*, en hommage à Descartes [2] qui généralisa l'usage des coordonnées en posant les bases de la géométrie analytique. L'égalité entre couples d'un même ensemble produit est définie par l'équivalence logique suivante

$$(a,b)=(c,d) \iff (a=c \text{ et } b=d).$$

Lorsque $E=F$, on note $E\times F=E^2$. Par extension, étant donnés un entier $n\geq 1$ et des ensembles $E_1,\dots,E_n$, on appelle produit de $E_1,\dots,E_n$ l'ensemble de tous les *n-uplets* $(x_1,\dots,x_n)$ tels que $x_1\in E_1,\dots,x_n\in E_n$, que l'on note $E_1\times\cdots\times E_n$, ou encore $\prod_{i=1}^n E_i$. Lorsque $E_1=\cdots=E_n=E$, l'ensemble produit résultant est noté E^n.

A.1.2 Relations

Nous allons à présent formaliser et généraliser la notion de relation précédemment introduite avec l'inclusion.

Définitions A.11 *Soit E et F deux ensembles non vides. Une* ***relation binaire****, ou* ***correspondance****,* $\mathscr{R}$ ***de*** *E* ***vers*** *F (****dans*** *E lorsque* $E=F$*) est définie par une partie R, appelée le* ***graphe*** *de la relation, de l'ensemble produit* $E\times F$*. Pour tout couple* (x,y) *appartenant à R, on dit que l'élément x de E est* ***en relation par*** $\mathscr{R}$ *avec l'élément y de F, ce que l'on note encore* $x\mathscr{R}y$*. Enfin, l'****ensemble de définition*** *de la relation* $\mathscr{R}$ *est la partie de E définie par*

$$\{x\in E\mid \exists y\in F,\ x\mathscr{R}y\}$$

et son ***ensemble image*** *est la partie de F définie par*

$$\{y\in F\mid \exists x\in E,\ x\mathscr{R}y\}.$$

Notons que l'on n'utilise généralement pas une notation ensembliste pour décrire une relation binaire mais plutôt la notation $x\mathscr{R}y$ introduite avec les dernières définitions.

Définition A.12 (relation composée) *Soit E, F et G trois ensembles non vides,* $\mathscr{R}$ *(resp.* $\mathscr{S}$*) une relation de E vers F (resp. de F vers G). On définit la* ***relation composée de*** $\mathscr{S}$ ***avec*** $\mathscr{R}$*, notée* $\mathscr{S}\circ\mathscr{R}$*, de E vers G par*

$$\forall(x,z)\in E\times G,\ (x\mathscr{S}\circ\mathscr{R}z \iff (\exists y\in F,\ x\mathscr{R}y \text{ et } y\mathscr{S}z)).$$

On a le résultat d'associativité suivant.

Proposition A.13 *Soit E, F, G et H des ensembles non vides et* $\mathscr{R}$*,* $\mathscr{S}$ *et* $\mathscr{T}$ *des relations, respectivement de E vers F, de F vers G et de G vers H. On a*

$$(\mathscr{T}\circ\mathscr{S})\circ\mathscr{R}=\mathscr{T}\circ(\mathscr{S}\circ\mathscr{R}).$$

1. Augustus De Morgan (27 juin 1806 - 18 mars 1871) était un mathématicien britannique. Il est considéré comme l'un des fondateurs de la logique moderne.

2. René Descartes (31 mars 1596 - 11 février 1650) était un mathématicien, physicien et philosophe français. Il introduisit la géométrie analytique et est considéré l'un des fondateurs de la philosophie moderne.

DÉMONSTRATION. On remarque tout d'abord que les relations $(\mathscr{T}\circ\mathscr{S})\circ\mathscr{R}$ et $\mathscr{T}\circ(\mathscr{S}\circ\mathscr{R})$ ont le même ensemble de départ (E) et d'arrivée (H). Pour un couple (x,t) de $E\times H$, on a alors

$$\begin{aligned} x(\mathscr{T}\circ\mathscr{S})\circ\mathscr{R}t &\iff (\exists y\in F,\ (x\mathscr{R}y \text{ et } y\mathscr{T}\circ\mathscr{S}t))\\ &\iff (\exists y\in F,\ \exists z\in G,\ (x\mathscr{R}y \text{ et } y\mathscr{S}z \text{ et } z\mathscr{T}t))\\ &\iff (\exists z\in G,\ (x\mathscr{S}\circ\mathscr{R}z \text{ et } z\mathscr{T}t))\\ &\iff x\mathscr{T}\circ(\mathscr{S}\circ\mathscr{R})t. \end{aligned}$$

□

Définition A.14 (relation réciproque) *Soit E et F deux ensembles non vides et $\mathscr{R}$ une relation de E vers F. On définit la* ***relation réciproque de*** *$\mathscr{R}$, notée $\mathscr{R}^{-1}$, de F vers E par*

$$\forall(x,y)\in E\times F,\ (y\mathscr{R}^{-1}x \iff x\mathscr{R}y).$$

Proposition A.15 *On a les assertions suivantes.*

1. *Pour toute relation $\mathscr{R}$, on a $(\mathscr{R}^{-1})^{-1}=\mathscr{R}$.*
2. *Soit E, F, G trois ensembles et $\mathscr{R}$ (resp. $\mathscr{S}$) une relation de E vers F (resp. de F vers G). On a $(\mathscr{S}\circ\mathscr{R})^{-1}=\mathscr{R}^{-1}\circ\mathscr{S}^{-1}$.*

DÉMONSTRATION.

1. C'est immédiat.
2. On a, $\forall(x,z)\in E\times G$,

$$z(\mathscr{S}\circ\mathscr{R})^{-1}x \iff x(\mathscr{S}\circ\mathscr{R})z \iff (\exists y\in F,\ (x\mathscr{R}y \text{ et } y\mathscr{S}z)) \iff (\exists y\in F,\ (z\mathscr{S}^{-1}y \text{ et } y\mathscr{R}^{-1}x)) \iff z\mathscr{R}^{-1}\circ\mathscr{S}^{-1}x.$$

□

Définitions A.16 *Une relation binaire $\mathscr{R}$ dans un ensemble E est dite*

- ***réflexive*** *si et seulement si $\forall x\in E,\ x\mathscr{R}x$,*
- ***symétrique*** *si et seulement si $\forall(x,y)\in E^2,\ (x\mathscr{R}y \implies y\mathscr{R}x)$,*
- ***antisymétrique*** *si et seulement si $\forall(x,y)\in E^2,\ ((x\mathscr{R}y \text{ et } y\mathscr{R}x) \implies x=y)$,*
- ***transitive*** *si et seulement si $\forall(x,y,z)\in E^3,\ ((x\mathscr{R}y \text{ et } y\mathscr{R}z) \implies x\mathscr{R}z)$.*

Définition A.17 (relation induite) *Soit E un ensemble, $\mathscr{R}$ une relation binaire sur E et A une partie de E. La relation binaire dans A, notée $\mathscr{R}_A$, définie par $(x\mathscr{R}_A y \iff x\mathscr{R}y)$, $\forall(x,y)\in A^2$, est appelée* ***relation induite par $\mathscr{R}$ sur A.***

Définition A.18 (relation d'équivalence) *Soit $\mathscr{R}$ une relation binaire dans un ensemble E. On dit que $\mathscr{R}$ est une* ***relation d'équivalence*** *si et seulement si elle est réflexive, symétrique et transitive.*

Étant donnée une relation d'équivalence, on identifie les éléments qui sont en relation en introduisant le concept de *classe d'équivalence*.

Définitions A.19 (classe d'équivalence et ensemble quotient) *Soit $\mathscr{R}$ une relation d'équivalence dans un ensemble E. Pour chaque x de E, on appelle* ***classe d'équivalence de x (modulo $\mathscr{R}$)*** *le sous-ensemble de E défini par $\mathscr{C}(x)=\{y\in E \mid x\mathscr{R}y\}$. Tout élément de $\mathscr{C}(x)$ est appelé un* ***représentant de la classe*** *$\mathscr{C}(x)$. L'ensemble des classes d'équivalence modulo $\mathscr{R}$ se nomme* ***ensemble quotient de E par $\mathscr{R}$*** *et se note $E/\mathscr{R}$.*

Théorème A.20 *À toute relation d'équivalence $\mathscr{R}$ dans un ensemble E correspond une partition de E en classes d'équivalence et réciproquement, toute partition de E définit sur E une relation d'équivalence $\mathscr{R}$, dont les classes coïncident avec les éléments de la partition donnée.*

DÉMONSTRATION. Soit $\mathscr{R}$ une relation d'équivalence dans E. Pour tout élément x de E, $\mathscr{C}(x)$ est non vide car x appartient à $\mathscr{C}(x)$. Soit un couple (x,y) de E^2 tel que $\mathscr{C}(x)\cap\mathscr{C}(y)\neq\varnothing$; il existe donc un élément z dans $\mathscr{C}(x)\cap\mathscr{C}(y)$. On a alors $x\mathscr{R}z$ et $y\mathscr{R}z$, d'où (par symétrie et transitivité de la relation) $x\mathscr{R}y$. On en déduit que $\mathscr{C}(x)\subset\mathscr{C}(y)$. Soit en effet t de $\mathscr{C}(x)$, on a $x\mathscr{R}t$ et $x\mathscr{R}y$, d'où $y\mathscr{R}t$ et t appartient à $\mathscr{C}(y)$. Les éléments x et y jouant des rôles symétriques, on a $\mathscr{C}(x)=\mathscr{C}(y)$. Puisque chaque élément x de E appartient à $\mathscr{C}(x)$, la réunion des éléments de $E/\mathscr{R}$ est E.

Réciproquement, soit $\mathscr{P}$ une partition de E et $\mathscr{R}$ la relation définie dans E par

$$\forall(x,y)\in E^2,\ (x\mathscr{R}y \iff (\exists P\in\mathscr{P},\ (x\in P \text{ et } y\in P))).$$

Par définition, il existe, pour chaque x de E, un élément P de $\mathscr{P}$ auquel x appartient, on a donc $x\mathscr{R}x$ et $\mathscr{R}$ est réflexive.

Pour tout (x,y) de E^2, on a

$$x\mathscr{R}y \iff (\exists P\in\mathscr{P},\ (x\in P \text{ et } y\in P)) \iff (\exists P\in\mathscr{P},\ (y\in P \text{ et } x\in P)) \iff y\mathscr{R}x,$$

et donc $\mathscr{R}$ est symétrique.

Soit $(x,y,z)\in E^3$ tel que $x\mathscr{R}y$ et $y\mathscr{R}z$. Il existe P et Q dans $\mathscr{P}$ tels que

$$((x\in P \text{ et } y\in P) \text{ et } (y\in Q \text{ et } z\in Q)).$$

Comme $P\cap Q\neq\emptyset$ et que $\mathscr{P}$ est une partition, on a $P=Q$, donc $(x\in P$ et $z\in P)$, d'où $x\mathscr{R}z$. Ainsi, $\mathscr{R}$ est transitive.

Enfin, soit x un élément de E. Il existe P de $\mathscr{P}$ tel que x appartienne à P et l'on a alors $\mathscr{C}(x)=P$. En effet, pour tout y de P, $(x\in P$ et $y\in P)$ donc $x\mathscr{R}y$.

Pour tout élément y de $\mathscr{C}(x)$, il existe Q appartenant à $\mathscr{P}$ tel que $(x\in Q$ et $y\in Q)$ et $Q=P$ (pour les mêmes raisons que précédemment) et donc $y\in P$. Ceci prouve que $E/\mathscr{R}\subset\mathscr{P}$.

Réciproquement, soit $P\in\mathscr{P}$. Il existe x appartenant à P et l'on a alors $\mathscr{C}(x)=\mathscr{P}$. Ceci montre que $\mathscr{P}\subset E/\mathscr{R}$. □

Définition A.21 (relation d'ordre) *Soit $\mathscr{R}$ une relation binaire dans un ensemble E. On dit que $\mathscr{R}$ est une **relation d'ordre** si et seulement si $\mathscr{R}$ est réflexive, antisymétrique et transitive.*

Une relation d'ordre est souvent notée $\leq$. Le couple $(E,\leq)$, où E désigne un ensemble et $\leq$ est une relation d'ordre, est appelé un *ensemble ordonné*. Ajoutons que la relation $(x\leq y$ et $x\neq y)$ est notée $x<y$.

Définitions A.22 (relation d'ordre total et d'ordre partiel) *Soit $(E,\leq)$ un ensemble ordonné. La relation $\leq$ est dite **relation d'ordre total** si deux éléments quelconques de E sont **comparables**,*

$$\forall (x,y)\in E^2,\ (x\leq y \text{ ou } y\leq x).$$

*Dans le cas contraire, l'ordre est dit **partiel**.*

Soit $(E,\leq)$ un ensemble (totalement) ordonné et A une partie de E. La relation induite par $\leq$ dans A est une relation d'ordre (total) appelée *relation d'ordre induite par $\leq$ sur A.*

L'introduction d'une relation d'ordre sur un ensemble rend certains éléments des parties de cet ensemble remarquables. Ils sont l'objet des définitions suivantes.

Définitions A.23 *Soit $(E,\leq)$ un ensemble ordonné et A une partie de E.*

- *Un élément x de E est appelé un **majorant** (resp. **minorant**) de A dans E si et seulement si*

$$\forall a\in A,\ a\leq x \quad (\text{resp. } \forall a\in A,\ x\leq a).$$

- *On dit que A est **majorée** (resp. **minorée**) dans E si et seulement si cette partie admet au moins un majorant (resp. minorant) dans E, c'est-à-dire*

$$\exists x\in E,\ \forall a\in A,\ a\leq x \quad (\text{resp. } \exists x\in E,\ \forall a\in A,\ x\leq a).$$

- *Un élément x de E est appelé un **plus grand** (resp. **plus petit**) élément de A si et seulement s'il appartient à A et majore (resp. minore) A, c'est-à-dire*

$$(x\in A \text{ et } (\forall a\in A,\ a\leq x)) \quad (\text{resp. } (x\in A \text{ et } (\forall a\in A,\ x\leq a))).$$

- *Un élément x de A est dit **maximal** (resp. **minimal**) si et seulement si*

$$\forall a\in A,\ (x\leq a \implies x=a) \quad (\text{resp. } \forall a\in A,\ (a\leq x \implies x=a)).$$

Définition A.24 (bornes supérieure et inférieure) *Soit $(E,\leq)$ un ensemble ordonné et A une partie de E. On dit qu'un élément M de E est la **borne supérieure de A dans E**, notée $\sup A$, si l'ensemble des majorants de A dans E admet M comme plus petit élément. Un élément m de E sera appelé la **borne inférieure de A dans E**, notée $\inf A$, si l'ensemble des minorants de A dans E admet m comme plus grand élément.*

A.1.3 Applications

Nous allons à présent nous intéresser à des relations particulières nommées *applications*.

Définitions A.25 *On appelle **fonction** d'un ensemble E dans un ensemble F une relation qui à un élément de E associe au plus un élément de F. L'ensemble des éléments de E auxquels une fonction associe exactement un élément dans F est appelé l'**ensemble**, ou le **domaine, de définition** de cette fonction.*

Définitions A.26 *Soit E et F deux ensembles et f une fonction de E dans F. Tout élément y de F associé par la fonction f à un élément x de E est appelé l'**image** de x par f, ce que l'on note $y = f(x)$, tandis que x est un **antécédent** de y par f. On dit encore que E (resp. F) est l'**ensemble de départ** (resp. l'**ensemble de d'arrivée**) de f. Enfin, le **graphe** de la fonction est l'ensemble des couples $(x, f(x))$ lorsque x parcourt E.*

Définition A.27 (application) *Une fonction de E dans F est une **application** si et seulement si son domaine de définition est égal à E.*

On utilise la notation $f : E \longrightarrow F$ pour indiquer que f est une application d'un ensemble E dans un ensemble F. La définition de l'égalité entre deux ensembles (voir la définition A.1) implique que deux applications f et g de E dans F sont égales si, pour chaque élément x de E, on a $f(x) = g(x)$.

Définition A.28 (restriction d'une application) *Soit E et F deux ensembles, f une application de E dans F et A une partie de E. On appelle **restriction de f à A** l'application, notée $f_{|A}$, définie par*

$$\begin{array}{rccc} f_{|A} : & A & \longrightarrow & F \\ & x & \mapsto & f(x). \end{array}$$

Définition A.29 (prolongement d'une application) *Soit E et F deux ensembles, f une application de E dans F et G un ensemble tel que $E \subset G$. On appelle **prolongement de f à G** toute application $\widetilde{f} : G \longrightarrow F$ telle que*

$$\forall x \in E,\ \widetilde{f}(x) = f(x).$$

Définition A.30 (stabilité par une application) *Une partie A d'un ensemble E est dite **stable par** une application f de E dans E si et seulement si on a $f(a) \in A$, $\forall a \in A$.*

Définition A.31 (surjectivité d'une application) *Une application f d'un ensemble E dans un ensemble F est dite **surjective** (on dit encore que f est une **surjection**) si et seulement si*

$$\forall y \in F,\ \exists x \in E,\ y = f(x),$$

c'est-à-dire si tout élément de F est l'image par f d'au moins un élément de E.

Définition A.32 (injectivité d'une application) *Une application f d'un ensemble E dans un ensemble F est dite **injective** (on dit encore que f est une **injection**) si et seulement si*

$$\forall (x_1, x_2) \in E^2,\ f(x_1) = f(x_2) \implies x_1 = x_2,$$

c'est-à-dire si deux éléments distincts de E ont des images distinctes.

Définition A.33 (bijectivité d'une application) *Une application est dite **bijective** (on dit encore qu'elle est une **bijection**) si et seulement si elle est à la fois surjective et injective.*

Une bijection d'un ensemble E dans lui-même est appelée une *permutation* et l'ensemble des permutations de E est noté $\mathfrak{S}(E)$.

Proposition A.34 *Une application f d'un ensemble E dans un ensemble F est bijective si et seulement si tout élément de F possède un unique antécédent par f dans E, c'est-à-dire*

$$\forall y \in F,\ \exists! x \in E,\ f(x) = y.$$

DÉMONSTRATION. Si l'application f est bijective, alors elle est surjective. Par conséquent, tout élément y appartenant à F admet au moins un antécédent x par f dans E. Supposons maintenant que y ait deux antécédents x_1 et x_2. On a alors $y = f(x_1) = f(x_2)$, d'où $x_1 = x_2$ puisque f est injective. On en déduit que y admet un seul antécédent.
Réciproquement, si tout élément y de F admet un unique antécédent x par f dans E, alors f est surjective de E dans F. Soit x_1 et x_2 des éléments de E tels que $f(x_1) = f(x_2)$. Posons $y = f(x_1) = f(x_2)$, alors x_1 et x_2 sont deux antécédents de y. Par unicité de l'antécédent, on a $x_1 = x_2$, ce qui prouve l'injectivité de f. L'application f est donc bijective de E dans F. □

Introduisons maintenant les notions d'*application composée* et d'*application réciproque*.

Définition A.35 (application composée) *Soit E, F et G trois ensembles, f une application de E dans F et g une application de F dans G. L'application $g \circ f$ de E dans G définie par $g \circ f(x) = g(f(x))$ est appelée* ***composée de g et de f***.

Pour pouvoir définir l'application composée $g \circ f$, il est nécessaire que l'ensemble de départ de g soit égal à l'ensemble d'arrivée de f. L'ordre de composition est également important. Même dans le cas où l'on peut composer dans les deux sens, on a en général $g \circ f \neq f \circ g$.

Proposition A.36 *Soit E, F, G et H quatre ensembles, f une application de E dans F, g une application de F dans G et h une application de G dans H. On a*

$$(h \circ g) \circ f = h \circ (g \circ f).$$

On se référera à la preuve de la proposition A.13 pour une démonstration de ce dernier résultat.

Proposition A.37 *La composée de deux injections (resp. surjections, resp. bijections) est une injection (resp. surjection, resp. bijection).*

DÉMONSTRATION. Soit f une application d'un ensemble E dans un ensemble F et g une application de F dans un ensemble G, que l'on suppose dans un premier temps f et g injectives. On a, pour tout couple (x_1, x_2) de E^2,

$$(g \circ f)(x_1) = (g \circ f)(x_2) \iff g(f(x_1)) = g(f(x_2)) \implies f(x_1) = f(x_2) \implies x_1 = x_2,$$

d'où $g \circ f$ est injective.

On suppose à présent que f et g sont simplement surjectives. Soit z un élément de G. Puisque l'application g est surjective, il existe un élément y de F tel que $z = g(y)$. L'application f étant surjective, il existe alors un élément x de E tel que $y = f(x)$. On a donc $z = g(f(x)) = (g \circ f)(x)$, ce qui montre que $g \circ f$ est surjective.

Enfin, on a, en se servant des deux assertions qui viennent d'être démontrées,

$$(f \text{ et } g \text{ bijectives}) \implies \begin{cases} f \text{ et } g \text{ injectives} \\ f \text{ et } g \text{ surjectives} \end{cases} \implies \begin{cases} g \circ f \text{ injective} \\ g \circ f \text{ surjective} \end{cases} \implies (g \circ f \text{ bijective}).$$

□

Proposition A.38 *Soit E, F et G trois ensembles, f une application de E dans F et g une application de F dans G. Si $g \circ f$ est injective (resp. surjective), alors f est injective (resp. g est surjective).*

DÉMONSTRATION. Supposons que $g \circ f$ est injective. On a, pour tout couple (x_1, x_2) de E^2,

$$f(x_1) = f(x_2) \implies g(f(x_1)) = g(f(x_2)) \iff (g \circ f)(x_1) = (g \circ f)(x_2) \implies x_1 = x_2,$$

ce qui montre que f est injective.

Supposons maintenant que $g \circ f$ surjective. Pour tout élément z de G, il alors existe un élément x de E tel que $z = (g \circ f)(x) = g(f(x))$. L'application g est donc surjective. □

Définition A.39 (application réciproque) *Soit f une application d'un ensemble E dans un ensemble F. On appelle* ***application réciproque*** *(ou* ***inverse***) ***de*** *f toute application g de F dans E telle que*

$$\forall x \in E,\ g(f(x)) = x, \forall y \in F,\ f(g(y)) = y.$$

Proposition A.40 *Toute application admet au plus une application réciproque.*

DÉMONSTRATION. Soit f une application d'un ensemble E dans un ensemble F, g_1 et g_2 deux applications de F dans E satisfaisant aux conditions de la définition A.39. En particulier, on a, pour tout élément y de F, $f(g_1(y)) = y$ et, pour tout élément x de E, $g_2(f(x)) = x$. En composant la première de ces relations par l'application g_2 et en posant $x = g_1(y)$ dans la seconde, il vient alors $g_2(y) = g_2(f(g_1(y))) = g_1(y)$. □

Si f est une application d'un ensemble E dans un ensemble F admettant application réciproque, on note f^{-1} cette dernière. Dans ce cas, l'application f^{-1} est elle-même inversible et l'on a que $(f^{-1})^{-1} = f$.

Proposition A.41 *Une application d'un ensemble E dans un ensemble F admet une application réciproque si et seulement si elle est bijective.*

DÉMONSTRATION. Soit f une application de E dans F admettant une application réciproque. Pour tout élément y de F, on a $f(f^{-1}(y)) = y$ et f est donc surjective. Soit deux éléments x_1 et x_2 de E tels que $f(x_1) = f(x_2)$. Il vient alors $x_1 = f^{-1}(f(x_1)) = f^{-1}(f(x_2)) = x_2$, dont on déduit que l'application f est injective.

Considérons maintenant une application f bijective et g l'application de F dans E définie de la façon suivante : pour tout élément y de F, on pose $g(y) = x$ où x est l'unique antécédent de y par f. On vérifie alors de manière immédiate que $f(g(y)) = y$, $\forall y \in F$, et $g(f(x)) = x$, $\forall x \in E$, d'où $g = f^{-1}$. □

La proposition suivante est une conséquence directe des propositions A.37 et A.15.

Proposition A.42 *Soit E, F et G trois ensembles, f une application de E dans F et g une application de F dans G, toutes deux bijectives. L'application $g \circ f$ est alors bijective et l'on a $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.*

Définition A.43 (involution) *Soit E un ensemble. On appelle **involution de** E toute application f de E dans E telle que $f \circ f = I_E$.*

Définitions A.44 (images directe et réciproque d'une partie par une application) *Soit E et F deux ensembles, A une partie de E, B une partie de F et f une application de E dans F. L'**image directe de A par f**, ou, plus simplement, l'**image de A par f**, notée $f(A)$, est le sous-ensemble de F contenant l'image des éléments de A par f,*

$$f(A) = \{y \in F \mid \exists x \in A,\ y = f(x)\}.$$

*L'**image réciproque de B par f**, notée $f^{-1}(B)$, est le sous-ensemble de E contenant les antécédents des éléments de B par f,*

$$f^{-1}(B) = \{x \in E \mid f(x) \in B\}.$$

Pour toute application f d'un ensemble E dans un ensemble F, il est toujours possible de définir $f^{-1}(B)$, même si l'application n'est pas bijective. Lorsque c'est cependant le cas, c'est-à-dire si f^{-1} existe, on pourra vérifier que l'image directe d'une partie B de F par f^{-1} est aussi l'image réciproque $f^{-1}(B)$ de B par f. En effet, dire que x est un élément de $f^{-1}(B)$ signifie que $f(x)$ appartient à B et réciproquement, si l'on pose $y = f(x)$, on aura $x = f^{-1}(y)$ avec y un élément de B, ce qui équivaut à dire que x appartient à $f^{-1}(B)$.

La proposition qui suit est utile en pratique pour déterminer si une application est surjective ou non.

Proposition A.45 *Soit f une application d'un ensemble E dans un ensemble F. Elle est surjective si et seulement si $f(E) = F$.*

DÉMONSTRATION. On a toujours $f(E) \subset F$. Par ailleurs, l'ensemble F est inclus dans $f(E)$ si et seulement si tout élément de F est l'image d'au moins un élément de E par l'application f, ce qui signifie que f est surjective. □

La définition suivante constitue une généralisation de la notion de *suite* que nous étudierons notamment dans dans la section B.2 consacrée aux *suites numériques*.

Définition A.46 (famille) *Soit E et I des ensembles. On appelle **famille d'éléments de** E toute application de I à valeurs dans E, les éléments de I étant appelés les **indices**.*

Une famille $(x_i)_{i \in I}$ est dite *finie* ou *infinie*, selon que l'ensemble I de ses indices est fini ou infini (voir la sous-section A.1.4). On note $(x_i)_{i \in I}$ la famille d'éléments x_i d'un ensemble E indexée par les éléments i d'un ensemble I. On veillera à ne pas confondre la famille $(x_i)_{i \in I}$ et l'ensemble $\{\{x_i\} \mid i \in I\}$, qui est l'ensemble image de l'application en question.

Définitions A.47 (réunion et intersection de parties) *Soit E un ensemble et $(A_i)_{i\in I}$ une famille de parties de E. La **réunion** (resp. l'**intersection**) de la famille $(A_i)_{i\in I}$, notée $\bigcup_{i\in I} A_i$ (resp. $\bigcap_{i\in I} A_i$) est définie par*

$$\bigcup_{i\in I} A_i = \{x \in E \mid \exists i \in I,\ x \in A_i\} \text{ (resp. } \bigcap_{i\in I} A_i = \{x \in E \mid \forall i \in I,\ x \in A_i\}).$$

Définition A.48 *Soit E un ensemble. Une famille $(A_i)_{i\in I}$ de parties de E est appelée **partition de** E si et seulement si*

- *aucun des ensembles A_i n'est vide, $\forall \in I$, $A_i \neq \emptyset$,*
- *les ensembles A_i sont disjoints deux à deux, $\forall (i,j) \in I^2$, $(i \neq j \implies A_i \cap A_j = \emptyset)$,*
- *la réunion des ensembles A_i est égale à E, $\bigcup_{i\in I} A_i = E$.*

Il est à retenir de cette définition que tout élément de un ensemble appartient à un unique élément de sa partition. On notera par ailleurs que cet énoncé est cohérent avec la définition A.5, car pour que la famille $(A_i)_{i\in I}$ soit une partition au sens ci-dessus, il faut, et il suffit, que l'ensemble image $\{A_i \mid i \in I\}$ soit une partition de E au sens de cette définition.

A.1.4 Cardinalité, ensembles finis et infinis

Nous terminons en rappelant quelques propriétés élémentaires relatives aux ensembles finis, souvent considérées comme intuitivement évidentes.

Définition A.49 (relation d'équipotence) *On dit qu'un ensemble E est **équipotent** à un ensemble F si et seulement s'il existe une bijection de E sur F.*

La relation d'équipotence constitue une relation d'équivalence entre ensembles. Elle va permettre de formaliser la *dénombrabilité* et la *finitude* d'un ensemble.

Définition A.50 (ensemble dénombrable) *On dit qu'un ensemble est **dénombrable** si et seulement s'il est équipotent à l'ensemble des entiers naturels $\mathbb{N}$.*

Définition A.51 (ensembles finis et infinis) *On dit qu'un ensemble E est **fini** si et seulement s'il existe un entier naturel n tel que E est équipotent à $\{1,\dots,n\}$. Il est dit **infini** si et seulement s'il n'est pas fini.*

Nous admettrons la proposition suivante, dont la preuve s'appuie sur les propriétés de l'ensemble des entiers naturels.

Proposition A.52 *Soit (n,p) un couple d'entiers naturels. On a les assertions suivantes.*

1. *Il existe une injection de $\{1,\dots,n\}$ dans $\{1,\dots,p\}$ si et seulement si $n \leq p$.*
2. *Il existe une surjection de $\{1,\dots,n\}$ sur $\{1,\dots,p\}$ si et seulement si $n \leq p$.*
3. *Il existe une bijection de $\{1,\dots,n\}$ dans $\{1,\dots,p\}$ si et seulement si $n = p$.*

La dernière assertion de cette proposition amène la définition suivante.

Définition A.53 (cardinal d'un ensemble) *Soit E un ensemble fini. Il existe alors un entier naturel n, appelé le **cardinal** de E et noté $\mathrm{card}(E)$, tel que E soit équipotent à $\{1,\dots,n\}$.*

Par convention, le cardinal de l'ensemble vide est égal à 0.

Proposition A.54 *Si E est un ensemble fini, toute partie F de E est finie, et l'on a $\mathrm{card}(F) \leq \mathrm{card}(E)$.*

Proposition A.55 *Si E et F sont deux ensembles finis, alors l'ensemble $E \cup F$ est fini et l'on a*

$$\mathrm{card}(E \cup F) + \mathrm{card}(E \cap F) = \mathrm{card}(E) + \mathrm{card}(F).$$

DÉMONSTRATION. Établissons tout d'abord un résultat préliminaire. Soit A et B deux ensembles finis disjoints ; notons $a = \mathrm{card}(A)$, $b = \mathrm{card}(B)$. Il existe des bijections $\alpha : \{1,\dots,a\} \implies A$ et $\beta : \{1,\dots,b\} \implies B$. Il est clair que l'application $\gamma : \{1,\dots,a+b\} \implies A \cup B$ définie par

$$\forall n \in \{1,\dots,a+b\},\ \gamma(n) = \begin{cases} \alpha(n) & \text{si } 1 \leq n < a \\ \beta(n-a) & \text{si } a+1 \leq n \leq a+b \end{cases}$$

est une bijection. Il en résulte que l'ensemble $A \cup B$ est fini et que $\text{card}(A \cup B) = \text{card}(A) + \text{card}(B)$.

En appliquant ce résultat aux ensembles E et $E\backslash F$, il vient que l'ensemble $E \cup F$ est fini et

$$\begin{aligned}\text{card}(E \cup F) + \text{card}(E \cap F) = \text{card}(E \cup (F\backslash E)) + \text{card}(E \cap F) &= (\text{card}(E) + \text{card}(F\backslash E)) + \text{card}(E \cap F) \\ &= \text{card}(E) + (\text{card}(F\backslash E) + \text{card}(E \cap F)) = \text{card}(E) + \text{card}(F).\end{aligned}$$

□

Corollaire A.56 *Soit E un ensemble fini et F une partie de E. Si $\text{card}(F) = \text{card}(E)$ alors $F = E$.*

DÉMONSTRATION. Si $\text{card}(F) = \text{card}(E)$, comme $\text{card}(E) = \text{card}(F) + \text{card}(E\backslash F)$, on en déduit $\text{card}(E\backslash F) = 0$, d'où $E\backslash F = \varnothing$, $E = F$. □

Cette dernière propriété est particulièrement importante. Nous en verrons une analogue portant sur des dimensions d'espaces vectoriels en algèbre linéaire.

Proposition A.57 *Soit E et F des ensembles finis ayant même cardinal et f une application de E dans F. Les assertions suivantes sont deux à deux équivalentes.*

i) L'application f est injective.

ii) L'application f est surjective.

iii) L'application f est bijective.

DÉMONSTRATION. Montrons tout d'abord que i implique ii et iii. Si l'application f est injective, alors la *corestriction* $f^{|_{f(E)}} : E \longrightarrow f(E)$ qui à un élément x de E associe $f(x)$ est bijective, donc $\text{card}(f(E)) = \text{card}(E) = \text{card}(F)$ et $f(E) = F$ d'après le corollaire A.56, d'où f est surjective et par conséquent bijective.

Prouvons à présent que ii implique i et iii. Supposons l'application f est surjective mais non injective. Il existe dans ce cas un couple (x_1, x_2) de E^2 tel que $x_1 \neq x_2$ et $f(x_1) \neq f(x_2)$. L'application $g : E\backslash\{x_2\} \longrightarrow F$ qui à un élément x de E différent de x_2 associe $f(x)$ est surjective, d'où $\text{card}(E\backslash\{x_2\}) \leq \text{card}(F)$. Mais on a $\text{card}(E\backslash\{x_2\}) = \text{card}(E) - 1$ et $\text{card}(E) = \text{card}(F)$, d'où une contradiction.

Enfin, l'assertion iii implique i et ii de manière triviale. □

A.2 Structures algébriques

Nous allons maintenant étudier des exemples de *structures*, c'est-à-dire des ensembles munis d'une ou de plusieurs « opérations » appelées *lois de composition* et satisfaisant à un certain nombre d'axiomes.

A.2.1 Lois de composition

Commençons par introduire les applications particulières que sont les lois de composition. Tout d'abord, étant donnés trois ensembles E, F et G non vides, toute application de l'ensemble produit $E \times F$ à valeurs dans l'ensemble G est appelée loi de composition de $E \times F$ dans G. Cependant, dans toute la suite, nous aurons systématiquement $E = F = G$ ou bien encore $E = G \neq F$. Ces deux cas particuliers de loi de composition sont l'objet des définitions suivantes.

Définition A.58 (loi de composition interne) *Soit E un ensemble non vide. On appelle **loi de composition interne sur** E toute application de $E \times E$ dans E.*

Les opérations d'addition et de multiplication sur l'ensemble des entiers naturels $\mathbb{N}$ sont deux exemples de loi de composition interne.

Définition A.59 (loi de composition externe) *Soit E et F des ensembles non vides. On appelle **loi de composition externe sur** E **à opérateurs dans** F toute application de $F \times E$ à valeurs dans E.*

On dit encore d'une telle loi qu'elle est une *action de l'ensemble F sur l'ensemble E*. Dans la définition ci-dessus, on remarque que l'on a choisi de placer le *domaine d'opérateurs F* en premier dans le produit $F \times E$, c'est-à-dire qu'on a considéré, de manière implicite, que la loi de composition était externe *à gauche*, mais des lois de composition externes *à droite* sont également possibles. Ajoutons que lorsque les *opérateurs externes*, c'est-à-dire les éléments

de l'ensemble F, sont des nombres réels ou complexes (ou, plus généralement, les éléments d'un *corps*), ceux-ci sont appelés *scalaires* et l'on a coutume de noter la loi de composition externe multiplicativement en utilisant le symbole « $\cdot$ ».

Donnons à présent un tout premier exemple de structure et énonçons quelques propriétés des lois de composition internes.

Définition A.60 (magma) *Soit E un ensemble non vide et $\star$ une loi de composition interne sur E. On appelle* ***magma*** *le couple $(E,\star)$.*

Définitions A.61 (associativité et commutativité d'une loi interne) *Soit $(E,\star)$ un magma. La loi $\star$ est dite* ***associative*** *(et le magma $(E,\star)$* ***associatif****) si*

$$\forall (x,y,z)\in E^3,\ (x\star y)\star z = x\star(y\star z).$$

Elle est ***commutative*** *(et le magma $(E,\star)$* ***commutatif****) si*

$$\forall (x,y)\in E^2,\ x\star y = y\star x.$$

Définition A.62 (élément neutre) *Soit $(E,\star)$ un magma. Un élément e de E est un élément* ***neutre*** *(resp.* ***neutre à gauche****, resp.* ***neutre à droite****) pour la loi $\star$ si*

$$\forall x\in E,\ e\star x = x\star e = x \text{ (resp. } e\star x = x,\text{ resp. } x\star e = x).$$

Un magma possédant un élément neutre est dit *unifère*.

Proposition A.63 *Si un magma possède un élément neutre alors ce dernier est unique.*

Un deuxième exemple de structure est fourni par la définition suivante.

Définition A.64 (monoïde) *On appelle* ***monoïde*** *un magma associatif unifère.*

Définition A.65 (élément symétrique) *Soit $(E,\star)$ un magma pour lequel la loi de composition interne $\star$ admet un élément neutre e. On dit qu'un élément x de E possède un* ***élément symétrique*** *(resp.* ***élément symétrique à gauche****, resp.* ***élément symétrique à droite****) pour la loi $\star$ s'il existe un élément y de E tel que*

$$x\star y = y\star x = e \text{ (resp. } y\star x = e,\text{ resp. } x\star y = e).$$

En général, un élément donné d'un magma unifère peut avoir plusieurs éléments symétriques à gauche ou à droite et même plusieurs éléments symétriques à gauche *et* à droite. Cependant, si l'on travaille sur un monoïde, c'est-à-dire si la loi de composition interne considérée est associative, et qu'un élément possède à la fois un élément symétrique à droite et un élément symétrique à gauche, ceux-ci sont égaux et l'élément symétrique est unique. Dans ce cas, l'élément symétrique d'un élément x est généralement noté $-x$ lorsque la loi de composition est l'addition, x^{-1} ou $\frac{1}{x}$ si c'est la multiplication.

Lorsque un ensemble est muni de deux lois de composition, une propriété particulièrement intéressante est la *distributivité*.

Définition A.66 (distributivité) *Soit E un ensemble non vide muni de deux lois de composition internes $\star$ et $\circ$. La loi $\star$ est dite* ***distributive à gauche*** *(resp.* ***distributive à droite****) par rapport à la loi $\circ$ si*

$$\forall (x,y,z)\in E^3,\ x\star(y\circ z) = (x\star y)\circ(x\star z) \text{ (resp. } (y\circ z)\star x = (y\star x)\circ(z\star x)).$$

On dit que la loi $\star$ est ***distributive*** *par rapport à la loi $\circ$ si elle est distributive à gauche et à droite par rapport à $\circ$. Ces définitions restent valables lorsque $\star$ est une loi de composition externe à opérateurs dans un ensemble F non vide et $\circ$ une loi de composition interne sur E, à condition que l'élément x appartienne à F.*

A.2.2 Structures de base

Dans cette section, nous introduisons des structures qui, comme les magmas, ne sont munies que de lois de composition internes et, comme les monoïdes, satisfont à des axiomes.

Groupes

Parmi les structures algébriques les plus simples se trouve la notion de *groupe*. Elle occupe une place centrale dans les mathématiques et en physique en raison du lien étroit qu'elle possède avec la notion de *symétrie*.

Définition A.67 (groupe) *On appelle* ***groupe*** *tout magma* $(E,\star)$ *vérifiant les propriétés suivantes :*

- *la loi* $\star$ *est associative :* $\forall (x,y,z)\in E^3,\ (x\star y)\star z = x\star(y\star z)$,
- *il existe un élément neutre pour la loi* $\star$ *:* $\exists e\in E,\ \forall x\in E,\ e\star x = x\star e = x$,
- *tout élément possède un élément symétrique pour la loi* $\star$ *:* $\forall x\in E,\ \exists y\in E,\ x\star y = y\star x = e$.

Lorsque la loi de composition interne d'un groupe est commutative, on dit que le groupe est *commutatif*, ou encore *abélien*.

Les ensembles $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ et $\mathbb{C}$ munis de l'addition usuelle sont des groupes. Il en est de même des ensembles $\mathbb{Q}^*$, $\mathbb{R}^*$ et $\mathbb{C}^*$ munis de la multiplication ou de l'ensemble des bijections d'un ensemble E dans lui-même muni de la composition des applications de E dans E.

Anneaux

Un autre exemple de structure jouant un rôle fondamental en mathématiques est celui des *anneaux*, qui interviennent notamment dans l'étude des équations algébriques et des nombres algébriques.

Définition A.68 (anneau) *On appelle* ***anneau*** *(**ring** en anglais) tout ensemble E non vide muni deux lois de composition internes* $+$ *et* $*$ *tel que*

- $(E,+)$ *est un groupe commutatif, c'est-à dire que*
  - *la loi* $+$ *est associative :* $\forall (x,y,z)\in E^3,\ x+(y+z)=(x+y)+z$,
  - *il existe un élément neutre, noté* 0_E*, pour la loi* $+$ *:* $\forall x\in E,\ x+0_E = 0_E+x = x$,
  - *tout élément possède un élément symétrique pour la loi* $+$ *:* $\forall x\in E,\ \exists(-x)\in E,\ x+(-x)=(-x)+x=0_E$,
- $(E,*)$ *est un monoïde, c'est-à dire que*
  - *la loi* $*$ *est associative :* $\forall (x,y,z)\in E^3,\ x*(y*z)=(x*y)*z$,
  - *il existe un élément neutre, noté* 1_E*, pour la loi* $*$ *:* $\forall x\in E,\ x*1_E = 1_E*x = x$,
- *la loi* $*$ *est distributive par rapport à la loi* $+$ *:* $\forall (x,y,z)\in E^3,\ x*(y+z)=x*y+x*z$ *et* $(y+z)*x = y*x+z*x$.

Les lois $+$ et $*$ sont traditionnellement appelées *addition* et *multiplication* (on remarquera que les notations utilisées dans cette définition pour les éléments unitaires pour l'addition et la multiplication, ainsi que pour le symétrique d'un élément pour l'addition, sont, bien purement conventionnelles, intuitives). Ajoutons que l'on parle d'anneau *commutatif* lorsque la multiplication est de plus commutative.

Les ensembles $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ et $\mathbb{C}$ munis de l'addition et de la multiplication usuelles sont des anneaux.

Corps

La structure algébrique qui nous servira en algèbre linéaire est le *corps*.

Définition A.69 (corps) *On appelle* ***corps*** *(**field** en anglais) tout anneau* $(E,+,*)$ *tel que* $(E\backslash\{0_E\},*)$ *est un groupe.*

On dit qu'un corps est *commutatif* si la multiplication est commutative.

Des exemples de corps commutatifs sont les ensembles $\mathbb{Q}$, $\mathbb{R}$ et $\mathbb{C}$ munis de l'addition et de la multiplication usuelles.

A.2.3 Structures à opérateurs externes

Nous allons maintenant nous intéresser à des structures possédant à la fois des lois de composition internes *et* externes. Elles peuvent être aussi bien considérées d'un point vue algébrique que géométrique.

Dans toute la suite de cette annexe, on notera simplement $\mathbb{K}$ un corps commutatif $(\mathbb{K},+,*)$ appelé le corps des *scalaires*, avec $\mathbb{K}=\mathbb{R}$ (corps des nombres réels) ou bien $\mathbb{K}=\mathbb{C}$ (corps des nombres complexes) et où les lois $+$ et $*$ sont respectivement l'addition et la multiplication usuelles.

Espaces vectoriels *

L'*espace vectoriel* (*vector space* en anglais) est la structure de base en algèbre linéaire. Elle permet, en autres choses, d'effectuer des *combinaisons linéaires* de ses éléments.

Définition A.70 (espace vectoriel) *Un **espace vectoriel sur un corps commutatif** $\mathbb{K}$ est un ensemble non vide E muni d'une loi de composition interne, appelée **addition** et notée $+$, et d'une loi de composition externe à opérateurs dans $\mathbb{K}$, appelée **multiplication par un scalaire** et notée $\cdot$, possédant les propriétés suivantes :*

- $(E,+)$ *est un groupe commutatif,*
- $\forall(\lambda,\mu)\in\mathbb{K}^2$ *et* $\forall x\in E$, $(\lambda+\mu)x=\lambda x+\mu x$,
- $\forall\lambda\in\mathbb{K}$ *et* $\forall(x,y)\in E^2$, $\lambda(x+y)=\lambda x+\lambda y$,
- $\forall(\lambda,\mu)\in\mathbb{K}^2$ *et* $\forall x\in E$, $\lambda(\mu x)=(\lambda\mu)x$,
- $\forall x\in E$, $1_{\mathbb{K}}x=x$,

le scalaire $1_{\mathbb{K}}$ étant l'élément unitaire du corps $\mathbb{K}$.
*Les éléments de d'un espace vectoriel sont appelés des **vecteurs**.*

Dans cette définition, on observera qu'on a employé, par abus, le même symbole « $+$ » pour les lois additives sur $\mathbb{K}$ et E. On a également omis d'écrire le symbole « $\cdot$ » lorsqu'on multiplie un vecteur par un scalaire. Ajoutons qu'on utilisera dans la suite la seule lettre E pour désigner un espace vectoriel $(E,+,\cdot)$, comme c'est souvent le cas dans la pratique.

Définition A.71 (sous-espace vectoriel) *On dit qu'une partie non vide F d'un espace vectoriel E est un **sous-espace vectoriel** de E si et seulement si*

$$\forall\lambda\in\mathbb{K},\ \forall(x,y)\in F^2,\ \lambda x+y\in F.$$

On dit encore qu'un sous-espace vectoriel d'un espace vectoriel E est un sous-ensemble de E stable par les lois de composition interne et externe dont est muni E.

proprietes ? (intersection de sev, somme ?)
petit plan :
- def. combinaison linéaire
- def. famille libre, génératrice, base
REPRENDRE

Définitions A.72 *Une famille de vecteurs $\{x_i\}_{i=1,\dots,p}$ d'un espace vectoriel E est dite **libre** si les vecteurs $x_1,\dots,x_p$ sont **linéairement indépendants**, c'est-à-dire si la relation*

$$\lambda_1x_1+\cdots+\lambda_px_p=\mathbf{0},$$

*où 0 est l'élément nul de E et $\lambda_i\in\mathbb{K}$, $i=1,\dots,p$, implique que $\lambda_1=\cdots=\lambda_p=0$. Dans le cas contraire, la famille est dite **liée**.*

En particulier, l'ensemble des combinaisons linéaires d'une famille $\{x_i\}_{i=1,\dots,p}$ de p vecteurs de E est un sous-espace vectoriel de E, appelé *sous-espace engendré* par la famille de vecteurs. On le note

$$\mathrm{Vect}\{x_1,\dots,x_p\}=\left\{\boldsymbol{v}=\lambda_1x_1+\cdots+\lambda_px_p,\ \text{avec }\lambda_i\in\mathbb{K},\ i=1,\dots,p\right\}.$$

La famille $\{\boldsymbol{v}_i\}_{i=1,\dots,p}$ est alors appelée *famille génératrice* de ce sous-espace.

On appelle *base* de l'espace vectoriel E toute famille libre et génératrice de E. Si la famille $\{e_i\}_{i=1,\dots,n}$ est une base de E, tout vecteur de E admet une décomposition unique de la forme

$$x=\sum_{i=1}^{n}\lambda_ie_i,\ \forall\boldsymbol{v}\in E,$$

les scalaires λ_i, $i=1,\dots,n$, étant appelés les *composantes* du vecteur $\boldsymbol{v}$ dans la base $\{e_i\}_{i=1,\dots,n}$. On a de plus les résultats suivants.

Théorème A.73 *Si E est un espace vectoriel de dimension finie n, alors toute famille libre (et donc toute base) est finie et de cardinal au plus égal à n.*

DÉMONSTRATION. On va montrer par récurrence sur $n \geq 1$ que si $\mathscr{G} = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$ est une famille génératrice de E et si $\mathscr{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n, \boldsymbol{f}_{n+1}\}$ est une famille de $n+1$ éléments de E, alors cette dernière famille est liée.
Pour $n = 1$, on a $\boldsymbol{f}_1 = \lambda_1 \boldsymbol{g}_1$ et $\boldsymbol{f}_2 = \lambda_2 \boldsymbol{g}_1$. On en déduit que $\mathscr{F}$ est liée, car ou bien $\boldsymbol{f}_1 = \boldsymbol{0}$, ou bien $\boldsymbol{f}_2 = \frac{\lambda_2}{\lambda_1} \boldsymbol{f}_1$.
On suppose maintenant $n \geq 2$. Il existe alors une famille $\{a_{ij}\}_{i=1,\ldots,n+1,\ j=1,\ldots,n}$ de scalaires telle que

$$\begin{array}{ccccccccc}
\boldsymbol{f}_1 & = & a_{11}\boldsymbol{g}_1 & + & \ldots & + & a_{1n-1}\boldsymbol{g}_{n-1} & + & a_{1n}\boldsymbol{g}_n, \\
\boldsymbol{f}_2 & = & a_{21}\boldsymbol{g}_1 & + & \ldots & + & a_{2n-1}\boldsymbol{g}_{n-1} & + & a_{2n}\boldsymbol{g}_n, \\
\vdots & & \vdots & & & & \vdots & & \vdots \\
\boldsymbol{f}_n & = & a_{n1}\boldsymbol{g}_1 & + & \ldots & + & a_{nn-1}\boldsymbol{g}_{n-1} & + & a_{nn}\boldsymbol{g}_n, \\
\boldsymbol{f}_{n+1} & = & a_{n+11}\boldsymbol{g}_1 & + & \ldots & + & a_{n+1n-1}\boldsymbol{g}_{n-1} & + & a_{n+1n}\boldsymbol{g}_n.
\end{array}$$

Si les coefficients a_{in}, $1 \leq i \leq n+1$, sont nuls, alors les vecteurs $\boldsymbol{f}_i$, $1 \leq i \leq n+1$, sont dans $\mathrm{Vect}\{\boldsymbol{g}_i\}_{i=1,\ldots,n-1}$; de l'hypothèse de récurrence, on déduit que la famille $\{\boldsymbol{f}_i\}_{i=1,\ldots,n}$ est liée et donc que $\mathscr{F}$ est liée.
Sinon, il existe un entier i compris entre 1 et $n+1$, disons $i = n+1$ tel que $a_{in} \neq 0$. On peut alors remplacer $\boldsymbol{g}_n$ par $\frac{1}{a_{n+1n}}\left(\boldsymbol{f}_{n+1} - \sum_{j=1}^{n-1} a_{n+1j}\boldsymbol{g}_j\right)$, de sorte que les vecteurs $\boldsymbol{h}_j = \boldsymbol{f}_j - \frac{a_{jn}}{a_{n+1n}}\boldsymbol{f}_{n+1}$, $1 \leq j \leq n$ sont encore dans $\mathrm{Vect}\{\boldsymbol{g}_i\}_{i=1,\ldots,n-1}$. Par hypothèse de récurrence, la famille $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n\}$ est liée : il existe des scalaires $\lambda_1, \ldots, \lambda_n$ non tous nuls tels que $\sum_{i=1}^n \lambda_i \boldsymbol{h}_i = \sum_{i=1}^n \lambda_i \boldsymbol{f}_i + \mu \boldsymbol{f}_{n+1} = 0_E$. On en déduit que $\mathscr{F}$ est liée. □

A PLACER QUELQUE PART :

Dans un espace vectoriel, toute famille génératrice contient au moins une base de l'espace vectoriel. Étant donnée une famille libre, il existe au moins une base qui la contient (« théorème de la base incomplète »)

Si I est un ensemble, l'ensemble $\mathbb{K}^I$ des applications de I dans $\mathbb{K}$ est naturellement muni d'une structure d'espace vectoriel. La famille de vecteurs $\{\boldsymbol{e}_i\}_{i\in I}$ de $\mathbb{K}^I$ définie par

$$(\boldsymbol{e}_i)_j = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}, i \in I, j \in I,$$

forme une base de $\mathbb{K}^I$ appelée *base canonique*.

- sous-espace engendre
- dimension

Définition A.74 (dimension d'un espace vectoriel) *Le cardinal d'une base quelconque d'un espace vectoriel E de dimension finie s'appelle la* ***dimension de*** *E et se note* $\dim E$.

Définition A.75 *Un espace vectoriel sur $\mathbb{K}$ est dit* ***de dimension finie*** *s'il admet une famille génératrice de cardinal fini. Sinon, il est dit* ***de dimension infinie****.*

Corollaire A.76 *Si E est un espace vectoriel de dimension finie, alors toutes ses bases sont finies et ont le même cardinal.*

DÉMONSTRATION. Si $\mathscr{B}$ et $\mathscr{B}'$ sont deux bases, alors $\mathscr{B}$ est libre et $\mathscr{B}'$ est génératrice, donc $\mathrm{card}\,\mathscr{B} \leq \mathrm{card}\,\mathscr{B}'$ par le théorème précédent. On obtient l'autre inégalité en échangeant $\mathscr{B}$ et $\mathscr{B}'$. □

Dans toute la suite, nous ne considérons que des espaces vectoriels de dimension finie.

Algèbres *

Définition A.77 (algèbre) *On appelle* algèbre sur un corps commutatif $\mathbb{K}$ *tout ensemble E non vide muni de deux lois de composition internes $+$ et $*$ et d'une loi de composition externe $\cdot$ à opérateurs dans $\mathbb{K}$ tels que*

- *$(E, +, *)$ est un anneau,*
- *$(E, +, \cdot)$ est un espace vectoriel sur $\mathbb{K}$,*
- *$\forall \lambda \in \mathbb{K}$ et $\forall (x, y) \in E^2$, $\lambda(x * y) = (\lambda x) * y = x * (\lambda y)$.*

Lorsque la loi $*$ est commutative, l'algèbre est dite *commutative*.

A.3 Matrices

Soit m et n deux entiers strictement positifs. Une *matrice* (*matrix*[3] en anglais) *A à n lignes et m colonnes à coefficients dans un corps* $\mathbb{K}$ est une application définie sur l'ensemble $\{1,\dots,n\}\times\{1,\dots,m\}$ à valeurs dans $\mathbb{K}$, représentée par le tableau suivant

$$A=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}.$$

Les mn scalaires a_{ij}, $i=1,\dots,n$, $j=1,\dots,m$, sont appelés les *coefficients* ou les *éléments* (*entries* ou *elements* en anglais) de la matrice A, le premier indice i étant celui de la *ligne* (*row* en anglais) de la matrice à laquelle appartient l'élément considéré et le second j étant celui de la *colonne* (*column* en anglais). Ainsi, l'ensemble des coefficients $a_{i1},\dots,a_{im}$ forme la i^{e} ligne de la matrice et l'ensemble $a_{1j},\dots,a_{nj}$ la j^{e} colonne. Les éléments d'une matrice A sont notés $(A)_{ij}$ ou, plus simplement, a_{ij} lorsque qu'aucune confusion ou ambiguïté n'est possible.

On note $M_{n,m}(\mathbb{K})$ l'ensemble des matrices à n lignes et m colonnes dont les coefficients appartiennent à $\mathbb{K}$. Une matrice est dite *réelle* ou *complexe* selon que ses éléments sont dans $\mathbb{R}$ ou $\mathbb{C}$. Si $m=n$, la matrice est dite *carrée d'ordre n* (*square matrix of order n* en anglais) et l'on note $M_n(\mathbb{K})$ l'ensemble des matrices correspondant. Lorsque $m\neq n$, on parle de matrice *rectangulaire* (*rectangular matrix* en anglais).

On appelle *diagonale principale* (*main diagonal* en anglais) d'une matrice A de $M_{n,m}(\mathbb{K})$ l'ensemble des coefficients a_{ii}, $i=1,\dots,\min(m,n)$. Cette diagonale divise la matrice en une partie *sur-diagonale*, composée des éléments dont l'indice de ligne est strictement inférieur à l'indice de colonne, et une partie *sous-diagonale* formée des éléments pour lesquels l'indice de ligne est strictement supérieur à l'indice de colonne.

Étant donné $A\in M_{n,n}(\mathbb{R})$, on note $A^\top\in M_{m,n}(\mathbb{R})$ la *matrice transposée*[4] (*the transpose of the matrix* en anglais) de A telle que

$$(A^\top)_{ij}=(A)_{ji},\ 1\leq i\leq m,\ 1\leq j\leq n.$$

On a alors $(A^\top)^\top=A$. De même, étant donné $A\in M_{n,m}(\mathbb{C})$, on note $A^*\in M_{m,n}(\mathbb{C})$ la *matrice adjointe* (*the conjugate transpose of the matrix* en anglais) de A telle que

$$(A^*)_{ij}=\overline{(A)_{ji}},\ 1\leq i\leq m,\ 1\leq j\leq n,$$

le scalaire $\overline{z}$ désignant le nombre complexe conjugué du nombre z, et on $(A^*)^*=A$.

On appelle *vecteur ligne* (*row vector* en anglais) (resp. *vecteur colonne* (*column vector* en anglais)) une matrice n'ayant qu'une ligne (resp. colonne). Nous supposerons toujours qu'un vecteur est un vecteur colonne, c'est-à-dire que l'on représentera le vecteur $\boldsymbol{v}$ de $\mathbb{K}^n$ dans la base $\{\boldsymbol{e}_i\}_{i=1\dots,n}$ par la matrice

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

et que le *vecteur transposé* $\boldsymbol{v}^\top$ (resp. *vecteur adjoint* $\boldsymbol{v}^*$) de $\boldsymbol{v}$ sera alors représenté par la matrice

$$\begin{pmatrix} v_1 & v_2 & \dots & v_n \end{pmatrix}\ (\text{resp. } \begin{pmatrix} \overline{v_1} & \overline{v_2} & \dots & \overline{v_n} \end{pmatrix}).$$

Enfin, dans les démonstrations, il sera parfois utile de considérer un ensemble constitué de lignes et de colonnes particulières d'une matrice. On introduit pour cette raison la notion de *sous-matrice*.

3. L'introduction de ce terme est attribuée à Sylvester, son plus ancien ancien usage connu étant dans la publication [Syl50], à la page 369, où il désigne un objet donnant naissance aux déterminants de sous-matrices carrées appelés mineurs : *"For this purpose we must commence, not with a square, but with an oblong arrangement of terms consisting, suppose, of m lines and n columns. This will not in itself represent a determinant, but is, as it were, a Matrix out of which we may form various systems of determinants by fixing upon a number p, and selecting at will p lines and p columns, the squares corresponding to which may be termed determinants of the pth order."*.

4. On peut aussi définir la matrice transposée d'une matrice à coefficients complexes, mais cette notion n'a en général que peu d'intérêt dans ce cas.

Définition A.78 (sous-matrice) *Soit A une matrice de $M_{n,m}(\mathbb{K})$. Soient $1 \leq i_1 < \cdots < i_p \leq n$ et $1 \leq j_1 < \cdots < j_q \leq m$ deux ensembles d'indices. La matrice S de $M_{p,q}(\mathbb{K})$ ayant pour coefficients*

$$s_{kl} = a_{i_k j_l},\ 1 \leq k \leq p,\ 1 \leq l \leq q,$$

*est appelée une **sous-matrice** (**submatrix** en anglais) de A.*

Il est courant d'associer à une matrice une partition en sous-matrices.

Définition A.79 (matrice par blocs) *Une matrice A de $M_{n,m}(\mathbb{K})$ est dite **par blocs** ou **partitionnée** (**block matrix** ou **partitioned matrix** en anglais) si elle s'écrit*

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1M} \\ A_{21} & A_{22} & \dots & A_{2M} \\ \vdots & \vdots & & \vdots \\ A_{N1} & A_{N2} & \dots & A_{NM} \end{pmatrix},$$

*où les **blocs** A_{IJ}, $1 \leq I \leq N$, $1 \leq J \leq M$, sont des sous-matrices de A.*

L'intérêt de telles partitions réside dans le fait que certaines opérations définies sur les matrices restent formellement les mêmes (sous réserve que les opérations entre sous-matrices soient possibles, on parle alors de *partitions compatibles* (*conformable partitions* en anglais)), les coefficients de la matrice étant remplacés par ses sous-matrices.

A.3.1 Opérations sur les matrices

Nous rappelons à présent quelques opérations essentielles définies sur les matrices.

Définition A.80 (égalité de matrices) *Soit A et B deux matrices de $M_{n,m}(\mathbb{K})$. On dit que A est **égale** à B si $a_{ij} = b_{ij}$ pour $i = 1, \dots, n$, $j = 1, \dots, m$.*

Définition A.81 (somme de matrices) *Soit A et B deux matrices de $M_{n,m}(\mathbb{K})$. On appelle **somme** des matrices A et B la matrice C de $M_{n,m}(\mathbb{K})$, notée $A+B$, dont les coefficients sont $c_{ij} = a_{ij} + b_{ij}$, $i = 1, \dots, n$, $j = 1, \dots, m$.*

L'élément neutre pour la somme de matrices est la *matrice nulle*, notée 0, dont les coefficients sont tous égaux à zéro. On rappelle que l'on a par ailleurs

$$\forall A \in M_{n,m}(\mathbb{K}),\ \forall B \in M_{n,m}(\mathbb{K}),\ (A+B)^\top = A^\top + B^\top \text{ et } (A+B)^* = A^* + B^*.$$

Définition A.82 (multiplication d'une matrice par un scalaire) *Soit A une matrice de $M_{n,m}(\mathbb{K})$ et λ un scalaire. Le résultat de la **multiplication de la matrice A par le scalaire λ** est la matrice C de $M_{n,m}(\mathbb{K})$, notée αA, dont les coefficients sont $c_{ij} = \lambda a_{ij}$, $i = 1, \dots, n$, $j = 1, \dots, m$.*

On a

$$\forall \alpha \in \mathbb{K},\ \forall A \in M_{n,m}(\mathbb{K}),\ (\alpha A)^\top = \alpha A^\top \text{ et } (\alpha A)^* = \overline{\alpha} A^*.$$

Muni des deux dernières opérations, l'ensemble $M_{n,m}(\mathbb{K})$ est un espace vectoriel sur $\mathbb{K}$ (la vérification est laissée en exercice). On appelle alors *base canonique de $M_{n,m}(\mathbb{K})$* l'ensemble des mn matrices E_{kl}, $k = 1, \dots, n$, $l = 1, \dots, m$, de $M_{n,m}(\mathbb{K})$ dont les éléments sont définis par

$$(E_{kl})_{ij} = \begin{cases} 0 & \text{si } i \neq k \text{ ou } j \neq l \\ 1 & \text{si } i = k \text{ et } j = l \end{cases},\ 1 \leq i \leq n,\ 1 \leq j \leq m.$$

Définition A.83 (produit de matrices) *Soit A une matrice de $M_{n,p}(\mathbb{K})$ et B une matrice de $M_{p,m}(\mathbb{K})$. Le **produit** (**matrix product** en anglais) des matrices A et B est la matrice C de $M_{n,m}(\mathbb{K})$, notée AB, dont les coefficients sont*

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{jk},\ i = 1, \dots, n, j = 1, \dots, m. \tag{A.1}$$

Le produit de matrices est associatif et distributif par rapport à la somme de matrices.

Dans le cas de matrices carrées, on dit que deux matrices A et B *commutent* (pour le produit de matrices) si $AB = BA$. Toujours dans ce cas, l'élément neutre pour le produit de matrices d'ordre n est la matrice carrée, appelée *matrice identité* (*identity matrix* en anglais), définie par

$$I_n = (\delta_{ij})_{1\leq i,j\leq n},$$

avec δ_{ij} le *symbole de Kronecker*[5],

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases} \tag{A.2}$$

Cette matrice est, par définition, la seule matrice d'ordre n telle que $AI_n = I_nA = A$ pour toute matrice A d'ordre n. Muni de la multiplication par un scalaire, de la somme et du produit de matrices l'ensemble $M_n(\mathbb{K})$ est une algèbre sur $\mathbb{K}$, en général non commutative comme le montre l'exemple suivant.

Exemple de non-commutativité du produit de matrices carrées. Soit $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ et $B = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$. On a

$$AB = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 1 & 4 \\ 0 & 0 \end{pmatrix} = BA.$$

Si A est une matrice d'ordre n et p un entier, on définit la matrice A^p comme étant le produit de A par elle-même répété p fois, en posant $A^1 = A$ et $A^0 = I_n$. On mentionne enfin que l'on a

$$\forall A \in M_{n,p}(\mathbb{K}),\ \forall B \in M_{p,m}(\mathbb{K}),\ (AB)^\top = B^\top A^\top \text{ et } (AB)^* = B^*A^*.$$

Indiquons que toutes les précédentes opérations s'étendent au cas de matrices par blocs, pourvu que la taille de chacun des blocs soit telle que l'opération considérée soit correctement définie. On a notamment le résultat suivant.

Lemme A.84 (produit de matrices par blocs) *Soient A et B deux matrices de tailles compatibles pour effectuer le produit AB. Si A admet une décomposition en blocs $(A_{IK})_{1\leq I\leq N,\ 1\leq K\leq P}$ de formats respectifs (r_I, s_K) et B admet une décomposition compatible en blocs $(B_{KJ})_{1\leq K\leq P,\ 1\leq J\leq M}$ de formats respectifs (s_K, t_J), alors le produit $C = AB$ peut aussi s'écrire comme une matrice par blocs $(C_{IJ})_{1\leq I\leq N,\ 1\leq J\leq M}$, de formats respectifs (r_I, t_J) et donnés par*

$$C_{IJ} = \sum_{K=1}^{P} A_{IK}B_{KJ},\ 1 \leq I \leq N,\ 1 \leq J \leq M.$$

Exemple. Soit les matrices A et B d'ordre n admettant les partitions compatibles

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \text{ et } B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

On a alors

$$AB = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}.$$

Terminons en mentionnant qu'il existe d'autres types de produit que le produit « usuel » de matrices introduit ci-dessus, comme le produit composante par composante de matrices ou encore le produit tensoriel de matrices.

Définition A.85 (« produit d'Hadamard » ou « produit de Schur ») *Soit A et B deux matrices de $M_{n,m}(\mathbb{K})$. Le **produit d'Hadamard** (**Hadamard product** en anglais) des matrices A et B est la matrice C de $M_{n,m}(\mathbb{K})$, encore notée $A \circ B$, dont les coefficients sont $c_{ij} = a_{ij}b_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.*

Définition A.86 (« produit de Kronecker ») *Soit A une matrice de $M_{n,m}(\mathbb{K})$ et B une matrice de $M_{p,q}(\mathbb{K})$. Le **produit de Kronecker** (**Kronecker product** en anglais) des matrices A et B est la matrice C, encore notée $A \otimes B$, de $M_{np,mq}(\mathbb{K})$, définie par blocs $(C_{IJ})_{1\leq I\leq n,\ 1\leq J\leq m}$ de format (p, q), donnés par*

$$C_{IJ} = a_{IJ}B,\ 1 \leq I \leq n,\ 1 \leq J \leq m.$$

5. Leopold Kronecker (7 décembre 1823 - 29 décembre 1891) était un mathématicien et logicien allemand. Il était persuadé que l'arithmétique et l'analyse doivent être fondées sur les « nombres entiers » et apporta d'importantes contributions en théorie des nombres algébriques, en théorie des équations et sur les fonctions elliptiques.

A.3.2 Liens entre applications linéaires et matrices

Dans cette sous-section, on va établir qu'une matrice est la représentation d'une *application linéaire* entre deux espaces vectoriels, chacun de dimension finie, relativement à des bases données de ces espaces. Pour ce faire, l'introduction de quelques notions est nécessaire.

Définition A.87 (application linéaire) *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$ et f une application de E dans F. On dit que f est une* ***application linéaire (linear map** en anglais) si*

$$\forall \lambda \in \mathbb{K},\ \forall (\boldsymbol{v}, \boldsymbol{w}) \in E^2,\ f(\lambda \boldsymbol{v} + \boldsymbol{w}) = \lambda f(\boldsymbol{v}) + f'\boldsymbol{w}).$$

L'ensemble des applications linéaires de E dans F est noté $\mathscr{L}(E,F)$. Lorsque $F = \mathbb{K}$, on parle de ***forme linéaire (linear form** en anglais)*.

Définitions A.88 *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$ et f une application de $\mathscr{L}(E,F)$. On appelle* ***noyau (kernel** ou **null space** en anglais) de f, et l'on note* $\ker(f)$*, l'ensemble*

$$\ker(f) = \{\boldsymbol{x} \in E \,|\, f(\boldsymbol{x}) = \boldsymbol{0}\}.$$

On dit que l'application f est ***injective** si* $\ker(f) = \{\boldsymbol{0}\}$.
On appelle ***image (range** en anglais) de f, et l'on note* $\mathrm{Im}(f)$*, l'ensemble*

$$\mathrm{Im}(f) = \{\boldsymbol{y} \in F \,|\, \exists \boldsymbol{x} \in E,\ \boldsymbol{y} = f\boldsymbol{x})\},$$

et le ***rang (rank** en anglais) de f est la dimension de* $\mathrm{Im}(f)$*. L'application f est dite* ***surjective** si* $\mathrm{Im}(f) = F$.
Enfin, on dit que f est ***bijective**, ou que c'est un* ***isomorphisme (isomorphism** en anglais), si elle est injective et surjective.*

Le résultat suivant, dit *théorème du rang* (*rank-nullity theorem* en anglais), lie les dimensions respectives du noyau et de l'image d'une application linéaire.

Théorème A.89 (« théorème du rang ») *Soit E et F deux espaces vectoriels de dimension finie sur un même corps $\mathbb{K}$. Pour toute application f de $\mathscr{L}(E,F)$, on a*

$$\dim(\ker(f)) + \dim(\mathrm{Im}(f)) = \dim(E).$$

DÉMONSTRATION. Notons $n = \dim(E)$. Le sous-espace vectoriel $\ker(f)$ de E admet au moins une base $\{\boldsymbol{e}_i\}_{i=1,\dots,p}$ que l'on peut compléter en une base $\{\boldsymbol{e}_i\}_{i=1,\dots,n}$ de E. Nous allons montrer que $\{f(\boldsymbol{e}_{p+1}),\dots,f(\boldsymbol{e}_n)\}$ est une base de $\mathrm{Im}(f)$. Les vecteurs $f(\boldsymbol{e}_i)$, $p+1 \le i \le n$, sont à l'évidence des éléments de $\mathrm{Im}(f)$. Soit des scalaires $\lambda_{p+1},\dots,\lambda_n$ tels que

$$\sum_{i=p+1}^{n} \lambda_i f(\boldsymbol{e}_i) = \boldsymbol{0}.$$

On a $f\left(\sum_{i=p+1}^{n} \lambda_i \boldsymbol{e}_i\right) = \boldsymbol{0}$, et $\sum_{i=p+1}^{n} \lambda_i \boldsymbol{e}_i$ appartient donc à $\ker(f)$. Il existe par conséquent des scalaires $\mu_1,\dots,\mu_p$ tels que $\sum_{i=p+1}^{n} \lambda_i \boldsymbol{e}_i = \sum_{i=1}^{p} \mu_i \boldsymbol{e}_i$, d'où $\mu_1 \boldsymbol{e}_1 + \cdots + \mu_p \boldsymbol{e}_p - \lambda_{p+1}\boldsymbol{e}_{p+1} - \cdots - \lambda_n \boldsymbol{e}_n = \boldsymbol{0}$. Comme la famille $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_p\}$ est libre, on en déduit que $\lambda_{p+1} = \cdots = \lambda_n = 0$, ce qui montre que $\{f(\boldsymbol{e}_{p+1}),\dots,f(\boldsymbol{e}_n)\}$ est libre.

Soit à présent un élément $\boldsymbol{y}$ de $\mathrm{Im}(f)$. Par définition, il existe un vecteur $\boldsymbol{x}$ dans E tel que $\boldsymbol{y} = f(\boldsymbol{x})$. Puisque la famille $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\}$ engendre E, on peut trouver une famille $\{\alpha_1,\dots,\alpha_n\}$ d'éléments de $\mathbb{K}$ telle que $\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{e}_i$. On a alors

$$\boldsymbol{y} = f(\boldsymbol{x}) = f\left(\sum_{i=1}^{n} \alpha_i \boldsymbol{e}_i\right) = \sum_{i=1}^{n} \alpha_i f(\boldsymbol{e}_i) = \sum_{i=p+1}^{n} \alpha_i f(\boldsymbol{e}_i),$$

les vecteurs $\boldsymbol{e}_i$, $1 \le i \le p$, appartenant au noyau de f. La famille $\{f(\boldsymbol{e}_{p+1}),\dots,f(\boldsymbol{e}_n)\}$ engendre donc $\mathrm{Im}(f)$, c'est une base de ce sous-espace de F et l'on conclut alors que

$$\dim(\mathrm{Im}(f)) = n - p = \dim E - \dim(\ker(f)).$$

□

Supposons à présent que E et F sont deux espaces vectoriels, tous deux de dimension finie avec $\dim(E) = m$ et $\dim(F) = n$. Soit des bases respectives $\{\boldsymbol{e}_i\}_{i=1,\dots,m}$ une base de E et $\{\boldsymbol{f}_i\}_{i=1,\dots,n}$ une base de F. Pour toute application linéaire f de E dans F, on peut écrire que

$$\forall j \in \{1,\dots,m\},\ f(\boldsymbol{e}_j) = \sum_{i=1}^{n} a_{ij} \boldsymbol{f}_i, \tag{A.3}$$

ce qui conduit à la définition suivante.

Définition A.90 (représentation matricielle d'une application linéaire) *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$, de dimensions respectivement égales à m et n. On appelle* ***représentation matricielle*** *de l'application linéaire f de $\mathscr{L}(E,F)$, relativement aux bases $\{\boldsymbol{e}_i\}_{i=1,\dots,m}$ et $\{\boldsymbol{f}_i\}_{i=1,\dots,n}$, la matrice A de $M_{n,m}(\mathbb{K})$ ayant pour coefficients les scalaires a_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$, définis de manière unique par les relations* (A.3).

Une application de $\mathscr{L}(E,F)$ étant complètement caractérisée par la donnée de la matrice A et d'une couple de bases, on en déduit que $\mathscr{L}(E,F)$ est isomorphe à $M_{n,m}(\mathbb{K})$. Cet isomorphisme n'est cependant pas intrinsèque, puisque la représentation matricielle dépend des bases respectivement choisies pour E et pour F.

Réciproquement, si on se donne une matrice, alors il existe une infinité de choix d'espaces vectoriels et de bases qui permettent de définir une infinité d'applications linéaires dont elle sera la représentation matricielle. Par commodité, on fait le choix « canonique » de considérer l'application linéaire de $\mathbb{K}^m$ dans $\mathbb{K}^n$, tous deux munis de leurs bases canoniques respectives, qui admet pour représentation cette matrice. On peut ainsi étendre aux matrices toutes les définitions précédemment introduites pour les applications linéaires.

Définitions A.91 (noyau, image et rang d'une matrice) *Soit A une matrice de $M_{n,m}(\mathbb{K})$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Le* ***noyau*** *de A est le sous-espace vectoriel de $\mathbb{K}^m$ défini par*

$$\ker(A) = \{\boldsymbol{x} \in \mathbb{K}^m \mid A\boldsymbol{x} = \boldsymbol{0}\}.$$

*L'**image** de A est le sous-espace vectoriel de $\mathbb{K}^n$ défini par*

$$\mathrm{Im}(A) = \{\boldsymbol{y} \in \mathbb{K}^n \mid \exists \boldsymbol{x} \in \mathbb{K}^m \text{ tel que } A\boldsymbol{x} = \boldsymbol{y}\},$$

*et le **rang** de A est la dimension de cette image,*

$$\mathrm{rang}(A) = \dim(\mathrm{Im}(A)).$$

En vertu du théorème du rang (voir le théorème A.89), on a, pour toute matrice A de $M_{n,m}(\mathbb{K})$, la relation

$$\dim(\ker(A)) + \mathrm{rang}(A) = m,$$

dont on déduit que $\mathrm{rang}(A) \leq \min(m,n)$, la matrice étant dite *de rang maximal* si $\mathrm{rang}(A) = \min(m,n)$.

A.3.3 Inverse d'une matrice

Définitions A.92 *Soit A une matrice d'ordre n. On dit que A est* ***inversible*** *(**invertible** ou **non-singular** en anglais) s'il existe une (unique) matrice, notée A^{-1} et appelée la **matrice inverse** (**inverse matrix** en anglais) de A, telle que $AA^{-1} = A^{-1}A = I_n$. Une matrice non inversible est dite **singulière** (**singular** en anglais).*

Il ressort de cette définition qu'une matrice A inversible est la matrice d'un endomorphisme bijectif. Par conséquent, une matrice A d'ordre n est inversible si et seulement si $\mathrm{rang}(A) = n$.

Si une matrice A est inversible, son inverse est évidemment inversible et $(A^{-1})^{-1} = A$. On rappelle par ailleurs que, si A et B sont deux matrices inversibles, on a les égalités suivantes :

$$(AB)^{-1} = B^{-1}A^{-1},\ (A^\top)^{-1} = (A^{-1})^\top,\ (A^*)^{-1} = (A^{-1})^* \text{ et, } \forall \alpha \in \mathbb{K}^*,\ (\alpha A)^{-1} = \frac{1}{\alpha}A^{-1}.$$

L'ensemble des matrices inversibles de $M_n(\mathbb{K})$, muni du produit matriciel, est appelé le *groupe général linéaire* (*general linear group* en anglais) et noté $\mathrm{GL}_n(\mathbb{R})$.

A.3.4 Trace et déterminant d'une matrice

Nous rappellons dans cette section les notions de *trace* et de *déterminant* d'une matrice carrée.

Définition A.93 (trace d'une matrice) *La **trace** (**trace** en anglais) d'une matrice A d'ordre n, notée* $\mathrm{tr}(A)$*, est la somme de ses coefficients diagonaux,*

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

On montre facilement les relations

$$\forall \alpha \in \mathbb{K},\ \forall A \in M_n(\mathbb{K}),\ \forall B \in M_n(\mathbb{K}),\ \mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B) \text{ et } \mathrm{tr}(AB) = \mathrm{tr}(BA),$$

la première prouvant que l'application de $M_n(\mathbb{K})$ dans $\mathbb{K}$ qui à une matrice A d'ordre n associe $\mathrm{tr}(A)$ est linéaire, la seconde ayant comme conséquence le fait que la trace d'une matrice est invariante par changement de base. En effet, pour toute matrice A et tout matrice inversible P de même ordre, on a

$$\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(P^{-1}PA) = \mathrm{tr}(A).$$

Définition A.94 (déterminant d'une matrice) *On appelle **déterminant** (**determinant** en anglais) d'une matrice A d'ordre n le scalaire défini par la **formule de Leibniz***[6]

$$\det(A) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{\sigma(i)i},$$

où $\varepsilon(\sigma)$ désigne la signature d'une permutation[7] *σ de $\mathfrak{S}_n$.*

Par propriété des permutations, on a $\det(A^\top) = \det(A)$ et $\det(A^*) = \overline{\det(A)}$, pour toute matrice A d'ordre n.

On peut voir le déterminant d'une matrice A d'ordre n comme une *forme multilinéaire* des n colonnes de cette matrice,

$$\det(A) = \det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n),$$

où les vecteurs $\boldsymbol{a}_j$, $j = 1, \ldots, n$, désignent les colonnes de A. Ainsi, multiplier une colonne (ou une ligne, puisque $\det(A) = \det(A^\top)$) de A par un scalaire α multiplie le déterminant par ce scalaire. On a notamment

$$\forall \alpha \in \mathbb{K},\ \forall A \in M_n(\mathbb{K}),\ \det(\alpha A) = \alpha^n \det(A).$$

Cette forme est de plus *alternée* : échanger deux colonnes (ou deux lignes) de A entre elles entraîne la multiplication de son déterminant par -1 et si deux colonnes (ou deux lignes) sont égales ou, plus généralement, si les colonnes (ou les lignes) de A vérifient une relation non triviale de dépendance linéaire, le déterminant de A est nul. En revanche, ajouter à une colonne (resp. ligne) une combinaison linéaire des autres colonnes (resp. lignes) ne modifie pas le déterminant. Ces propriétés expliquent à elles seules le rôle essentiel que joue le déterminant en algèbre linéaire.

On rappelle enfin que le déterminant est un *morphisme de groupes*, c'est-à-dire une application entre deux groupes respectant la structure de ces groupes, du groupe linéaire des matrices inversibles de $M_n(\mathbb{K})$ dans $\mathbb{K}^*$ muni de la multiplication. Ainsi, si A et B sont deux matrices d'ordre n, on a

$$\det(AB) = \det(BA) = \det(A)\det(B),$$

et, si A est inversible,

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

Définition A.95 (déterminant extrait d'une matrice) *Soit A une matrice de $M_{n,m}(\mathbb{K})$ et p un entier strictement positif inférieur à m et à n. On appelle **déterminant extrait de A d'ordre p** le déterminant de n'importe quelle matrice d'ordre q obtenue à partir de A en éliminant $n-p$ lignes et $m-p$ colonnes.*

La démonstration du résultat suivant est immédiate.

6. Gottfried Wilhelm von Leibniz (1er juillet 1646 - 14 novembre 1716) était un philosophe, mathématicien (et plus généralement scientifique), bibliothécaire, diplomate et homme de loi allemand. Il inventa, indépendamment de Newton, le calcul intégral et différentiel et introduisit plusieurs notations mathématiques en usage aujourd'hui.

7. Rappelons qu'une permutation est une bijection d'un ensemble dans lui-même (voir la sous-section A.1.3). On note $\mathfrak{S}_n$ le groupe (pour la loi de composition $\circ$) des permutations de l'ensemble $\{1, \ldots, n\}$, avec $n \in \mathbb{N}$. La *signature* d'une permutation σ de $\mathfrak{S}_n$ est le nombre, égal à 1 ou -1, défini par

$$\varepsilon(\sigma) = \prod_{1 \le i < j \le n} \frac{\sigma(i) - \sigma(j)}{i - j}.$$

Proposition A.96 *Le rang d'une matrice A de $M_{n,m}(\mathbb{K})$ est égal à l'ordre maximal des déterminants extraits non nuls de A.*

On déduit de cette caractérisation et des propriétés du déterminant que $\text{rang}(A) = \text{rang}(A^\top) = \text{rang}(A^*)$.

Définitions A.97 (mineur, cofacteur et comatrice) *Soit A une matrice d'ordre n. On appelle **mineur** (**minor** en anglais) associé à l'élément a_{ij}, $1 \leq i, j \leq n$, de A le déterminant d'ordre $n-1$ de la matrice obtenue par suppression de la i^e et de la j^e colonne de A. On appelle **cofacteur** (**cofactor** en anglais) associé à ce même élément le scalaire*

$$\text{cof}_{ij}(A) = (-1)^{i+j} \begin{vmatrix} a_{11} & \dots & a_{1j-1} & a_{1j+1} & \dots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-11} & \dots & a_{i-1j-1} & a_{i-1j+1} & \dots & a_{i-1n} \\ a_{i+11} & \dots & a_{i+1j-1} & a_{i+1j+1} & \dots & a_{i+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj-1} & a_{nj+1} & \dots & a_{nn} \end{vmatrix}.$$

*Enfin, on appelle **matrice des cofacteurs**, ou **comatrice**, (**cofactor matrix** ou **comatrix** en anglais) de A la matrice d'ordre n constituée de l'ensemble des cofacteurs de A,*

$$\text{com}(A) = \left(\text{cof}_{ij}(A)\right)_{1 \leq i,j \leq n}.$$

DEFINITIONS *mineur principal* et *mineur principal dominant*

On remarque que si A est une matrice d'ordre n, α un scalaire et E_{ij}, $(i,j) \in \{1,\dots,n\}^2$, un vecteur de la base canonique de $M_n(\mathbb{K})$, on a, par multilinéarité du déterminant,

$$\det(A + \alpha E_{ij}) = \det(A) + \alpha \begin{vmatrix} a_{11} & \dots & a_{1j-1} & 0 & a_{1j+1} \dots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ a_{i-11} & \dots & a_{i-1j-1} & 0 & a_{i-1j+1} \dots & a_{i-1n} \\ a_{i1} & \dots & a_{ij-1} & 1 & a_{ij+1} \dots & a_{in} \\ a_{i+11} & \dots & a_{i+1j-1} & 0 & a_{i+1j+1} \dots & a_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nj-1} & 0 & a_{nj+1} \dots & a_{nn} \end{vmatrix} = \det(A) + \alpha \, \text{cof}_{ij}(A).$$

Cette observation conduit à une méthode récursive de calcul d'un déterminant d'ordre n par développement, ramenant ce calcul à celui de n déterminants d'ordre $n-1$, et ainsi de suite.

Proposition A.98 (« formule de Laplace ») *Soit A une matrice d'ordre n. On a*

$$\forall (i,j) \in \{1,\dots,n\}^2, \ \det(A) = \sum_{k=1}^{n} a_{ik} \, \text{cof}_{ik}(A) = \sum_{k=1}^{n} a_{kj} \, \text{cof}_{kj}(A).$$

DÉMONSTRATION. Quitte à transposer la matrice, il suffit de prouver la formule du développement par rapport à une colonne. On considère alors la matrice, de déterminant nul, obtenue en remplaçant la j^e colonne de A, $j \in \{1,\dots,n\}$, par une colonnes de zéros. Pour passer de cette matrice à A, on doit lui ajouter les n matrices $a_{ij} E_{ij}$, $i = 1,\dots,n$. On en déduit que pour passer du déterminant (nul) de cette matrice à celui de A, on doit lui ajouter les n termes $a_{ij} \, \text{cof}_{ij}$, $i = 1,\dots,n$, d'où le résultat. □

Proposition A.99 *Soit A une matrice d'ordre n. On a*

$$A(\text{com}(A))^\top = (\text{com}(A))^\top A = \det(A) \, I_n.$$

DÉMONSTRATION. Considérons la matrice, de déterminant nul, obtenue en remplaçant la j^e colonne de A, $j \in \{1,\dots,n\}$, par une colonnes de zéros et ajoutons lui les n matrices $a_{ik} E_{ik}$, $i = 1,\dots,n$, avec $k \in \{1,\dots,n\}$ et $k \neq j$. La matrice résultante est également de déterminant nul, puisque deux de ses colonnes sont identiques. Ceci signifie que

$$\forall (j,k) \in \{1,\dots,n\}^2, \ j \neq k, \ \sum_{i=1}^{n} a_{ik} \, \text{cof}_{ij}(A) = 0.$$

En ajoutant le cas $k = j$, on trouve

$$\forall (j,k) \in \{1,\dots,n\}^2,\ \sum_{i=1}^{n} a_{ik}\,\mathrm{cof}_{ij}(A) = \det(A)\,\delta_{jk},$$

ce qu'on traduit matriciellement par $A(\mathrm{com}(A))^\top = \det(A)\,I_n$. La seconde formule découle du fait que $\det(A^\top) = \det(A)$. □

Lorsque la matrice A est inversible, on a obtenu une formule pour son inverse,

$$A^{-1} = \frac{1}{\det(A)}\,(\mathrm{com}(A))^\top, \tag{A.4}$$

qui ne nécessite que des calculs de déterminants.

A.3.5 Valeurs et vecteurs propres

Les *valeurs propres* (*eigenvalues* en anglais) d'une matrice A d'ordre n sont les racines dans $\mathbb{K}$ du *polynôme caractéristique* (*characteristic polynomial* en anglais) associé à A, défini [8] par

$$p_A(\lambda) = \det(A - \lambda\, I_n).$$

Les valeurs propres de A sont donc les scalaires λ tels que le noyau de la matrice $A - \lambda\, I_n$ n'est pas réduit à $\{\mathbf{0}\}$. Le *spectre* (*spectrum* en anglais) de A, noté $\sigma_{\mathbb{K}}(A)$, est l'ensemble des valeurs propres de A dans $\mathbb{K}$ et, si $\sigma_{\mathbb{K}}(A) \neq \emptyset$, le *rayon spectral* (*spectral radius* en anglais) de A est le réel positif défini par

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \sigma_{\mathbb{K}}(A)\}.$$

Notons au passage que $\sigma_{\mathbb{K}}(A^\top) = \sigma_{\mathbb{K}}(A)$, puisque $\det(A^\top - \lambda\, I_n) = \det((A - \lambda\, I_n)^\top) = \det(A - \lambda\, I_n)$.

Dans le reste de cette sous-section, nous allons supposer que la matrice A d'ordre n, choisie de façon arbitraire, possède toujours n valeurs propres λ_i, $i = 1,\dots,n$, distinctes ou confondues (celles-ci étant alors comptées avec leur multiplicité), ce qui implique que le corps $\mathbb{K}$ est algébriquement clos [9] et donc que $\mathbb{K} = \mathbb{C}$. Dans ce cas, on a les propriétés suivantes :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i \text{ et } \det(A) = \prod_{i=1}^{n} \lambda_i,$$

la seconde impliquant que la matrice A est singulière dès que l'une de ses valeurs propres est nulle.

À toute valeur propre λ d'une matrice A est associé au moins un vecteur non nul $\boldsymbol{v}$ tel que

$$A\boldsymbol{v} = \lambda\,\boldsymbol{v},$$

appelé *vecteur propre* (*eigenvector* en anglais) de la matrice A correspondant à la valeur propre λ. Le sous-espace vectoriel $\ker(A - \lambda\, I_n)$, constitué de la réunion de l'ensemble des vecteurs propres associés à la valeur propre λ et du vecteur nul, est le *sous-espace propre* (*eigenspace* en anglais) correspondant à la valeur propre λ. Sa dimension définit la *multiplicité géométrique* (*geometric multiplicity* en anglais) de λ, qui ne peut jamais être supérieure à la *multiplicité algébrique* (*algebraic multiplicity* en anglais) de λ, c'est-à-dire la multiplicité de λ en tant que racine du polynôme caractéristique. Une valeur propre ayant une multiplicité géométrique inférieure à sa multiplicité algébrique est dite *défective* (*defective* en anglais).

Terminons cette sous-section par un résultat qui nous sera utile par la suite.

Lemme A.100 *Soit m et n deux entiers naturels non nuls et B et C deux matrices appartenant respectivement à $M_{n,m}(\mathbb{C})$ et $M_{m,n}(\mathbb{C})$. Alors, les valeurs propres non nulles des matrices BC et CB sont les mêmes.*

DÉMONSTRATION. Pour toute valeur propre λ de la matrice BC, il existe un vecteur $\boldsymbol{v}$ non nul de $\mathbb{C}^n$ tel que $BC\boldsymbol{v} = \lambda\,\boldsymbol{v}$. Si ce vecteur appartient de plus au noyau de la matrice C, on en déduit que $\lambda\,\boldsymbol{v} = 0$ et donc que $\lambda = 0$. Par conséquent, tout vecteur propre $\boldsymbol{v}$ de BC associé à une valeur propre λ non nulle ne peut appartenir à $\ker(C)$, et l'égalité $CBC\boldsymbol{v} = \lambda\, C\boldsymbol{v}$ implique alors que λ est également une valeur propre de la matrice CB, associée au vecteur propre $C\boldsymbol{v}$. □

8. On définit parfois le polynôme caractéristique comme étant égal à $\det(\lambda\, I_n - A)$, de manière à que le coefficient du terme de plus haut degré du polynôme soit toujours égal à 1. Ce n'est pas le cas avec le choix fait ici puisque l'on a $\det(A - \lambda\, I_n) = (-1)^n \det(\lambda\, I_n - A)$.

9. On rappelle qu'un corps commutatif $\mathbb{K}$ est dit *algébriquement clos* (*algebraically closed* en anglais) si tout polynôme de degré supérieur ou égal à un, à coefficients dans $\mathbb{K}$, admet (au moins) une racine dans $\mathbb{K}$.

A.3.6 Quelques matrices particulières

Nombre de méthodes numériques tirent parti des propriétés remarquables que présentent certaines classes de matrices recensées ci-après.

Matrice diagonale

Les matrices *diagonales* (*diagonal matrices* en anglais) jouent un rôle central en algèbre linéaire. D'un point de vue informatique, leur manipulation et leur stockage sont particulièrement aisés.

Définition A.101 (matrice diagonale) *Une matrice A d'ordre n est dite* ***diagonale*** *si on a $a_{ij} = 0$ pour les couples d'indices (i, j) de $\{1, \ldots, n\}^2$ tels que $i \neq j$.*

La démonstration du lemme suivant est laissée au lecteur.

Lemme A.102 *La somme et le produit de deux matrices diagonales sont des matrices diagonales. Le déterminant d'une matrice diagonale est égal au produit de ses éléments diagonaux. Une matrice diagonale A est donc inversible si et seulement si tous ses éléments diagonaux sont non nuls et, le cas échéant, son inverse est une matrice diagonale dont les éléments diagonaux sont les inverses des éléments diagonaux correspondants de A.*

Matrice triangulaire

Les matrices *triangulaires* (*triangular matrices* en anglais) forment une classe de matrices utilisée très couramment dans les méthodes numériques en raison de la facilité de résolution d'un système linéaire dont la matrice est triangulaire (voir la section 2.3).

Définition A.103 (matrice triangulaire) *On dit qu'une matrice A d'ordre n est* ***triangulaire supérieure*** *(resp.* ***inférieure****) si on a $a_{ij} = 0$ pour les couples d'indices (i, j) de $\{1, \ldots, n\}^2$ tels que $i > j$ (resp. $i < j$).*

Une matrice à la fois triangulaire supérieure et inférieure est une matrice diagonale. On vérifie par ailleurs facilement que la matrice transposée d'une matrice triangulaire supérieure est une matrice triangulaire inférieure, et vice versa.

La démonstration du lemme suivant est laissée en exercice.

Lemme A.104 *Soit A une matrice d'ordre n triangulaire supérieure (resp. inférieure). Son déterminant est égal au produit de ses termes diagonaux et elle est donc inversible si et seulement si ces derniers sont tous non nuls. Dans ce cas, son inverse est aussi une matrice triangulaire supérieure (resp. inférieure) dont les éléments diagonaux sont les inverses des éléments diagonaux de A. Soit B une autre matrice d'ordre n triangulaire supérieure (resp. inférieure). La somme $A + B$ et le produit AB sont des matrices triangulaires supérieures (resp. inférieures) dont les éléments diagonaux sont respectivement la somme et le produit des éléments diagonaux correspondants de A et B.*

Matrice bande

Une *matrice bande* (*band matrix* en anglais) est une matrice carrée dont les coefficients non nuls sont localisés dans une « bande » autour de la diagonale principale. Plus précisément, on a la définition suivante.

Définition A.105 (matrice bande) *Soit n un entier strictement positif. On dit qu'une matrice A de $M_n(\mathbb{R})$ est une* ***matrice bande*** *s'il existe des entiers positifs p et q strictement inférieurs à n tels que $a_{ij} = 0$ pour tous les couples d'entiers (i, j) de $\{1, \ldots, n\}^2$ tels que $i - j > p$ ou $j - i > q$. La* ***largeur de bande*** *(****bandwidth*** *en anglais) de la matrice vaut $p + q + 1$, avec p éléments a priori non nuls à gauche de la diagonale et q éléments à droite sur chaque ligne.*

Matrice à diagonale dominante

Les *matrices à diagonale dominante* (*diagonally dominant matrices* en anglais) possèdent des propriétés remarquables pour les différentes méthodes de résolution de systèmes linéaires présentées aux chapitres 2 et 3.

Définition A.106 *On dit qu'une matrice A d'ordre n est* ***à diagonale dominante par lignes*** *(respectivement* ***par colonnes****) si*

$$\forall i \in \{1,\dots,n\},\ |a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \ \ (\textit{resp. } |a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|).$$

On dit que A est ***à diagonale strictement dominante*** *(par lignes ou par colonnes respectivement) si ces inégalités sont strictes.*

Les matrices à diagonale strictement dominante possèdent la particularité d'être inversibles, comme le montre le résultat suivant, initialement [10] prouvé par Lévy [Lév81] et Desplanques [Des87].

Théorème A.107 *Une matrice A d'ordre n à diagonale strictement dominante (par lignes ou par colonnes) est inversible.*

DÉMONSTRATION. Supposons que A est une matrice à diagonale strictement dominante par lignes et prouvons l'assertion par l'absurde. Si A est non inversible, alors son noyau n'est pas réduit à zéro et il existe un vecteur $\boldsymbol{x}$ de $\mathbb{R}^n$ non nul tel que $A\boldsymbol{x} = \mathbf{0}$. Ceci implique que

$$\forall i \in \{1,\dots,n\},\ \sum_{j=1}^{n} a_{ij}\, x_j = 0.$$

Le vecteur $\boldsymbol{x}$ étant non nul, il existe un indice i_0 dans $\{1,\dots,n\}$ tel que $0 \neq |x_{i_0}| = \max_{1\leq i\leq n} |x_i|$ et l'on a alors

$$-a_{i_0 i_0} x_{i_0} = \sum_{\substack{j=1\\ j\neq i_0}}^{n} a_{i_0 j} x_j,$$

d'où

$$|a_{i_0 i_0}| \leq \sum_{\substack{j=1\\ j\neq i_0}}^{n} |a_{i_0 j}| \frac{|x_j|}{|x_{i_0}|} \leq \sum_{\substack{j=1\\ j\neq i_0}}^{n} |a_{i_0 j}|,$$

ce qui contredit le fait que A est à diagonale strictement dominante par lignes.

Si la matrice A est à diagonale strictement dominante par colonnes, on montre de la même manière que sa transposée $A^\top$, qui est une matrice à diagonale strictement dominante par lignes, est inversible et on utilise que $\det(A^\top) = \det(A)$. □

Matrice symétrique ou hermitienne

Définitions A.108 *Soit A une matrice carrée de $M_n(\mathbb{R})$. On dit que A est* ***symétrique*** *(****symmetric*** *en anglais) si $A = A^\top$,* ***antisymétrique*** *(****skew-symmetric*** *en anglais) si $A = -A^\top$ et* ***orthogonale*** *(****orthogonal*** *en anglais) si $A^\top A = AA^\top = I_n$.*

Définitions A.109 *Soit A une matrice carrée de $M_n(\mathbb{C})$. On dit que A est* ***hermitienne*** *(****Hermitian*** *en anglais) si $A = A^*$,* ***unitaire*** *(****unitary*** *en anglais) si $A^*A = AA^* = I_n$ et* ***normale*** *(****normal*** *en anglais) si $A^*A = AA^*$.*

On notera que les coefficients diagonaux d'une matrice hermitienne sont réels. On déduit aussi immédiatement de ces dernières définitions et de la définition A.92 qu'une matrice orthogonale est telle que $A^{-1} = A^\top$ et qu'une matrice unitaire est telle que $A^{-1} = A^*$.

M-matrice

Définition A.110 *Une matrice A de $M_n(\mathbb{R})$ est une* ***M-matrice*** *(****M-matrix*** *en anglais) si ses coefficients extra-diagonaux sont négatifs ou nuls et que l'on peut l'écrire sous la forme $A = s\, I_n - B$, où les coefficients de la matrice B sont négatifs ou nuls et le réel s est supérieur ou égal au rayon spectral de B.*

Il découle de cette définition et du théorème de Perron–Frobenius qu'une M-matrice est inversible si et seulement si le réel s de l'énoncé est strictement supérieur au rayon spectral $\rho(B)$. Dans ce cas, les coefficients diagonaux de la M-matrice sont strictement positifs.

Pour les M-matrices inversibles, il existe de nombreuses caractérisations équivalentes à la définition ci-dessus (voir [Ple77] par exemple).

10. Ce théorème semble avoir été redécouvert de nombreuses fois de manière totalement indépendante. On trouvera dans l'article [Tau49] une liste de références le concernant.

Proposition A.111 *Une matrice A d'ordre n dont les coefficients extra-diagonaux sont négatifs ou nuls. est une M-matrice inversible si et seulement si l'une des assertions suivantes est vérifiée :*

- *Tous les mineurs principaux de A sont strictement positifs.*
- *Tous les mineurs principaux dominants de A sont strictement positifs.*
- *Toute valeur propre réelle de A est strictement positive.*
- *La matrice A est inversible et les coefficients de son inverse sont positifs.*
- *On peut décomposer A sous la forme $A = M - N$, avec M inversible, dont l'inverse est à coefficients positifs, et N à coefficients positifs, telles que $\rho(M^{-1}N)$.*

Matrice réductible

Définition A.112 *Une matrice A de $M_n(\mathbb{K})$ est dite **réductible** (**reducible** en anglais) s'il existe une partition non triviale (I, J) de $\{1, \ldots, n\}$ telle que*

$$\forall (i,j) \in I \times J,\ a_{ij} = 0.$$

*Dans le cas contraire, la matrice A est dite **irréductible** (**irreducible** en anglais).*

De manière équivalente, une matrice A de $M_n(\mathbb{K})$ est réductible s'il existe une matrice de permutation P et des matrices B, C et D, appartenant respectivement à $M_r(\mathbb{K})$, $M_{r,n-r}(\mathbb{K})$ et $M_{n-r}(\mathbb{K})$, avec $1 \leq r \leq n-1$, telles que l'on a

$$PAP^\top = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}.$$

La résolution d'un système linéaire dont la matrice est réductible peut ainsi se ramener (ou *se réduire*) à celle de deux systèmes de taille moindre.

A.3.7 Matrices équivalentes et matrices semblables

Commençons par la définition suivante.

Définition A.113 (matrices équivalentes) *Soit m et n deux entiers naturels non nuls. Deux matrices A et B à m lignes et n colonnes sont dites **équivalentes** s'il existe deux matrices inversibles P et Q, respectivement d'ordre m et n, telles que $B = PAQ$.*

L'équivalence entre matrices au sens de cette définition est effectivement une relation d'équivalence et deux matrices sont équivalentes si et seulement si elles représentent une même application linéaire dans des bases différentes. De même, deux matrices sont équivalentes si et seulement si elles ont même rang.

Définition A.114 (matrices semblables) *On dit que deux matrices A et B d'ordre n sont **semblables** s'il existe une matrice P inversible telle que*

$$A = PBP^{-1}.$$

On dit que deux matrices A et B sont *unitairement* (resp. *orthogonalement*) semblables si la matrice P de la définition est unitaire (resp. orthogonale). Deux matrices sont semblables si et seulement si elles représentent un même endomorphisme dans deux bases différentes. La matrice P de la définition est donc une matrice de passage et on en déduit que deux matrices semblables possèdent le même rang, la même trace, le même déterminant et le même polynôme caractéristique (et donc le même spectre). Ces applications sont appelées *invariants de similitude*. Enfin, s'il ne faut pas confondre la notion de matrices semblables avec celle de matrices équivalentes, on voit que deux matrices semblables sont équivalentes.

L'exploitation de la similitude entre matrices permet entre autres choses de réduire la complexité du problème de l'évaluation des valeurs propres d'une matrice. En effet, si l'on sait transformer une matrice donnée en une matrice semblable diagonale ou triangulaire, le calcul des valeurs propres devient alors immédiat. On a notamment le théorème suivant [11].

Théorème A.115 (« décomposition de Schur ») *Soit une matrice A carrée. Il existe une matrice U unitaire telle que la matrice U^*AU soit triangulaire supérieure avec pour coefficients diagonaux les valeurs propres de A.*

11. Dans la démonstration de ce résultat, on fait appel à plusieurs notions abordées dans la section A.4.

DÉMONSTRATION. Le théorème affirme qu'il existe une matrice triangulaire unitairement semblable à la matrice A. Les éléments diagonaux d'une matrice triangulaire étant ses valeurs propres et deux matrices semblables ayant le même spectre, les éléments diagonaux de U^*AU sont bien les valeurs propres de A.

Le résultat est prouvé par récurrence sur l'ordre n de la matrice. Il est clairement vrai pour $n=1$ et on le suppose également vérifié pour une matrice d'ordre $n-1$, avec $n \geq 2$. Soit λ_1 une valeur propre d'une matrice A d'ordre n et soit $\boldsymbol{u}_1$ une vecteur propre associé normalisé, c'est-à-dire tel que $\|\boldsymbol{u}_1\|_2 = 1$. Ayant fait le choix de $n-1$ vecteurs pour obtenir une base orthonormée $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ de $\mathbb{C}^n$, la matrice U_n, ayant pour colonnes les vecteurs $\boldsymbol{u}_j$, $j = 1, \ldots, n$, est unitaire et on a

$$U_n^* A U_n = \begin{pmatrix} \lambda_1 & s_{12} & \ldots & s_{in} \\ 0 & & & \\ \vdots & & S_{n-1} & \\ 0 & & & \end{pmatrix},$$

où $s_{1j} = (\boldsymbol{u}_1, A\boldsymbol{u}_j)$, $j = 2, \ldots, n$, et où le bloc S_{n-1} est une matrice d'ordre $n-1$. Soit à présent U_{n-1} une matrice unitaire telle que $U_{n-1}^* S_{n-1} U_{n-1}$ soit une matrice triangulaire supérieure et soit

$$\widetilde{U}_{n-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & & & \\ \vdots & & U_{n-1} & \\ 0 & & & \end{pmatrix}.$$

La matrice $\widetilde{U}_{n-1}$ est unitaire et, par suite, $U_n \widetilde{U}_{n-1}$ également. On obtient par ailleurs

$$(U_n \widetilde{U}_{n-1})^* A (U_n \widetilde{U}_{n-1}) = \widetilde{U}_{n-1}^* (U_n^* A U_n) \widetilde{U}_{n-1} = \begin{pmatrix} \lambda_1 & t_{12} & \ldots & t_{in} \\ 0 & & & \\ \vdots & & U_{n-1}^* S_{n-1} U_{n-1} & \\ 0 & & & \end{pmatrix},$$

avec $\begin{pmatrix} \lambda_1 & t_{12} & \ldots & t_{in} \end{pmatrix} = \begin{pmatrix} \lambda_1 & s_{12} & \ldots & s_{in} \end{pmatrix} \widetilde{U}_{n-1}$, ce qui achève la preuve. □

La décomposition de Schur constitue un premier exemple de *réduction de matrice*, c'est-à-dire de transformation d'une matrice par changement de base en une matrice ayant une structure particulière (une matrice triangulaire supérieure en l'occurrence). Nous aborderons plus loin un autre de type de réduction, portant le nom de *diagonalisation*, dans laquelle on se ramène à une matrice diagonale.

Indiquons également que, parmi les différents résultats qu'implique cette décomposition, il y a en particulier le fait que toute matrice hermitienne A est unitairement semblable à une matrice diagonale réelle, les colonnes de la matrice U étant des vecteurs propres de A. Cette observation est le point de départ de la *méthode de Jacobi* pour le calcul approché des valeurs propres d'une matrice réelle symétrique (voir la section 4.5.1).

A.3.8 Matrice associée à une forme bilinéaire *

Une matrice n'est pas qu'une représentation d'une application linéaire ; elle peut également être associée à une *forme bilinéaire*. Pour le voir, commençons par définir ce dernier objet et énoncer quelques propriétés qui lui sont relatives.

Définition A.116 (application bilinéaire) *Soit E, F et G trois espaces vectoriels sur un même corps $\mathbb{K}$. Une application de $E \times F$ à valeurs dans G est dite* ***bilinéaire*** *(**bilinear** en anglais) si elle est linéaire en chacune de ses variables. L'ensemble des applications bilinéaires de $E \times F$ dans G est noté $\mathscr{L}(E, F; G)$. Lorsque $G = \mathbb{K}$, on parle de* ***forme bilinéaire*** *(**bilinear form** en anglais).*

Définition A.117 (symétrie d'une application bilinéaire) *Soit E et G deux espaces vectoriels sur un corps $\mathbb{K}$. Une application bilinéaire b sur $E \times E$ à valeurs dans G est dite* ***symétrique*** *(resp. **antisymétrique**) si elle vérifie*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in E \times E,\ b(\boldsymbol{x}, \boldsymbol{y}) = b(\boldsymbol{y}, \boldsymbol{x}) \text{ (resp. } b(\boldsymbol{x}, \boldsymbol{y}) = -b(\boldsymbol{y}, \boldsymbol{x})).$$

Définition A.118 (non dégénérescence d'une forme bilinéaire) *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$. Une forme bilinéaire b sur $E \times F$ est dite* ***non dégénérée à gauche*** *(**left nondegenerate** en anglais) si la* ***forme linéaire associée à gauche*** *$L(b)$,définie par*

$$\begin{array}{rcl} L(b) & : & E \to \mathscr{L}(F, \mathbb{K}) \\ & & \boldsymbol{x} \mapsto (\boldsymbol{y} \mapsto b(\boldsymbol{x}, \boldsymbol{y})) \end{array}\ ,$$

est injective. Elle est ***non dégénérée à droite*** *(**right nondegenerate** en anglais) si la* ***forme linéaire associée à droite*** *$R(b)$, définie par*

$$\begin{array}{rcl} R(b) & : & F \to \mathscr{L}(E,\mathbb{K}) \\ & & \boldsymbol{y} \mapsto (\boldsymbol{x} \mapsto b(\boldsymbol{x},\boldsymbol{y})) \end{array} \text{ ,}$$

est injective.

Résultat : E et F de dimensions finies et égales : non dégénérée à gauche = non dégénérée à droite, donner notion de rang d'une forme bilinéaire dans ce cas

La preuve du résultat suivant est immédiate.

Lemme A.119 *Pour une forme bilinéaire symétrique, la non dégénérescence à gauche et équivalente à la non dégénérescence à droite.*

Définition A.120 (forme bilinéaire définie) *Soit E un espace vectoriel sur un corps $\mathbb{K}$. Une forme bilinéaire b sur $E \times E$ est dite* ***définie positive*** *(resp.* ***définie négative****) si*

$$\forall \boldsymbol{x} \in E,\ \boldsymbol{x} \neq \boldsymbol{0},\ b(\boldsymbol{x},\boldsymbol{x}) > 0 \text{ (resp. } b(\boldsymbol{x},\boldsymbol{x}) < 0).$$

Lemme A.121 *Toute forme bilinéaire définie positive (ou définie négative) est non dégénérée.*

PREUVE

Définition A.122 (orthogonalité) *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$ et b une forme bilinéaire sur $E \times F$. On dit qu'un vecteur $\boldsymbol{x}$ de E et qu'un vecteur $\boldsymbol{y}$ de F sont* ***orthogonaux relativement à la forme*** *b si $b(\boldsymbol{x},\boldsymbol{y}) = 0$.*
On appelle ***orthogonal relativement à la forme*** *b d'une partie G de E, et l'on note $G^\perp$, l'ensemble des vecteurs de F orthogonaux à tous les éléments de G. On définit de façon similaire l'orthogonal d'une partie de F.*

On a, de manière immédiate, que l'orthogonal d'une partie est un sous-espace vectoriel.
$G^{\perp\perp}$ double orthogonal d'un sous-espace G de E est un sous-espace de E. On a toujours $G \subset G^{\perp\perp}$ (preuve en note ?)
Si E et F ont même dimension finie et b est une forme bilinéaire non dégénérée, on a, pour toute partie G de E ou F, $G = G^{\perp\perp}$
corollaire : E et F ont même dimension finie et b est une forme bilinéaire non dégénérée, G et H deux sev de E ou F, on a $(G+H)^\perp = G^\perp \cap H^\perp$ et $(G \cap H)^\perp = G^\perp + H^\perp$
theoreme : E et F de dimension n et b est une forme bilinéaire non dégénérée, pour toute partie G de E ou F, on a $\dim(G) + \dim(G^\perp) = n$
Si $E = F$, les sous-espaces G et $G^\perp$, bien que de dimensions complémentaires, ne sont néanmoins pas nécessairement en somme complémentaire
CNS (theoreme) : $G \cap G^\perp = \boldsymbol{0} \Leftrightarrow E = G + G^\perp \Leftrightarrow$ la restriction de b à G est non dégénérée
un vecteur est dit isotrope s'il est orthogonal à lui-même

Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$, de dimensions respectivement égales à m et n, et soit $\{\boldsymbol{e}_i\}_{i=1,\dots,m}$ et $\{\boldsymbol{f}_i\}_{i=1,\dots,n}$ des bases respectives de ces espaces. Pour tous vecteurs $\boldsymbol{x}$ de E et $\boldsymbol{y}$ de F, on a $\boldsymbol{x} = \sum_{i=1}^m x_i \boldsymbol{e}_i$ et $\boldsymbol{y} = \sum_{j=1}^m y_j \boldsymbol{f}_j$ et l'on peut donc écrire, pour toute forme bilinéaire b sur $E \times F$,

$$b(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^m \sum_{j=1}^n x_i y_j\, b(\boldsymbol{e}_i, \boldsymbol{f}_j).$$

Réciproquement, on peut construire une forme bilinéaire b sur $E \times F$ à partir d'une matrice M de $M_{m,n}(\mathbb{K})$ donnée en posant

$$\forall \boldsymbol{x} \in E,\ \forall \boldsymbol{y} \in F,\ b(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^m \sum_{j=1}^n x_i y_j\, m_{ij}.$$

Il existe donc un isomorphisme, dépendant du choix des bases, entre l'ensemble des formes bilinéaires sur $E \times F$ et $M_{m,n}(\mathbb{K})$ et l'on a la définition suivante.

Définition A.123 (matrice d'une forme bilinéaire) *Soit E et F deux espaces vectoriels sur un même corps $\mathbb{K}$, de dimensions respectivement égales à m et n, et b une forme bilinéaire sur $E \times F$. On appelle* ***matrice associée à*** *b* ***relativement aux bases*** *$\{\boldsymbol{e}_i\}_{i=1,\dots,m}$* ***et*** *$\{\boldsymbol{f}_i\}_{i=1,\dots,n}$ la matrice M à m lignes et n colonnes ayant pour coefficients*

$$m_{ij} = b(\boldsymbol{e}_i, \boldsymbol{f}_j)\ i = 1, \dots, m,\ j = 1, \dots, n.$$

Diverses propriétés dont :

Lorsque M est une matrice carrée, $E = F$ et $\boldsymbol{e}_i = \boldsymbol{f}_i$, une matrice symétrique est associée à une forme bilinéaire symétrique.

Une forme bilinéaire sur E est non dégénérée si et seulement si le déterminant de sa matrice dans une base quelconque est non nul.

formule de changement de base lorsque $E = F$

Conséquence (deux matrices sont congrues si et seulement si elles sont associées à une même forme bilinéaire dans deux bases différentes) :

Définition A.124 (matrices congrues) *Deux matrices A et B réelles (resp. complexes) d'ordre n sont dites* ***congrues*** *(****congruent*** *en anglais) s'il existe une matrice inversible P telle que*

$$A = P^\top BP \text{ (resp. } A = P^* BP).$$

La congruence entre matrices est une relation d'équivalence. Deux matrices congrues ont le même rang.

Définition A.125 (forme quadratique) *Soit E un espace vectoriel sur un corps $\mathbb{K}$. Une application q de E dans $\mathbb{K}$ est une* ***forme quadratique*** *(****quadratic form*** *en anglais) s'il existe une forme bilinéaire b sur $E \times E$ telle que*

$$\forall \boldsymbol{x} \in E,\ q(\boldsymbol{x}) = b(\boldsymbol{x}, \boldsymbol{x}).$$

Si toute forme bilinéaire sur une espace vectoriel E donne lieu à une forme quadratique associée, une forme quadratique donnée peut être associée à plusieurs formes bilinéaires. On a néanmoins le résultat suivant.

Proposition et définition A.126 (forme polaire d'une forme quadratique) *Soit E un espace vectoriel sur un corps $\mathbb{K}$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, et q une forme quadratique sur E. Parmi toutes les formes bilinéaires dont la forme quadratique associée est q, il en existe une unique qui est symétrique. On appelle une telle forme b la* ***forme polaire*** *de q, et celle-ci est donnée par*

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in E \times E,\ b(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left(q(\boldsymbol{x} + \boldsymbol{y}) - q(\boldsymbol{x}) - q(\boldsymbol{y})\right).$$

DÉMONSTRATION. A ECRIRE □

note : plus généralement vrai dans tout corps de caractéristique différente de deux

Réduction de Gauss d'une forme quadratique

Définition A.127 (signature d'une forme quadratique) *A ECRIRE (E défini sur $\mathbb{R}$*

Théorème A.128 (« loi d'inertie de Sylvester [12] » [Syl52]) *ENONCER*

Diagonalisation d'une forme quadratique ? ? ? (nécessite notion de base duale)

A.3.9 Diagonalisation des matrices *

Définition A.129 (matrice diagonalisable) *Une matrice carrée A est dite* ***diagonalisable*** *s'il existe une matrice inversible P et une matrice diagonale D telles que*

$$A = PDP^{-1}.$$

12. James Joseph Sylvester (3 septembre 1814 - 13 mars 1897) était un mathématicien et géomètre anglais. Il travailla sur les formes algébriques, en particulier sur les formes quadratiques et leurs invariants, et la théorie des déterminants. On lui doit l'introduction de nombreux objets, notions et notations mathématiques, comme le *discriminant* ou la *fonction indicatrice d'Euler*.

Il ressort de cette définition qu'une matrice A d'ordre n est diagonalisable si elle est semblable à une matrice diagonale. Dans ce cas, les éléments diagonaux de la matrice $P^{-1}AP$ sont les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ de A et la j^{e} colonne de la matrice P, $1 \leq j \leq n$, est formée des composantes (relativement à la base considérée) d'un vecteur propre associé à λ_j.

Proposition A.130 *REPRENDRE Une matrice est diagonalisable si et seulement ses vecteurs propres forment une base./une matrice est diagonalisable lorsque ses valeurs propres sont distinctes ou que leurs multiplicités algébrique et géométrique coïncident.*

Les matrices symétriques et hermitiennes vérifient un résultat de diagonalisation remarquable, que nous énonçons sans démonstration.

Théorème A.131 (diagonalisation des matrices symétriques et hermitiennes) *Soit A une matrice réelle symétrique (resp. complexe hermitienne) d'ordre n. Alors, il existe une matrice orthogonale (resp. unitaire) P telle que la matrice P^TAP (resp. P^*AP) soit une matrice diagonale. Les éléments diagonaux de cette matrice sont les valeurs propres de A et sont réels.*

A.3.10 Décomposition en valeurs singulières *

Il existe une manière plus générale que la diagonalisation pour réduire une matrice sous une forme diagonale, ce dernier mot prenant une forme adaptée lorsque la matrice n'est pas carrée, il s'agit de la *décomposition en valeurs singulières* (*singular value decomposition* en anglais). Avant d'introduire cette dernière, commençons par donner un résultat technique dont la démonstration est laissée au lecteur.

Lemme A.132 *Soit m et n deux entiers naturels non nuls et A une matrice de $M_{n,m}(\mathbb{C})$. Alors, la matrice A^*A est hermitienne d'ordre n et ses valeurs propres sont réelles et positives.*

En vertu de ce lemme, la définition suivante a bien un sens.

Définition A.133 (valeurs singulières d'une matrice) *On appelle **valeurs singulières (singular values** en anglais) d'une matrice A de $M_{n,m}(\mathbb{C})$ les racines carrées positives des valeurs propres de la matrice A^*A.*

On peut à présent énoncer le résultat central de cette sous-section.

Théorème A.134 (décomposition en valeurs singulières d'une matrice) *Pour toute matrice A de $M_{n,m}(\mathbb{C})$ de rang r, il existe deux matrices unitaires U (d'ordre n) et V (d'ordre m) et une matrice Σ de $M_{n,m}(\mathbb{R})$ telles que*

$$A = U\Sigma V^*,$$

avec

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix},\ D = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix},$$

*les réels σ_i, $i = 1, \ldots, r$ étant les valeurs singulières non nulles de A. Cette décomposition s'appelle la **décomposition en valeurs singulières (singular value decomposition** en anglais, abrégé en SVD) de A.*

DÉMONSTRATION. On peut supposer sans perte de généralité que $n \geq m$ (dans le cas contraire, il suffit de considérer la décomposition de A^*). Puisque la matrice A^*A est hermitienne, il existe, en vertu du théorème A.131, une matrice unitaire V d'ordre m telle que la matrice V^*A^*AV est une matrice diagonale dont les éléments diagonaux sont les valeurs propres de A^*A, ces dernières étant positives ou nulles. En effet, si λ est une valeur propre de A^*A et que $\boldsymbol{v}$ est un vecteur propre associé, *i.e.* $A\boldsymbol{v} = \lambda\,\boldsymbol{v}$, $\boldsymbol{v} \neq \boldsymbol{0}$, on a

$$\|A\boldsymbol{v}\|_2{}^2 = \boldsymbol{v}^*A^*A\boldsymbol{v} = \lambda\,\boldsymbol{v}^*\boldsymbol{v} = \lambda\,\|\boldsymbol{v}\|_2{}^2.$$

Le rang de A^*A étant égal à r, on peut numéroter les valeurs singulières de A de manière à avoir $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$, puis choisir la matrice V de manière à avoir

$$V^*A^*AV = \begin{pmatrix} {\sigma_1}^2 & & & & & \\ & \ddots & & & & \\ & & {\sigma_r}^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}.$$

En notant $\boldsymbol{v}_j$, $j = 1, \ldots, m$, les colonnes de V, on déduit de l'égalité ci-dessus que les vecteurs $A\boldsymbol{v}_j$, $j = 1, \ldots, m$, de $\mathbb{C}^n$ sont orthogonaux deux à deux et tels que $\|A\boldsymbol{v}_j\|_2 = \sigma_j$, $1 \leq j \leq r$, et $A\boldsymbol{v}_j = \boldsymbol{0}$, $r+1 \leq j \leq m$. En complétant la famille de vecteurs normalisés $\boldsymbol{u}_j = {\sigma_j}^{-1} A\boldsymbol{v}_j$, $1 \leq j \leq r$, pour former une base orthonormée de $\mathbb{C}^n$, on obtient alors une matrice unitaire U d'ordre n ayant pour colonnes les vecteurs $\boldsymbol{u}_j$, $1 \leq j \leq n$, de cette base, qui est telle que $AV = U\Sigma$ par construction. A FINIR □

quelques remarques : Dans le cas réel, les matrices U et V sont orthogonales, pas d'unicité de la décomposition, caractérisation des valeurs singulières d'une matrice normale, interprétation géométrique (image par la matrice A de la sphère unité dans l'orthogonal du noyau de A est un ellipsoïde dans l'image de A, dont les axes...) et applications

La décomposition en valeurs singulières permet le calcul effectif du *pseudo-inverse* (*pseudoinverse* ou *generalized inverse* en anglais) *de Moore* [13]*–Penrose* [14] [Moo20, Bje51, Pen55] d'une matrice de rang non nul qui est un objet généralisant la notion d'inverse à des matrices rectangulaires, ou carrées mais non inversibles [15]. On pose pour cela, pour toute matrice A de $M_{n,m}(\mathbb{C})$,

$$A^\dagger = U\Sigma^\dagger V^*, \text{ avec } \Sigma^\dagger = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

A.4 Normes et produits scalaires

La notion de norme est particulièrement utile en algèbre linéaire numérique pour quantifier l'erreur de l'approximation de la solution d'un système linéaire par une méthode itérative (voir le chapitre 3), auquel cas on fait appel à une norme dite *vectorielle* sur $\mathbb{C}^n$ (ou $\mathbb{R}^n$), ou bien effectuer des analyses d'erreur *a priori* des méthodes directes de résolution de systèmes linéaires (voir le chapitre 2), qui utilisent des normes dites *matricielles* définies sur $M_n(\mathbb{C})$ (ou $M_n(\mathbb{R})$).

A.4.1 Définitions générales

Nous rappelons dans cette section plusieurs définitions et propriétés à caractère général relatives aux notions de norme (*norm* en anglais) et de produit scalaire (*scalar product* en anglais) sur un espace vectoriel.

Définition A.135 (norme) *Soit E un espace vectoriel sur le corps $\mathbb{K}$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. On dit qu'une application $\|\cdot\|$ de E dans $\mathbb{R}$ est une* ***norme*** *sur E si elle vérifie*

- *une propriété de* ***séparation****, $\|\boldsymbol{v}\| = 0$ si et seulement si $\boldsymbol{v} = \boldsymbol{0}$,*
- *une propriété d'****homogénéité*** *(****homogeneity*** *en anglais), $\forall \alpha \in \mathbb{K}$, $\forall \boldsymbol{v} \in E$, $\|\alpha\, \boldsymbol{v}\| = |\alpha|\, \|\boldsymbol{v}\|$,*
- *une propriété de* ***sous-additivité*** *(****subadditivity*** *en anglais), encore appelée* ***inégalité triangulaire*** *(****triangle inequality*** *en anglais), c'est-à-dire*

$$\forall \boldsymbol{u} \in E,\ \forall \boldsymbol{v} \in E,\ \|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|.$$

On appelle *espace vectoriel normé* (*normed vector space* en anglais) un espace vectoriel muni d'une norme. C'est un cas particulier d'*espace métrique* (*metric space* en anglais) dans lequel la distance entre deux éléments est donnée par

$$\forall \boldsymbol{u} \in E,\ \forall \boldsymbol{v} \in E,\ d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|.$$

13. Eliakim Hastings Moore (26 janvier 1862 - 30 décembre 1932) était un mathématicien américain. Il travailla en algèbre, en géométrie algébrique, en théorie des nombres ainsi qu'en théorie des équations intégrales, et s'intéressa aux fondements de la géométrie et de l'analyse. Ses contributions à l'axiomatisation sont considérées comme l'un des points de départ des métamathématiques et de la théorie des modèles.

14. Sir Roger Penrose (né le 8 août 1931) est un physicien et mathématicien britannique. Il est connu pour ses travaux en physique mathématique et plus particulièrement ses contributions à la théorie de la relativité générale appliquée à la cosmologie et à l'étude des trous noirs.

15. Étant donné une matrice A, à coefficients réels ou complexes, possédant n lignes et m colonnes, son pseudo-inverse est l'unique matrice $A^\dagger$ à m lignes et n colonnes satisfaisant les conditions suivantes :

1. $AA^\dagger A = A$,
2. $A^\dagger A A^\dagger = A^\dagger$,
3. $(AA^\dagger)^* = AA^\dagger$,
4. $(A^\dagger A)^* = A^\dagger A$.

Définition A.136 (normes équivalentes) *Soit E un espace vectoriel sur le corps $\mathbb{K}$, avec $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$. On dit que deux normes $\|\cdot\|_*$ et $\|\cdot\|_{**}$ sur E sont* ***équivalentes*** *(**equivalent** en anglais) lorsqu'il existe deux constantes strictement positives c et C telles que*

$$\forall \boldsymbol{v}\in E,\ c\,\|\boldsymbol{v}\|_* \leq \|\boldsymbol{v}\|_{**} \leq C\,\|\boldsymbol{v}\|_*.$$

Le résultat fondamental concernant l'équivalence des normes est le suivant.

Théorème A.137 *Soit E un espace vectoriel sur le corps $\mathbb{K}$, avec $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$. Si l'espace E est de dimension finie, alors toutes les normes sur E sont équivalentes. S'il est de dimension infinie, alors il existe des normes qui ne sont pas équivalentes.*

Définition A.138 (produit scalaire) *Un* ***produit scalaire*** *(resp. **produit scalaire hermitien**) sur un espace vectoriel E sur $\mathbb{R}$ (resp. $\mathbb{C}$) est une application de $E\times E$ dans $\mathbb{R}$ (resp. $\mathbb{C}$) notée $(\cdot\,,\cdot)$ possédant les propriétés suivantes :*

- *elle est* ***bilinéaire*** *(resp. **sesquilinéaire**), c'est-à-dire linéaire par rapport à la première variable*

$$\forall \alpha\in\mathbb{R}\ \text{(resp. }\mathbb{C}),\ \forall(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\in E^3,\ (\alpha\,\boldsymbol{u}+\boldsymbol{v},\boldsymbol{w})=\alpha\,(\boldsymbol{u},\boldsymbol{w})+(\boldsymbol{u},\boldsymbol{v}),$$

  et linéaire (resp. antilinéaire) par rapport à la seconde

$$\forall \alpha\in\mathbb{R}\ \text{(resp. }\mathbb{C}),\ \forall(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\in E^3,$$
$$(\boldsymbol{u},\alpha\,\boldsymbol{v}+\boldsymbol{w})=\alpha\,(\boldsymbol{u},\boldsymbol{v})+(\boldsymbol{u},\boldsymbol{w})\ \text{(resp. }(\boldsymbol{u},\alpha\,\boldsymbol{v}+\boldsymbol{w})=\overline{\alpha}\,(\boldsymbol{u},\boldsymbol{v})+(\boldsymbol{u},\boldsymbol{w})),$$

- *elle est* ***symétrique*** *(resp. **à symétrie hermitienne**), c'est-à-dire*

$$\forall(\boldsymbol{u},\boldsymbol{v})\in E^2,\ (\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{v},\boldsymbol{u})\ \text{(resp. }(\boldsymbol{u},\boldsymbol{v})=\overline{(\boldsymbol{v},\boldsymbol{u})}),$$

- *elle est* ***définie positive**, c'est-à-dire*

$$\forall \boldsymbol{v}\in E,\ (\boldsymbol{v},\boldsymbol{v})\geq 0,\ \text{et } (\boldsymbol{v},\boldsymbol{v})=0 \text{ si et seulement si } \boldsymbol{v}=\boldsymbol{0}.$$

Définitions A.139 (espace préhilbertien et espace euclidien) *On appelle* ***espace préhilbertien*** *(**pre-Hilbert space** en anglais) tout espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni d'un produit scalaire. Si l'espace vectoriel est réel et de dimension finie, on parle d'***espace euclidien*** (**Euclidean space** en anglais).*

À tout produit scalaire, on peut associer une norme comme le montre le résultat suivant.

Théorème A.140 *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ et $(\cdot\,,\cdot)$ un produit scalaire sur E. L'application de E dans $\mathbb{R}$ définie par*

$$\forall \boldsymbol{v}\in E,\ \|\boldsymbol{v}\|=\sqrt{(\boldsymbol{v},\boldsymbol{v})}$$

est une norme sur E, appelée ***norme induite*** *par le produit scalaire $(\cdot\,,\cdot)$.*

DÉMONSTRATION. Il s'agit de montrer que l'application ainsi définie possède toutes les propriétés d'une norme énoncées dans la définition A.135. La seule de ses propriétés non évidente est l'inégalité triangulaire, que l'on va ici démontrer dans le cas complexe, le cas réel s'en déduisant trivialement. Pour tous vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ de E, on a

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2=\|\boldsymbol{u}\|^2+(\boldsymbol{u},\boldsymbol{v})+(\boldsymbol{v},\boldsymbol{u})+\|\boldsymbol{v}\|^2=\|\boldsymbol{u}\|^2+(\boldsymbol{u},\boldsymbol{v})+\overline{(\boldsymbol{u},\boldsymbol{v})}+\|\boldsymbol{v}\|^2=\|\boldsymbol{u}\|^2+2\,\mathrm{Re}((\boldsymbol{u},\boldsymbol{v}))+\|\boldsymbol{v}\|^2.$$

En vertu de l'inégalité de Cauchy–Schwarz (voir le lemme A.141), on obtient alors

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2\leq\|\boldsymbol{u}\|^2+2\,\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|+\|\boldsymbol{v}\|^2=(\|\boldsymbol{u}\|+\|\boldsymbol{v}\|)^2.$$

□

Lemme A.141 (« inégalité de Cauchy–(Bunyakovskii [16] –)Schwarz [17] » [Cau21, Bou59, Sch88]) *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni du produit scalaire $(\cdot\,,\cdot)$. On a*

$$\forall \boldsymbol{u}\in E,\ \forall \boldsymbol{v}\in E,\ |(\boldsymbol{u},\boldsymbol{v})|\leq\|\boldsymbol{u}\|\|\boldsymbol{v}\|,$$

où l'on a noté $\|\cdot\|$ la norme induite par le produit scalaire, avec égalité si et seulement si les vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ sont linéairement dépendants.

16. Viktor Yakovlevich Bunyakovskii (Виктор Яковлевич Буняковский en russe, 16 décembre 1804 - 12 décembre 1889) était un mathématicien russe qui travailla principalement en géométrie, en théorie des nombres et en mécanique théorique. Il est connu pour sa publication en 1859 de l'inégalité de Cauchy–Schwarz sous forme fonctionnelle, soit près de trente ans avant sa redécouverte par Schwarz.

17. Karl Hermann Amandus Schwarz (25 janvier 1843 - 30 novembre 1921) était un mathématicien allemand. Ses travaux, sur des sujets allant de la théorie des fonctions à la géométrie différentielle en passant par le calcul des variations, furent marqués par une forte interaction entre l'analyse et la géométrie.

DÉMONSTRATION. Soit $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de E. La démonstration du résultat dans le cas complexe pouvant se ramener au cas réel en multipliant $\boldsymbol{u}$ par un scalaire de la forme $e^{i\theta}$, avec θ un réel, de manière à ce que le produit $(e^{i\theta}\boldsymbol{u}, \boldsymbol{v})$ soit réel, il suffit d'établir l'inégalité dans le cas réel.

On considère pour cela l'application qui à tout réel t associe $\|\boldsymbol{u} - t\boldsymbol{v}\|$. On a, par propriétés du produit scalaire,

$$\forall t \in \mathbb{R},\ 0 \leq \|\boldsymbol{u} - t\,\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + 2t\,(\boldsymbol{u}, \boldsymbol{v}) + t^2\|\boldsymbol{v})\|^2.$$

Le polynôme ci-dessus étant du second ordre et positif sur $\mathbb{R}$, son discriminant doit être négatif, c'est-à-dire

$$4\,|(\boldsymbol{u}, \boldsymbol{v})|^2 \leq 4\,\|\boldsymbol{u}\|^2\|\boldsymbol{v}\|^2,$$

d'où l'inégalité annoncée. En outre, on a égalité lorsque le discriminant est nul, ce qui signifie que le polynôme possède alors une unique racine réelle et qu'il existe un réel λ tel que $\boldsymbol{u} = \lambda\,\boldsymbol{v}$. □

Proposition A.142 (« identité du parallélogramme [18] ») *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$, muni de la norme $\|\cdot\|$ induite par le produit scalaire $(\cdot\,,\cdot)$. On a l'égalité suivante :*

$$\forall \boldsymbol{u} \in E,\ \forall \boldsymbol{v} \in E,\ \|\boldsymbol{u} + \boldsymbol{v}\|^2 + \|\boldsymbol{u} - \boldsymbol{v}\|^2 = 2\left(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2\right). \tag{A.5}$$

DÉMONSTRATION. Pour tous vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ de E, développons les quantités $\|\boldsymbol{u} + \boldsymbol{v}\|^2$ et $\|\boldsymbol{u} - \boldsymbol{v}\|^2$. Par bilinéarité (ou sesquilinéarité) du produit scalaire, il vient

$$\|\boldsymbol{u} + \boldsymbol{v}\|^2 = (\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{u} + \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{u}) + (\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{v}, \boldsymbol{u}) + (\boldsymbol{v}, \boldsymbol{v}) = \|\boldsymbol{u}\|^2 + (\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{v}, \boldsymbol{u}) + \|\boldsymbol{v}\|^2,$$

et

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 = (\boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{u}) - (\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{v}, \boldsymbol{u}) + (\boldsymbol{v}, \boldsymbol{v}) = \|\boldsymbol{u}\|^2 - (\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{v}, \boldsymbol{u}) + \|\boldsymbol{v}\|^2.$$

Il suffit alors de sommer ces deux égalités pour obtenir le résultat. □

L'identité du parallélogramme (*parallelogram law* en anglais) fournit une condition nécessaire et suffisante d'existence d'un produit scalaire (éventuellement hermitien) dont dérive la norme considérée. On a en effet le résultat suivant.

Théorème A.143 (« théorème de Fréchet–Jordan–von Neumann » [Fré35, JvN35]) *Un espace vectoriel normé $(E, \|\cdot\|)$ est un espace préhilbertien si et seulement si la norme $\|\cdot\|$ satisfait l'identité du parallélogramme* (A.5).

Définition A.144 *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni d'un produit scalaire $(\cdot\,,\cdot)$. On dit qu'une famille finie de vecteurs $\{\boldsymbol{u}_i\}_{i=1,\ldots,m}$, $2 \leq m \leq n$, de E est dite* ***orthonormale*** *(pour $(\cdot\,,\cdot)$) si ses éléments vérifient*

$$(\boldsymbol{u}_i, \boldsymbol{u}_j) = \delta_{ij},\ 1 \leq i, j \leq m.$$

A.4.2 Produits scalaires et normes vectoriels

Nous nous intéressons maintenant aux produits scalaires et normes définis sur l'espace vectoriel de dimension finie $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, $n \in \mathbb{N}^*$.

L'application $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ définie par

$$(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{v}^\top \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{v} = \sum_{i=1}^{n} u_i v_i \text{ si } \mathbb{K} = \mathbb{R},$$

$$(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{v}^* \boldsymbol{u} = \overline{\boldsymbol{u}^* \boldsymbol{v}} = \sum_{i=1}^{n} u_i \overline{v_i} \text{ si } \mathbb{K} = \mathbb{C},$$

est appelée *produit scalaire canonique* (et *produit scalaire euclidien* lorsque $\mathbb{K} = \mathbb{R}$). La norme induite par ce produit scalaire, appelée *norme euclidienne* dans le cas réel ou plus généralement *norme* 2, est

$$\|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^* \boldsymbol{v}} = \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2}.$$

18. Le nom de cette égalité provient du fait qu'elle traduit que la somme des carrés des longueurs des côtés d'un parallélogramme est égale à la somme des carrés des longueurs de ses diagonales.

On rappelle que les matrices orthogonales (resp. unitaires) préservent le produit scalaire canonique sur $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) et donc sa norme induite. On a en effet, pour toute matrice orthogonale (resp. unitaire) U,

$$(U\boldsymbol{u}, U\boldsymbol{v}) = (U^\top U\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{v}) \text{ (resp. } U\boldsymbol{u}, U\boldsymbol{v}) = (U^* U\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, \boldsymbol{v})),\ \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n \text{ (resp. } \mathbb{C}^n).$$

D'autres normes couramment utilisées en analyse numérique sont la *norme* 1 (encore appelée *taxicab norm* ou *Manhattan norm* en anglais)

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|,$$

et la *norme « ∞ »*

$$\|\boldsymbol{v}\|_\infty = \max_{1\le i\le n} |v_i|.$$

Plus généralement, on peut définir une *norme p* pour tout entier naturel p non nul et on a le résultat suivant.

Théorème A.145 *Pour tout nombre $1 \le p \le +\infty$, l'application $\|\cdot\|_p$ définie sur $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, par*

$$\forall \boldsymbol{v} \in \mathbb{K}^n,\ \|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p},\ 1 \le p < +\infty,\ \|\boldsymbol{v}\|_\infty = \max_{1\le i\le n} |v_i|,$$

est une norme appelée ***norme de Hölder*** [19].

DÉMONSTRATION. Pour $p = 1$ ou $p = +\infty$, la preuve est immédiate. On va donc considérer que p est strictement compris entre 1 et $+\infty$. Dans ce cas, on désigne par q le nombre réel tel que

$$\frac{1}{p} + \frac{1}{q} = 1.$$

On va maintenant établir que, si α et β sont positifs, alors on a

$$\alpha\beta \le \frac{\alpha^p}{p} + \frac{\beta^q}{q}.$$

Le cas $\alpha\beta = 0$ étant trivial, on suppose que $\alpha > 0$ et $\beta > 0$. On a alors

$$\alpha\beta = e^{\left(\frac{1}{p}(p\,\ln(\alpha)) + \frac{1}{q}(q\,\ln(\beta))\right)} \le \frac{1}{p}\, e^{p\,\ln(\alpha)} + \frac{1}{q}\, e^{q\,\ln(\beta)} = \frac{\alpha^p}{p} + \frac{\beta^q}{q},$$

par convexité de l'exponentielle. Soit $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de $\mathbb{K}^n$. D'après l'inégalité ci-dessus, on a

$$\frac{|u_i \overline{v_i}|}{\|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q} \le \frac{1}{p}\frac{|u_i|^p}{\|\boldsymbol{u}\|_p} + \frac{1}{q}\frac{|v_i|^q}{\|\boldsymbol{v}\|_q},\ 1 \le i \le n,$$

d'où, par sommation,

$$\sum_{i=1}^{n} |u_i \overline{v_i}| \le \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q. \tag{A.6}$$

Pour établir que l'application $\|\cdot\|_p$ est une norme, il suffit à présent de prouver qu'elle vérifie l'inégalité triangulaire, les autres propriétés étant évidentes. Pour cela, on écrit que

$$(|u_i| + |v_i|)^p = |u_i|\,(|u_i| + |v_i|)^{p-1} + |v_i|\,(|u_i| + |v_i|)^{p-1},\ 1 \le i \le n.$$

En sommant et en utilisant l'inégalité précédemment établie, on obtient

$$\sum_{i=1}^{n} (|u_i| + |v_i|)^p \le \left(\|\boldsymbol{u}\|_p + \|\boldsymbol{v}\|_p\right)\left(\sum_{i=1}^{n} (|u_i| + |v_i|)^{(p-1)q}\right)^{1/q}.$$

19. Otto Ludwig Hölder (22 décembre 1859 - 29 août 1937) était un mathématicien allemand. Il est connu pour plusieurs découvertes aujourd'hui associées à son nom, au nombre desquelles l'*inégalité de Hölder*, qui est fondamentale à l'étude des espaces de fonctions L^p, le *théorème de Hölder*, qui implique que la fonction gamma ne satisfait aucune équation différentielle algébrique dont les coefficients sont des fonctions rationnelles, ou encore la *condition de Hölder*, qui est une condition suffisante pour qu'une application définie entre deux espaces métriques soit uniformément continue.

L'inégalité triangulaire découle alors de la relation $(p-1)q = p$. □

On déduit de (A.6) l'inégalité suivante

$$\forall (p,q) \in [1,+\infty]^2, \text{ avec } \frac{1}{p}+\frac{1}{q}=1, \ \forall (\boldsymbol{u},\boldsymbol{v}) \in (\mathbb{K}^n)^2, \ |(\boldsymbol{u},\boldsymbol{v})| \leq \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q,$$

appelée *inégalité de Hölder*, tandis que l'inégalité triangulaire

$$\forall (\boldsymbol{u},\boldsymbol{v}) \in (\mathbb{K}^n)^2, \ \|\boldsymbol{u}+\boldsymbol{v}\|_p \leq \|\boldsymbol{u}\|_p + \|\boldsymbol{v}\|_p,$$

porte le nom d'*inégalité de Minkowski* [20].

On rappelle enfin que dans un espace vectoriel de dimension finie sur un corps complet (comme $\mathbb{R}$ ou $\mathbb{C}$) toutes les normes sont équivalentes. Sur $\mathbb{R}^n$ ou $\mathbb{C}^n$, on a ainsi, pour $0 < p < q$ et tout vecteur $\boldsymbol{v}$,

$$\|\boldsymbol{v}\|_q \leq \|\boldsymbol{v}\|_p \leq n^{\frac{1}{p}-\frac{1}{q}} \|\boldsymbol{v}\|_q,$$

par utilisation de l'inégalité de Hölder, les constantes étant optimales.

Nous aurons besoin des définitions suivantes dans la suite.

Définition A.146 (norme duale d'une norme vectorielle) *Soit $\|\cdot\|$ une norme définie sur $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. La fonction définie par*

$$\forall \boldsymbol{v} \in \mathbb{K}^n, \ \|\boldsymbol{v}\|_D = \sup_{\|\boldsymbol{u}\|=1} \mathrm{Re}(\boldsymbol{v}^*\boldsymbol{u}) = \sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{v}^*\boldsymbol{u}|,$$

est appelée la ***norme duale de*** $\|\cdot\|$.

On note que cette fonction est bien définie : l'ensemble $\{\boldsymbol{u} \in \mathbb{K}^n \,|\, \|\boldsymbol{u}\| = 1\}$ étant compact et l'application $\boldsymbol{v} \mapsto |\boldsymbol{v}^*\boldsymbol{u}|$ étant continue sur $\mathbb{K}^n$ pour tout vecteur $\boldsymbol{u}$ de $\mathbb{K}^n$ fixé, le borne supérieure de la définition est atteinte en vertu d'une généralisation du théorème des bornes (voir le théorème B.88). On observe par ailleurs que la norme duale est bien une norme. Les propriétés de positivité et d'homogénéité sont en effet évidentes. Pour montrer celle de séparation, on utilise l'homogénéité pour écrire que, pour tout vecteur $\boldsymbol{v}$ non nul,

$$\|\boldsymbol{v}\|_D = \sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{v}^*\boldsymbol{u}| \geq \left| \boldsymbol{v}^* \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \right| = \frac{\|\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|} > 0.$$

L'obtention de l'inégalité triangulaire est tout aussi immédiate. Pour tous vecteurs $\boldsymbol{v}$ et $\boldsymbol{w}$ de $\mathbb{K}^n$, on a

$$\|\boldsymbol{v}+\boldsymbol{w}\|_D = \sup_{\|\boldsymbol{u}\|=1} |(\boldsymbol{v}+\boldsymbol{w})^*\boldsymbol{u}| \leq \sup_{\|\boldsymbol{u}\|=1} (|\boldsymbol{v}^*\boldsymbol{u}| + |\boldsymbol{w}^*\boldsymbol{u}|) \leq \sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{v}^*\boldsymbol{u}| + \sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{w}^*\boldsymbol{u}| = \|\boldsymbol{v}\|_D + \|\boldsymbol{w}\|_D.$$

Définition A.147 (dual d'un vecteur) *Soit $\|\cdot\|$ une norme définie sur $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, et $\boldsymbol{u}$ un vecteur non nul de $\mathbb{K}^n$. L'ensemble*

$$\{\boldsymbol{v} \in \mathbb{K}^n \,|\, \|\boldsymbol{v}\|_D \|\boldsymbol{u}\| = |\boldsymbol{v}^*\boldsymbol{u}| = 1\},$$

est le ***dual du vecteur $\boldsymbol{u}$ par rapport à la norme*** $\|\cdot\|$.

Il découle du *théorème de dualité* (voir le théorème 5.5.9 dans [HJ13]) que, pour toute norme vectorielle, tout vecteur non nul possède un dual non vide, qui peut être constitué d'un ou de plusieurs vecteurs.

La notion de produit scalaire sur $\mathbb{R}^n$ et $\mathbb{C}^n$ étant introduite, nous pouvons maintenant considérer celle de matrice *symétrique définie positive*, dont les propriétés sont particulièrement intéressantes pour les méthodes de résolution de systèmes linéaires étudiées dans les chapitres 2 et 3.

Définition A.148 (matrice définie positive) *Une matrice d'ordre n est dite* ***définie positive*** *sur $\mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, si*

$$\forall \boldsymbol{v} \in \mathbb{K}^n, \ (A\boldsymbol{v},\boldsymbol{v}) \geq 0,$$

avec $(A\boldsymbol{v},\boldsymbol{v}) = 0$ si et seulement si $\boldsymbol{v} = \boldsymbol{0}$.

20. Hermann Minkowski (22 juin 1864 - 12 janvier 1909) était un mathématicien et physicien théoricien allemand. Il créa la *géométrie des nombres* pour résoudre des problèmes difficiles en théorie des nombres et ses travaux sur la notion d'un continuum espace-temps à quatre dimensions furent à la base de la théorie de la relativité générale.

On voit immédiatement qu'une matrice est définie positive si elle est associée à une forme bilinéaire (ou sesquilinéaire) définie positive.

On notera que les matrices définies positives sur $\mathbb{R}^n$ ne sont pas nécessairement symétriques. On peut cependant prouver qu'une matrice réelle A est définie positive sur $\mathbb{R}^n$ si et seulement si sa *partie symétrique*, qui est la matrice $\frac{1}{2}(A+A^\top)$, est définie positive sur $\mathbb{R}^n$. Plus généralement, les résultats suivants montrent qu'une matrice définie positive à coefficients complexes est nécessairement hermitienne, ce qui nous amène à ne considérer dans la suite que des matrices définies positives symétriques ou hermitiennes.

Proposition A.149 *Soit A une matrice de $M_n(\mathbb{C})$ (resp. $M_n(\mathbb{R})$). Si, pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, la quantité $(A\boldsymbol{v},\boldsymbol{v})$ est réelle, alors A est une matrice hermitienne (resp. symétrique).*

DÉMONSTRATION. Si la quantité $(A\boldsymbol{v},\boldsymbol{v})$ est réelle pour tout vecteur de $\mathbb{C}^n$, alors $(A\boldsymbol{v},\boldsymbol{v}) = \overline{(A\boldsymbol{v},\boldsymbol{v})}$, c'est-à-dire

$$\sum_{i=1}^n\sum_{j=1}^n \overline{a_{ij}v_j}v_i = \sum_{i=1}^n\sum_{j=1}^n \overline{\overline{a_{ij}v_j}v_i} = \sum_{i=1}^n\sum_{j=1}^n a_{ij}v_j\overline{v_i} = \sum_{i=1}^n\sum_{j=1}^n a_{ji}v_i\overline{v_j},$$

ce qui implique

$$\forall \boldsymbol{v}\in\mathbb{C}^n,\ \sum_{i=1}^n\sum_{j=1}^n(\overline{a_{ij}}-a_{ji})v_i\overline{v_j} = 0.$$

Par des choix appropriés du vecteur $\boldsymbol{v}$, on en déduit que $\overline{a_{ij}} = a_{ji}$, pour tous i, j dans $\{1,\dots,n\}$. □

Proposition A.150 *Une matrice d'ordre n est définie positive sur $\mathbb{C}^n$ si et seulement si elle est hermitienne et ses valeurs propres sont strictement positives.*

DÉMONSTRATION. Soit A une matrice définie positive. On sait alors, d'après la précédente proposition, qu'elle est hermitienne et il existe donc une matrice unitaire U telle que la matrice U^*AU est diagonale, avec pour coefficients diagonaux les valeurs propres λ_i, $i=1,\dots,n$, de A. En posant $\boldsymbol{v} = U\boldsymbol{w}$ pour tout vecteur de $\mathbb{C}^n$, on obtient

$$(A\boldsymbol{v},\boldsymbol{v}) = (AU\boldsymbol{w},U\boldsymbol{w}) = (U^*AU\boldsymbol{w},\boldsymbol{w}) = \sum_{i=1}^n \overline{\lambda_i}\,|w_i|^2 = \sum_{i=1}^n \lambda_i\,|w_i|^2\,.$$

En choisissant successivement $\boldsymbol{w} = \boldsymbol{e}_i$, avec $i=1,\dots,n$, on trouve que $0 < (A\boldsymbol{e}_i,\boldsymbol{e}_i) = \lambda_i$. La réciproque est immédiate, puisque si la matrice A est hermitienne, alors il existe une base orthonormée de $\mathbb{C}^n$ formée de ses vecteurs propres. □

On a en particulier qu'une matrice définie positive est inversible.

Le résultat classique suivant fournit une caractérisation utile des matrices symétriques (ou hermitienne) définies positives.

Théorème A.151 (« critère de Sylvester ») *Une matrice A symétrique ou hermitienne d'ordre n est définie positive si et seulement tous ses mineurs principaux dominants sont strictement positifs, c'est-à-dire si toutes les sous-matrices principales*

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix},\ 1\le k\le n,$$

extraites de A ont un déterminant strictement positif.

DÉMONSTRATION. On démontre le théorème dans le cas réel, l'extension au cas complexe ne posant aucune difficulté, par récurrence sur l'ordre n de la matrice. Dans toute la preuve, la notation $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ désigne le produit scalaire euclidien sur $\mathbb{R}^n$.

Pour $n=1$, la matrice A est un nombre réel, $A=(a_{11})$, et, pour tout réel x, $\langle Ax,x\rangle_{\mathbb{R}} = a_{11}\,x^2$ est par conséquent positif si et seulement si $a_{11}>0$, a_{11} étant par ailleurs le seul mineur principal.
Supposons maintenant le résultat vrai pour des matrices symétriques d'ordre $n-1$, $n\ge 2$, et prouvons-le pour celles d'ordre n. Soit A une telle matrice. On note respectivement λ_i et $\boldsymbol{v}_i$, $1\le i\le n$ les valeurs et vecteurs propres de A, l'ensemble $\{\boldsymbol{v}_i\}_{1\le i\le n}$ formant par ailleurs une base orthonormée de $\mathbb{R}^n$.

Puisque $\langle A\boldsymbol{x},\boldsymbol{x}\rangle_{\mathbb{R}^n} > 0$ pour tout vecteur $\boldsymbol{x}$ non nul de $\mathbb{R}^n$, ceci est donc en particulier vrai pour tous les vecteurs de la forme

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}.$$

Observons alors que

$$\left\langle A\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix} \right\rangle_{\mathbb{R}^n} = \left\langle A_{n-1}\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \right\rangle_{\mathbb{R}^{n-1}}$$

Par conséquent, la matrice A_{n-1} est définie positive et tous ses mineurs principaux dominants, qui ne sont autres que les $n-1$ mineurs principaux dominants de A, sont strictement positifs. Le fait que A soit définie positive impliquant que ses valeurs propres sont strictement positives, on a que $\det(A) = \prod_{i=1}^{n} \lambda_i > 0$ et l'on vient donc de montrer le sens direct de l'équivalence.

Réciproquement, si tous les mineurs principaux dominants de A sont strictement positifs, on applique l'hypothèse de récurrence pour en déduire que la sous-matrice A_{n-1} est définie positive. Comme $\det(A) > 0$, on a l'alternative suivante : soit toutes les valeurs propres de A sont strictement positives (et donc A est définie positive), soit au moins deux d'entre elles, λ_i et λ_j, sont strictement négatives. Dans ce dernier cas, il existe au moins une combinaison linéaire $\alpha \boldsymbol{v}_i + \beta \boldsymbol{v}_j$, avec α et β tous deux non nuls, ayant zéro pour dernière composante. Puisqu'on a démontré que A_{n-1} était définie positive, il s'ensuit que $\left\langle A(\alpha \boldsymbol{v}_i + \beta \boldsymbol{v}_j), \alpha \boldsymbol{v}_i + \beta \boldsymbol{v}_j \right\rangle_{\mathbb{R}^n} > 0$. Mais, on a par ailleurs

$$\left\langle A(\alpha \boldsymbol{v}_i + \beta \boldsymbol{v}_j), \alpha \boldsymbol{v}_i + \beta \boldsymbol{v}_j \right\rangle_{\mathbb{R}^n} = \alpha^2 \lambda_i + \beta^2 \lambda_j < 0,$$

d'où une contradiction. □

A.4.3 Normes de matrices *

Nous introduisons dans cette section des normes sur les espaces de matrices. En plus des propriétés habituelles d'une norme, on demande généralement qu'une norme de matrices satisfasse à une propriété de *sous-multiplicativité* qui la rend intéressante en pratique [21]. On parle dans ce cas de norme *matricielle*.

Dans toute la suite, on ne va considérer que des matrices à coefficients complexes, mais les résultats s'appliquent aussi bien à des matrices à coefficients réels, en remplaçant le cas échéant les mots « complexe », « hermitien » et « unitaire » par « réel », « symétrique » et « orthogonale », respectivement.

Définition A.152 (normes compatibles) *Soit m et n deux entiers naturels non nuls. On dit que trois normes, toutes notées $\|\cdot\|$ et respectivement définies sur $\mathbb{C}^n$, $M_{n,m}(\mathbb{C})$ et $\mathbb{C}^m$, sont* ***compatibles*** *si*

$$\forall A \in M_{n,m}(\mathbb{C}),\ \forall \boldsymbol{v} \in \mathbb{C}^m,\ \|A\boldsymbol{v}\| \leq \|A\|\,\|\boldsymbol{v}\|.$$

Définition A.153 (norme consistante) *On dit qu'une norme $\|\cdot\|$, définie sur $M_{n,m}(\mathbb{C})$ pour toutes valeurs de m et n dans $\mathbb{N}^*$, est* ***consistante*** *si elle vérifie la propriété de* ***sous-multiplicativité*** *(**submultiplicativity** en anglais)*

$$\|AB\| \leq \|A\|\,\|B\| \tag{A.7}$$

dès que le produit de matrices AB a un sens.

Définition A.154 (norme matricielle) *Une* ***norme matricielle*** *(**matrix norm** en anglais) est une application de $M_{n,m}(\mathbb{C})$ dans $\mathbb{R}$, définie pour toutes valeurs de m et n dans $\mathbb{N}^*$, vérifiant les propriétés d'une norme (voir la définition A.135) et la propriété de sous-multiplicativité* (A.7).

Il est important de remarquer que toutes les normes ne sont pas des normes matricielles comme le montre l'exemple suivant, tiré de [GV13].

Exemple de norme de matrice non consistante. La norme $\|\cdot\|_{\max}$, définie sur $M_{n,m}(\mathbb{C})$ par

$$\forall A \in M_{n,m}(\mathbb{C}),\ \|A\|_{\max} = \max_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \left|a_{ij}\right|,$$

ne satisfait pas à la propriété de sous-multiplicativité (A.7), puisque pour la matrice

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

on a $2 = \|A^2\|_{\max} > \|A\|_{\max}^2 = 1$.

21. Sur $M_n(\mathbb{K})$, une telle norme est alors une *norme d'algèbre*.

Il existe toujours une norme vectorielle avec laquelle une norme matricielle donnée est compatible. En effet, étant donnée une norme matricielle $\|\cdot\|$ sur $M_{n,m}(\mathbb{C})$ dans $\mathbb{R}$, définie pour toutes valeurs de m et n dans $\mathbb{N}^*$, et un vecteur $\boldsymbol{u}$ de $M_{m,1}(\mathbb{C})$ non nul, il suffit de définir une telle norme vectorielle par

$$\forall \boldsymbol{w} \in \mathbb{C}^n,\ \|\boldsymbol{w}\| = \|\boldsymbol{w}\,\boldsymbol{u}^*\|.$$

On déduit alors de la propriété (A.7) que

$$\forall A \in M_{n,m}(\mathbb{C}),\ \forall \boldsymbol{v} \in \mathbb{C}^m,\ \|A\boldsymbol{v}\| = \|A\boldsymbol{v}\,\boldsymbol{u}^*\| \leq \|A\|\|\boldsymbol{v}\,\boldsymbol{u}^*\| = \|A\|\|\boldsymbol{v}\|.$$

Exemple de norme matricielle compatible avec la norme vectorielle euclidienne. L'application définie par

$$\forall A \in M_{n,m}(\mathbb{C}),\ \|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|a_{ij}\right|^2} = \sqrt{\mathrm{tr}(AA^*)}, \tag{A.8}$$

est une norme matricielle (la démonstration est laissée en exercice), dite *norme de Frobenius*, compatible avec la norme vectorielle euclidienne $\|\cdot\|_2$, car on a

$$\|A\boldsymbol{v}\|_2{}^2 = \sum_{i=1}^{n}\left|\sum_{j=1}^{m}a_{ij}\,x_j\right|^2 \leq \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left|a_{ij}\right|^2\sum_{j=1}^{m}\left|x_j\right|^2\right) = \|A\|_F{}^2\|\boldsymbol{v}\|_2{}^2.$$

L'application trace étant un invariant de similitude, on remarque que l'on a encore

$$\|A\|_F = \sqrt{\sum_{i=1}^{\min(m,n)}\sigma_i{}^2},$$

où les réels σ_i, $1 \leq i \leq \min(m,n)$, sont les valeurs singulières de la matrice A.

Proposition A.155 (norme de matrice subordonnée) *Soit m et n deux entiers naturels non nuls. Étant donné deux normes vectorielles $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$ sur $\mathbb{C}^m$ et $\mathbb{C}^n$ respectivement, l'application $\|\cdot\|_{\alpha,\beta}$ de $M_{n,m}(\mathbb{C})$ dans $\mathbb{R}$ définie par*

$$\|A\|_{\alpha,\beta} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^m\\ \boldsymbol{v}\neq\boldsymbol{0}}}\frac{\|A\boldsymbol{v}\|_\beta}{\|\boldsymbol{v}\|_\alpha} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^m\\ \|\boldsymbol{v}\|_\alpha\leq 1}}\|A\boldsymbol{v}\|_\beta = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^m\\ \|\boldsymbol{v}\|_\alpha=1}}\|A\boldsymbol{v}\|_\beta, \tag{A.9}$$

*est une norme de matrice dite **subordonnée (subordinate matrix norm** en anglais) aux normes $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$.*

DÉMONSTRATION. On remarque tout d'abord que la quantité $\|A\|_{\alpha,\beta}$ est bien définie pour tout matrice A de $M_{n,m}(\mathbb{C})$: ceci découle de la continuité de l'application de $\mathbb{C}^m$ dans $\mathbb{R}$ qui à un vecteur $\boldsymbol{v}$ associe $\|A\boldsymbol{v}\|_\beta$ sur la sphère unité, qui est compacte car l'espace est de dimension finie. La vérification des propriétés satisfaites par une norme est alors immédiate. □

On dit encore que la norme $\|\cdot\|_{\alpha,\beta}$ est *induite* par les normes vectorielles $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$. Une norme subordonnée de matrice est un cas particulier de *norme d'opérateur*. La propriété de sous-multiplicativité (A.7) n'est généralement [22] pas vérifiée par une norme subordonnée, mais on a en revanche

$$\forall A, B \in M_n(\mathbb{C}),\ \|AB\|_{\alpha,\beta} \leq \|A\|_{\gamma,\beta}\|B\|_{\alpha,\gamma},$$

pour toute norme vectorielle $\|\cdot\|_\gamma$ sur $M_{n,1}(\mathbb{C})$ A VOIR, PREUVE ?. Une norme subordonnée $\|\cdot\|_{\alpha,\beta}$ est compatible avec les normes qui l'induisent puisqu'on a, par définition,

$$\forall A \in M_{m,n}(\mathbb{C}),\ \forall \boldsymbol{v} \in M_{n,1}(\mathbb{C}),\ \boldsymbol{v} \neq \boldsymbol{0},\ \|A\|_{\alpha,\beta} \geq \frac{\|A\boldsymbol{v}\|_\beta}{\|\boldsymbol{v}\|_\alpha}.$$

22. Par exemple, le choix $\alpha = 1$ et $\beta = \infty$ conduit à $\|A\|_{1,\infty} = \max_{1\leq i,j\leq n}\left|a_{ij}\right|$, $\forall A \in M_n(\mathbb{C})$, qui est la norme notée $\|\cdot\|_{\max}$ introduite plus haut. Pour tout matrice A d'ordre n et tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, on a en effet

$$\|A\boldsymbol{v}\|_\infty = \max_{1\leq i\leq n}\left|\sum_{j=1}^{n}a_{ij}v_j\right| \leq \left(\max_{1\leq i,j\leq n}\left|a_{ij}\right|\right)\|\boldsymbol{v}\|_1,$$

avec égalité pour le vecteur de composantes $v_i = 0$ pour $i \neq j_0$, $v_{j_0} = 1$, où j_0 est un indice vérifiant $\max_{1\leq i,j\leq n}\left|a_{ij}\right| = \max_{1\leq i\leq n}\left|a_{ij_0}\right|$.

Ceci implique de manière immédiate qu'une norme subordonnée est sous-multiplicative lorsque [23] $\alpha = \beta$. Dans toute la suite de cette annexe et dans l'ensemble du cours, nous nous restreignons à des normes subordonnées sur $M_n(\mathbb{C})$, $n \in \mathbb{N}^*$, pour lesquelles $\alpha = \beta = p$ avec $p \geq 1$, que nous noterons $\|\cdot\|_p$, ou bien encore $\|\cdot\|$ lorsqu'aucune précision n'est nécessaire.

On déduit de la définition (A.9) que $\|I_n\| = 1$ pour toute norme matricielle subordonnée sur $M_n(\mathbb{C})$. Un exemple de norme matricielle sur $M_n(\mathbb{C})$ n'étant pas subordonnée à une norme vectorielle sur $\mathbb{C}^n$ est la norme de Frobenius, puisque l'on a $\|I_n\|_F = \sqrt{n}$.

La proposition suivante donne des formules pour le calcul des normes subordonnées aux normes vectorielles $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$.

Proposition A.156 *Soit A une matrice d'ordre n. On a*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} \left|a_{ij}\right|, \tag{A.10}$$

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \|A^*\|_2, \tag{A.11}$$

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \left|a_{ij}\right|. \tag{A.12}$$

Par ailleurs, la norme $\|\cdot\|_2$ est invariante par transformation unitaire et si A est une matrice normale, alors

$$\|A\|_2 = \rho(A).$$

DÉMONSTRATION. Pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, on a

$$\|A\boldsymbol{v}\|_1 = \sum_{i=1}^{n} \left|\sum_{j=1}^{n} a_{ij} v_j\right| \leq \sum_{j=1}^{n} \left|v_j\right| \sum_{i=1}^{n} \left|a_{ij}\right| \leq \left(\max_{1\leq j\leq n} \sum_{i=1}^{n} \left|a_{ij}\right|\right) \|\boldsymbol{v}\|_1.$$

Pour montrer (A.10), on construit un vecteur $\boldsymbol{v}$ non nul (qui dépend de la matrice A) de sorte à avoir une égalité dans l'inégalité ci-dessus. Il suffit pour cela de considérer le vecteur de composantes

$$v_i = 0 \text{ pour } i \neq j_0,\ v_{j_0} = 1,$$

où j_0 est un indice vérifiant

$$\max_{1\leq j\leq n} \sum_{i=1}^{n} \left|a_{ij}\right| = \sum_{i=1}^{n} \left|a_{ij_0}\right|.$$

De la même manière, on prouve (A.12) en écrivant

$$\|A\boldsymbol{v}\|_\infty = \max_{1\leq i\leq n} \left|\sum_{j=1}^{n} a_{ij} v_j\right| \leq \left(\max_{1\leq i\leq n} \sum_{j=1}^{n} \left|a_{ij}\right|\right) \|\boldsymbol{v}\|_\infty,$$

et en choisissant le vecteur $\boldsymbol{v}$ tel que

$$v_j = \frac{\overline{a_{i_0 j}}}{\left|a_{i_0 j}\right|} \text{ si } a_{i_0 j} \neq 0,\ v_j = 1 \text{ sinon},$$

avec i_0 un indice vérifiant

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} \left|a_{ij}\right| = \sum_{j=1}^{n} \left|a_{i_0 j}\right|.$$

On prouve à présent (A.11). La matrice A^*A étant hermitienne, il existe (voir le théorème A.131) une matrice unitaire U telle que la matrice U^*A^*AU est une matrice diagonale dont les éléments sont les valeurs propres, par ailleurs positives, μ_i, $i = 1, \ldots, n$, de A^*A. En posant $\boldsymbol{w} = U^*\boldsymbol{v}$, on a alors

$$\|A\|_2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \sqrt{\frac{(A^*A\boldsymbol{v}, \boldsymbol{v})}{(\boldsymbol{v}, \boldsymbol{v})}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\frac{(U^*A^*AU\boldsymbol{w}, \boldsymbol{w})}{(\boldsymbol{w}, \boldsymbol{w})}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\sum_{i=1}^{n} \mu_i \frac{|w_i|^2}{\sum_{j=1}^{n} \left|w_j\right|^2}} = \sqrt{\max_{1\leq i\leq n} \mu_i}.$$

23. Observons que, bien que l'on ait $\alpha = \beta$, la définition (A.9) utilise deux normes vectorielles notées $\|\cdot\|_\alpha$, l'une au numérateur définie sur $\mathbb{C}^m$, l'autre au dénominateur définie sur $\mathbb{C}^n$. On a, de fait, implicitement supposé qu'on désignait par $\|\cdot\|_\alpha$ une *famille* de normes vectorielles, chacune définie sur $\mathbb{C}^d$ pour tout d dans $\mathbb{N}^*$.

D'autre part, en utilisant l'inégalité de Cauchy–Schwarz, on trouve, pour tout vecteur $\boldsymbol{v}$ non nul,

$$\frac{\|A\boldsymbol{v}\|_2^{\,2}}{\|\boldsymbol{v}\|_2^{\,2}} = \frac{(A^*A\boldsymbol{v},\boldsymbol{v})}{\|\boldsymbol{v}\|_2^{\,2}} \leq \frac{\|A^*A\boldsymbol{v}\|_2\|\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2^{\,2}} \leq \|A^*A\|_2 \leq \|A^*\|_2\|A\|_2,$$

d'où $\|A\|_2 \leq \|A^*\|_2$. En appliquant cette inégalité à A^*, on obtient l'égalité $\|A\|_2 = \|A^*\|_2 = \sqrt{\rho(AA^*)}$.

On montre ensuite l'invariance de la norme $\|\cdot\|_2$ par transformation unitaire, c'est-à-dire que $\|UA\|_2 = \|AU\|_2 = \|A\|_2$ pour toute matrice unitaire U et toute matrice A. Puisque $U^*U = I_n$, on a

$$\|UA\|_2^{\,2} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|UA\boldsymbol{v}\|_2^{\,2}}{\|\boldsymbol{v}\|_2^{\,2}} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{(U^*UA\boldsymbol{v},A\boldsymbol{v})}{\|\boldsymbol{v}\|_2^{\,2}} = \|A\|_2.$$

Le changement de variable $\boldsymbol{u} = U\boldsymbol{v}$ vérifiant $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2$, on a par ailleurs

$$\|AU\|_2^{\,2} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|AU\boldsymbol{v}\|_2^{\,2}}{\|\boldsymbol{v}\|_2^{\,2}} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2^{\,2}}{\|U^*\boldsymbol{u}\|_2^{\,2}} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2^{\,2}}{\|\boldsymbol{u}\|_2^{\,2}} = \|A\|_2.$$

Enfin, si A est une matrice normale, alors elle est diagonalisable dans une base orthonormée de vecteurs propres (voir le théorème A.131 et on a $A = UDU^*$, avec U une matrice unitaire et D une matrice diagonale ayant pour éléments les valeurs propres de A, d'où

$$\|A\|_2 = \|UDU^*\|_2 = \|D\|_2 = \rho(A).$$

□

Cette proposition amène quelques remarques. Tout d'abord, il est clair à l'examen de la démonstration ci-dessus que les expressions trouvées pour $\|A\|_1$, $\|A\|_2$ et $\|A\|_\infty$, avec A une matrice d'ordre n, sont encore valables pour une matrice rectangulaire. On observe aussi que $\|A\|_1 = \|A^*\|_\infty$ et que l'on a $\|A\|_1 = \|A\|_\infty$ et $\|A\|_2 = \rho(A)$ si A est une matrice hermitienne (donc normale). Par ailleurs, si U est une matrice unitaire (donc normale), on a $\|U\|_2 = \rho(I_n) = 1$. Enfin, la quantité $\|A\|_2$ n'est autre que la plus grande valeur singulière (voir la sous-section A.3.10) de la matrice A et son calcul pratique est donc beaucoup plus difficile et coûteux que celui de $\|A\|_1$ ou $\|A\|_\infty$. Cette propriété donne son nom à la norme $\|\cdot\|_2$, qui est dite *spectrale* (*spectral norm* en anglais). L'invariance unitaire de cette norme, également vérifiée par la norme de Frobenius a des implications importantes pour l'analyse d'erreur des méthodes numériques utilisées en algèbre linéaire, car elle signifie que la multiplication par une matrice unitaire n'amplifie pas les erreurs déjà présentes dans une matrice. Par exemple, si A est une matrice d'ordre n entachée d'une erreur E et Q est une matrice unitaire, alors $Q(A+E)q^* = QAQ^* + F$, où $\|F\|_2 = \|QEQ^*\|_2 = \|E\|_2$.

Comme toutes les normes définies sur un espace vectoriel de dimension finie sur un corps complet, les normes sur $M_{n,m}(\mathbb{C})$ sont équivalentes. Le tableau A.1 donne les constantes d'équivalence entre les normes les plus utilisées en pratique.

		q			
		1	2	∞	F
p	1	1	$\sqrt{n}$	n	$\sqrt{n}$
	2	$\sqrt{m}$	1	$\sqrt{n}$	1
	∞	m	$\sqrt{m}$	1	$\sqrt{m}$
	F	$\sqrt{m}$	$\sqrt{\text{rang}(A)}$	$\sqrt{n}$	1

TABLE A.1 – A REVOIR Constantes C_{pq} telles que, $\forall A \in M_{n,m}(\mathbb{C})$, $\|A\|_p \leq C_{pq}\|A\|_q$.

Enfin, si l'on a montré qu'il existait des normes matricielles et des matrices A pour lesquelles on a l'égalité $\|A\| = \rho(A)$, il faut insister sur le fait que le rayon spectral n'est pas une norme [24]. On peut néanmoins prouver que l'on peut toujours approcher le rayon spectral d'une matrice donnée d'aussi près que souhaité par valeurs supérieures, à l'aide d'une norme matricielle convenablement choisie. Ce résultat est fondamental pour l'étude de la convergence des suites de matrices (voir le théorème A.159).

24. Par exemple, toute matrice triangulaire non nulle dont les coefficients diagonaux sont nuls a un rayon spectral égal à zéro.

Théorème A.157 *Soit A une matrice carrée et $\|\cdot\|$ une norme matricielle. Alors, on a*

$$\rho(A) \leq \|A\|.$$

D'autre part, étant donné une matrice A et un nombre réel strictement positif ε, il existe une norme matricielle subordonnée $\|\cdot\|_\varepsilon$ telle que

$$\|A\|_\varepsilon \leq \rho(A) + \varepsilon.$$

DÉMONSTRATION. Si λ est une valeur propre de A, il existe un vecteur propre $\boldsymbol{v} \neq \boldsymbol{0}$ associé, tel que $A\boldsymbol{v} = \lambda\,\boldsymbol{v}$. Soit $\boldsymbol{w}$ un vecteur tel que la matrice $\boldsymbol{v}\boldsymbol{w}^*$ ne soit pas nulle. On a alors

$$|\lambda|\ \|\boldsymbol{v}\boldsymbol{w}^*\| = \|\lambda\,\boldsymbol{v}\boldsymbol{w}^*\| = \|A\boldsymbol{v}\boldsymbol{w}^*\| \leq \|A\|\,\|\boldsymbol{v}\boldsymbol{w}^*\|,$$

d'après la propriété de sous-multiplicativité d'une norme matricielle, et donc $|\lambda| \leq \|A\|$. Cette dernière inégalité étant vraie pour toute valeur propre de A, elle l'est en particulier quand $|\lambda|$ est égal au rayon spectral de la matrice et la première inégalité du théorème se trouve démontrée.

Supposons à présent que la matrice A est d'ordre n. Toute matrice à coefficients complexe étant trigonalisable, il existe une matrice unitaire U telle que $T = U^{-1}AU$ soit triangulaire (supérieure par exemple) et que les éléments diagonaux de T soient les valeurs propres de A. À tout réel $\delta > 0$, on définit la matrice diagonale D_δ telle que $d_{ii} = \delta^{i-1}$, $i = 1, \ldots, n$. Étant donné $\varepsilon > 0$, on peut choisir δ suffisamment petit pour que les éléments extradiagonaux de la matrice $(UD_\delta)^{-1}A(UD_\delta) = (D_\delta)^{-1}TD_\delta$ soient aussi petits, par exemple de façon à avoir

$$\sum_{j=i+1}^{n} \delta^{j-i} \left|t_{ij}\right| \leq \varepsilon,\ 1 \leq i \leq n-1.$$

On a alors

$$\|(UD_\delta)^{-1}A(UD_\delta)\|_\infty = \max_{1\leq i\leq n} \sum_{j=i}^{n} \delta^{j-i} \left|t_{ij}\right| \leq \rho(A) + \varepsilon.$$

Il reste à vérifier que l'application qui à une matrice B d'ordre n associe $\|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$ est une norme matricielle (qui dépend de A et de ε), ce qui est immédiat puisque c'est la norme subordonnée à la norme vectorielle $\|(UD_\delta)^{-1}\cdot\|_\infty$. □

On notera que la première inégalité du théorème A.157 est plus généralement vraie pour toute norme sur $M_n(\mathbb{C})$ compatible avec une norme sur $M_{n,1}(\mathbb{C})$, ce qui est trivialement le cas pour les normes subordonnées.

Théorème A.158 *Soit $\|\cdot\|$ une norme matricielle subordonnée et A une matrice d'ordre n vérifiant $\|A\| < 1$. Alors la matrice $I_n - A$ est inversible et on a les inégalités*

$$\frac{1}{1+\|A\|} \leq \left\|(I_n - A)^{-1}\right\| \leq \frac{1}{1-\|A\|}.$$

Par ailleurs, si une matrice de la forme $I_n - A$ est singulière, alors on a nécessairement $\|A\| \geq 1$ pour toute norme matricielle $\|\cdot\|$.

DÉMONSTRATION. On remarque que $(I_n - A)\boldsymbol{v} = \boldsymbol{0}$ implique que $\|A\boldsymbol{v}\| = \|\boldsymbol{v}\|$. D'autre part, puisque $\|A\| < 1$, on a, si $\boldsymbol{v} \neq \boldsymbol{0}$ et par définition d'une norme matricielle subordonnée, $\|A\boldsymbol{v}\| < \|\boldsymbol{v}\|$. On en déduit que, si $(I_n - A)\boldsymbol{v} = \boldsymbol{0}$, alors $\boldsymbol{v} = \boldsymbol{0}$ et la matrice $I_n - A$ est donc inversible.

On a par ailleurs

$$1 = \|I_n\| \leq \|I_n - A\| \left\|(I_n - A)^{-1}\right\| \leq (1 + \|A\|) \left\|(I_n - A)^{-1}\right\|,$$

dont on déduit la première inégalité. La matrice $I_n - A$ étant inversible, on peut écrire

$$(I_n - A)^{-1} = I_n + A(I_n - A)^{-1},$$

d'où

$$\left\|(I_n - A)^{-1}\right\| \leq 1 + \|A\| \left\|(I_n - A)^{-1}\right\|,$$

ce qui conduit à la seconde inégalité.

Enfin, dire que la matrice $I_n - A$ est singulière signifie que -1 est valeur propre de A et donc que $\rho(A) \geq 1$. On se sert alors du théorème A.157 pour conclure. □

On dit qu'une suite $\left\{A^{(k)}\right\}_{k\in\mathbb{N}}$ de matrices de $M_{n,m}(\mathbb{C})$ *converge* vers une matrice A de $M_{n,m}(\mathbb{C})$ si

$$\lim_{k\to+\infty} \left\|A^{(k)} - A\right\| = 0$$

pour une norme de matrice (le choix de la norme importe peu en raison de l'équivalence des normes sur $M_{n,m}(\mathbb{C})$). Le résultat qui suit donne des conditions nécessaires et suffisantes pour que la suite formée des puissances successives d'une matrice carrée converge vers la matrice nulle. Il fournit un critère fondamental de convergence pour les méthodes itératives de résolution de systèmes linéaires introduites dans le chapitre 3.

Théorème A.159 *Soit A une matrice carrée. Les conditions suivantes sont équivalentes.*

i) $\lim_{k\to+\infty} A^k = 0$,

ii) $\lim_{k\to+\infty} A^k \boldsymbol{v} = \boldsymbol{0}$ *pour tout vecteur* $\boldsymbol{v}$,

iii) $\rho(A) < 1$,

iv) $\|A\| < 1$ *pour au moins une norme subordonnée* $\|\cdot\|$.

DÉMONSTRATION. Prouvons que i implique ii. Soit $\|\cdot\|$ une norme vectorielle et $\|\cdot\|$ la norme matricielle subordonnée lui correspondant. Pour tout vecteur $\boldsymbol{v}$, on a l'inégalité

$$\|A^k \boldsymbol{v}\| \leq \|A^k\| \, \|\boldsymbol{v}\|,$$

qui montre que $\lim_{k\to+\infty} A^k \boldsymbol{v} = \boldsymbol{0}$. Montrons ensuite que ii implique iii. Si $\rho(A) \geq 1$, alors il existe λ une valeur propre de A et $\boldsymbol{v} \neq \boldsymbol{0}$ un vecteur propre associé tels que

$$A\boldsymbol{v} = \lambda \boldsymbol{v} \text{ et } |\lambda| \geq 1.$$

La suite $(A^k \boldsymbol{v})_{k\in\mathbb{N}}$ ne peut donc converger vers $\boldsymbol{0}$, puisque $A^k \boldsymbol{v} = \lambda^k \boldsymbol{v}$. Le fait que iii implique iv est une conséquence immédiate du théorème A.157. Il reste à montrer que iv implique i. Il suffit pour cela d'utiliser l'inégalité

$$\forall k \in \mathbb{N}, \ \|A^k\| \leq \|A\|^k,$$

vérifiée par la norme subordonnée de l'énoncé. □

On déduit de ce théorème un résultat sur la convergence d'une série géométrique remarquable de matrice, dite *série de Neumann*.

Corollaire A.160 *Soit A une matrice carrée d'ordre n telle que* $\lim_{k\to+\infty} A^k = 0$. *Alors, la matrice* $I_n - A$ *est inversible et on a*

$$\sum_{i=1}^{+\infty} A^k = (I_n - A)^{-1}.$$

DÉMONSTRATION. On sait d'après le théorème A.159 que $\rho(A) < 1$ si $\lim_{k\to+\infty} A^k = 0$, la matrice $I_n - A$ est donc inversible. En considérant l'identité

$$(I_n - A)(I_n + A + \cdots + A^k) = I_n + A^{k+1}$$

et en faisant tendre k vers l'infini, on obtient alors l'identité recherchée. □

Nous pouvons maintenant prouver le résultat suivant, qui précise un peu plus le lien existant entre le rayon spectral et la norme d'une matrice.

Théorème A.161 (« formule de Gelfand[25] **» [Gel41])** *Soit A une matrice carrée et* $\|\cdot\|$ *une norme matricielle. On a*

$$\rho(A) = \lim_{k\to+\infty} \|A^k\|^{1/k}.$$

DÉMONSTRATION. Puisque $\rho(A) \leq \|A\|$ d'après le théorème A.157 et comme $\rho(A) = \left(\rho(A^k)\right)^{1/k}$, on sait déjà que

$$\rho(A) \leq \|A^k\|^{1/k}, \ \forall k \in \mathbb{N}.$$

Soit $\varepsilon > 0$ donné. La matrice

$$A_\varepsilon = \frac{A}{\rho(A) + \varepsilon}$$

25. Israïl Moiseevich Gelfand (Изра́иль Моисе́евич Ге́льфанд en russe, 2 septembre 1913 - 5 octobre 2009) était un mathématicien russe. Ses contributions aux mathématiques furent diverses et de nombreux résultats sont associés à son nom, notamment en théorie des groupes, en théorie des représentations et en analyse fonctionnelle.

vérifie $\rho(A_\varepsilon) < 1$ et on déduit du théorème A.159 que $\lim_{k\to+\infty} A_\varepsilon^k = 0$. Par conséquent, il existe un entier l, dépendant de ε, tel que

$$k \geq l \implies \|A_\varepsilon^k\| = \frac{\|A^k\|}{(\rho(A)+\varepsilon)^k} \leq 1.$$

Ceci implique que

$$k \geq l \implies \|A_\varepsilon^k\|^{1/k} \leq \rho(A)+\varepsilon,$$

et démontre donc l'égalité cherchée. □

A.5 Systèmes linéaires

Soit m et n deux entiers strictement positifs. Résoudre un *système linéaire de n équations à m inconnues* (*system of n linear equations with m unknowns* en anglais), à coefficients dans un corps $\mathbb{K}$ consiste à trouver la ou les solutions, s'il en existe, de l'équation matricielle

$$A\boldsymbol{x} = \boldsymbol{b},$$

où A est une matrice de $M_{n,m}(\mathbb{K})$, appelée *matrice du système*, $\boldsymbol{b}$ est un vecteur de $\mathbb{K}^n$, appelé *second membre du système*, et $\boldsymbol{x}$ est un vecteur de $\mathbb{K}^m$, appelé *inconnue du système*. On dit que le vecteur $\boldsymbol{x}$ est *solution du système* ci-dessus si ces composantes vérifient les n équations

$$\sum_{j=1}^{m} a_{ij}\, x_k = b_i,\ i = 1,\dots,n.$$

Enfin, le système linéaire est dit *compatible* s'il admet au moins une solution, *incompatible* sinon, et *homogène* si son second membre est nul.

Dans cette section, nous rappelons des résultats sur l'existence et l'unicité éventuelle des solutions de systèmes linéaires et leur détermination.

A.5.1 Systèmes linéaires carrés

Considérons pour commencer des systèmes ayant un même nombre d'équations et d'inconnues, c'est-à-dire tels que $m = n$. Le système est alors dit *carré*, par analogie avec la « forme » de sa matrice. Dans ce cas, l'inversibilité de la matrice du système fournit un critère très simple d'existence et d'unicité de la solution.

Théorème A.162 *Si A est une matrice inversible, alors il existe une unique solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$. Si A n'est pas inversible, alors soit le second membre $\boldsymbol{b}$ appartient à l'image de A et il existe alors une infinité de solutions du système qui diffèrent deux à deux par un élément du noyau de A, soit le second membre n'appartient pas à l'image de A, auquel cas il n'y a pas de solution.*

La démonstration de ce résultat est évidente et laissée au lecteur. Si ce dernier théorème ne donne pas de forme explicite de la solution permettant son calcul, cette dernière peut s'exprimer à l'aide des formules suivantes.

Proposition A.163 (« règle de Cramer » [Cra50]) *On désigne par les vecteurs $\boldsymbol{a}_j$, $j = 1,\dots,n$, de $M_{n,1}(\mathbb{K})$ les colonnes d'une matrice inversible A de $M_n(\mathbb{K})$. Les composante de la solution du système $A\boldsymbol{x} = \boldsymbol{b}$, avec $\boldsymbol{b}$ un vecteur de $M_{n,1}(\mathbb{K})$, sont données par*

$$\forall i \in \{1,\dots,n\},\ x_i = \frac{\det(\boldsymbol{a}_1,\dots,\boldsymbol{a}_{i-1},\boldsymbol{b},\boldsymbol{a}_{i+1},\dots,\boldsymbol{a}_n)}{\det(A)}.$$

DÉMONSTRATION. Le déterminant étant une forme multilinéaire alternée, on a

$$\forall (i,j) \in \{1,\dots,n\}^2,\ i \neq j,\ \forall (\lambda,\mu) \in \mathbb{K}^2,\ \det(\boldsymbol{a}_1,\dots,\boldsymbol{a}_{i-1},\lambda\,\boldsymbol{a}_i + \mu\,\boldsymbol{a}_j,\boldsymbol{a}_{i+1},\dots,\boldsymbol{a}_n) = \lambda\,\det(A).$$

Or, si le vecteur $\boldsymbol{x}$ est solution de $A\boldsymbol{x} = \boldsymbol{b}$, ses composantes sont les composantes du vecteur $\boldsymbol{b}$ dans la base de $M_{n,1}(\mathbb{K})$ formée par les colonnes de A, c'est-à-dire

$$\boldsymbol{b} = \sum_{j=1}^{n} x_j\,\boldsymbol{a}_j.$$

On en déduit que

$$\forall i \in \{1,\dots,n\},\ \det\Big(\boldsymbol{a}_1,\dots,\boldsymbol{a}_{i-1},\sum\nolimits_{j=1}^{n} x_j\,\boldsymbol{a}_j,\boldsymbol{a}_{i+1},\dots,\boldsymbol{a}_n\Big) = \det(\boldsymbol{a}_1,\dots,\boldsymbol{a}_{i-1},\boldsymbol{b},\boldsymbol{a}_{i+1},\dots,\boldsymbol{a}_n) = x_i\ \det(A),$$

d'où la formule. □

Par extension, on appelle parfois *système de Cramer* tout système d'équations linéaires dont la matrice est inversible.

Aussi séduisante qu'elle puisse sembler, la règle de Cramer s'avère généralement inefficace en pratique si le système possède plus de deux ou trois équations. L'évaluation des déterminants intervenant dans les formules nécessite en effet bien trop d'opérations si l'on applique une méthode récursive naïve [26] de calcul du déterminant.

A.5.2 Systèmes linéaires sur ou sous-déterminés

Considérons maintenant des systèmes linéaires n'ayant pas le même nombre d'équations et d'inconnues, c'est-à-dire tels que $m \neq n$. Dans ce cas, la matrice du système est rectangulaire. Lorsque $n < m$, on dit que le système est *sous-déterminé* : il y a plus d'inconnues que d'équations, ce qui donne, heuristiquement, plus de « liberté » pour l'existence de solutions. Si $n > m$, on dit que le système est *sur-déterminé* : il y a moins d'inconnues que d'équations, ce qui restreint cette fois-ci les possibilités d'existence de solutions. On a le résultat fondamental suivant dont la démonstration est laissée en exercice.

Théorème A.164 *Le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ possède une solution si et seulement si le second membre $\boldsymbol{b}$ appartient à l'image de la matrice A. La solution est unique si est seulement si le noyau de A est réduit au vecteur nul. Deux solutions du système diffèrent par un élément du noyau de A.*

Le résultat suivant est obtenu par simple application du théorème du rang (voir le théorème A.89).

Lemme A.165 *Si $n < m$, alors $\dim\ker(A) \geq m - n \geq 1$, et s'il existe une solution au système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, il en existe une infinité.*

A.5.3 Systèmes linéaires sous forme échelonnée

Nous abordons maintenant le cas de systèmes linéaires dont les matrices sont *sous forme échelonnée*. S'intéresser à ce type particulier de systèmes est de toute première importance, puisque l'enjeu de méthodes de résolution comme la méthode d'élimination de Gauss (voir la section 2.4) est de ramener un système linéaire quelconque à un système sous forme échelonnée équivalent (c'est-à-dire ayant le même ensemble de solutions), plus simple à résoudre.

Définition A.166 (matrice sous forme échelonnée) *Une matrice A de $M_{n,m}(\mathbb{K})$ est dite **sous forme échelonnée** (**in row echelon form** en anglais) s'il existe un entier r, $1 \leq r \leq \min(m,n)$ et une suite d'entiers $1 \leq j_1 < j_2 < \ldots < j_r \leq m$ tels que*

- *$a_{ij_i} \neq 0$ pour $1 \leq i \leq r$, et $a_{ij} = 0$ pour $1 \leq i \leq r$ et $1 \leq j < j_i$ ($i \geq 2$ si $j_1 = 1$), c'est-à-dire que les coefficients a_{ij_i}, appelés **pivots**, sont les premiers coefficients non nuls des r premières lignes,*
- *$a_{ij} = 0$ pour $r < i \leq n$ et $1 \leq j \leq m$, c'est-à-dire que toutes les lignes après les r premières sont nulles.*

*Une telle matrice A est par ailleurs dite **sous forme échelonnée réduite** (**in reduced row echelon form** en anglais) si tous ses pivots valent 1 et si les autres coefficients des colonnes contenant un pivot sont nuls.*

Exemple de matrice sous forme échelonnée. La matrice

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & -1 & 5 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

est une matrice sous forme échelonnée dont les pivots sont 1, 2 et 3.

26. On peut observer que l'utilisation de la méthode d'élimination de Gauss (voir la section 2.4 du chapitre 2) permet la résolution du système linéaire en un nombre d'opérations arithmétiques essentiellement égal à celui de l'évaluation par ce même procédé d'un seul des $n+1$ déterminants requis par la règle.

On déduit immédiatement de la définition précédente que le rang d'une matrice sous forme échelonnée est égal au nombre r de pivots. Dans un système linéaire de n équations à m inconnues dont la matrice est sous forme échelonnée, les inconnues $x_{j_1}, \ldots, x_{j_r}$ sont dites *principales* et les $n-r$ inconnues restantes sont appelées *secondaires*.

Considérons à présent la résolution d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, de n équations à m inconnues et de rang r, sous forme échelonnée. Commençons par discuter de la compatibilité de ce système. Tout d'abord, si $r = n$, le système linéaire est compatible et ses équations sont linéairement indépendantes. Sinon, c'est-à-dire si $r < n$, les $n-r$ dernières lignes de la matrice A sont nulles et le système linéaire n'est donc compatible que si les $n-r$ dernières composantes du vecteur $\boldsymbol{b}$ sont également nulles, ce qui revient à vérifier $n-r$ *conditions de compatibilité*.

Parlons à présent de la résolution effective du système lorsque ce dernier est compatible. Plusieurs cas de figure se présentent.

- Si $r = m = n$, le système est de Cramer et admet une unique solution. Le système échelonné est alors triangulaire (supérieur) et se résout par des substitutions successives (voir la section 2.3).
- Si $r = m < n$, la solution existe, puisque le système est supposé satisfaire aux $n-r$ conditions de compatibilité, et unique. On l'obtient en résolvant le système linéaire équivalent

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \ldots & + & a_{1r}\,x_r & = & b_1 \\ & & a_{21}\,x_2 & + & \ldots & + & a_{2r}\,x_r & = & b_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & a_{rr}\,x_r & = & b_r \end{array}$$

  par des substitutions successives comme dans le cas précédent.
- Enfin, si $r < m \leq n$ et le système est compatible, on commence par faire « passer » les inconnues secondaires dans les membres de droite du système. Ceci se traduit matriciellement par la réécriture du système sous la forme

$$A_P \boldsymbol{x}_P = \boldsymbol{b} - A_S \boldsymbol{x}_S,$$

  où A_P est une sous-matrice extraite de A à n lignes et r colonnes, constituée des colonnes de A qui contiennent un pivot, $\boldsymbol{x}_P$ est un vecteur de $\mathbb{K}^r$ ayant pour composantes les inconnues principales, A_S est une sous-matrice extraite de A à n lignes et $m-r$ colonnes, constituée des colonnes de A ne contenant pas de pivot, et $\boldsymbol{x}_S$ est un vecteur de $\mathbb{K}^{m-r}$ ayant pour composantes les inconnues secondaires. Ce dernier système permet d'obtenir de manière unique les inconnues principales en fonction des inconnues secondaires, qui jouent alors le rôle de paramètres. Dans ce cas, le système admet une infinité de solutions, qui sont chacune la somme d'une solution particulière de $A\boldsymbol{x} = \boldsymbol{b}$ et d'une solution du système homogène $A\boldsymbol{x} = \boldsymbol{0}$ (c'est-à-dire un élément du noyau de A).

  Une solution particulière $\boldsymbol{s}_0$ du système est obtenue, par exemple, en complétant la solution du système $A_P \boldsymbol{x}_{P_0} = \boldsymbol{b}$, que l'on résout de la même façon que dans le cas précédent, par des zéros pour obtenir un vecteur de $\mathbb{K}^m$ (ceci revient à fixer la valeur de toutes les inconnues secondaires à zéro), *i.e.*,

$$\boldsymbol{s}_0 = \begin{pmatrix} \boldsymbol{x}_{P_0} \\ \boldsymbol{0} \end{pmatrix}.$$

  On détermine ensuite une base du noyau de A en résolvant les $m-r$ systèmes linéaires $A_P \boldsymbol{x}_{P_k} = \boldsymbol{b} - A_S \boldsymbol{e}_k^{(m-r)}$, $1 \leq k \leq n-r$, où $\boldsymbol{e}_k^{(m-r)}$ désigne le k^{e} vecteur de la base canonique de $\mathbb{K}^{m-r}$ (ceci revient à fixer la valeur de la k^{e} inconnue secondaire à 1 et celles des autres à zéro), le vecteur de base $\boldsymbol{x}_k$ correspondant étant

$$\boldsymbol{s}_k = \begin{pmatrix} \boldsymbol{x}_{P_k} \\ \boldsymbol{e}_k^{(m-r)} \end{pmatrix}.$$

  La solution générale du système est alors de la forme

$$\boldsymbol{x} = \boldsymbol{s}_0 + \sum_{k=1}^{m-r} c_k \, \boldsymbol{s}_k,$$

  avec les c_k, $1 \leq k \leq m-r$, des scalaires.

A.5.4 Conditionnement d'une matrice

La résolution d'un système linéaire par les méthodes numériques des chapitres 2 et 3 est sujette à des erreurs d'arrondis dont l'accumulation peut détériorer notablement la précision de la solution obtenue. Afin de mesurer la sensibilité de la solution $\boldsymbol{x}$ d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ par rapport à des perturbations des données A et $\boldsymbol{b}$, on utilise une quantité appelée *conditionnement*, introduite pour la première fois explicitement par Turing [Tur48] dans le cas de la norme de Frobenius. C'est un cas particulier de la notion générale de conditionnement d'un problème définie dans la sous-section 1.4.2.

Définition A.167 (conditionnement d'une matrice) *Soit $\|\cdot\|$ une norme matricielle. Pour toute matrice inversible A d'ordre n, on appelle* ***conditionnement de*** *A* ***relativement à la norme*** *$\|\cdot\|$ le nombre*

$$\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|.$$

La valeur du conditionnement d'une matrice dépendant en général de la norme matricielle choisie, on a coutume de signaler celle-ci en ajoutant un indice dans la notation, par exemple $\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$. On note que l'on a toujours $\mathrm{cond}(A) \geq 1$ pour une norme matricielle subordonnée (et $\mathrm{cond}_F(A) \geq \sqrt{n}$), puisque $\|A\|\,\|A^{-1}\| \geq \|AA^{-1}\| = \|I_n\| = 1$ (et $\|I_n\|_F = \sqrt{n}$). D'autres propriétés évidentes du conditionnement sont rassemblées dans le résultat suivant.

Théorème A.168 *Soit A une matrice inversible d'ordre n.*

1. *On a* $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$ *et* $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$ *pour tout scalaire α non nul.*
2. *On a*
$$\mathrm{cond}_2(A) = \frac{\mu_n}{\mu_1},$$
où μ_1 et μ_n désignent respectivement la plus petite et la plus grande des valeurs singulières de A.
3. *Si A est une matrice normale, on a*
$$\mathrm{cond}_2(A) = \frac{\max\limits_{1\leq i\leq n} |\lambda_i|}{\min\limits_{1\leq i\leq n} |\lambda_i|} = \rho(A)\rho(A^{-1}),$$
où les scalaires λ_i, $1 \leq i \leq n$, sont les valeurs propres de A.
4. *Si A est une matrice unitaire ou orthogonale, son conditionnement* $\mathrm{cond}_2(A)$ *vaut* 1.
5. *Le conditionnement* $\mathrm{cond}_2(A)$ *est invariant par transformation unitaire (ou orthogonale),*
$$UU^* = I_n \implies \mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^*AU).$$

DÉMONSTRATION.

1. Les égalités découlent de la définition du conditionnement et des propriétés de la norme.
2. On a établi dans la proposition A.156 que $\|A\|_2 = \sqrt{\rho(A^*A)}$ et, d'après la définition des valeurs singulières de A, on a donc $\|A\|_2 = \mu_n$. Par ailleurs, on voit que
$$\|A^{-1}\|_2 = \sqrt{\rho((A^{-1})^*A^{-1})} = \sqrt{\rho(A^{-1}(A^{-1})^*)} = \sqrt{\rho((A^*A)^{-1})} = \frac{1}{\mu_1},$$
ce qui démontre le résultat.
3. La propriété résulte de l'égalité $\|A\|_2 = \rho(A)$ vérifiée par les matrices normales (voir encore la proposition A.156).
4. Le résultat découle de l'égalité $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(I_n)} = 1$.
5. La propriété est une conséquence de l'invariance par transformation unitaire de la norme $\|\cdot\|_2$ (voir une nouvelle fois la proposition A.156).

□

La proposition ci-dessous montre que plus le conditionnement d'une matrice est grand, plus la solution d'un système linéaire qui lui est associé est sensible aux perturbations des données.

Proposition A.169 *Soit A une matrice inversible d'ordre n, $\boldsymbol{b}$ un vecteur non nul de taille correspondante et $\|\cdot\|$ une norme matricielle subordonnée. Si $\boldsymbol{x}$ et $\boldsymbol{x}+\delta\boldsymbol{x}$ sont les solutions respectives des systèmes linéaires $A\boldsymbol{x}=\boldsymbol{b}$ et $A(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$, avec $\delta\boldsymbol{b}$ un vecteur de taille n, on a*

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leq \mathrm{cond}(A)\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \tag{A.13}$$

Si $\boldsymbol{x}$ et $\boldsymbol{x}+\delta\boldsymbol{x}$ sont les solutions respectives[27] *des systèmes linéaires $A\boldsymbol{x}=\boldsymbol{b}$ et $(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}$, avec δA une matrice d'ordre n, on a*

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}+\delta\boldsymbol{x}\|}\leq \mathrm{cond}(A)\,\frac{\|\delta A\|}{\|A\|}. \tag{A.14}$$

De plus, ces deux inégalités sont optimales, dans le sens où, pour toute matrice A donnée, on peut trouver des vecteurs $\boldsymbol{b}$ et $\delta\boldsymbol{b}$ (resp. une matrice δA et un vecteur $\boldsymbol{b}$) non nuls tels que l'on a une égalité.

DÉMONSTRATION. On remarque que le vecteur $\delta\boldsymbol{x}$ est donné par $\delta\boldsymbol{x}=A^{-1}\delta\boldsymbol{b}$, d'où $\|\delta\boldsymbol{x}\|\leq\|A^{-1}\|\,\|\delta\boldsymbol{b}\|$. Comme on a par ailleurs $\|\boldsymbol{b}\|\leq\|A\|\,\|\boldsymbol{x}\|$, on en déduit la première inégalité. Son optimalité découle de la définition d'une norme matricielle subordonnée ; pour toute matrice A d'ordre n, il existe des vecteurs $\delta\boldsymbol{b}$ et $\boldsymbol{x}$ non nuls tels que $\|A^{-1}\delta\boldsymbol{b}\|=\|A^{-1}\|\,\|\delta\boldsymbol{b}\|$ et $\|A\boldsymbol{x}\|=\|A\|\,\|\boldsymbol{x}\|$ (voir la démonstration de la proposition A.155).
Pour la seconde inégalité, on tire de $A\delta\boldsymbol{x}+\delta A(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{0}$ la majoration $\|\delta\boldsymbol{x}\|\leq\|A^{-1}\|\,\|\delta A\|\,\|\boldsymbol{x}+\delta\boldsymbol{x}\|$, qui donne le résultat. Pour prouver que l'inégalité obtenue est la meilleure possible, on considère un vecteur $\boldsymbol{y}$ non nul tel que $\|A^{-1}\boldsymbol{y}\|=\|A^{-1}\|\,\|\boldsymbol{y}\|$ et un scalaire η non nul n'appartenant pas au spectre de la matrice A. On pose alors $\delta A=-\eta I$, $\delta\boldsymbol{x}=\eta A^{-1}\boldsymbol{y}$, $\boldsymbol{x}+\delta\boldsymbol{x}=\boldsymbol{y}$ et $\boldsymbol{b}=(A-\eta I_n)\boldsymbol{y}$ (ce dernier vecteur étant non nul puisque $A-\eta I_n$ est inversible), qui vérifient $A\boldsymbol{x}=\boldsymbol{b}$, $(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}$ et $\|\delta\boldsymbol{x}\|=|\eta|\,\|A^{-1}\boldsymbol{y}\|=\|\delta A\|\|A^{-1}\|\|\boldsymbol{x}+\delta\boldsymbol{x}\|$. □

Il peut sembler étrange d'un point de vue théorique[28] que l'erreur $\delta\boldsymbol{x}$ sur la solution majorée dans (A.14) soit mesurée relativement à $\boldsymbol{x}+\delta\boldsymbol{x}$. Il possible d'obtenir un résultat comparable à (A.13) en faisant l'hypothèse (parfaitement loisible car on étudie l'influence de petites perturbations) que la matrice δA est telle que l'on ait

$$\|A^{-1}\|\|\delta A\|<1. \tag{A.15}$$

Dans ce cas, on déduit du théorème A.158 que la matrice $A+\delta A=A(I_n+A^{-1}\delta A)$ est inversible et il vient alors

$$\delta\boldsymbol{x}=-(I_n+A^{-1}\delta A)^{-1}A^{-1}\delta A\boldsymbol{x},$$

soit encore, en faisant appel une nouvelle fois au théorème A.158,

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leq\|(I_n+A^{-1}\delta A)^{-1}\|\|A^{-1}\|\|\delta A\|\leq\frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\|\delta A\|=\frac{\mathrm{cond}(A)}{1-\|A^{-1}\|\|\delta A\|}\,\frac{\|\delta A\|}{\|A\|},$$

qui est bien un inégalité de la forme voulue.

Malgré leur optimalité, toutes ces inégalités sont, en général, pessimistes. Elles conduisent néanmoins à l'introduction d'une terminologie courante, issue de la notion générale de conditionnement d'un problème (voir la section 1.4.2), visant à traduire le fait que la résolution numérique d'un système linéaire donné peut être particulièrement sensible aux erreurs d'arrondis, ce qui conduit à d'importantes erreurs sur la solution calculée. Ainsi, on dit qu'une matrice inversible est *bien conditionnée* (relativement à une norme matricielle subordonnée) si son conditionnement est proche de l'unité. Au contraire, elle est dite *mal conditionnée* si son conditionnement est très grand devant 1. Les matrices unitaires (ou orthogonales) étant très bien conditionnées relativement à la norme spectrale (voir le théorème A.168), on comprend la place privilégiée qu'elles occupent dans diverses méthodes numériques matricielles.

Terminons par une interprétation géométrique du conditionnement d'une matrice, liée à la condition (A.15) garantissant que la matrice perturbée $A+\delta A$ est non singulière et qui tend à compléter la dénomination que nous

27. On notera qu'il n'est pas nécessaire de supposer que la matrice $A+\delta A$ est inversible pour établir l'inégalité, mais simplement que le système linéaire associé possède au moins une solution.

28. En pratique, une telle majoration est en revanche très utile, puisque c'est la solution calculée $\boldsymbol{x}+\delta\boldsymbol{x}$, et non la solution exacte $\boldsymbol{x}$, que l'on connaît effectivement.

venons d'introduire. On définit la *distance d'une matrice A d'ordre n à l'ensemble Σ_n des matrices singulières d'ordre n relativement à une norme matricielle* $\|\cdot\|$ par

$$\operatorname{dist}(A, \Sigma_n) = \min\{\|\delta A\| \mid \delta A \in M_n(\mathbb{C}),\ A + \delta A \in \Sigma_n\}.$$

On a le théorème suivant, qui découle d'un résultat[29] de Eckart et Young [EY36] dans le cas de la norme subordonnée à la norme euclidienne et établi par Kahan [Kah66] (qui l'attribue à Gastinel) dans le cas d'une norme matricielle subordonnée quelconque.

Théorème A.170 *Soit A une matrice d'ordre n inversible et $\|\cdot\|$ une norme matricielle subordonnée. On a*

$$\operatorname{dist}(A, \Sigma_n) = \frac{\|A\|}{\operatorname{cond}(A)}.$$

DÉMONSTRATION. Si la matrice $A + \delta A$ est singulière, alors il existe un vecteur x de $\mathbb{C}^n$ non nul tel que $(A + \delta A)\boldsymbol{x} = \mathbf{0}$. On a alors, en notant également $\|\cdot\|$ la norme vectorielle à laquelle la norme matricielle $\|\cdot\|$ est subordonnée,

$$\|\boldsymbol{x}\| = \|A^{-1}\delta A\boldsymbol{x}\| \leq \|A^{-1}\|\|\delta A\|\|\boldsymbol{x}\|,$$

d'où

$$\|\delta A\| \geq \frac{1}{\|A^{-1}\|} = \frac{\|A\|}{\operatorname{cond}(A)}.$$

Pour montrer qu'il existe une perturbation δA donnant lieu à une égalité dans cette majoration, on considère un vecteur $\boldsymbol{y}$ de $\mathbb{C}^n$ tel que $\|A^{-1}\boldsymbol{y}\| = \|A^{-1}\|\|\boldsymbol{y}\| \neq 0$ et l'on pose $\delta A = -\boldsymbol{y}\boldsymbol{w}^*$, où $\boldsymbol{w}$ est un élément du dual de $A^{-1}\boldsymbol{y}$ par rapport à la norme $\|\cdot\|$ (voir la définition A.147). On a alors $(A + \delta A)A^{-1}\boldsymbol{y} = \mathbf{0}$, la matrice $A + \delta A$ est donc singulière, et, en vertu de (A.9), il vient

$$\|\delta A\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\mathbf{0}}} \frac{\|\boldsymbol{y}\boldsymbol{w}^*\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \|\boldsymbol{y}\| \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\mathbf{0}}} \frac{|\boldsymbol{w}^*\boldsymbol{v}|}{\|\boldsymbol{v}\|} = \|\boldsymbol{y}\|\|\boldsymbol{w}^*\| = \frac{\|\boldsymbol{y}\|}{\|A^{-1}\boldsymbol{y}\|} = \frac{1}{\|A^{-1}\|}.$$

□

On peut donc voir une matrice mal conditionnée comme « presque » singulière, ce qui n'est pas sans conséquence pour la résolution numérique de tout système linéaire lui étant associé. En effet, en raison des erreurs d'arrondis, la matrice intervenant dans les calculs est une matrice *perturbée* et donc potentiellement singulière. Ainsi, même lorsque le second membre du système linéaire considéré n'est pas perturbé, la solution obtenue peut être très différente de la solution recherchée.

A.6 Note sur l'annexe

La décomposition en valeurs singulières fut essentiellement introduite par Beltrami[30] et Jordan[31]. Le lecteur intéressé trouvera de nombreux détails historiques sur ce sujet dans l'article [Ste93].

Références

[Bje51] A. BJERHAMMAR. Rectangular reciprocal matrices, with special reference to geodetic calculations. *Bull. Géodésique*, 20(1):188–220, 1951. DOI: 10.1007/BF02526278.

[Bou59] V. BOUNIAKOWSKY. Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies. *Mem. Acad. Imp. Sci. St.-Pétersb. (7)*, 1(9) :1-18, 1859.

[Cau21] A.-L. CAUCHY. *Cours d'analyse de l'École Royale Polytechnique. Première partie. Analyse algébrique*. Imprimerie Royale, 1821. Note II. Sur les formules qui résultent de l'emploi du signe > ou <, et sur les moyennes entre plusieurs quantités.

[Cra50] G. CRAMER. *Introduction à l'analyse des lignes courbes algébriques*. Frères Cramer et Claude Philibert, 1750.

29. Le résultat en question montre que $\operatorname{dist}_F(A, \Sigma_n)$ est égal à la plus petite valeur singulière de la matrice A, dont on déduit que $\operatorname{dist}_F(A, \Sigma_n) = \operatorname{dist}_2(A, \Sigma_n) = \|A\|_2(\operatorname{cond}_2(A))^{-1}$.

30. Eugenio Beltrami (16 novembre 1835 - 18 février 1900) était un mathématicien italien, célèbre pour ses travaux en géométrie différentielle et en physique mathématique.

31. Marie Ennemond Camille Jordan (5 janvier 1838 - 22 janvier 1922) était un mathématicien français, connu à la fois pour son travail fondamental en théorie des groupes et pour son influent *cours d'analyse*.

[Des87] J. DESPLANQUES. Théorème d'algèbre. *J. Math. Spéc. (3)*, 1 :12-13, 1887.

[EY36] C. ECKART and G. YOUNG. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. DOI: 10.1007/BF02288367.

[Fré35] M. FRÉCHET. Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert. *Ann. Math. (2)*, 36(3) :705-718, 1935. DOI : 10.2307/1968652.

[Gel41] I. GELFAND. Normierte Ringe. *Rec. Math. [Mat. Sbornik] N.S.*, 9(51)(1):3–24, 1941.

[GV13] G. H. GOLUB and C. F. VAN LOAN. *Matrix computations*. Johns Hopkins University Press, fourth edition, 2013.

[HJ13] R. A. HORN and C. R. JOHNSON. *Matrix analysis*. Cambridge University Press, second edition, 2013. DOI: 10.1017/CBO9781139020411.

[JvN35] P. JORDAN and J. von NEUMANN. On inner products in linear, metric spaces. *Ann. Math. (2)*, 36(3):719–723, 1935. DOI: 10.2307/1968653.

[Kah66] W. KAHAN. Numerical linear algebra. *Canad. Math. Bull.*, 9(6):757–801, 1966. DOI: 10.4153/CMB-1966-083-2.

[Lév81] L. LÉVY. Sur la possibilité de l'équilibre électrique. *C. R. Acad. Sci. Paris*, 93 :706-708, 1881.

[Moo20] E. H. MOORE. On the reciprocal of the general algebraic matrix (abstract). *Bull. Amer. Math. Soc.*, 26(9):385–396, 1920. DOI: 10.1090/S0002-9904-1920-03322-7.

[Pen55] R. PENROSE. A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.*, 51(3):406–413, 1955. DOI: 10.1017/S0305004100030401.

[Ple77] R. J. PLEMMONS. M-matrix characterizations. I – nonsingular M-matrices. *Linear Algebra Appl.*, 18(2):175–188, 1977. DOI: 10.1016/0024-3795(77)90073-8.

[Sch88] H. A. SCHWARZ. Über ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung. *Acta Soc. Sci. Fennicae*, 15:318–362, 1888.

[Ste93] G. W. STEWART. On the early history of the singular value decomposition. *SIAM Rev.*, 35(4):551–566, 1993. DOI: 10.1137/1035134.

[Syl50] J. J. SYLVESTER. XLVII. Additions to the articles in the September number of this journal, "On a new class of theorems," and on Pascal's theorem. *Philos. Mag. (3)*, 37(251):363–370, 1850. DOI: 10.1080/14786445008646629.

[Syl52] J. J. SYLVESTER. XIX. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *Philos. Mag. (4)*, 4(23):138–142, 1852. DOI: 10.1080/14786445208647087.

[Tau49] O. TAUSSKY. A recurring theorem on determinants. *Amer. Math. Monthly*, 56(10):672–676, 1949. DOI: 10.2307/2305561.

[Tur48] A. M. TURING. Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.*, 1(1):287–308, 1948. DOI: 10.1093/qjmam/1.1.287.

Annexe B

Rappels et compléments d'analyse

Cette annexe est consacrée à des rappels de quelques notions et résultats d'analyse, en incluant la plupart du temps leurs démonstrations, auxquels il est fait appel dans les chapitres 5, 6 et 7.

B.1 Nombres réels

Afin de ne pas entrer dans les détails de sa construction (au moyen des *coupures de Dedekind* [1], qui utilise la théorie des ensembles, par exemple) que le lecteur pourra trouver dans tout bon manuel d'analyse, nous admettons l'existence et l'unicité de l'*ensemble des nombres réels* $\mathbb{R}$, muni des lois internes d'*addition*, notée $+$, et de *multiplication*, notée [2] $\cdot$, et d'une *relation binaire* [3] notée $\leq$, vérifiant les propriétés suivantes :

1. $(\mathbb{R}, +, \cdot)$ est un *corps commutatif*,
2. $(\mathbb{R}, +, \cdot, \leq)$ est un *corps totalement ordonné*,
3. Toute partie non vide et majorée de $\mathbb{R}$ admet une *borne supérieure* dans $\mathbb{R}$.

Rappelons que la propriété 1. signifie que $(\mathbb{R}, +)$ (resp. $(\mathbb{R}^*, \cdot)$ avec $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$) est un *groupe commutatif*, c'est-à-dire que

i) la loi $+$ (resp. $\cdot$) est *commutative* : $\forall (x, y) \in \mathbb{R}^2$, $x + y = b + y$ (resp. $\forall (x, y) \in \mathbb{R}^2$, $xy = yx$),

ii) la loi $+$ (resp. $\cdot$) est *associative* : $\forall (x, y, z) \in \mathbb{R}^3$, $(x+y)+z = x+(y+z)$ (resp. $\forall (x, y, z) \in \mathbb{R}^3$, $(xy)z = x(yz)$),

iii) la loi $+$ (resp. $\cdot$) admet un *élément neutre* : $\forall x \in \mathbb{R}$, $x + 0 = 0 + x = x$ (resp. $\forall x \in \mathbb{R}$, $x1 = 1x = x$),

iv) tout élément de $\mathbb{R}$ (resp. $\mathbb{R}^*$) admet un *symétrique* pour la loi $+$ (resp. $\cdot$) : $\forall x \in \mathbb{R}$, $\exists -x \in \mathbb{R}$, $x + (-x) = (-x) + x = 0$ (resp. $\forall x \in \mathbb{R}^*$, $\exists \frac{1}{x} \in \mathbb{R}$, $x \frac{1}{x} = \frac{1}{x} x = 1$),

et que la multiplication est *distributive* par rapport à l'addition :

$$\forall (x, y, z) \in \mathbb{R}^3,\ x(y + z) = xy + xz.$$

La propriété 2. signifie pour sa part que la relation $\leq$ est une *relation d'ordre total* dans $\mathbb{R}$, c'est-à-dire qu'elle est

i) *réflexive* : $\forall x \in \mathbb{R}$, $x \leq x$,

ii) *antisymétrique* : $\forall (x, y) \in \mathbb{R}^2$, $(x \leq y \text{ et } y \leq x) \implies x = y$,

iii) *transitive* : $\forall (x, y, z) \in \mathbb{R}^3$, $(x \leq y \text{ et } y \leq z) \implies x \leq z$,

iv) *totale* : $\forall (x, y) \in \mathbb{R}^2$, $(x \leq y \text{ ou } y \leq x)$,

et qu'elle est de plus *compatible* avec l'addition et la multiplication, c'est-à-dire que

1. Julius Wilhelm Richard Dedekind (6 octobre 1831 - 12 février 1916) était un mathématicien allemand. Il réalisa des travaux de première importance en algèbre (en introduisant notamment la théorie des anneaux) et en théorie algébrique des nombres.

2. Dans la pratique, on omet souvent d'écrire le symbole « $\cdot$ ». C'est ce que nous faisons ici.

3. On rappelle qu'une relation binaire $\mathscr{R}$ d'un ensemble non vide E vers un ensemble non vide F est définie par une partie G de $E \times F$. Si $(x, y) \in G$, on dit que x est en relation avec y et on note $x \mathscr{R} y$. Dans le cas particulier où $E = F$, on dit que $\mathscr{R}$ est une relation binaire définie sur, ou dans, E.

i) $\forall (x,y,z) \in \mathbb{R}^3,\ x \leq y \implies x+z \leq y+z$,

ii) $\forall (x,y,z) \in \mathbb{R}^3,\ (x \leq y \text{ et } 0 \leq z) \implies xz \leq yz$.

Pour $(x,y) \in \mathbb{R}^2$, $x < y$ signifie $x \leq y$ et $x \neq y$. On peut également noter $y \geq x$ (resp. $y > x$) au lieu de $x \leq y$ (resp. $x < y$).

Nous reviendrons dans la suite sur la propriété 3., qui est appelée l'*axiome de la borne supérieure*.

Les éléments de $\mathbb{R}$ sont appelés les *nombres réels* et l'on note $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$, $\mathbb{R}_- = \{x \in \mathbb{R} \mid x \leq 0\}$, $\mathbb{R}_+^* = \mathbb{R}_+ \backslash \{0\}$ et $\mathbb{R}_-^* = \mathbb{R}_- \backslash \{0\}$.

B.1.1 Majorant et minorant

Les définitions et propositions suivantes particularisent des notions introduites sur les parties d'un ensemble ordonné (voir les définitions A.23) aux cas de parties de $\mathbb{R}$.

Définitions B.1 *Soit A une partie de $\mathbb{R}$ et x un nombre réel. On dit que x est*

- *un **majorant** (resp. **minorant**) de A dans $\mathbb{R}$ si et seulement s'il est supérieur (resp. inférieur) ou égal à tous les éléments de A :*
$$\forall a \in A,\ a \leq x \text{ (resp. } x \leq a),$$
- *un **plus grand élément** (resp. **plus petit élément**) de A dans $\mathbb{R}$ si et seulement si c'est un majorant (resp. minorant) de A dans $\mathbb{R}$ appartenant à A.*

La partie A est dite *majorée* (resp. *minorée*) dans $\mathbb{R}$ si et seulement si elle possède au moins un majorant (resp. minorant) et *bornée* si et seulement si elle est à la fois majorée et minorée.

Proposition B.2 *Soit A une partie de $\mathbb{R}$. Si A admet un plus grand (resp. petit) élément, celui-ci est unique et on le note* $\max(A)$ *(resp.* $\min(A)$*).*

DÉMONSTRATION. Soit x et x' deux plus grands éléments de A. Puisque x' appartient à A et x est un plus grand élément de A, on a $x' \leq x$. De la même manière, x appartient à A et x' est un plus grand élément de A, d'où $x \leq x'$. L'antisymétrie de la relation $\leq$ implique alors que $x = x'$. □

Définitions B.3 *On appelle **borne supérieure** (resp. **borne inférieure**) de A dans $\mathbb{R}$ le plus petit des majorants (resp. le plus grand des minorants) de A dans $\mathbb{R}$, s'il existe. Elle est alors unique et notée* $\sup(A)$ *(resp.* $\inf(A)$*).*

Si A possède un plus grand (resp. petit) élément $\max(A)$ (resp. $\min(A)$), alors $\max(A) = \sup(A)$ (resp. $\min(A) = \inf(A)$).

Rappelons enfin une propriété que nous admettrons, à savoir l'axiome de la borne supérieure.

Proposition B.4 (« axiome de la borne supérieure ») *Toute partie non vide et majorée de $\mathbb{R}$ admet une borne supérieure dans $\mathbb{R}$.*

En considérant l'ensemble des opposés des éléments de la partie envisagée, on obtient à partir de cet axiome le résultat suivant.

Proposition B.5 *Toute partie non vide et minorée de $\mathbb{R}$ admet une borne inférieure dans $\mathbb{R}$.*

B.1.2 Propriétés des nombres réels

Propriété d'Archimède

Une conséquence de la proposition B.4 est que l'ensemble des nombres réels satisfait la *propriété d'Archimède* [4].

4. Archimède de Syracuse (Αρχιμήδης en grec, 287 av. J.-C. - 212 av. J.-C.) était un physicien, mathématicien et ingénieur grec de l'Antiquité. Scientifique de grande envergure, il inventa la poulie, la roue dentée, la vis sans fin ainsi que des machines de guerre pour repousser les romains durant le siège de Syracuse. En physique, on lui doit en particulier les premières lois de l'hydrostatique et une étude précise sur l'équilibre des surfaces planes. En mathématiques, il établit notamment de nombreuses formules relatives aux mesures des surfaces et des volumes qui font de lui un précurseur du calcul intégral.

Théorème B.6 *L'ensemble $\mathbb{R}$ est un corps* ***archimédien****, c'est-à-dire qu'il vérifie la propriété suivante :*

$$\forall x \in \mathbb{R}_+^*,\ \forall y \in \mathbb{R}_+^*,\ \exists n \in \mathbb{N}^*,\ ny > x.$$

DÉMONSTRATION. Soit $x \in \mathbb{R}_+^*$ et $y \in \mathbb{R}_+^*$. Vérifions par l'absurde que x n'est pas un majorant de l'ensemble $E = \{ny \mid n \in \mathbb{N}^*\}$. Supposons donc que x est un majorant de E. L'ensemble E étant une partie non vide de $\mathbb{R}$ et de plus majorée par x, celui-ci admet, d'après la proposition B.4, une borne supérieure réelle que l'on note M. Comme M est le plus petit des majorants de E, le réel $M-y$ n'est pas un majorant de E, ce qui signifie qu'il existe un entier relatif n tel que $ny > M-y$, et donc $(n+1)y > M$. Or, $(n+1)y \in E$, ce qui contredit le fait que $M = \sup E$. On en déduit que le réel donné x ne majore pas E et, par suite, qu'on peut trouver un élément n de E qui vérifie $ny > x$. □

Partie entière d'un nombre réel

En appliquant le précédent théorème avec $y = 1$, on voit que, pour tout nombre réel x, l'ensemble $\{n \in \mathbb{Z} \mid n \leq x\}$ est une partie majorée et non vide de $\mathbb{Z}$, qui admet par conséquent un plus grand élément. Nous obtenons ainsi le résultat suivant.

Proposition et définition B.7 *Pour tout réel x, il existe un unique entier relatif n vérifiant $n \leq x < n+1$. Cet entier est appelé* ***partie entière (par défaut) de*** *x et noté $\mathrm{E}(x)$, ou encore*[5] *$\lfloor x \rfloor$.*

Exemples. On a $\mathrm{E}(\pi) = 3$, $\mathrm{E}\left(\frac{3}{2}\right) = 1$, $\mathrm{E}\left(-\frac{3}{2}\right) = -2$.

Valeur absolue d'un nombre réel

Définition B.8 *On appelle* ***valeur absolue*** *du réel x, et on note $|x|$, le réel défini par $|x| = \max\{x, -x\}$.*

Le soin est laissé[6] au lecteur de démontrer la proposition ci-dessous.

Proposition B.9 *La valeur absolue possède les propriétés suivantes.*

i) $\forall x \in \mathbb{R},\ |x| = 0 \iff x = 0$.

ii) $\forall (x,y) \in \mathbb{R}^2,\ |x| \leq y \iff -y \leq x \leq y$.

iii) $\forall (x,y) \in \mathbb{R}^2,\ |x| \geq y \iff (x \geq y \text{ ou } x \leq -y)$.

iv) $\forall (x,y) \in \mathbb{R}^2,\ |xy| = |x||y|$.

v) $\forall x \in \mathbb{R},\ \forall y \in \mathbb{R}^*,\ \left|\frac{x}{y}\right| = \frac{|x|}{|y|}$.

vi) $\forall (x,y) \in \mathbb{R}^2,\ |x+y| \leq |x| + |y|$ ***(inégalité triangulaire)*** *et*

$$\forall n \in \mathbb{N}^*,\ \forall (x_1, \ldots, x_n) \in \mathbb{R}^n,\ \left|\sum_{i=1}^{n} x_i\right| \leq \sum_{i=1}^{n} |x_i|.$$

vii) $\forall (x,y) \in \mathbb{R}^2,\ ||x| - |y|| \leq |x-y|$ *(deuxième inégalité triangulaire).*

5. Cette dernière notation, d'origine anglo-saxonne, possède l'avantage de permettre la différentiation entre la partie entière par défaut (*floor* en anglais) $\lfloor x \rfloor$ d'un nombre x et la *partie entière par excès* (*ceiling* en anglais) $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$ de ce même nombre.

6. La valeur absolue étant positive, il suffit, pour prouver l'inégalité triangulaire, de remarquer que, étant toutes positives, de comparer les carrés des membres de l'inégalité. On a ainsi

$$(|x| + |y|)^2 = |x|^2 + |y|^2 + 2|x||y| = x^2 + y^2 + 2|x||y|,\ \forall (x,y) \in \mathbb{R}^2,$$

d'où $|x+y|^2 \leq (|x| + |y|)^2$ puisque $xy \leq |x||y|$.

Densité de $\mathbb{Q}$ et de $\mathbb{R}\backslash\mathbb{Q}$ dans $\mathbb{R}$

Nous examinons à présent une propriété topologique de l'ensemble $\mathbb{Q}$ vu comme une partie de $\mathbb{R}$, ainsi que son complémentaire $\mathbb{R}\backslash\mathbb{Q}$, l'ensemble de *nombres irrationnels*.

Définition B.10 *Une partie D de $\mathbb{R}$ est dite **dense** dans $\mathbb{R}$ si et seulement si*

$$\forall (x,y)\in\mathbb{R}^2,\ (x<y \implies (\exists d\in D,\ x<d<y)).$$

Théorème B.11 *Les ensembles $\mathbb{Q}$ et $\mathbb{R}\backslash\mathbb{Q}$ sont denses dans $\mathbb{R}$.*

DÉMONSTRATION. Soit $(x,y)\in\mathbb{R}^2$, tel que $x<y$, et $\varepsilon = y-x>0$. Puisque $\mathbb{R}$ est archimédien, il existe $n\in\mathbb{N}^*$ tel que $n\varepsilon>1$, c'est-à-dire $\frac{1}{n}<\varepsilon$. En notant $m=E(nx)+1$, on obtient $m-1\leq nx<m$, d'où $x<\frac{m}{n}\leq x+\frac{1}{n}<x+\varepsilon=y$, ce qui prouve le premier point. Pour démontrer le second, on sait, d'après la densité de $\mathbb{Q}$ dans $\mathbb{R}$, qu'il existe un nombre rationnel q tel que $\frac{x}{\sqrt{2}}<q<\frac{y}{\sqrt{2}}$, d'où $x<q\sqrt{2}<y$ avec $q\sqrt{2}\in\mathbb{R}\backslash\mathbb{Q}$. □

B.1.3 Intervalles

Définition B.12 *Une partie I de $\mathbb{R}$ est un **intervalle** si, dès qu'elle contient deux réels, elle contient tous les réels intermédiaires, c'est-à-dire*

$$\forall (a,b)\in I^2,\ \forall x\in\mathbb{R},\ (a\leq x\leq b \implies x\in I).$$

L'ensemble vide $\emptyset$ et tout singleton $\{x\}$, avec x un nombre réel, sont des intervalles, puisque ces deux types d'ensemble vérifient la définition ci-dessus. La propriété de la borne supérieure permet par ailleurs de classifier tout intervalle I contenant au moins deux éléments.

Tout d'abord, si un tel intervalle est à la fois majoré et minoré, il possède une borne supérieure, que l'on désigne par $b=\sup(I)$, et une borne inférieure, $a=\inf(I)$. On a alors, $\forall x\in I,\ a\leq x\leq b$ et donc $I\subset\{x\in\mathbb{R}\mid a\leq x\leq b\}$. Réciproquement, soit x un réel tel que $a<x<b$; x n'est pas un majorant (resp. minorant) de I, donc il existe un réel z (resp. y) tel que $z>x$ (resp. $x>y$). Par conséquent, les nombres y et z appartiennent à I et on a $y<x<z$. En utilisant la définition ci-dessus, il vient que x appartient à I, qui contient donc tous les éléments compris entre a et b. Selon que les réels a et b appartiennent eux-mêmes à I ou non, on peut avoir

- $I=\{x\in\mathbb{R}\mid a\leq x\leq b\}=[a,b]$ et l'intervalle est dit *fermé borné* ou encore appelé *segment*,
- $I=\{x\in\mathbb{R}\mid a<x\leq b\}=]a,b]$ et l'intervalle est dit *borné semi-ouvert à gauche*,
- $I=\{x\in\mathbb{R}\mid a\leq x<b\}=[a,b[$ et l'intervalle est dit *borné semi-ouvert à droite*,
- $I=\{x\in\mathbb{R}\mid a<x<b\}=]a,b[$ et l'intervalle est dit *borné ouvert*.

D'autre part, si l'intervalle I est minoré et non majoré, il admet une borne inférieure que l'on note a et $\forall x\in I$, $x\geq a$. Réciproquement, soit x un réel tel que $x>a$; x n'est ni un majorant, ni un minorant de I, donc il existe y et z appartenant à A tels que $y<x<z$ et par conséquent $x\in I$. Selon que le réel a appartient ou non à I, on peut avoir

- $I=\{x\in\mathbb{R}\mid x\geq a\}=[a,+\infty[$ et l'intervalle est dit *fermé non majoré*,
- $I=\{x\in\mathbb{R}\mid x>a\}=]a,+\infty[$, l'intervalle est dit *ouvert non majoré*.

De la même façon, si l'intervalle I est majoré et non minoré, on peut avoir

- $I=\{x\in\mathbb{R}\mid x\leq b\}=]-\infty,b]$, l'intervalle est dit *fermé non minoré*,
- $I=\{x\in\mathbb{R}\mid x<b\}=]-\infty,b[$, l'intervalle est dit *ouvert non minoré*.

Enfin, dans le cas où I est non majoré et non minoré, un réel x quelconque n'est ni minorant, ni majorant de I, il existe donc des réels y et z appartenant à I tels que $y<x<z$ et $x\in I$. On a par conséquent $I=\mathbb{R}$.

En définitive, tout intervalle de $\mathbb{R}$ est de l'un des onze types que nous venons d'énoncer.

B.1.4 Droite numérique achevée

Définition B.13 *On appelle **droite numérique achevée**, et l'on note $\overline{\mathbb{R}}$, l'ensemble $\mathbb{R}\cup\{-\infty,+\infty\}$, où $-\infty$ et $+\infty$ sont deux éléments non réels, sur lequel sont prolongées la structure algébrique et la relation d'ordre total définies sur $\mathbb{R}$.*

La relation $\leq$ est étendue à $\overline{\mathbb{R}}$ de la manière suivante

$$\forall x\in\mathbb{R},\ -\infty<x<+\infty\ ;\ -\infty\leq-\infty,\ +\infty\leq+\infty.$$

On remarquera que l'addition et la multiplication ne sont définies que partiellement sur $\overline{\mathbb{R}}$; on a en effet

$$\forall x \in \mathbb{R},\ x+(+\infty)=+\infty,\ x+(-\infty)=-\infty,\ (+\infty)+(+\infty)=+\infty,\ (-\infty)+(-\infty)=-\infty,$$

d'une part et

$$\begin{array}{l}\forall x>0,\ x(+\infty)=(+\infty)x=+\infty,\ x(-\infty)=(-\infty)x=-\infty,\\ \forall x<0,\ x(+\infty)=(+\infty)x=-\infty,\ x(-\infty)=(-\infty)x=+\infty,\\ (+\infty)(+\infty)=(-\infty)(-\infty)=+\infty,\ (+\infty)(-\infty)=(-\infty)(+\infty)=-\infty,\end{array}$$

d'autre part, mais les sommes $(+\infty)+(-\infty)$ et $(-\infty)+(+\infty)$ et les produits $0(+\infty)$, $(+\infty)0$, $0(-\infty)$ et $(-\infty)0$ n'ont pas de sens. L'ensemble $\overline{\mathbb{R}}$ n'est donc pas un anneau.

Nous terminons par un résultat admis, analogue à la proposition B.4.

Proposition B.14 *Toute partie non vide de $\overline{\mathbb{R}}$ admet une borne supérieure et une borne inférieure dans $\overline{\mathbb{R}}$.*

B.2 Suites numériques

Une *suite numérique* est une application de $\mathbb{N}$ dans un corps $\mathbb{K}$, avec $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$. Plutôt que de noter

$$\begin{array}{rccc} u: & \mathbb{N} & \to & \mathbb{K} \\ & n & \mapsto & u(n)\end{array}$$

on emploie les notations $(u_n)_{n\in\mathbb{N}}$, $(u_n)_{n\geq 0}$ ou encore $(u_n)_n$. Pour chaque entier n, le nombre u_n est appelé le n^e *terme* de la suite[7]. L'ensemble des suites numériques est noté $\mathscr{F}(\mathbb{N},\mathbb{K})$ ou $\mathbb{K}^{\mathbb{N}}$. Une *suite réelle* (resp. *complexe*) est une suite numérique telle que

$$\forall n\in\mathbb{N},\ u_n\in\mathbb{R}\ (\text{resp. }\mathbb{C}).$$

On rappelle dans cette section plusieurs définitions et résultats importants sur les suites numériques, en mettant particulièrement l'accent sur le cas des suites réelles, car celles-ci sont au cœur des différentes méthodes itératives présentées dans le cours. Dans toute la suite, le symbole $|\cdot|$ désignera soit la valeur absolue soit le module d'un nombre selon que $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$.

B.2.1 Premières définitions et propriétés

Définitions B.15 *Une suite numérique $(u_n)_{n\in\mathbb{N}}$ est dite **constante** si et seulement si*

$$\forall n\in\mathbb{N},\ u_{n+1}=u_n.$$

*Elle est dite **stationnaire** si et seulement si elle est constante à partir d'un certain rang, c'est-à-dire si*

$$\exists N\in\mathbb{N},\ \forall n\in\mathbb{N},\ (n\geq N\implies u_{n+1}=u_n).$$

*Enfin, la suite est dite **périodique** si et seulement s'il existe un entier p strictement positif tel que*

$$\forall n\in\mathbb{N},\ u_{n+p}=u_n.$$

Définition B.16 *Une suite numérique $(u_n)_{n\in\mathbb{N}}$ est dite **bornée** si et seulement s'il existe un réel positif M tel que*

$$\forall n\in\mathbb{N},\ |u_n|\leq M.$$

Définitions B.17 *Une suite réelle $(u_n)_{n\in\mathbb{N}}$ est dite **majorée** (resp. **minorée**) si et seulement s'il existe un réel M (resp. m), appelé **majorant** (resp. **minorant**) de la suite $(u_n)_{n\in\mathbb{N}}$, tel que*

$$\forall n\in\mathbb{N},\ u_n\leq M\ (\textit{resp. } u_n\geq m).$$

On voit qu'une suite réelle est bornée si et seulement si elle est majorée et minorée.

7. On veillera à ne pas confondre la suite $(u_n)_{n\in\mathbb{N}}$ et son terme général u_n.

Opérations sur les suites

On peut définir sur l'ensemble des suites numériques $\mathbb{K}^{\mathbb{N}}$ une *addition*,

$$(w_n)_{n\in\mathbb{N}} = (u_n)_{n\in\mathbb{N}} + (v_n)_{n\in\mathbb{N}} \Leftrightarrow \forall n \in \mathbb{N},\ w_n = u_n + v_n,$$

une *multiplication interne*,

$$(w_n)_{n\in\mathbb{N}} = (u_n)_{n\in\mathbb{N}}(v_n)_{n\in\mathbb{N}} \Leftrightarrow \forall n \in \mathbb{N},\ w_n = u_n v_n,$$

et tune *multiplication externe par les scalaires*,

$$\forall \lambda \in \mathbb{K},\ (w_n)_{n\in\mathbb{N}} = \lambda(u_n)_{n\in\mathbb{N}} \Leftrightarrow \forall n \in \mathbb{N},\ w_n = \lambda u_n.$$

L'addition est commutative, associative et admet pour élément neutre la suite constante nulle. Tout suite $(u_n)_{n\in\mathbb{N}}$ a une suite opposée $(-u_n)_{n\in\mathbb{N}}$ et $(\mathbb{K}^{\mathbb{N}}, +)$ est donc un groupe commutatif. La multiplication interne commutative, associative et admet pour élément neutre la suite constante égale à 1. Elle est distributive par rapport à l'addition. Il existe cependant des suites non nulles n'ayant pas d'inverse et $(\mathbb{K}^{\mathbb{N}}, \cdot)$ n'est par conséquent pas un groupe.

Suites réelles monotones

Définitions B.18 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite réelle. On dit que $(u_n)_{n\in\mathbb{N}}$ est*

- ***croissante** (resp. **décroissante**) si et seulement si*

$$\forall n \in \mathbb{N},\ u_n \leq u_{n+1}\ (\textit{resp. } u_{n+1} \leq u_n),$$

- ***strictement croissante** (resp. **strictement décroissante**) si et seulement si*

$$\forall n \in \mathbb{N},\ u_n < u_{n+1}\ (\textit{resp. } u_{n+1} < u_n),$$

- ***monotone** si et seulement si elle est croissante ou décroissante,*
- ***strictement monotone** si et seulement si elle est strictement croissante ou strictement décroissante.*

Définition B.19 (suite extraite) *Soit une suite $(u_n)_{n\in\mathbb{N}}$. On appelle **suite extraite de** $(u_n)_{n\in\mathbb{N}}$ toute suite $(u_{\sigma(n)})_{n\in\mathbb{N}}$, où $\sigma : \mathbb{N} \to \mathbb{N}$ est une application strictement croissante, appelée **extractrice**.*

On donne aussi le nom de *sous-suite de* $(u_n)_{n\in\mathbb{N}}$ à toute suite extraite de $(u_n)_{n\in\mathbb{N}}$. On montre aisément, par récurrence, que pour toute extractrice σ, on a

$$\forall n \in \mathbb{N},\ \sigma(n) \geq n.$$

Exemples. Les suites $(u_{2n})_{n\in\mathbb{N}}$, $(u_{2n+1})_{n\in\mathbb{N}}$ et $(u_{n^2})_{n\in\mathbb{N}}$ sont toutes trois des suites extraites de $(u_n)_{n\in\mathbb{N}}$.

Définition B.20 (suite de Cauchy) *Une suite $(u_n)_{n\in\mathbb{N}}$ est dite **de Cauchy** si et seulement si*

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \ \forall (p,q) \in \mathbb{N}^2,\ (p \geq N,\ q \geq N \implies \left|u_p - u_q\right| \leq \varepsilon).$$

Proposition B.21 *Toute suite de Cauchy est bornée.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite de Cauchy. On a

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \ \forall (p,q) \in \mathbb{N}^2,\ (p \geq N,\ q \geq N \implies \left|u_p - u_q\right| \leq \varepsilon).$$

Fixons une valeur particulière pour ε, par exemple $\varepsilon = 1$. Il existe alors un entier N' tel que $(p \geq N',\ q \geq N' \implies \left|u_p - u_q\right| \leq 1)$, ou encore, pour $q = N'$, $(p \geq N' \implies \left|u_p - u_{N'}\right| \leq 1)$, c'est-à-dire

$$\forall n \in \mathbb{N},\ (n \geq N' \implies u_{N'} - 1 \leq u_n \leq u_{N'} + 1).$$

Notons $a = \min\{u_0, \ldots, u_{N'-1}, u_{N'} - 1\}$ et $b = \max\{u_0, \ldots, u_{N'-1}, u_{N'} + 1\}$. Nous avons alors

$$\forall n \in \mathbb{N},\ a \leq u_n \leq b,$$

et, par suite, $(u_n)_{n\in\mathbb{N}}$ est bornée. □

B.2.2 Convergence d'une suite

Définitions B.22 *On dit qu'une suite numérique* $(u_n)_{n\in\mathbb{N}}$ ***est convergente*** *si et seulement si*

$$\exists l \in \mathbb{K},\ \forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \in \mathbb{N}, (n \geq N \implies |u_n - l| \leq \varepsilon).$$

On dit encore que la suite $(u_n)_{n\in\mathbb{N}}$ ***converge vers*** l *ou* ***tend vers*** l*. Le scalaire* l *est appelé* ***limite*** *de la suite et l'on note* $\lim\limits_{n\to+\infty} u_n = l$.

En revanche, on dit qu'une suite numérique $(u_n)_{n\in\mathbb{N}}$ ***diverge*** *ou* ***est divergente*** *si et seulement si elle ne converge pas, c'est-à-dire*

$$\forall l \in \mathbb{K},\ \exists \varepsilon > 0,\ \forall N \in \mathbb{N},\ \exists n \in \mathbb{N}, (n \geq N \text{ et } |u_n - l| \geq \varepsilon).$$

Proposition B.23 *La limite d'une suite, si elle existe, est unique.*

DÉMONSTRATION. Supposons que $(u_n)_{n\in\mathbb{N}}$ converge à la fois vers l et vers l', avec $l \neq l'$. Posons $\varepsilon = \frac{1}{3}\left|l - l'\right|$. Par définition de la convergence, il existe des entiers N et N' tels que

$$\forall n \in \mathbb{N},\ (n \geq N \implies |u_n - l| \leq \varepsilon) \text{ et } (n \geq N' \implies \left|u_n - l'\right| \leq \varepsilon).$$

Soit $n \geq \max(N, N')$, nous avons alors $|u_n - l| \leq \varepsilon$ et $|u_n - l'| \leq \varepsilon$, d'où

$$\left|l - l'\right| \leq |l - u_n| + \left|u_n - l'\right| \leq 2\varepsilon = \frac{2}{3}\left|l - l'\right|,$$

ce qui est absurde. □

La notion de limite d'une suite que l'on vient d'introduire peut être étendue de la manière suivante dans le cas d'une suite réelle.

Définition B.24 *Soit* $(u_n)_{n\in\mathbb{N}}$ *une suite réelle. On dit que* $(u_n)_{n\in\mathbb{N}}$ ***tend vers*** $+\infty$ *(resp.* ***tend vers*** $-\infty$*) si et seulement si*

$$\forall A \in \mathbb{R},\ \exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies u_n \geq A) \text{ (resp. } (n \geq N \implies u_n \leq A)).$$

On note alors $\lim\limits_{n\to+\infty} u_n = +\infty$ *(resp.* $\lim\limits_{n\to+\infty} u_n = -\infty$*).*

La limite d'une suite étant définie, nous pouvons introduire la notion de *suites adjacentes*.

Définition B.25 (suites adjacentes) *Deux suites réelles* $(u_n)_{n\in\mathbb{N}}$ *et* $(v_n)_{n\in\mathbb{N}}$ *sont dites* ***adjacentes*** *si et seulement si*

i) l'une est croissante et l'autre est décroissante,

ii) $\lim\limits_{n\to+\infty} (u_n - v_n) = 0.$

L'intérêt des suites adjacentes provient de la propriété suivante [8].

Proposition B.26 *Deux suites adjacentes convergent et ont même limite.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites adjacentes. Supposons $(u_n)_{n\in\mathbb{N}}$ croissante et $(v_n)_{n\in\mathbb{N}}$ décroissante. La suite $(u_n - v_n)_{n\in\mathbb{N}}$ est donc croissante et tend par hypothèse vers 0, on en déduit que c'est une suite négative et, $\forall n \in \mathbb{N}$, $u_n \leq v_n$. De plus, $\forall n \in \mathbb{N}$, $u_0 \leq u_n$ et $v_n \leq v_0$. En combinant ces inégalités, nous obtenons

$$\forall n \in \mathbb{N},\ u_0 \leq u_n \leq v_n \leq v_0.$$

La suite $(u_n)_{n\in\mathbb{N}}$ est alors croissante et majorée par v_0, c'est donc une suite convergente. De même, la suite $(v_n)_{n\in\mathbb{N}}$ est décroissante et minorée par u_0, donc $(v_n)_{n\in\mathbb{N}}$ est convergente.
D'autre part, on a $\lim\limits_{n\to+\infty} (u_n - v_n) = 0$ et, comme les deux suites sont convergentes, on en déduit $\lim\limits_{n\to+\infty} u_n = \lim\limits_{n\to+\infty} v_n$. □

Ce dernier résultat permet d'établir le théorème suivant.

8. On observera dans la preuve que des suites adjacentes fournissent également un encadrement aussi précis que souhaité de leur limite puisqu'on a

$$u_n \leq u_{n+1} \leq l \leq v_{n+1} \leq v_n,\ \forall n \in \mathbb{N}.$$

Théorème B.27 (« théorème des segments emboîtés ») *Soit $([a_n, b_n])_{n\in\mathbb{N}}$ une suite de segments emboîtés (c'est-à-dire, $\forall n \in \mathbb{N}$, $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$) telle que $\lim_{n\to+\infty}(b_n - a_n) = 0$. Alors l'intersection $\bigcap_{n\in\mathbb{N}}[a_n, b_n]$ est un singleton.*

DÉMONSTRATION. Notons que les suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont adjacentes. On en déduit que qu'elles convergent vers une même limite l et l'on a $\forall n \in \mathbb{N}$, $a_n \leq l \leq b_n$ donc $l \in [a_n, b_n]$ pour tout entier n, et, par suite, $l \in \bigcap_{n\in\mathbb{N}}[a_n, b_n]$. D'autre part si $l' \in \bigcap_{n\in\mathbb{N}}[a_n, b_n]$, alors $l' \in [a_n, b_n]$ pour tout entier n. Comme on a également $l \in [a_n, b_n]$ pour tout entier n, nous obtenons $\forall n \in \mathbb{N}$, $b_n - a_n \geq |l - l'|$. En faisant tendre n vers l'infini, nous trouvons $l = l'$. En conclusion, on a $\bigcap_{n\in\mathbb{N}}[a_n, b_n] = \{l\}$. □

Propriétés des suites convergentes

Proposition B.28 *Toute suite numérique convergente est bornée.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite convergente de limite l et ε un réel strictement positif. Il existe alors un entier N tel que, pour tout $n \in \mathbb{N}$, si $n \geq N$ alors $|u_n - l| \leq \varepsilon$. L'ensemble $\{|u_n| \mid n \geq N\}$ est donc majoré par $|l| + \varepsilon$. Il vient par conséquent

$$\forall n \in \mathbb{N},\ |u_n| \leq \max\{|u_0|, \ldots, |u_{N-1}|, |l| + \varepsilon\},$$

et la suite $(u_n)_{n\in\mathbb{N}}$ est donc bornée. □

La réciproque de cette proposition est fausse, comme le montre l'exemple de la suite définie par $u_n = (-1)^n$ pour tout entier $n \geq 0$.

Proposition B.29 *Si une suite numérique $(u_n)_{n\in\mathbb{N}}$ est convergente, toute suite extraite de $(u_n)_{n\in\mathbb{N}}$ est convergente et tend vers la même limite.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite convergente de limite l. Nous avons alors

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies |u_n - l| \leq \varepsilon).$$

Soit σ une extractrice. Nous savons que, pour tout entier naturel n, $\sigma(n) \geq n$, donc si $n \geq N$ alors $\sigma(n) \geq \sigma(N) \geq N$ et, par suite, $|u_{\sigma(n)} - l| \leq \varepsilon$. On en conclut que

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies \left|u_{\sigma(n)} - l\right| \leq \varepsilon),$$

et donc que la suite extraite $(u_{\sigma(n)})_{n\in\mathbb{N}}$ converge vers l. □

La contraposée de cette dernière proposition permet de montrer qu'une suite diverge : il suffit pour cela d'en extraire deux suites qui convergent vers deux limites différentes.

Exemple. Soit $(u_n)_{n\in\mathbb{N}}$ la suite définie par $u_n = (-1)^n$ pour tout entier $n \geq 0$. La suite extraite $(u_{2n})_{n\in\mathbb{N}}$ a pour limite 1 et la suite $(u_{2n+1})_{n\in\mathbb{N}}$ a pour limite -1. Cette suite diverge donc.

En revanche, sa réciproque est en général fausse ; en effet, on peut trouver des suites divergentes qui admettent pourtant deux suites extraites convergeant vers une même limite.

Exemple. Soit la suite réelle définie par $u_n = \cos\left((n + (-1)^n)\frac{\pi}{3}\right)$ pour tout entier positif n. Nous avons $u_{6n} = \cos\left((6n+1)\frac{\pi}{3}\right) = \cos\left(2n\pi + \frac{\pi}{3}\right) = \frac{1}{2}$ et $u_{6n+4} = \cos\left((6n+5)\frac{\pi}{3}\right) = \cos\left(2n\pi + \frac{5\pi}{3}\right) = \frac{1}{2}$, mais la suite $(u_n)_{n\in\mathbb{N}}$ est divergente car $u_{3n+6} = \cos\left((6n+2)\frac{\pi}{3}\right) = \cos\left(2n\pi + \frac{2\pi}{3}\right) = -\frac{1}{2}$.

Cependant, on a le résultat suivant.

Proposition B.30 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite numérique et l un scalaire. Pour que $(u_n)_{n\in\mathbb{N}}$ converge vers l, il faut et il suffit que les suites $(u_{2n})_{n\in\mathbb{N}}$ et $(u_{2n+1})_{n\in\mathbb{N}}$ convergent toutes deux vers l.*

DÉMONSTRATION. Supposons que les suites extraites $(u_{2n})_{n\in\mathbb{N}}$ et $(u_{2n+1})_{n\in\mathbb{N}}$ convergent vers une limite l. Soit ε un réel strictement positif. Il existe des entiers N et N' tels que, pour tout entier n,

$$(n \geq N \implies |u_{2n} - l| \leq \varepsilon) \text{ et } (n \geq N' \implies |u_{2n+1} - l| \leq \varepsilon).$$

Notons $N'' = \max(2N, 2N' + 1)$ et soit $p \in \mathbb{N}$ tel que $p \geq N''$. Si p est pair, il existe n tel que $p = 2n$. Dans ce cas, nous avons $2n \geq 2N$ donc $n \geq N$, d'où

$$\left|u_p - l\right| = |u_{2n} - l| \leq \varepsilon.$$

Si p est impair, il existe n tel que $p = 2n+1$. Nous avons alors $2n+1 \geq 2N'+1$ donc $n \geq N'$, d'où

$$\left|u_p - l\right| = |u_{2n+1} - l| \leq \varepsilon.$$

Ceci montre que la suite $(u_p)_{p\in\mathbb{N}}$ converge vers l. □

Lorsqu'une suite numérique diverge, il peut exister des points auprès desquels s'accumulent une infinité de termes de la suite. On introduit alors la notion suivante.

Définition B.31 (valeur d'adhérence d'une suite) *Soit $(u_n)_{n\in\mathbb{N}}$ une suite numérique et a un scalaire. On dit que a est une **valeur d'adhérence** de la suite $(u_n)_{n\in\mathbb{N}}$ s'il existe une sous-suite de $(u_n)_{n\in\mathbb{N}}$ qui converge vers a.*

Proposition B.32 *Toute suite numérique convergente est de Cauchy.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite convergente de limite l et ε un réel strictement positif. Il existe alors un entier N tel que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_n - l| \leq \frac{\varepsilon}{2}\right).$$

Soit p et q deux entiers supérieurs à N ; nous avons alors

$$\left|u_p - u_q\right| \leq \left|u_p - l\right| + \left|u_q - l\right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

d'où

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall (p,q) \in \mathbb{N}^2,\ (p \geq N,\ q \geq N \implies \left|u_p - u_q\right| \leq \varepsilon),$$

et la suite $(u_n)_{n\in\mathbb{N}}$ est une suite de Cauchy. □

Nous montrerons plus loin que la réciproque de cette proposition est vraie. Concluons maintenant cette section par quelques propriétés d'ordre des suites réelles convergentes.

Proposition B.33 *Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles convergentes. Si $u_n \leq v_n$ pour tout entier n, alors*

$$\lim_{n\to+\infty} u_n \leq \lim_{n\to+\infty} v_n.$$

DÉMONSTRATION. Supposons que $\lim_{n\to+\infty} u_n > \lim_{n\to+\infty} v_n$. Nous avons alors $\lim_{n\to+\infty} (u_n - v_n) > 0$, ce qui entraîne

$$\exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies (u_n - v_n) > 0),$$

ce qui contredit l'hypothèse. □

Même si l'on a $u_n < v_n$ pour tout $n \in \mathbb{N}$, on peut avoir $\lim_{n\to+\infty} u_n \leq \lim_{n\to+\infty} v_n$ car le passage à la limite élargit les inégalités, comme l'illustre l'exemple suivant.

Exemple. Soit les suites définies par

$$\forall n \in \mathbb{N},\ u_n = 1 - \frac{1}{n+1} \text{ et } v_n = 1 + \frac{1}{n+1}.$$

Nous avons, pour tout entier naturel n, $u_n < v_n$ et $\lim_{n\to+\infty} u_n = \lim_{n\to+\infty} v_n = 1$.

Proposition B.34 (« théorème des gendarmes ») *Soit $(u_n)_{n\in\mathbb{N}}$, $(v_n)_{n\in\mathbb{N}}$ et $(w_n)_{n\in\mathbb{N}}$ trois suites réelles telles que*

$$\exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies u_n \leq v_n \leq w_n).$$

Si $(u_n)_{n\in\mathbb{N}}$ et $(w_n)_{n\in\mathbb{N}}$ convergent vers la même limite l, alors la suite $(v_n)_{n\in\mathbb{N}}$ converge aussi vers l.

DÉMONSTRATION. Soit ε un réel strictement positif. Puisque $(u_n)_{n\in\mathbb{N}}$ et $(w_n)_{n\in\mathbb{N}}$ convergent toutes deux vers l, il existe des entiers N et N' tels que

$$\forall n \in \mathbb{N},\ (n \geq N \implies |u_n - l| \leq \varepsilon) \text{ et } (n \geq N' \implies |w_n - l| \leq \varepsilon).$$

En notant $N'' = \max(N, N')$, nous avons

$$\forall n \in \mathbb{N},\ n \geq N'' \implies \left\{\begin{array}{l} u_n \leq v_n \leq w_n \\ |u_n - l| \leq \varepsilon \\ |w_n - l| \leq \varepsilon \end{array}\right. \implies -\varepsilon \leq u_n - l \leq v_n - l \leq w_n - l \leq \varepsilon \implies |v_n - l| \leq \varepsilon,$$

ce qui montre la convergence de $(v_n)_{n\in\mathbb{N}}$ vers l. □

Exemple. Étant donné un réel x, soit la suite $(u_n)_{n\in\mathbb{N}}$ définie par

$$\forall n \in \mathbb{N}^*,\ u_n = \frac{\cos(x)}{n}.$$

On a $-1 \leq \cos x \leq 1$, d'où, $\forall n \in \mathbb{N}^*$, $-\frac{1}{n} \leq u_n \leq \frac{1}{n}$. Comme $\lim\limits_{n\to+\infty}\left(-\frac{1}{n}\right) = \lim\limits_{n\to+\infty}\frac{1}{n} = 0$, on a finalement $\lim\limits_{n\to+\infty} u_n = 0$.

Proposition B.35 *Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles telles que $u_n \leq v_n$ pour tout entier n. On a les assertions suivantes.*

i) Si $\lim\limits_{n\to+\infty} u_n = +\infty$ alors $\lim\limits_{n\to+\infty} v_n = +\infty$.

ii) Si $\lim\limits_{n\to+\infty} v_n = -\infty$ alors $\lim\limits_{n\to+\infty} u_n = -\infty$.

DÉMONSTRATION. Prouvons i. Soit $A \in \mathbb{R}$. Comme $\lim\limits_{n\to+\infty} u_n = +\infty$, il existe un entier N tel que

$$\forall n \in \mathbb{N},\ (n \geq N \implies u_n \geq A),$$

et, *a fortiori*, $v_n \geq A$ d'après l'hypothèse, ce qui implique que $\lim\limits_{n\to+\infty} v_n = +\infty$.
La preuve de ii s'obtient de manière analogue à celle de i. □

Propriétés algébriques des suites convergentes

Proposition B.36 *Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites numériques, $\lambda \in \mathbb{K}$ et $(l, l') \in \mathbb{K}^2$. On a les assertions suivantes.*

i) $\lim\limits_{n\to+\infty} u_n = l \implies \lim\limits_{n\to+\infty} |u_n| = |l|$.

ii) $\lim\limits_{n\to+\infty} u_n = l$ *et* $\lim\limits_{n\to+\infty} v_n = l' \implies \lim\limits_{n\to+\infty} (u_n + v_n) = l + l'$.

iii) $\lim\limits_{n\to+\infty} u_n = l \implies \lim\limits_{n\to+\infty} \lambda u_n = \lambda l$,

iv) $\lim\limits_{n\to+\infty} u_n = 0$ *et* $(v_n)_{n\in\mathbb{N}}$ *bornée* $\implies \lim\limits_{n\to+\infty} u_n v_n = 0$.

v) $\lim\limits_{n\to+\infty} u_n = l$ *et* $\lim\limits_{n\to+\infty} v_n = l' \implies \lim\limits_{n\to+\infty} u_n v_n = ll'$.

vi) $\lim\limits_{n\to+\infty} v_n = l'$ *et* $l' \neq 0 \implies \frac{1}{v_n}$ *est défini à partir d'un certain rang et* $\lim\limits_{n\to+\infty} \frac{1}{v_n} = \frac{1}{l'}$.

vii) $\lim\limits_{n\to+\infty} u_n = l$ *et* $\lim\limits_{n\to+\infty} v_n = l'$ *et* $l' \neq 0 \implies \frac{u_n}{v_n}$ *est défini à partir d'un certain rang et* $\lim\limits_{n\to+\infty} \frac{u_n}{v_n} = \frac{l}{l'}$.

DÉMONSTRATION.

i) Soit $\varepsilon > 0$. Puisque la suite $(u_n)_{n\in\mathbb{N}}$ tend vers l, il existe $N \in \mathbb{N}$ tel que

$$\forall n \in \mathbb{N},\ (n \geq N \implies |u_n - l| \leq \varepsilon).$$

Or, on a, $\forall n \in \mathbb{N}$, l'inégalité $||u_n| - |l|| \leq |u_n - l|$, dont on déduit que

$$\forall n \in \mathbb{N},\ (n \geq N \implies ||u_n| - |l|| \leq \varepsilon),$$

et donc $\lim\limits_{n\to+\infty} |u_n| = |l|$.

ii) Soit $\varepsilon > 0$. Puisque les suites $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ convergent respectivement vers l et l', il existe des entiers N et N' tels que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_n - l| \leq \frac{\varepsilon}{2}\right) \text{ et } \left(n \geq N' \implies \left|v_n - l'\right| \leq \frac{\varepsilon}{2}\right).$$

En notant $N'' = \max(N, N')$, nous avons

$$\forall n \in \mathbb{N},\ \left(n \geq N'' \implies \left|(u_n + v_n) - (l + l')\right| = \left|(u_n - l) + (v_n - l')\right| \leq |u_n - l| + \left|v_n - l'\right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\right),$$

d'où $\lim\limits_{n\to+\infty} (u_n + v_n) = l + l'$.

iii) Soit $\varepsilon > 0$. Puisque la suite $(u_n)_{n\in\mathbb{N}}$ tend vers l, il existe $N \in \mathbb{N}$ tel que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_n - l| \leq \frac{\varepsilon}{|\lambda|+1}\right),$$

d'où, $\forall n \in \mathbb{N},\ \left(n \geq N \implies |\lambda u_n - \lambda l| = |\lambda|\,|u_n - l| \leq \dfrac{|\lambda|\,\varepsilon}{|\lambda|+1} \leq \varepsilon\right)$, et donc $\lim\limits_{n\to+\infty} \lambda u_n = \lambda l$.

iv) Par hypothèse, il existe $M \in \mathbb{R}_+$ tel que, $\forall n \in \mathbb{N}$, $|v_n| \leq M$. Soit $\varepsilon > 0$. Puisque la suite $(u_n)_{n\in\mathbb{N}}$ tend vers 0, il existe $N \in \mathbb{N}$ tel que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_n| \leq \frac{\varepsilon}{M+1}\right).$$

Nous avons alors, $\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_n v_n| = |u_n|\,|v_n| \leq \dfrac{M\varepsilon}{M+1} \leq \varepsilon\right)$ et donc $\lim\limits_{n\to+\infty} u_n v_n = 0$.

v) Notons, pour tout $n \in \mathbb{N}$, $w_n = u_n - l$. Nous avons

$$\forall n \in \mathbb{N},\ u_n v_n = (w_n + l)v_n = w_n v_n + l v_n.$$

D'après iii, $\lim\limits_{n\to+\infty} l v_n = ll'$. D'autre part, $\lim\limits_{n\to+\infty} w_n = 0$ et $(v_n)_{n\in\mathbb{N}}$ est bornée puisque $(v_n)_{n\in\mathbb{N}}$ est convergente, donc $\lim\limits_{n\to+\infty} w_n v_n = 0$ d'après iv. Finalement, on a $\lim\limits_{n\to+\infty} u_n v_n = ll'$ d'après ii.

vi) Puisque $(v_n)_{n\in\mathbb{N}}$ converge vers l', la suite $(|v_n|)_{n\in\mathbb{N}}$ converge vers $|l'|$. Il existe donc un entier N tel que, pour tout entier n, $\left(n \geq N \implies |v_n| \geq \dfrac{|l'|}{2}\right)$ (il suffit de choisir $\varepsilon = \dfrac{|l'|}{2}$). En particulier, $\forall n \in \mathbb{N}$, $(n \geq N \implies v_n \neq 0)$, et la suite $\left(\dfrac{1}{v_n}\right)_{n\in\mathbb{N}}$ est donc définie. Nous avons alors, pour tout entier n tel que $n \geq N$

$$0 \leq \left|\frac{1}{v_n} - \frac{1}{l'}\right| = \frac{|v_n - l'|}{|v_n|\,|l'|} \leq \frac{2}{|l'|^2}\left|v_n - l'\right|.$$

Comme $\lim\limits_{n\to+\infty} v_n = l'$, on en déduit que $\lim\limits_{n\to+\infty}\left(\dfrac{2}{|l'|^2}\left|v_n - l'\right|\right) = 0$, puis que $\lim\limits_{n\to+\infty}\left|\dfrac{1}{v_n} - \dfrac{1}{l'}\right| = 0$, soit encore $\lim\limits_{n\to+\infty} \dfrac{1}{v_n} = \dfrac{1}{l'}$.

vii) Il suffit d'appliquer v et vi en remarquant que $\dfrac{u_n}{v_n} = u_n \dfrac{1}{v_n}$.

□

Proposition B.37 *Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles. On a les assertions suivantes.*

i) Si $\lim\limits_{n\to+\infty} u_n = +\infty$ et $(v_n)_{n\in\mathbb{N}}$ est minorée, alors $\lim\limits_{n\to+\infty}(u_n + v_n) = +\infty$.
En particulier, on a
— *$\lim\limits_{n\to+\infty} u_n = +\infty$ et $\lim\limits_{n\to+\infty} v_n = +\infty \implies \lim\limits_{n\to+\infty}(u_n + v_n) = +\infty$,*
— *$\lim\limits_{n\to+\infty} u_n = +\infty$ et $\lim\limits_{n\to+\infty} v_n = l' \implies \lim\limits_{n\to+\infty}(u_n + v_n) = +\infty$.*

ii) Si $\lim\limits_{n\to+\infty} u_n = +\infty$ et $\left(\exists C \in \mathbb{R}_+^*,\ \exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies v_n \geq C)\right)$, alors $\lim\limits_{n\to+\infty} u_n v_n = +\infty$.
En particulier, on a
— *$\lim\limits_{n\to+\infty} u_n = +\infty$ et $\lim\limits_{n\to+\infty} v_n = +\infty \implies \lim\limits_{n\to+\infty} u_n v_n = +\infty$,*
— *$\lim\limits_{n\to+\infty} u_n = +\infty$ et $\lim\limits_{n\to+\infty} v_n = l' \in \mathbb{R}_+^* \implies \lim\limits_{n\to+\infty} u_n v_n = +\infty$.*

iii) $\lim\limits_{n\to+\infty} u_n = +\infty \implies \lim\limits_{n\to+\infty} \dfrac{1}{u_n} = 0$.

iv) Si $\lim\limits_{n\to+\infty} u_n = 0$ et si $(\exists N \in \mathbb{N},\ \forall n \in \mathbb{N},\ (n \geq N \implies u_n > 0))$, alors $\lim\limits_{n\to+\infty} \dfrac{1}{u_n} = +\infty$.

DÉMONSTRATION.

i) Par hypothèse, il existe $m \in \mathbb{R}$ tel que

$$\forall n \in \mathbb{N},\ v_n \geq m.$$

Soit $A > 0$. Puisque $\lim\limits_{n\to+\infty} u_n = +\infty$, il existe $N \in \mathbb{N}$ tel que, $\forall n \in \mathbb{N}$, $(n \geq N \implies u_n \geq A - m)$. Nous avons alors

$$\forall n \in \mathbb{N},\ (n \geq N \implies u_n + v_n \geq (A-m) + m),$$

et donc $\lim\limits_{n\to+\infty}(u_n + v_n) = +\infty$.

ii) Soit $A>0$. Puisque $\lim_{n\to+\infty} u_n = +\infty$, il existe $N' \in \mathbb{N}$ tel que, $\forall n \in \mathbb{N}$, $\left(n \geq N' \implies u_n \geq \frac{A}{C}\right)$.
En notant $N'' = \max(N, N')$, nous avons alors

$$\forall n \in \mathbb{N},\ \left(n \geq N'' \implies \left(u_n \geq \frac{A}{C} \text{ et } v_n \geq C\right) \implies u_n v_n \geq A\right),$$

et donc $\lim_{n\to+\infty} u_n v_n = +\infty$.

iii) Soit $\varepsilon > 0$. Puisque $\lim_{n\to+\infty} u_n = +\infty$, il existe $N \in \mathbb{N}$ tel que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies u_n \geq \frac{1}{\varepsilon}\right),$$

d'où $\lim_{n\to+\infty} \frac{1}{u_n} = 0$.

iv) Soit $A>0$. Puisque $\lim_{n\to+\infty} u_n = 0$, il existe $N' \in \mathbb{N}$ tel que, $\forall n \in \mathbb{N}$, $\left(n \geq N' \implies |u_n| \geq \frac{1}{A}\right)$.
En notant $N'' = \max(N, N')$, nous avons alors

$$\forall n \in \mathbb{N},\ \left(n \geq N'' \implies \left(|u_n| \geq \frac{1}{A} \text{ et } u_n \geq 0\right) \implies \frac{1}{u_n} \geq A\right),$$

et donc $\lim_{n\to+\infty} \frac{1}{u_n} = +\infty$.

□

Certains des résultats des deux dernières propositions sont résumés dans les tableaux B.1 et B.2.

$(v_n)_{n\in\mathbb{N}}$ \ $(u_n)_{n\in\mathbb{N}}$	l	$+\infty$	$-\infty$
l'	$l+l'$	$+\infty$	$-\infty$
$+\infty$	$+\infty$	$+\infty$	FI
$-\infty$	$-\infty$	FI	$-\infty$

TABLE B.1 – limites possibles pour la suite $(u_n + v_n)_{n\in\mathbb{N}}$ en fonction des limites respectives des suites réelles $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$.

Dans ceux-ci, les lettres « FI » correspondent à une *forme indéterminée* qu'il faudra chercher à lever. Différents cas sont possibles ; pour un produit de suites, la limite peut

— être finie, par exemple, pour tout $n \in \mathbb{N}^*$, $u_n = n$ et $v_n = \frac{1}{n}$, $\lim_{n\to+\infty} u_n v_n = 1$,

— être infinie, par exemple, pour tout $n \in \mathbb{N}$, $u_n = n^2$ et $v_n = \frac{1}{n}$, $\lim_{n\to+\infty} u_n v_n = +\infty$,

— ne pas exister, par exemple, pour tout $n \in \mathbb{N}^*$, $u_n = n$ et $v_n = \frac{(\sin n)^2}{n}$, alors $u_n v_n = (\sin n)^2$ qui ne possède pas de limite.

$(v_n)_{n\in\mathbb{N}}$ \ $(u_n)_{n\in\mathbb{N}}$	$l>0$	$l<0$	$l=0$	$+\infty$	$-\infty$
$l'>0$	ll'	ll'	0	$+\infty$	$-\infty$
$l'<0$	ll'	ll'	0	$-\infty$	$+\infty$
$l'=0$	0	0	0	FI	FI
$+\infty$	$+\infty$	$-\infty$	FI	$+\infty$	$-\infty$
$-\infty$	$-\infty$	$+\infty$	FI	$-\infty$	$+\infty$

TABLE B.2 – limites possibles pour la suite $(u_n v_n)_{n\in\mathbb{N}}$ en fonction des limites respectives des suites réelles $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$.

B.2.3 Existence de limite

Nous rassemblons dans cette section plusieurs résultats relatifs à l'existence de la limite d'une suite numérique.

Proposition B.38 *i) Toute suite réelle croissante et majorée est convergente.*

ii) Toute suite réelle décroissante et minorée est convergente.

DÉMONSTRATION.

i) Soit $(u_n)_{n\in\mathbb{N}}$ une suite réelle croissante et majorée. L'ensemble $\{u_k \mid k \in \mathbb{N}\}$ des termes de la suite est une partie de $\mathbb{R}$ non vide et majorée, qui admet donc une borne supérieure, notée l. On a alors $u_n \leq l$, pour tout entier n, et pour tout réel ε strictement positif, $l-\varepsilon$ n'est pas un majorant de l'ensemble $\{u_k \mid k \in \mathbb{N}\}$. Il existe alors $N \in \mathbb{N}$ tel que

$$l - \varepsilon \leq u_N \leq l.$$

La suite $(u_n)_{n\in\mathbb{N}}$ étant croissante, on en déduit

$$\forall n \in \mathbb{N},\ (n \geq N \implies l - \varepsilon \leq u_N \leq u_n \leq l),$$

et donc $|x_n - l| \leq \varepsilon$. On en conclut que $(u_n)_{n\in\mathbb{N}}$ converge vers l.

ii) Il suffit d'appliquer le résultat précédent à la suite $(-u_n)_{n\in\mathbb{N}}$.

□

Proposition B.39 *i) Toute suite réelle croissante et non majorée tend vers $+\infty$.*

ii) Toute suite réelle décroissante et non minorée tend vers $-\infty$.

DÉMONSTRATION.

i) Soit $(u_n)_{n\in\mathbb{N}}$ une suite réelle croissante non majorée. L'ensemble $\{u_k \mid k \in \mathbb{N}\}$ des termes de la suite est une partie de $\mathbb{R}$ non majorée, et donc, quel que soit $A > 0$, il existe un entier N tel que $u_N > A$. La suite $(u_n)_{n\in\mathbb{N}}$ étant croissante, on en déduit

$$\forall n \in \mathbb{N},\ (n \geq N \implies u_n \geq u_N > A),$$

d'où $\lim\limits_{n\to+\infty} u_n = +\infty$.

ii) Il suffit d'appliquer le résultat précédent à la suite $(-u_n)_{n\in\mathbb{N}}$.

□

Théorème B.40 (« théorème de Bolzano [9]–Weierstrass » [Bol17]) *De toute suite réelle bornée, on peut extraire une suite convergente (on dit encore que toute suite réelle bornée admet au moins une valeur d'adhérence).*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite réelle bornée. Il existe alors deux réels a_0 et b_0 tels que, pour tout entier n, $a_0 \leq u_n \leq b_0$. Il est clair que $\{k \in \mathbb{N} \mid u_k \in [a_0, b_0]\} = \mathbb{N}$ est infini.
Soit à présent $n \in \mathbb{N}$; nous supposons défini le couple $(a_n, b_n) \in \mathbb{R}^2$ tel que $a_n \leq b_n$, $\{k \in \mathbb{N} \mid u_k \in [a_n, b_n]\}$ est infini et $b_n - a_n = \frac{1}{2^n}(b_0 - a_0)$. En considérant alors le milieu $\frac{a_n+b_n}{2}$ de l'intervalle fermé $[a_n, b_n]$, il est clair que l'un des deux intervalles $\left[a_n, \frac{a_n+b_n}{2}\right]$, $\left[\frac{a_n+b_n}{2}, b_n\right]$ est tel que l'ensemble des entiers k tels que u_k soit dans cet intervalle est infini. Il existe donc $(a_{n+1}, b_{n+1}) \in \mathbb{R}^2$ tel que $a_{n+1} \leq b_{n+1}$, $\{k \in \mathbb{N} \mid u_k \in [a_{n+1}, b_{n+1}]\}$ est infini, $b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) = \frac{1}{2^{n+1}}(b_0 - a_0)$. Il est alors évident que les intervalles $[a_n, b_n]$, $n \in \mathbb{N}$, forment une suite de segments emboîtés dont la longueur tend vers 0. On en déduit du théorème B.27 qu'ils ont un seul point commun $l \in \mathbb{R}$, qui est la limite commune de $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$.
D'autre part, il est aisé de construire une extractrice σ telle que $\sigma(0) = 0$ et telle qu'il existe, pour tout entier n, un entier k tel que si $k > \sigma(n)$ alors $u_k \in [a_n, b_n]$ et $\sigma(n+1) = k$. Les inégalités $a_n \leq u_{\sigma(n)} \leq b_n$, valables pour tout entier n, montrent alors que la suite $(u_{\sigma(n)})_{n\in\mathbb{N}}$ tend vers l. □

Théorème B.41 *Toute suite de Cauchy à valeurs réelles est convergente (on dit que $\mathbb{R}$ est **complet**).*

9. Bernardus Placidus Johann Nepomuk Bolzano (5 octobre 1781 - 18 décembre 1848) était un mathématicien, théologien et philosophe bohémien de langue allemande. Ses travaux portèrent essentiellement sur les fonctions et la théorie des nombres et il est considéré comme un des fondateurs de la logique moderne.

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}}$ une suite réelle de Cauchy. D'après la proposition B.21, nous savons que $(u_n)_{n\in\mathbb{N}}$ est bornée. Il existe alors, en vertu du théorème de Bolzano–Weierstrass (voir le théorème B.40), une suite extraite $(u_{\sigma(n)})_{n\in\mathbb{N}}$ qui converge vers une limite l. Montrons que la suite $(u_n)_{n\in\mathbb{N}}$ converge vers cette limite.
Soit $\varepsilon > 0$. Il existe des entiers N et N' tels que

$$\forall n \in \mathbb{N},\ \left(n \geq N \implies |u_{\sigma(n)} - l| \leq \frac{\varepsilon}{2}\right) \text{ (car } (u_{\sigma(n)})_{n\in\mathbb{N}} \text{ converge vers } l)$$

et

$$\forall (p,q) \in \mathbb{N}^2,\ \left(p \geq N',\ q \geq N' \implies \left|u_p - u_q\right| \leq \frac{\varepsilon}{2}\right) \text{ (car } (u_n)_{n\in\mathbb{N}} \text{ est une suite de Cauchy).}$$

Notons $N'' = \max(N, N')$. Si $n \geq N''$, nous avons d'une part $\sigma(n) \geq n \geq N'$, d'où $\left|u_n - u_{\sigma(n)}\right| \leq \frac{\varepsilon}{2}$, et d'autre part $n \geq N$, d'où $\left|u_{\sigma(n)} - l\right| \leq \frac{\varepsilon}{2}$. En combinant ces deux inégalités, nous obtenons

$$\forall n \in \mathbb{N},\ \left(n \geq N'' \implies |u_n - l| \leq \left|u_n - u_{\sigma(n)}\right| + \left|u_{\sigma(n)} - l\right| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\right),$$

ce qui permet de conclure que la suite $(u_n)_{n\in\mathbb{N}}$ est convergente. □

En alliant ce théorème à la proposition B.32, nous en déduisons qu'une suite réelle converge si et seulement si elle est une suite de Cauchy. Ce résultat reste vrai si la suite est complexe.

Nous terminons cette section avec un résultat relatif à la suite des moyennes arithmétiques des premiers termes d'une suite convergente.

Définition B.42 (« moyenne de Cesàro [10] ») *Soit une suite $(u_n)_{n\in\mathbb{N}^*}$. On appelle* ***moyenne de Cesàro*** *de $(u_n)_{n\in\mathbb{N}^*}$ la suite $(s_n)_{n\in\mathbb{N}^*}$ de terme général*

$$\forall n \in \mathbb{N}^*,\ s_n = \frac{1}{n}\sum_{k=1}^{n} u_k.$$

Théorème B.43 (« lemme de Cesàro » [Ces90]) *La moyenne de Cesàro d'une suite convergente de limite l converge vers l.*

DÉMONSTRATION. Soit $(u_n)_{n\in\mathbb{N}^*}$ une suite convergente de limite l et $(s_n)_{n\in\mathbb{N}^*}$ sa moyenne de Cesàro. Soit $\varepsilon > 0$. Il existe un entier naturel non nul N tel que

$$\forall n \in \mathbb{N}^*,\ \left(n \geq N \implies |u_n - l| \leq \frac{\varepsilon}{2}\right).$$

Pour $n \geq N$, on a donc

$$|s_n - l| = \left|\frac{1}{n}\sum_{k=1}^{n}(u_k - l)\right| \leq \frac{1}{n}\sum_{k=1}^{n}|u_k - l| \leq \frac{1}{n}\sum_{k=1}^{N-1}|u_k - l| + \frac{1}{n}\sum_{k=N}^{n}|u_k - l| \leq \frac{1}{n}\sum_{k=1}^{N-1}|u_k - l| + \frac{n-N+1}{n}\,\frac{\varepsilon}{2}.$$

La somme $\sum_{k=1}^{N-1}|u_k - l|$ étant finie, on a

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^{N-1}|u_k - l| = 0,$$

et il existe donc un entier naturel non nul N_{bis} tel que

$$\forall n \in \mathbb{N}^*,\ \left(n \geq N_{\text{bis}} \implies \frac{1}{n}\sum_{k=1}^{N-1}|u_k - l| \leq \frac{\varepsilon}{2}\right).$$

Notons $N_{\text{ter}} = \max(N, N_{\text{bis}})$. Il vient alors

$$\forall n \in \mathbb{N}^*,\ \left(n \geq N_{\text{ter}} \implies |s_n - l| \leq \frac{\varepsilon}{2} + \frac{n-N+1}{n}\,\frac{\varepsilon}{2} \leq \varepsilon\right),$$

ce qui montre que $(s_n)_{n\in\mathbb{N}^*}$ converge vers l. □

B.2.4 Quelques suites particulières

Nous concluons ces rappels sur les suites numériques par l'étude de suites remarquables.

10. Ernesto Cesàro (12 mars 1859 - 12 septembre 1906) était un mathématicien italien, connu pour ses contributions à la géométrie différentielle et son procédé de sommation des séries divergentes.

Suite arithmétique

Définition B.44 *Une suite numérique $(u_n)_{n\in\mathbb{N}}$ est dite* ***arithmétique de raison*** *r si et seulement s'il existe un scalaire r tel que, pour tout entier naturel n, $u_{n+1} = u_n + r$.*

Si $(u_n)_{n\in\mathbb{N}}$ est une suite arithmétique de raison r, on a

$$\forall n \in \mathbb{N},\ u_n = u_0 + nr.$$

Si $r = 0$, la suite est constante. Lorsque $\mathbb{K} = \mathbb{R}$, la suite est strictement croissante si $r > 0$ et strictement décroissante si $r < 0$.

Proposition B.45 (somme des termes d'une suite arithmétique) *La somme des m premiers termes d'une suite arithmétique $(u_n)_{n\in\mathbb{N}}$ de raison r est*

$$\forall m \in \mathbb{N}^*,\ S_m = \sum_{k=0}^{m-1} u_k = m\,u_0 + \frac{m(m-1)}{2} r = \frac{m}{2}(u_0 + u_{m-1}).$$

La preuve de cette proposition est laissée en exercice.

Suite géométrique

Définition B.46 *Une suite numérique $(u_n)_{n\in\mathbb{N}}$ est dite* ***géométrique de raison*** *q si et seulement s'il existe un scalaire q tel que, pour tout entier naturel n, $u_{n+1} = q\,u_n$.*

Si $(u_n)_{n\in\mathbb{N}}$ est une suite géométrique de raison q, on a

$$\forall n \in \mathbb{N},\ u_n = q^n u_0.$$

La suite $(u_n)_{n\in\mathbb{N}}$ est constante si $q = 1$ et stationnaire en 0 (à partir de $n = 1$) si $q = 0$. Lorsque $\mathbb{K} = \mathbb{R}$ et $q > 0$, la suite $(u_n)_{n\in\mathbb{N}}$ est monotone et garde un signe constant, alors que si $q < 0$, pour tout entier n, les termes u_n et u_{n+1} sont de signes contraires et la suite n'est donc pas monotone.

Proposition B.47 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite géométrique réelle de premier terme $u_0 \in \mathbb{R}^*$ et de raison $q \in \mathbb{R}$. On a les assertions suivantes.*

i) Si $|q| < 1$, $\lim\limits_{n\to+\infty} u_n = 0$.

ii) Si $q > 1$, $\lim\limits_{n\to+\infty} u_n = \begin{cases} +\infty & si\ u_0 > 0 \\ -\infty & si\ u_0 < 0 \end{cases}$.

iii) Si $q = 1$, $\lim\limits_{n\to+\infty} u_n = u_0$.

iv) Si $q \leq -1$, $(u_n)_{n\in\mathbb{N}}$ n'a pas de limite.

DÉMONSTRATION.

i) Si $r = 0$, la suite est stationnaire en 0 à partir du rang $n = 1$. Sinon $0 < |q| < 1$ implique que $\dfrac{1}{|q|} > 1$ et il existe alors un réel h strictement positif tel que $\dfrac{1}{|q|} = 1 + h$. On a donc

$$\forall n \in \mathbb{N}^*,\ \left(\frac{1}{|q|}\right)^n = (1+h)^n = \sum_{k=0}^{n} C_n^k h^k \geq 1 + nh,$$

en utilisant la formule du binôme de Newton. Par ailleurs, on a $|u_n| = q^n\,|u_0| \leq \dfrac{|u_0|}{1+nh}$, d'où $\lim\limits_{n\to+\infty} u_n = 0$.

ii) La démonstration est similaire à celle de i.

Les preuves de iii et iv sont évidentes. □

Proposition B.48 (somme des termes d'une suite géométrique) *La somme des m premiers termes d'une suite géométrique $(u_n)_{n\in\mathbb{N}}$ de raison q est*

$$\forall m \in \mathbb{N}^*,\ S_m = \sum_{k=0}^{m-1} u_k = u_0 \sum_{k=0}^{m-1} q^k = u_0 \frac{1-q^m}{1-q}\ si\ q \neq 1,\ S_m = m\,u_0\ si\ q = 1.$$

Suite arithmético-géométrique

Définition B.49 *Une suite $(u_n)_{n\in\mathbb{N}}$ est dite **arithmético-géométrique** si et seulement s'il existe des scalaires q et r tels que, pour tout entier naturel n, $u_{n+1} = q\,u_n + r$.*

Remarquons qu'une telle suite est arithmétique si $q = 1$ et géométrique si $r = 0$.

Méthode d'étude d'une suite arithmético-géométrique par utilisation de point fixe. Supposons $q \neq 1$. Soit α l'unique scalaire vérifiant $\alpha = q\,\alpha + r$, c'est-à-dire $\alpha = \dfrac{r}{1-q}$ (on dit que α est un *point fixe* de la fonction $f(x) = q\,x + r$). Nous avons

$$\forall n \in \mathbb{N}^*,\ u_n - \alpha = q\,u_{n-1} + r - (a\alpha + b) = q(u_{n-1} - \alpha).$$

La suite $(u_n - \alpha)_{n\in\mathbb{N}}$ est donc une suite géométrique de raison q. Ceci implique que $u_n - \alpha = q^n(u_0 - \alpha)$, soit encore

$$\forall n \in \mathbb{N},\ u_n = q^n(u_0 - \alpha) + \alpha.$$

On en déduit que, si $u_0 = \alpha$, la suite $(u_n)_{n\in\mathbb{N}}$ est constante et vaut α. Si $u_0 \neq \alpha$, on a alors

$$\lim_{n\to+\infty} u_n = \alpha \text{ si } |q| < 1 \text{ et } \lim_{n\to+\infty} |u_n| = +\infty \text{ si } |q| > 1.$$

Suite définie par récurrence

Définition B.50 *Soit I un intervalle fermé de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une application. On suppose que $f(I) \subset I$. La suite réelle $(u_n)_{n\in\mathbb{N}}$ définie par $u_0 \in I$ et la relation de récurrence*

$$\forall n \in \mathbb{N},\ u_{n+1} = f(u_n), \tag{B.1}$$

*est appelée **suite récurrente** (d'ordre un).*

Cette suite est bien définie car, pour tout entier naturel n, on a $u_n \in I$ et $f(I) \subset I$. Si f est une application affine à coefficients constants, la suite récurrente est une suite arithmético-géométrique.

Pour étudier une suite récurrente du type $u_{n+1} = f(u_n)$, on a recours aux propriétés élémentaires des applications continues (voir la section B.3) et des applications dérivables (voir la section B.3.5).

Détermination du sens de variation d'une suite récurrente. Soit f une application monotone sur l'intervalle fermé I et $(u_n)_{n\in\mathbb{N}}$ une suite définie par (B.1). Si f est croissante, on a

$$\forall n \in \mathbb{N}^*,\ u_{n+1} - u_n = f(u_n) - f(u_{n-1}),$$

et la différence $u_{n+1} - u_n$ a le même signe que $u_1 - u_0 = f(u_0) - u_0$. Ainsi, la suite $(u_n)_{n\in\mathbb{N}}$ est monotone et son sens de variation dépend de la position relative de u_0 et u_1. Il reste à voir si la suite est minorée, majorée.

Si l'application f est décroissante, on remarque que, pour tout entier naturel n, $u_{n+1} - u_n$ a le signe opposé de $u_n - u_{n-1}$. Il faut alors étudier les suites extraites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$. Pour tout entier naturel p, on a

$$u_{2p+2} = f(u_{2p+1}) = (f \circ f)(u_{2p}) \text{ et } u_{2p+3} = f(u_{2p+2}) = (f \circ f)(u_{2p+1}).$$

Par décroissance de f, l'application composée $f \circ f$ est croissante et les suites extraites $(u_{2p})_{p\in\mathbb{N}}$ et $(u_{2p+1})_{p\in\mathbb{N}}$ sont donc toutes deux monotones et de sens de variation contraires.

Détermination de la limite d'une suite récurrente. Soit f une application continue sur l'intervalle I et $(u_n)_{n\in\mathbb{N}}$ une suite définie par (B.1). Si la suite $(u_n)_{n\in\mathbb{N}}$ converge vers le réel l appartenant à I, alors, en faisant tendre n vers l'infini dans la relation de récurrence, on obtient que le réel l est un point fixe de l'application f. Pour déterminer les seules limites possibles d'une suite récurrente $(u_n)_{n\in\mathbb{N}}$ de type $u_{n+1} = f(u_n)$, on doit donc chercher à résoudre [11] l'équation $f(l) = l$ sur l'intervalle I.

11. Ce problème possède au moins une solution en vertu du théorème 5.7.

B.3 Fonctions d'une variable réelle *

Cette section est consacrée aux *fonctions numériques d'une variable réelle*, c'est-à-dire aux applications définies sur une partie D de $\mathbb{R}$ et à valeurs dans le corps $\mathbb{K}$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. L'ensemble D, appelé le *domaine de définition* de la fonction, est généralement une réunion d'intervalles non vides de $\mathbb{R}$. L'étude d'une fonction s'effectuant cependant intervalle par intervalle, nous nous restreindrons parfois à des applications dont les ensembles de définition sont contenus dans un intervalle I non vide de $\mathbb{R}$.

B.3.1 Généralités sur les fonctions

Soit D une partie non vide de $\mathbb{R}$. On désigne par $\mathscr{F}(D,\mathbb{K})$ l'ensemble des applications définies sur I à valeurs dans $\mathbb{K}$. Pour toute fonction f de $\mathscr{F}(D,\mathbb{K})$, on note A FAIRE

Si f et g sont deux éléments de $\mathscr{F}(D,\mathbb{K})$, alors on a

$$f = g \Leftrightarrow \forall x \in D,\ f(x) = g(x).$$

Opérations sur les fonctions

L'ensemble $\mathscr{F}(D,\mathbb{K})$ est muni de deux lois internes, une addition

$$\forall f,g \in \mathscr{F}(D,\mathbb{K}),\ \forall x \in D,\ (f+g)(x) = f(x) + g(x),$$

et une multiplication

$$\forall f,g \in \mathscr{F}(D,\mathbb{K}),\ \forall x \in D,\ (fg)(x) = f(x)g(x),$$

et d'une loi externe qui est une multiplication par les scalaires,

$$\forall \lambda \in \mathbb{K},\ \forall f \in \mathscr{F}(D,\mathbb{K}),\ \forall x \in D,\ (\lambda f)(x) = \lambda f(x).$$

AJOUTER composition

Relation d'ordre pour les fonctions réelles

Lorsque les fonctions considérées sont à valeur réelles, il est possible d'utiliser la relation d'ordre total usuelle sur $\mathbb{R}$ pour comparer certaines fonctions entre elles.

Définition B.51 *On définit dans l'ensemble $\mathscr{F}(D,\mathbb{R})$ une relation, notée $\leq$, par*

$$\forall f,g \in \mathscr{F}(D,\mathbb{R}),\ (f \leq g \Leftrightarrow (\forall x \in D,\ f(x) \leq g(x))).$$

Les résultats suivants sont immédiats.

Proposition B.52 *La relation $\leq$ est une relation d'ordre sur $\mathscr{F}(D,\mathbb{R})$, compatible avec l'addition,*

$$\forall f,g,h \in \mathscr{F}(D,\mathbb{R}),\ (f \leq g \Rightarrow f + h \leq g + h).$$

On a de plus

$$\forall f,g,h \in \mathscr{F}(D,\mathbb{R}),\ (f \leq g \text{ et } 0 \leq h \Rightarrow fh \leq gh).$$

On notera que l'ordre introduit sur $\mathscr{F}(D,\mathbb{R})$ par la relation $\leq$ définie ci-dessus n'est plus total dès que la partie D n'est pas réduite à un point. Supposons en effet que D contienne deux éléments distincts a et b et considérons deux applications f et g de I dans $\mathbb{R}$ définies par

$$\forall x \in D,\ f(x) = \begin{cases} 1 & \text{si } x = a \\ 0 & \text{si } x \neq a \end{cases},\ g(x) = \begin{cases} 1 & \text{si } x = b \\ 0 & \text{si } x \neq b \end{cases}.$$

On n'a alors ni $f \leq g$ (car $g(a) < f(a)$), ni $g \leq f$ (car $f(b) < g(b)$). On dit dans ce cas que f et g ne sont pas *comparables pour* $\leq$.

B.3.2 Propriétés globales des fonctions

COMPLETER

Parité

On dit qu'une partie D de $\mathbb{R}$ est *symétrique par rapport à* 0 si elle vérifie

$$\forall x \in D,\ -x \in D.$$

Définitions B.53 *Soit D une partie de $\mathbb{R}$ symétrique par rapport à 0 et f une fonction de $\mathscr{F}(D,\mathbb{K})$. On dit que f est* ***paire*** *(resp.* ***impaire****) si et seulement si*

$$\forall x \in D,\ f(-x) = f(x) \text{ (resp. } f(-x) = -f(x)).$$

Périodicité

Définition B.54 *Soit D une partie de $\mathbb{R}$, f une fonction de $\mathscr{F}(D,\mathbb{K})$ et T un réel strictement positif. L'application f est dite* ***périodique de période*** *T si et seulement si*

$$\forall x \in D,\ x + T \in D \text{ et } f(x+T) = f(x).$$

Monotonie

Définition B.55 *Soit D une partie de $\mathbb{R}$ et f une fonction de $\mathscr{F}(D,\mathbb{R})$. On dit que f est*

- ***croissante*** *(resp.* ***décroissante****) si et seulement si*

$$\forall (x,y) \in D^2,\ (x \le y \Rightarrow f(x) \le f(y)) \text{ (resp. } (x \le y \Rightarrow f(x) \ge f(y))),$$

- ***strictement croissante*** *(resp.* ***strictement décroissante****) si et seulement si*

$$\forall (x,y) \in D^2,\ (x < y \Rightarrow f(x) < f(y)) \text{ (resp. } (x \le y \Rightarrow f(x) > f(y))),$$

- ***monotone*** *si et seulement si elle est croissante ou décroissante,*
- ***strictement monotone*** *si et seulement si elle est strictement croissante ou strictement décroissante.*

Les résultats de la proposition suivante s'obtiennent de manière immédiate.

Proposition B.56 *Soit D une partie de $\mathbb{R}$ et f et g deux fonctions de $\mathscr{F}(D,\mathbb{R})$.*

i) Si f et g sont croissantes, alors $f+g$ est croissante.

ii) Si f est croissante et $\lambda \in \mathbb{R}_+$, alors λf est croissante.

iii) Si f et g sont croissantes et positives, alors fg est croissante.

iv) Si f et g sont croissantes et si $f(D) \subset J$, alors l'application composée $g \circ f \in \mathscr{F}(D,\mathbb{R})$ est croissante.

Majoration, minoration

Définition B.57 *Soit D une partie de $\mathbb{R}$. Une fonction numérique f de $\mathscr{F}(D,\mathbb{K})$ est dite* ***bornée*** *si et seulement s'il existe un réel positif M tel que*

$$\forall x \in D,\ |f(x)| \le M.$$

Définitions B.58 *Soit D une partie de $\mathbb{R}$. Une fonction f de $\mathscr{F}(D,\mathbb{R})$ est dite* ***majorée*** *si et seulement s'il existe un réel M tel que*

$$\forall x \in D,\ f(x) \le M,$$

minorée *si et seulement s'il existe un réel m tel que*

$$\forall x \in D,\ m \le f(x).$$

Proposition et définition B.59 *Si l'application $f : D \to \mathbb{R}$ est majorée (resp. minorée), alors $f(D)$ admet une borne supérieure (resp. inférieure) dans $\mathbb{R}$, appelée* ***borne supérieure*** *(resp.* ***inférieure*****) de** *f et notée $\sup_{x\in I} f(x)$ (resp. $\inf_{x\in I} f(x)$).*

Cette proposition résulte directement de l'axiome de la borne supérieure (voir la proposition B.4). Ainsi, par définition, $\sup_{x\in I} f(x) = \sup(\{f(x) \mid x \in I\}) = \sup(f(I))$.

Convexité et concavité

B.3.3 Limites

Dans cette section, la lettre I désigne un intervalle de $\mathbb{R}$, non vide et non réduit à un point. Nous commençons par rappeler la notion de *voisinage d'un point*, qui sera employée à plusieurs reprises.

Définitions B.60 *Soit une propriété dépendant du point x de I. On dit que cette propriété est vraie* ***au voisinage d'un point*** *a* ***de*** *I si elle est vraie sur l'intersection de I avec un intervalle non vide, ouvert et centré en a.*

Dans le cas où l'intervalle I est non majoré (resp. non minoré), on dit que la propriété est ***vraie au voisinage de*** *$+\infty$ (resp. $-\infty$) s'il existe un réel M tel qu'elle est vraie sur l'intersection de I avec un intervalle $]M, +\infty[$ (resp. $]-\infty, M[$).*

Limite d'une fonction en un point

Définition B.61 *Soit a un point de I, f une fonction définie sur I, sauf peut-être au point a, et à valeurs dans $\mathbb{K}$, et l un scalaire. On dit que f* ***admet*** *l* ***pour limite en*** *a si et seulement si*

$$\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in I,\ (0 < |x-a| \leq \alpha \Rightarrow |f(x)-l| \leq \varepsilon).$$

On voit que le fait que f ne soit pas définie en a n'empêche pas de considérer sa limite en ce point. On dit alors que f admet une limite lorsque x tend vers a *par valeurs différentes*. Lorsqu'une fonction a pour limite l en a, on dit encore qu'elle *admet une limite finie en a*.

Définitions B.62 *Soit a un point de I et f une fonction définie sur I, sauf peut-être au point a, et à valeurs réelles. On dit que f admet $+\infty$ (resp. $-\infty$)* ***pour limite en*** *a si et seulement si*

$$\forall A \in \mathbb{R},\ \exists \alpha > 0,\ \forall x \in I,\ (0 < |x-a| \leq \alpha \Rightarrow f(x) \geq A) \text{ (resp. } (0 < |x-a| \leq \alpha \Rightarrow f(x) \leq A)).$$

Proposition B.63 *Si une application admet une limite finie en un point, alors celle-ci est unique.*

DÉMONSTRATION. Raisonnons par l'absurde et supposons que l'application $f : I \to \mathbb{K}$ admet l et l' appartenant à $\mathbb{K}$ pour limites en un point a de I, avec $l \neq l'$. Posons $\varepsilon = \frac{1}{3}\left|l-l'\right|$. Il existe $\alpha > 0$ et $\alpha' > 0$ tels que

$$\forall x \in I,\ (0 < |x-a| \leq \alpha \Rightarrow |f(x)-l| \leq \varepsilon) \text{ et } (0 < |x-a| \leq \alpha' \Rightarrow \left|f(x)-l'\right| \leq \varepsilon).$$

Alors, pour tout x de I tel que $0 < |x-a| \leq \min(\alpha, \alpha')$, nous avons

$$\left|l-l'\right| \leq |l-f(x)| + \left|f(x)-l'\right| \leq 2\varepsilon = \frac{2}{3}\left|l-l'\right|,$$

d'où une contradiction. □

En vertu de cette unicité, si l'application f admet l pour limite en a, on dit que *l est la limite de f en a* et l'on note

$$\lim_{x\to a} f(x) = l \text{ ou } f(x) \underset{x\to a}{\longrightarrow} l.$$

Il est possible d'étendre les définitions de limites finie et infinie si la fonction f est définie sur un intervalle non majoré ou non minoré.

Définitions B.64 *Soit f une fonction de $\mathscr{F}(I,\mathbb{K})$ et l un scalaire. Si l'intervalle I admet $+\infty$ (resp. $-\infty$) comme extrémité, on dit que f* ***admet*** *l* ***pour limite en*** *$+\infty$ (resp. $-\infty$) si et seulement si*

$$\forall \varepsilon > 0,\ \exists A \in \mathbb{R},\ \forall x \in I,\ (x \geq A \Rightarrow |f(x)-l| \leq \varepsilon) \text{ (resp. } (x \leq A \Rightarrow |f(x)-l| \leq \varepsilon)),$$

et l'on note $\lim\limits_{x\to+\infty} f(x) = l$ *(resp.* $\lim\limits_{x\to-\infty} f(x) = l$*).*

Définitions B.65 *Soit f une fonction de $\mathscr{F}(I,\mathbb{R})$. Si I admet $+\infty$ comme extrémité, on dit que f* ***admet*** *$+\infty$ (resp. $-\infty$)* ***pour limite en*** *$+\infty$ si et seulement si*

$$\forall A \in \mathbb{R},\ \exists A' \in \mathbb{R},\ \forall x \in I,\ (x \geq A' \Rightarrow f(x) \geq A) \text{ (resp. } (x \geq A' \Rightarrow f(x) \leq A)),$$

et l'on note $\lim\limits_{x\to+\infty} f(x) = +\infty$ *(resp.* $\lim\limits_{x\to+\infty} f(x) = -\infty$*). Si I admet $-\infty$ comme extrémité, on dit que f* ***admet*** *$+\infty$ (resp. $-\infty$)* ***pour limite en*** *$-\infty$ si et seulement si*

$$\forall A \in \mathbb{R},\ \exists A' \in \mathbb{R},\ \forall x \in I,\ (x \leq A' \Rightarrow f(x) \geq A) \text{ (resp. } (x \leq A' \Rightarrow f(x) \leq A)),$$

et l'on note $\lim\limits_{x\to-\infty} f(x) = +\infty$ *(resp.* $\lim\limits_{x\to-\infty} f(x) = -\infty$*).*

Afin d'unifier la présentation des définitions et résultats, nous considérons dans toute la suite le point a en tant qu'élément de $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$.

Proposition B.66 *Soit f une fonction de $\mathscr{F}(I,\mathbb{K})$. Si f admet une limite finie en un point a de I, alors f est bornée au voisinage de a.*

DÉMONSTRATION. Supposons $a \in \mathbb{R}$, les cas $a = +\infty$ et $a = -\infty$ étant analogues.
Il existe $\alpha > 0$ tel que l'on a

$$\forall x \in I,\ (0 < |x-a| \leq \alpha \Rightarrow |f(x)-l| \leq 1 \Rightarrow |f(x)| \leq |f(x)-l| + |l| \leq 1 + |l|),$$

et donc f est bornée au voisinage de a. □

Limite à droite, limite à gauche

Définition B.67 *Soit f une fonction de $\mathscr{F}(I,\mathbb{K})$ et a un réel appartenant à l'intervalle I. On dit que f* ***admet une limite à droite*** *(resp.* ***à gauche)*** ***en*** *a si et seulement si la restriction de f à $I \cap]a,+\infty[$ (resp. $]-\infty,a[\cap I$), notée $f_{|I\cap]a,+\infty[}$ (resp. $f_{|]-\infty,a[\cap I}$), admet une limite en a.*

Lorsqu'une fonction f admet l pour limite à droite (resp. à gauche) en a, on note

$$\lim_{\substack{x\to a\\ x>a}} f(x) = l \text{ ou } \lim_{x\to a^+} f(x) = l \text{ (resp. } underset x < a \lim_{x\to a} f(x) = l \text{ ou } \lim_{x\to a^-} f(x) = l),$$

et l'on a

$$\lim_{x\to a^+} f(x) = l \Leftrightarrow (\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in I,\ (0 < x-a \leq \alpha \Rightarrow |f(x)-l| \leq \varepsilon))$$
$$(\text{resp. } \lim_{x\to a^-} f(x) = l \Leftrightarrow (\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in I,\ (0 < a-x \leq \alpha \Rightarrow |f(x)-l| \leq \varepsilon))).$$

On dit également que f tend vers l lorsque x tend vers a par *valeurs supérieures* (resp. *inférieures*).

Exemple. Considérons la fonction $x \mapsto \dfrac{|x|}{x}$ définie sur $\mathbb{R}^*$. Si $x > 0$, $f(x) = 1$ et l'on a $\lim\limits_{x\to0^+} f(x) = 1$. Si $x < 0$, alors $f(x) = -1$ et l'on a $\lim\limits_{x\to0^-} f(x) = -1$.

Caractérisation séquentielle de la limite

Proposition B.68 *Pour qu'une application f de $\mathscr{F}(I,\mathbb{K})$ admette un scalaire l pour limite en un point a de l'intervalle I, il faut et il suffit que, pour toute suite $(u_n)_{n\in\mathbb{N}}$ d'éléments de I ayant a pour limite, on ait*

$$\lim_{n\to+\infty} f(u_n) = l.$$

DÉMONSTRATION. Faisons l'hypothèse que $a\in\mathbb{R}$, les cas $a=+\infty$ et $a=-\infty$ étant analogues. Supposons tout d'abord que f admet l pour limite en a. Soit $(u_n)_{n\in\mathbb{N}}$ une suite dans I, telle que $\lim\limits_{n\to+\infty} u_n = a$, et $\varepsilon>0$. Puisque $\lim\limits_{x\to a} f(x) = l$, il existe $\alpha>0$ tel que

$$\forall x\in I,\ (0<|x-a|\leq\alpha\Rightarrow|f(x)-l|\leq\varepsilon).$$

Puisque $(u_n)_{n\in\mathbb{N}}$ tend vers a, il existe $N\in\mathbb{N}$ tel que

$$\forall n\in\mathbb{N},\ (n\geq N\Rightarrow|u_n-a|\leq\alpha).$$

On a alors

$$\forall n\in\mathbb{N},\ (n\geq N\Rightarrow|u_n-a|\leq\alpha\Rightarrow|f(u_n)-l|\leq\varepsilon),$$

d'où $\lim\limits_{n\to+\infty} f(u_n) = l$.

Supposons à présent que f n'admet pas l pour limite en a. Il existe donc $\varepsilon>0$ tel que

$$\forall\alpha>0,\ \exists x\in I,\ (0<|x-a|\leq\alpha \text{ et } |f(x)-l|>\varepsilon).$$

En particulier, en prenant $\alpha=\dfrac{1}{n}$ pour tout n de $\mathbb{N}^*$, il existe $u_n\in I$ tel que

$$|u_n-a|\leq\frac{1}{n} \text{ et } |f(u_n)-l|>\varepsilon.$$

On constate alors que la suite $(u_n)_{n\in\mathbb{N}}$ dans I ainsi construite satisfait $\lim\limits_{n\to+\infty} u_n = a$, mais est telle que $(f(u_n))_{n\in\mathbb{N}}$ ne converge pas vers l. □

Passage à la limite dans une inégalité

Proposition B.69 *Soit f et g deux fonctions de $\mathscr{F}(I,\mathbb{R})$ admettant une limite en un point a de l'intervalle I. Si l'on a $f(x)\leq g(x)$ au voisinage de a, alors*

$$\lim_{x\to a} f(x) \leq \lim_{x\to a} g(x).$$

DÉMONSTRATION. Supposons que f et g tendent respectivement vers l et l' lorsque x tend vers a. Soit $\varepsilon>0$. Il existe $\alpha>0$ et $\alpha'>0$ tels que

$$\forall x\in I,\ \left(0<|x-a|\leq\alpha\Rightarrow l-\frac{\varepsilon}{2}\leq f(x)\leq l+\frac{\varepsilon}{2}\right) \text{ et } \left(0<|x-a|\leq\alpha'\Rightarrow l'-\frac{\varepsilon}{2}\leq g(x)\leq l'+\frac{\varepsilon}{2}\right).$$

Nous avons donc, pour tout $x\in I$ tel que $0<|x-a|\leq\min(\alpha,\alpha')$,

$$l-\frac{\varepsilon}{2}\leq f(x)\leq g(x)\leq l+\frac{\varepsilon}{2},$$

d'où $l-l'\leq\varepsilon$. On a finalement $l\leq l'$, car le choix de ε est arbitraire. □

Théorème d'encadrement

Proposition B.70 *Soit f, g et h trois fonctions de $\mathscr{F}(I,\mathbb{R})$ telles que $f(x)\leq g(x)\leq h(x)$ au voisinage d'un point a de I. Si f et h admettent une même limite l en a, alors g admet l pour limite en a.*

DÉMONSTRATION. Supposons $a\in\mathbb{R}$, les cas $a=+\infty$ et $a=-\infty$ étant analogues.
Soit $\varepsilon>0$. Puisque f et h admettent l pour limite en a, il existe $\alpha>0$ et $\alpha'>0$ tels que

$$\forall x\in I,\ (0<|x-a|\leq\alpha\Rightarrow|f(x)-l|\leq\varepsilon) \text{ et } (0<|x-a|\leq\alpha'\Rightarrow|h(x)-l|\leq\varepsilon).$$

Nous avons donc, pour tout $x\in I$ tel que $0<|x-a|\leq\min(\alpha,\alpha')$,

$$((|f(x)-l|\leq\varepsilon \text{ et } |h(x)-l|\leq\varepsilon)\Rightarrow-\varepsilon\leq f(x)-l\leq g(x)-l\leq h(x)-l\leq\varepsilon\Rightarrow|g(x)-l|\leq\varepsilon).$$

Donc g admet l pour limite en a. □

Ce théorème d'encadrement s'avère très utile en pratique puisqu'il permet notamment de conclure à l'existence d'une limite.

Opérations algébriques sur les limites

Proposition B.71 *Soit f et g deux fonctions de $\mathscr{F}(I,\mathbb{K})$, a un point de I, λ, l et l' trois scalaires. On a*

i) $\lim_{x\to a} f(x) = l \Rightarrow \lim_{x\to a} |f(x)| = |l|$,

ii) $\lim_{x\to a} f(x) = l$ *et* $\lim_{x\to a} g(x) = l' \Rightarrow \lim_{x\to a}(f(x)+g(x)) = l+l'$,

iii) $\lim_{x\to a} f(x) = l \Rightarrow \lim_{x\to a} \lambda f(x) = \lambda l$,

iv) $\lim_{x\to a} f(x) = 0$ *et* g *est bornée au voisinage de* $a \Rightarrow \lim_{x\to a} f(x)g(x) = 0$,

v) $\lim_{x\to a} f(x) = l$ *et* $\lim_{x\to a} g(x) = l' \Rightarrow \lim_{x\to a} f(x)g(x) = ll'$,

vi) $\lim_{x\to a} g(x) = l'$ *et* $l' \neq 0 \Rightarrow \lim_{x\to a} \dfrac{1}{g(x)} = \dfrac{1}{l'}$,

vii) $\lim_{x\to a} f(x) = l$ *et* $\lim_{x\to a} g(x) = l'$ *et* $l' \neq 0 \Rightarrow \lim_{x\to a} \dfrac{f(x)}{g(x)} = \dfrac{l}{l'}$.

DÉMONSTRATION. Supposons $a \in I \subset \mathbb{R}$, les cas $a = +\infty$ et $a = -\infty$ étant analogues.

i) Soit $\varepsilon > 0$. Puisque $\lim_{x\to a} f(x) = l$, il existe $\alpha > 0$ tel que

$$\forall x \in I,\ (|x-a| \leq \alpha \Rightarrow |f(x)-l| \leq \varepsilon).$$

Comme $\forall x \in I,\ ||f(x)|-|l|| \leq |f(x)-l|$, on déduit

$$\forall x \in I,\ (|x-a| \leq \alpha \Rightarrow ||f(x)|-|l|| \leq \varepsilon),$$

et finalement $\lim_{x\to a} |f(x)| = |l|$.

ii) Soit $\varepsilon > 0$. Puisque $\lim_{x\to a} f(x) = l$ et $\lim_{x\to a} g(x) = l'$, il existe $\alpha > 0$ et $\alpha' > 0$ tels que

$$\forall x \in I,\ \left(|x-a| \leq \alpha \Rightarrow |f(x)-l| \leq \frac{\varepsilon}{2}\right) \text{ et } \left(|x-a| \leq \alpha' \Rightarrow \left|g(x)-l'\right| \leq \frac{\varepsilon}{2}\right)$$

En notant $\alpha'' = \min(\alpha,\alpha') > 0$, nous avons, $\forall x \in I$,

$$\left(|x-a| \leq \alpha'' \Rightarrow (|f(x)-l| \leq \frac{\varepsilon}{2} \text{ et } \left|g(x)-l'\right| \leq \frac{\varepsilon}{2}\right) \Rightarrow \left|(f(x)+g(x))-(l+l')\right| \leq |f(x)-l| + \left|g(x)-l'\right| \leq \varepsilon),$$

d'où $\lim_{x\to a}(f(x)+g(x)) = l+l'$.

iii) Soit $\varepsilon > 0$. Puisque $\lim_{x\to a} f(x) = l$, il existe $\alpha > 0$ tel que

$$\forall x \in I,\ \left(|x-a| \leq \alpha \Rightarrow |f(x)-l| \leq \frac{\varepsilon}{|\lambda|+1}\right),$$

d'où $\forall x \in I,\ \left(|x-a| \leq \alpha \Rightarrow |\lambda f(x)-\lambda l| = |\lambda|\,|f(x)-l| \leq \dfrac{|\lambda|\,\varepsilon}{|\lambda|+1} \leq \varepsilon\right)$, et donc $\lim_{x\to a} \lambda f(x) = \lambda l$.

iv) Par hypothèse, il existe $\alpha > 0$ et $C \in \mathbb{R}_+$ tels que

$$\forall x \in I,\ (|x-a| \leq \alpha \Rightarrow |g(x)| \leq C).$$

Soit $\varepsilon > 0$. Puisque $\lim_{x\to a} f(x) = 0$, il existe $\alpha' > 0$ tel que

$$\forall x \in I,\ \left(|x-a| \leq \alpha' \Rightarrow |f(x)| \leq \frac{\varepsilon}{C+1}\right).$$

En notant $\alpha'' = \min(\alpha,\alpha') > 0$, nous avons alors

$$\forall x \in I,\ \left(|x-a| \leq \alpha'' \Rightarrow |f(x)g(x)| = |f(x)|\,|g(x)| \leq \frac{C\varepsilon}{C+1} \leq \varepsilon\right),$$

et donc $\lim_{x\to a} f(x)g(x) = 0$.

v) Notons h l'application de I dans $\mathbb{K}$ telle que $h(x) = f(x) - l$, $\forall x \in I$. Nous avons

$$\forall x \in I,\ f(x)g(x) = h(x)g(x) + l\,g(x).$$

D'après iii, $\lim_{x\to a} l\,g(x) = ll'$. D'autre part, $\lim_{x\to a} h(x) = 0$, donc, d'après iv, $\lim_{x\to a} h(x)g(x) = 0$, puisque g est bornée au voisinage de a. Finalement, on a $\lim_{x\to a} f(x)g(x) = ll'$ d'après ii.

vi) Puisque $\lim_{x\to a} g(x) = l'$, on a, d'après i, $\lim_{x\to a} |g(x)| = |l'|$. Comme $|l'| > 0$, il existe $\alpha > 0$ tel que

$$\forall x \in I,\ \left(|x-a| \le \alpha \Rightarrow |g(x)| > \frac{|l'|}{2}\right),$$

En particulier, $\forall x \in I$, $(|x-a| \le \alpha \Rightarrow g(x) \neq 0)$. La fonction $\left(\frac{1}{g}\right)$ est donc définie, au moins sur $I \cap]a-\alpha, a+\alpha[$. Nous avons alors, pour tout x de $I \cap]a-\alpha, a+\alpha[$,

$$0 \le \left|\frac{1}{g(x)} - \frac{1}{l'}\right| = \frac{|g(x)-l'|}{|g(x)|\,|l'|} \le \frac{2}{|l'|^2}\left|g(x)-l'\right|.$$

Comme $\lim_{x\to a} g(x) = l'$, on en déduit que $\lim_{x\to a}\left(\frac{2}{|l'|^2}\left|g(x)-l'\right|\right) = 0$, puis que $\lim_{x\to a}\left|\frac{1}{g(x)} - \frac{1}{l'}\right| = 0$, soit encore $\lim_{x\to a}\frac{1}{g(x)} = \frac{1}{l'}$.

vii) Il suffit d'appliquer v et vi en remarquant que $\frac{f}{g} = f\frac{1}{g}$.

□

Proposition B.72 *Soit f et g deux fonctions de $\mathscr{F}(I, \mathbb{K})$ et a un point de I.*

i) Si $\lim_{x\to a} f(x) = +\infty$ et si g est minorée au voisinage de a, alors

$$\lim_{x\to a}(f(x) + g(x)) = +\infty.$$

ii) Si $\lim_{x\to a} f(x) = +\infty$ et si g est minorée au voisinage de a par une constante strictement positive, alors

$$\lim_{x\to a} f(x)g(x) = +\infty.$$

DÉMONSTRATION. Les preuves sont analogues à celles de i et ii dans la proposition B.37. □

Composition des limites

Proposition B.73 *Soit f un fonction de $\mathscr{F}(I, \mathbb{R})$, J un intervalle de $\mathbb{R}$ tel que $f(I) \subset J$, g un fonction de $\mathscr{F}(J, \mathbb{R})$, a un point de I, b un point de J et l un élément de $\mathbb{R}$. Si f admet b pour limite en a et g admet l pour limite en b, alors la fonction composée $g \circ f$ admet l pour limite en a.*

DÉMONSTRATION. Supposons $a \in \mathbb{R}$, $b \in \mathbb{R}$ et $l \in \mathbb{R}$, les autres cas étant analogues.
Soit $\varepsilon > 0$. Puisque $\lim_{y\to b} g(y) = l$, il existe $\alpha > 0$ tel que

$$\forall y \in J,\ (|y-b| \le \alpha \Rightarrow |g(y)-l| \le \varepsilon).$$

Puis, comme $\lim_{x\to a} f(x) = b$, il existe $\alpha' > 0$ tel que

$$\forall x \in I,\ (|x-a| \le \alpha' \Rightarrow |f(x)-b| \le \alpha_1).$$

Nous avons alors

$$\forall x \in I,\ (|x-a| \le \alpha' \Rightarrow |f(x)-b| \le \alpha \Rightarrow |g(f(x))-l| \le \varepsilon),$$

d'où $\lim_{x\to a} g \circ f(x) = l$. □

Cas des fonctions monotones

Nous considérons dans cette sous-section des fonctions à valeurs réelles.

Théorème B.74 *Soit a et b deux éléments de $\overline{\mathbb{R}}$, tels que $a < b$, et f une application croissante définie sur l'intervalle $]a,b[$.*

i) Si f est majorée, alors f admet une limite finie en b et $\lim_{x\to b} f(x) = \sup_{x\in]a,b[} f(x)$.

ii) Si f n'est pas majorée, alors f admet $+\infty$ pour limite en b.

DÉMONSTRATION.

i) La partie $f(]a,b[)$ de $\mathbb{R}$ est non vide et majorée, elle admet par conséquent une borne supérieure l dans $\mathbb{R}$. Soit $\varepsilon > 0$. Puisque $l-\varepsilon$ n'est pas un majorant de $f(]a,b[)$ dans $\mathbb{R}$, il existe $x_0 \in]a,b[$ tel que $l-\varepsilon \le f(x_0) \le l$. Alors, pour tout $x \in]a,b[$, nous avons
$$x_0 \le x \Rightarrow f(x_0) \le f(x) \Rightarrow l-\varepsilon \le f(x) \Rightarrow |f(x)-l| \le \varepsilon.$$
Supposons $b \in \mathbb{R}$ (le cas $b = +\infty$ étant analogue). En posant $\alpha = b - x_0 > 0$, nous avons ainsi $\forall x \in]a,b[$, $(0 < b-x \le \alpha \Rightarrow |f(x)-l| \le \varepsilon)$, d'où $\lim_{x\to b} f(x) = l$.

ii) Soit $A \in \mathbb{R}$. Puisque f n'est pas majorée, il existe $x_0 \in]a,b[$ tel que $f(x_0) \ge A$. Alors, pour $x \in]a,b[$, nous avons
$$x_0 \le x \Rightarrow f(x_0) \le f(x) \Rightarrow f(x) \ge A.$$
Supposons $b \in \mathbb{R}$ (le cas $b = +\infty$ étant analogue). En posant $\alpha = b - x_0 > 0$, nous avons ainsi $\forall x \in]a,b[$, $(0 < b-x \le \alpha \Rightarrow |f(x)| \ge A)$, d'où $\lim_{x\to b} f(x) = +\infty$.

□

Lorsque b appartient à $\mathbb{R}$, on peut parler de limite à gauche en b dans le théorème précédent. On déduit de ce dernier résultat qu'une application croissante admet toujours une limite, finie ou infinie, en b. Un résultat analogue est obtenu pour les applications décroissantes en considérant $-f$ dans cette démonstration.

B.3.4 Continuité

Dans cette section, la lettre I désigne un intervalle de $\mathbb{R}$, non vide et non réduit à un point.

Continuité en un point

Définition B.75 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$ et a un point de I. On dit que f* ***est continue en*** *a si et seulement si*
$$\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall x \in I,\ (|x-a| \le \alpha \Rightarrow |f(x)-f(a)| \le \varepsilon).$$

À la différence de la notion de limite, on ne parle de continuité qu'en des points où la fonction est définie. On dit que *f est discontinue en a* si et seulement si f n'est pas continue en a, qui est alors appelé un *point de discontinuité de f*.

Définition B.76 *On dit qu'une application f admet une* ***discontinuité de première espèce en*** *a si et seulement si elle n'est pas continue en a et possède une limite à droite et une limite à gauche en a. Lorsque f n'est pas continue et n'admet pas de discontinuité de première espèce en a, on dit qu'elle admet une* ***discontinuité de seconde espèce en*** *a.*

La démonstration de la proposition suivante est immédiate.

Proposition B.77 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$ et a un point de I. Pour que f soit continue en a, il faut et il suffit qu'elle admette $f(a)$ pour limite en a.*

Proposition B.78 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$ et a un point de I. Si f est continue en a, alors f est bornée au voisinage de a.*

DÉMONSTRATION. Si l'application f est continue en a, alors il existe $\alpha > 0$ tel que
$$\forall x \in I, (|x-a| \le \alpha \Rightarrow |f(x)-f(a)| \le 1 \Rightarrow |f(x)| \le |f(x)-f(a)| + |f(a)| \le |f(a)| + 1).$$
La fonction f est donc bornée au voisinage de a. □

Continuité à droite, continuité à gauche

Définition B.79 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$ et a un point de I. On dit que f est **continue à droite** (resp. **à gauche**) **en** a si et seulement si la restriction de f à $I\cap]a,+\infty[$ (resp. $]-\infty,a[\cap I$) est continue en a.*

On déduit de cette définition qu'une application est continue en un point si et seulement si elle est continue à droite et à gauche en ce point.

Exemple. Considérons l'application qui à tout réel x associe la partie entière de x, $E(x)$. Pour tout entier naturel n, cette application est continue à droite en n, mais n'est pas continue à gauche en n.

Caractérisation séquentielle de la continuité

Comme dans le cas de la limite, on peut définir la notion de continuité en se servant de suites réelles.

Proposition B.80 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$ et a un point de I. Alors, la fonction f est continue en a si et seulement si, pour toute suite $(u_n)_{n\in\mathbb{N}}$ d'éléments de I tendant vers a, la suite $(f(u_n))_{n\in\mathbb{N}}$ tend vers $f(a)$.*

DÉMONSTRATION. La preuve découle directement de la proposition B.77 et de la caractérisation séquentielle de la limite (voir la proposition B.68). □

Prolongement par continuité

Définition B.81 *Soit f une fonction définie sur I, sauf en un point a de I, et admettant une limite finie l en a. On appelle **prolongement par continuité de f en** a la fonction g, définie sur l'intervalle I par*

$$g(x)=\begin{cases} l & \text{si } x=a \\ f(x) & \text{sinon} \end{cases}.$$

Cette fonction est, par définition, continue en a.

Il est facile de vérifier que lorsqu'une fonction admet un prolongement par continuité en un point, celui-ci est unique. S'il n'y a pas de risque d'ambiguïté, on désigne alors la fonction et son prolongement par le même symbole. Notons qu'on peut aussi définir un prolongement par continuité à droite de a ou à gauche de a.

Continuité sur un intervalle

Définition B.82 *Soit f une application définie sur I à valeurs dans $\mathbb{K}$. On dit que f est **continue sur** I si et seulement si f est continue en tout point de I.*

On note $\mathscr{C}(I,\mathbb{K})$ l'ensemble des applications de I dans $\mathbb{K}$ qui sont continues sur I.

Continuité par morceaux

Définition B.83 *Soit $[a,b]$ un intervalle borné non vide de $\mathbb{R}$ et f une application définie sur $[a,b]$ à valeurs dans $\mathbb{K}$. On dit que la fonction f est **continue par morceaux sur** $[a,b]$ si et seulement s'il existe un entier n non nul et des points $a_0,\dots,a_n$ de $[a,b]$ vérifiant $a=a_0<\cdots<a_n=b$ tels que, pour tout entier i compris entre 0 et $n-1$, f soit continue sur $]a_i,a_{i+1}[$ et admette une limite finie à droite en a_i et une limite finie à gauche en a_{i+1}.*

Opérations algébriques sur les applications continues

Proposition B.84 *Soit f et g des applications définies sur I à valeurs dans $\mathbb{K}$, λ un scalaire et a un point de I.*

i) Si f est continue en a, alors $|f|$ est continue en a.

ii) Si f et g sont continues en a, alors $f+g$ est continue en a.

iii) Si f est continue en a, alors λf est continue en a.

iv) Si f et g sont continues en a, alors fg est continue en a.

v) Si g est continue en a et si $g(a) \neq 0$, alors $\dfrac{1}{g}$ est continue en a.

vi) Si f et g sont continues en a et si $g(a) \neq 0$, alors $\dfrac{f}{g}$ est continue en a.

DÉMONSTRATION. Les preuves sont similaires à celles de la proposition B.71. □

Proposition B.85 *Soit J un intervalle de $\mathbb{R}$, f une application définie sur I à valeurs réelles telle que $f(I) \subset J$, g une application définie sur J à valeurs dans $\mathbb{K}$ et a un point de I. Si f est continue en a et g est continue en $f(a)$, alors $g \circ f$ est continue en a.*

DÉMONSTRATION. La preuve est analogue à celle de la proposition B.73. □

De ces deux propositions, nous déduisons aisément des résultats de continuité globale sur l'intervalle I.

Proposition B.86 *Soit f et g des applications définies sur I à valeurs dans $\mathbb{K}$ et λ un scalaire.*

i) Si f est continue sur I, alors $|f|$ est continue sur I.

ii) Si f et g sont continues sur I, alors $f+g$ est continue sur I.

iii) Si f est continue sur I, alors λf est continue sur I.

iv) Si f et g sont continues sur I, alors fg est continue sur I.

v) Si g est continue sur I et si $(\forall x \in I,\ g(x) \neq 0$, alors $\dfrac{1}{g}$ est continue sur I.

vi) Si f et g sont continues sur I et si $(\forall x \in I,\ g(x) \neq 0$, alors $\dfrac{f}{g}$ est continue sur I.

Proposition B.87 *Soit J un intervalle de $\mathbb{R}$, f une application définie sur I à valeurs réelles telle que $f(I) \subset J$ et g une application définie sur J à valeurs dans $\mathbb{K}$. Si f est continue sur I et g est continue sur $f(I)$, alors $g \circ f$ est continue sur I.*

Théorèmes des bornes et des valeurs intermédiaires

Les théorèmes qui suivent constituent tous deux des résultats fondamentaux de la théorie des fonctions réelles d'une variable réelle.

Théorème B.88 (« théorème des bornes ») *Toute application réelle définie et continue sur un intervalle non vide de $\mathbb{R}$ est bornée et atteint ses bornes.*

DÉMONSTRATION. REPRENDRE en introduisant f et $[a,b]$
Montrons que la fonction f est bornée en raisonnant par l'absurde. Supposons que f est non majorée. Il existe alors une suite $(x_n)_{n\in\mathbb{N}}$ d'éléments de $[a,b]$ telle que, pour chaque entier n, on a $f(x_n) > n$. Puisque la suite est bornée, il existe, d'après le théorème de Bolzano–Weierstrass (voir le théorème B.40), une suite extraite de $(x_n)_{n\in\mathbb{N}}$, notée $(x_{\sigma(n)})_{n\in\mathbb{N}}$, qui converge vers un point c de $[a,b]$. Puisque f est continue sur $[a,b]$, on déduit que la suite $(f(x_{\sigma(n)}))_{n\in\mathbb{N}}$ tend vers $f(c)$. Mais d'autre part, on a $\forall n \in \mathbb{N}$, $f(x_{\sigma(n)}) > \sigma(n) \leq n$, donc $\lim\limits_{n\to+\infty} f(x_{\sigma(n)}) = +\infty$, d'où une contradiction. L'application f est donc majorée. En appliquant ce résultat à $-f$ au lieu de f, on en déduit que f est minorée. Finalement, f est bornée.
Montrons à présent que f atteint ses bornes. Notons $M = \sup\limits_{x\in[a,b]} f(x)$. Pour chaque entier $n \geq 1$, il existe un réel x_n dans $[a,b]$ tel que

$$M - \frac{1}{n} < f(x_n) \leq M.$$

La suite $(x_n)_{n\in\mathbb{N}}$ ainsi construite étant bornée, on peut en extraire, en vertu du théorème de Bolzano–Weierstrass, une sous-suite de $(x_n)_{n\in\mathbb{N}}$, notée $(x_{\tau(n)})_{n\in\mathbb{N}}$, qui converge vers un élément d de $[a,b]$. Puisque f est continue sur $[a,b]$, on en déduit que la suite $(f(x_{\tau(n)}))_{n\in\mathbb{N}}$ tend vers $f(d)$. D'autre part, on a

$$\forall n \in \mathbb{N}^*,\ M - \frac{1}{\tau(n)} < f(x_{\tau(n)}) \leq M,$$

d'où, par passage à la limite, $M = f(d)$. Ceci montre que M est atteint par f : $\exists d \in [a,b]$, $M = f(d)$. En appliquant ce résultat à $-f$ au lieu de f, on montre que f atteint aussi $\inf\limits_{x\in[a,b]} f(x)$. □

Dans une formulation due à Weierstrass, ce dernier théorème affirme qu'une fonction à valeurs réelles continue sur un ensemble compact y atteint son maximum et son minimum.

Théorème B.89 (« théorème des valeurs intermédiaires ») *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une application définie et continue sur $[a,b]$ à valeurs dans $\mathbb{R}$. Alors, pour tout réel y compris entre $f(a)$ et $f(b)$, il existe (au moins) un réel c dans $[a,b]$ tel que $f(c)=y$.*

DÉMONSTRATION. Si $y=f(a)$ ou $y=f(b)$, le résultat est immédiat. Dans toute la suite, on peut supposer que $f(a)<f(b)$, quitte à poser $g=-f$ si $f(a)>f(b)$. Soit donc $y\in]f(a),f(b)[$ et considérons l'ensemble $E=\{x\in[a,b]\mid f(x)\leq y\}$; E est une partie de $\mathbb{R}$ non vide (car $a\in E$) et majorée (par b), qui admet donc une borne supérieure, notée c. Nous allons montrer que $f(c)=y$.
Par définition de la borne supérieure, il existe une suite $(x_n)_{n\in\mathbb{N}}$ d'éléments de E telle que $\lim\limits_{n\to+\infty}x_n=c$. L'application f étant continue en c, on a $\lim\limits_{n\to+\infty}f(x_n)=f(c)$. Or, pour tout $n\in\mathbb{N}$, $f(x_n)\leq y$ donc $f(c)\leq y$. D'autre part, $f(b)>y$, donc $c\neq b$. Pour tout $x\in]c,b[$, $f(x)>y$ donc $\lim\limits_{x\to c,\ x>c}f(x)=f(c)\geq y$ d'où $f(c)=y$. □

Corollaire B.90 *L'image d'un intervalle par une application continue à valeurs réelles est un intervalle.*

DÉMONSTRATION. Soit I un intervalle de $\mathbb{R}$ et f une application continue sur I à valeurs dans $\mathbb{R}$. D'après le théorème des valeurs intermédiaires (voir le théorème B.89), l'ensemble $f(I)$ est un intervalle de $\mathbb{R}$. □

Exemples. Considérons l'intervalle $I=]0,1]$ et la fonction $f:x\mapsto\dfrac{1}{x}$. L'application f est continue sur I et $f(I)=[1,+\infty[$. L'intervalle I est borné alors que $f(I)$ n'est pas borné. Soit à présent $I=]0,2\pi[$ et $f:x\mapsto\sin x$. L'application f est continue sur I et $f(I)=[-1,1]$. Dans ce cas, l'intervalle I est ouvert alors que $f(I)$ est fermé.

Comme le montrent les exemples ci-dessus, le caractère ouvert, fermé ou borné d'un intervalle n'est pas toujours conservé par une application. On a cependant le résultat suivant.

Corollaire B.91 *Soit I un intervalle de $[a,b]$ un intervalle non vide de $\mathbb{R}$ et non réduit à un point et f une fonction définie sur $[a,b]$ à valeurs réelles. Si f est continue, alors l'ensemble $f([a,b])$ est un segment de $\mathbb{R}$.*

DÉMONSTRATION. D'après le corollaire précédent, l'ensemble $f([a,b])$ est un intervalle. D'après le théorème des bornes (voir le théorème B.88), c'est une partie bornée de $\mathbb{R}$ qui contient ses bornes. □

Application réciproque d'une application continue strictement monotone

REPRENDRE

Théorème B.92 (« théorème de la bijection ») *Soit f une application continue et strictement monotone sur I ; on note $\widetilde{f}:I\to f(I)$ l'application qui à tout x de I associe $f(x)$. Alors*

i) L'ensemble $f(I)$ est un intervalle dont les bornes sont les limites de f aux bornes de I.

ii) L'application $\widetilde{f}$ est bijective.

iii) La bijection réciproque $\widetilde{f}^{-1}$ est continue sur $f(I)$ et strictement monotone de même sens que f.

DÉMONSTRATION. Supposons, par exemple, que f est strictement croissante et posons $I=]a,b[$, avec $(a,b)\in\overline{\mathbb{R}}^2$ tel que $a<b$.

i) Pour tout x de $]a,b[$, on a $\lim\limits_{t\to a}f(t)<f(x)<\lim\limits_{t\to b}f(t)$, donc $f(I)\subset]\lim\limits_{t\to a}f(t),\lim\limits_{t\to b}f(t)[$. Réciproquement, soit $y\in]\lim\limits_{x\to a}f(x),\lim\limits_{x\to b}f(x)[$; y n'est ni un majorant, ni un minorant de $f(I)$, il existe donc des éléments x_1 et x_2 de I tels que $f(x_1)<y<f(x_2)$. D'après le théorème des valeurs intermédiaires (voir le théorème B.89), il existe $x_0\in I$ tel que $f(x_0)=y$, d'où $y\in f(I)$. Nous en concluons $f(I)=]\lim\limits_{x\to a}f(x),\lim\limits_{x\to b}f(x)[$.

ii) Par définition, $\widetilde{f}$ est une surjection. Montrons qu'elle est également injective. Soit x_1 et x_2 des éléments de I tels que $\widetilde{f}(x_1)=\widetilde{f}(x_2)$. Si $x_1<x_2$, alors $\widetilde{f}(x_1)<\widetilde{f}(x_2)$ et si $x_1>x_2$, alors $\widetilde{f}(x_1)>\widetilde{f}(x_2)$, d'où une contradiction dans les deux cas. Par conséquent, $x_1=x_2$ et $\widetilde{f}$ est injective.

iii) Soit y_1 et y_2 appartenant à $f(I)$, tels que $y_1<y_2$. Posons $x_1=\widetilde{f}^{-1}(y_1)$ et $x_2=\widetilde{f}^{-1}(y_2)$. Si $x_1\geq x_2$, alors $\widetilde{f}(x_1)\geq\widetilde{f}(x_2)$, comme f est croissante, soit encore $y_1\geq y_2$, ce qui est absurde, donc $x_1<x_2$ et $\widetilde{f}^{-1}$ est strictement croissante.
Montrons qu'elle est également continue. Soit $y_0\in f(I)$; posons $x_0=\widetilde{f}^{-1}(y_0)$ et donnons-nous un réel $\varepsilon>0$ tel que $x_0-\varepsilon$ et $x_0+\varepsilon$ appartiennent à I. Posons alors $y_1=\widetilde{f}(x_0-\varepsilon)$ et $y_2=\widetilde{f}(x_0-\varepsilon)$. L'application f étant strictement croissante, on a $y_1<y_0<y_2$ et, pour tout $y\in]y_1,y_2[$, $\widetilde{f}^{-1}(y_1)<\widetilde{f}^{-1}(y)<\widetilde{f}^{-1}(y_2)$, c'est-à-dire $x_0-\varepsilon<\widetilde{f}^{-1}(y)<x_0+\varepsilon$. Il existe donc $\beta>0$ tel que $|y-y_0|\leq\beta\Rightarrow\left|\widetilde{f}^{-1}(y)-\widetilde{f}^{-1}(y_0)\right|\leq\varepsilon$. Par conséquent, $\widetilde{f}$ est continue en y_0.

□

Exemple de fonction réciproque. La fonction sinus est continue et strictement croissante sur l'intervalle $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Elle induit donc une bijection de $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ sur $[-1,1]$. Sa bijection réciproque est appelée *arc sinus* et notée arcsin. C'est une fonction continue strictement croissante de $[-1,1]$ sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

Continuité uniforme

Nous introduisons à présent une notion de continuité plus forte que celle donnée dans la définition B.82.

Définition B.93 *Soit f une fonction définie sur I à valeurs dans $\mathbb{R}$. On dit que f est **uniformément continue sur** I si et seulement si*

$$\forall \varepsilon > 0,\ \exists \alpha > 0,\ \forall (x,x') \in I^2,\ (\left|x-x'\right| \leq \alpha \Rightarrow \left|f(x)-f(x')\right| \leq \varepsilon).$$

Le qualificatif d'« uniforme » signifie que le choix de α en fonction de ε ne dépend pas du point considéré : il est le même sur tout l'intervalle I.

Exemples. L'application qui à tout réel x associe $|x|$ est uniformément continue sur $\mathbb{R}$, mais celle qui à tout réel x associe x^2 ne l'est pas.

La preuve du résultat suivant est immédiate.

Proposition B.94 *Si f est uniformément continue sur I, alors f est continue sur I.*

REPRENDRE Comme on l'a vu précédemment, il existe des fonctions continues non uniformément continues. Cependant, lorsque I est un segment de $\mathbb{R}$, c'est-à-dire un intervalle fermé et borné, nous disposons du résultat suivant.

Théorème B.95 (« théorème de Heine [12] ») *Toute fonction continue sur un segment $[a,b]$ de $\mathbb{R}$ est uniformément continue sur $[a,b]$.*

DÉMONSTRATION. Raisonnons par l'absurde. Soit f une fonction continue et non uniformément continue sur $[a,b]$. Il existe donc $\varepsilon > 0$ tel que

$$\forall \alpha > 0,\ \exists (x,x') \in [a,b]^2,\ \left|x-x'\right| \leq \alpha \text{ et } \left|f(x)-f(x')\right| > \varepsilon.$$

En particulier, en prenant $\alpha = \dfrac{1}{n}$, il existe $(x_n, x'_n) \in [a,b]^2$ tel que

$$\forall n \in \mathbb{N}^*,\ \left|x_n - x'_n\right| \leq \frac{1}{n} \text{ et } \left|f(x_n)-f(x'_n)\right| > \varepsilon.$$

La suite $(x_n)_{n\in\mathbb{N}^*}$ étant bornée, elle admet, en vertu du théorème de Bolzano–Weierstrass (voir le théorème B.40), une sous-suite, notée $(x_{\sigma(n)})_{n\in\mathbb{N}^*}$, convergente vers un réel, noté l, appartenant à $[a,b]$. Comme

$$\forall n \in \mathbb{N}^*,\ \left|x_{\sigma(n)} - x'_{\sigma(n)}\right| \leq \frac{1}{\sigma(n)} \leq \frac{1}{n},$$

on déduit que la suite extraite $(x'_{\sigma(n)})_{n\in\mathbb{N}^*}$ converge aussi vers l. L'application f étant continue en l, les suites $(f(x_{\sigma(n)}))_{n\in\mathbb{N}^*}$ et $(f(x'_{\sigma(n)}))_{n\in\mathbb{N}^*}$ convergent vers $f(l)$ et, par conséquent, $\lim\limits_{n\to+\infty} \left|f(x_{\sigma(n)})-f(x'_{\sigma(n)})\right| = 0$, ce qui contredit le fait que $\left|f(x_{\sigma(n)})-f(x'_{\sigma(n)})\right| > \varepsilon$. □

Applications lipschitziennes

Nous allons à présent introduire une propriété de régularité des applications plus forte que la notion de continuité.

Définition B.96 *Soit f une fonction définie sur I à valeurs réelles et un réel k strictement positif. On dit que f est **lipschitzienne** si et seulement si*

$$\forall (x,x') \in I^2,\ \left|f(x)-f(x')\right| \leq k\left|x-x'\right|.$$

12. Heinrich Eduard Heine (15 mars 1821 - 21 octobre 1881) était un mathématicien allemand. Il est célèbre pour ses résultats en analyse réelle et sur les fonctions spéciales.

Lorsque $k \in]0,1[$, on dit que l'application f est *contractante*. Le plus petit réel k tel que f soit k-lipschitzienne est appelé la *constante de Lipschitz* [13] *de* f.

Proposition B.97 *Une application lipschitzienne est uniformément continue.*

DÉMONSTRATION. Supposons f k-lipschitzienne ($k \in \mathbb{R}_+^*$) et soit $\varepsilon > 0$. Si $k = 0$, f est constante sur I et donc uniformément continue sur I. Si $k > 0$, en prenant $\alpha = \dfrac{\varepsilon}{k}$, nous obtenons

$$\forall (x,x') \in I^2,\ (\left|x - x'\right| \leq \alpha \Rightarrow \left|f(x) - f(x')\right| \leq \varepsilon),$$

ce qui montre que f est uniformément continue sur I. □

B.3.5 Dérivabilité *

Dans cette section, $\mathbb{K}$ désigne un corps ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$), I un intervalle de $\mathbb{R}$ non vide et non réduit à un point et $\mathscr{F}(I,\mathbb{K})$ est l'ensemble des applications définies sur I et à valeurs dans $\mathbb{K}$.

Dérivabilité en un point

Définition B.98 *Soit $f \in \mathscr{F}(I,\mathbb{K})$ et $a \in I$. On dit que f est **dérivable en** a si et seulement si $\lim\limits_{x\to a} \dfrac{f(x)-f(a)}{x-a}$ existe et est finie ; cette limite est alors appelée **dérivée de** f **en** a et notée $f'(a)$.*

En posant $h = x - a$, on obtient une autre écriture, très souvent employée,

$$f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h},$$

dans laquelle le rapport $\dfrac{f(a+h)-f(a)}{h}$ s'appelle le *taux d'accroissement de f entre a et $a+h$*. On note aussi parfois $\dfrac{\mathrm{d}f}{\mathrm{d}x}(a)$ au lieu de $f'(a)$.

Définition B.99 *Soit $f \in \mathscr{F}(I,\mathbb{K})$ et $a \in I$.*

*i) On dit que f est **dérivable à droite en** a si et seulement si $\lim\limits_{x\to a^+} \dfrac{f(x)-f(a)}{x-a}$ existe et est finie ; cette limite est alors appelée **dérivée de f à droite en** a et notée $f'_d(a)$.*

*ii) On dit que f est **dérivable à gauche en** a si et seulement si $\lim\limits_{x\to a^-} \dfrac{f(x)-f(a)}{x-a}$ existe et est finie ; cette limite est alors appelée **dérivée de f à gauche en** a et notée $f'_g(a)$.*

Exemple. L'application $x \mapsto |x|$ de $\mathbb{R}$ dans $\mathbb{R}$ est dérivable à gauche en 0 et dérivable à droite en 0, et $f'_g(0) = -1$, $f'_d(0) = 1$.

Le résultat suivant est immédiat.

Proposition B.100 *Soit $f \in \mathscr{F}(I,\mathbb{K})$ et $a \in I$. Pour que l'application f soit dérivable en a, il faut et il suffit que f soit dérivable à gauche et à droite en a et que $f'_g(a) = f'_d(a)$. Dans ces conditions, on a $f'(a) = f'_g(a) = f'_d(a)$.*

Proposition B.101 *Soit $f \in \mathscr{F}(I,\mathbb{K})$ et $a \in I$. Si l'application f est dérivable en a, alors elle est continue en a.*

DÉMONSTRATION. On sait d'une part que

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = f'(a).$$

D'autre part, on a

$$\lim_{x\to a} f(x) - f(a) = \left(\lim_{x\to a} x - a\right)\left(\lim_{x\to a} \frac{f(x)-f(a)}{x-a}\right),$$

d'où $\lim\limits_{x\to a} f(x) = f(a)$. □

13. Rudolph Otto Sigismund Lipschitz (14 mai 1832 - 7 octobre 1903) était un mathématicien allemand. Son travail s'étend sur des domaines aussi variés que la théorie des nombres, l'analyse, la géométrie différentielle et la mécanique classique.

Remarque. La réciproque de cette proposition est fausse : une application peut être continue en a sans pour autant être dérivable en ce point. Par exemple, l'application $x \mapsto |x|$ de $\mathbb{R}$ dans $\mathbb{R}$, déjà étudiée plus haut, est continue en 0 sans y être dérivable.

Propriétés algébriques des fonctions dérivables en un point

Théorème B.102 *Soit $a \in I$, $\lambda \in \mathbb{K}$ et $f, g \in \mathscr{F}(I, \mathbb{K})$ deux applications dérivables en a. Alors on a*

i) $f+g$ *est dérivable en* a *et* $(f+g)'(a) = f'(a) + g'(a)$.

ii) λf *est dérivable en* a *et* $(\lambda f)'(a) = \lambda f'(a)$.

iii) fg *est dérivable en* a *et* $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$.

iv) Si $g(a) \neq 0$, $\dfrac{f}{g}$ *est dérivable en* a *et* $\left(\dfrac{f}{g}\right)'(a) = \dfrac{f'(a)g(a) - f(a)g'(a)}{g(a)^2}$.

DÉMONSTRATION.

i) On a
$$\frac{(f+g)(x) - (f+g)(a)}{x-a} = \frac{f(x)-f(a)}{x-a} + \frac{g(x)-g(a)}{x-a} \underset{x\to a}{\longrightarrow} f'(a) + g'(a).$$

ii) On a
$$\frac{(\lambda f)(x) - (\lambda f)(a)}{x-a} = \lambda \frac{f(x)-f(a)}{x-a} \underset{x\to a}{\longrightarrow} \lambda f'(a).$$

iii) On a
$$\frac{(fg)(x) - (fg)(a)}{x-a} = \frac{f(x)g(x) - f(a)g(a)}{x-a} = \frac{(f(x)-f(a))g(a)}{x-a} + \frac{f(a)(g(x)-g(a))}{x-a}.$$
Puisque f et g sont dérivables en a, ces applications sont continues en a, donc $\lim\limits_{x\to a} f(x) = f(a)$ et $\lim\limits_{x\to a} g(x) = g(a)$, d'où $\lim\limits_{x\to a} \dfrac{(f+g)(x) - (f+g)(a)}{x-a} = f'(a)g(a) + f(a)g'(a)$.

iv) Puisque $g(a) \neq 0$ et que g est continue en a, on a, au voisinage de a, $g(x) \neq 0$. La fonction $\dfrac{1}{g}$ est alors définie au voisinage de a. De plus, on a
$$\frac{1}{x-a}\left(\left(\frac{1}{g}\right)(x) - \left(\frac{1}{g}\right)(a)\right) = \frac{1}{x-a}\left(\frac{1}{g(x)} - \frac{1}{g(a)}\right) = \frac{g(a)-g(x)}{x-a}\frac{1}{g(x)g(a)},$$
d'où $\lim\limits_{x\to a} \dfrac{1}{x-a}\left(\left(\dfrac{1}{g}\right)(x) - \left(\dfrac{1}{g}\right)(a)\right) = -\dfrac{g'(a)}{g(a)^2}$. Le résultat se déduit alors de 2) en utilisant que $\dfrac{f}{g} = f\dfrac{1}{g}$.

□

Dérivée d'une composée de fonctions

Théorème B.103 *Soit J un intervalle de $\mathbb{R}$, $f : I \to \mathbb{K}$ telle que $f(I) \subset J$, $g : J \to \mathbb{K}$ et $a \in I$. Si f est dérivable en a et si g est dérivable en $f(a)$, alors $g \circ f$ est dérivable en a et $(g\circ f)'(a) = f'(a)g'(f(a))$.*

DÉMONSTRATION. On a
$$\frac{(g\circ f)(x) - (g\circ f)(a)}{x-a} = \frac{g(f(x)) - g(f(a))}{x-a} = \frac{g(f(x)) - g(f(a))}{f(x)-f(a)}\frac{f(x)-f(a)}{x-a},$$
d'où
$$\lim_{x\to a} \frac{(g\circ f)(x) - (g\circ f)(a)}{x-a} = \left(\lim_{x\to a} \frac{g(f(x)) - g(f(a))}{f(x)-f(a)}\right)\left(\lim_{x\to a} \frac{f(x)-f(a)}{x-a}\right) = g'(f(a))f'(a).$$

□

Dérivée d'une fonction réciproque

Théorème B.104 *Soit $a \in I$ et $f : I \to \mathbb{R}$ une application continue et strictement monotone sur I, dérivable en a et telle que $f'(a) \neq 0$; on note $\widetilde{f} : I \to f(I)$ l'application qui à tout x de I associe $f(x)$. Alors la fonction réciproque $\widetilde{f}^{-1}$ est dérivable en $f(a)$ et l'on a $(\widetilde{f}^{-1})'(f(a)) = \dfrac{1}{f'(a)}$.*

DÉMONSTRATION. D'après le théorème B.92, l'application $\widetilde{f} : I \to f(I)$ est bijective et sa bijection réciproque $\widetilde{f}^{-1}$ est strictement monotone, de même sens que f, et continue sur $f(I)$. Pour tout y de $f(I) \setminus \{f(a)\}$, on a alors

$$\frac{\widetilde{f}^{-1}(y) - \widetilde{f}^{-1}(f(a))}{y - f(a)} = \frac{\widetilde{f}^{-1}(y) - a}{f(\widetilde{f}^{-1}(y)) - f(a)}.$$

Comme f est dérivable en a, de dérivée non nulle en ce point, et que $\lim\limits_{y \to f(a)} \widetilde{f}^{-1}(y) = a$, on obtient, après composition des limites,

$$\lim_{y \to f(a)} \frac{\widetilde{f}^{-1}(y) - a}{f(\widetilde{f}^{-1}(y)) - f(a)} = \frac{1}{f'(a)}.$$

L'application $\widetilde{f}^{-1}$ est donc dérivable en $f(a)$. □

Exemples. On sait que la restriction de la fonction tangente à l'intervalle $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$ est une bijection continue de cet intervalle sur $\mathbb{R}$. Sa dérivée, la fonction $x \mapsto 1 + \tan^2 x$, ne s'annule pas. Sa bijection réciproque arc tangente est donc dérivable sur $\mathbb{R}$, de dérivée

$$(\arctan x)' = \frac{1}{1 + \tan^2(\arctan x)} = \frac{1}{1 + x^2}, \ \forall x \in \mathbb{R}.$$

Application dérivée

Définition B.105 *Soit $f \in \mathscr{F}(I, \mathbb{K})$. On appelle **dérivée de** f l'application qui à chaque x de I tel que $f'(x)$ existe associe $f'(x)$.*

Dérivées successives

Définitions B.106 *Soit $f \in \mathscr{F}(I, \mathbb{K})$. On définit les **dérivées successives de** f par récurrence pour tout n de $\mathbb{N}^*$:*

- *pour $a \in I$, $f^{(n)}(a)$ est, si elle existe, la dérivée de $f^{(n-1)}$ en a,*
- *$f^{(n)}$ est l'application dérivée de $f^{(n-1)}$.*

*On appelle **dérivée n^e de f en** a le réel $f^{(n)}(a)$ et **application dérivée n^e de** f l'application $x \mapsto f^{(n)}(x)$.*
*On dit que f est n **fois dérivable sur** I si et seulement si $f^{(n)}$ est définie sur I. Enfin, on dit que f est **indéfiniment dérivable sur** I si et seulement si f est n fois dérivable sur I pour tout entier positif n.*

Par convention, $f^{(0)} = f$ et l'on note aussi $\dfrac{\mathrm{d}^n f}{\mathrm{d}x^n}$ au lieu de $f^{(n)}$. Comme on l'a déjà vu, on écrit souvent $f' = f^{(1)}$ et, de la même manière, $f'' = f^{(2)}$ et $f''' = f^{(3)}$.

Proposition B.107 *Soit $\lambda \in \mathbb{K}$, $n \in \mathbb{N}^*$ et $f, g : I \to \mathbb{K}$ des applications n fois dérivables sur l'intervalle I. On a*

i) $f + g$ est n fois dérivable sur I et $(f+g)^{(n)} = f^{(n)} + g^{(n)}$.

ii) λf est n fois dérivable sur I et $(\lambda f)^{(n)} = \lambda f^{(n)}$.

iii) fg est n fois dérivable sur I et on a la formule, dite de Leibniz, suivante

$$(fg)^{(n)} = \sum_{k=0}^{n} C_n^k f^{(k)} g^{(n-k)}.$$

iv) Si $(\forall x \in I,\ g(x) \neq 0)$, alors $\dfrac{f}{g}$ est n fois dérivable.

DÉMONSTRATION. Tous ces résultats se démontrent par récurrence sur n. Nous laissons au lecteur les preuves (aisées) de 1) et de 2).

iii) Le cas $n=1$ a été traité dans le théorème B.102. Supposons la propriété vraie au rang $n>1$. Soit f et g deux fonctions de I dans $\mathbb{K}$, $(n+1)$ fois dérivables sur I. D'après l'hypothèse de récurrence, fg est n fois dérivable sur I et

$$(fg)^{(n)}=\sum_{k=0}^{n}C_n^k f^{(k)}g^{(n-k)}.$$

Ainsi, $(fg)^{(n)}$ apparaît comme somme de produits d'applications dérivables sur I et est donc dérivable sur I. On a alors

$$\begin{aligned}\left((fg)^{(n)}\right)'&=\left(\sum_{k=0}^{n}C_n^k f^{(k)}g^{(n-k)}\right)'\\&=\sum_{k=0}^{n}C_n^k f^{(k+1)}g^{(n-k)}+\sum_{k=0}^{n}C_n^k f^{(k)}g^{(n-k+1)}\\&=\sum_{k=1}^{n+1}C_n^{k-1} f^{(k)}g^{(n-k+1)}+\sum_{k=0}^{n}C_n^k f^{(k)}g^{(n-k+1)}\\&=f^{(n+1)}g+\sum_{k=1}^{n}\left(C_n^{k-1}+C_n^k\right)f^{(k)}g^{(n-k+1)}+fg^{(n+1)}\\&=f^{(n+1)}g+\sum_{k=1}^{n}C_{n+1}^k f^{(k)}g^{(n-k+1)}+fg^{(n+1)}\\&=\sum_{k=0}^{n+1}C_{n+1}^k f^{(k)}g^{(n+1-k)}.\end{aligned}$$

i) Le cas $n=1$ a déjà été vu dans le théorème B.102. Supposons la propriété vraie au rang $n>1$. Soit f et g deux fonctions de I dans $\mathbb{K}$, $(n+1)$ fois dérivables sur I et telles que $(\forall x\in I,\ g(x)\neq 0)$. L'application $\dfrac{f}{g}$ étant dérivable sur I, nous avons

$$\left(\frac{f}{g}\right)'=\frac{f'g-fg'}{g^2}.$$

Puisque f, f', g et g' sont n fois dérivables sur I, $f'g-fg'$ et g^2 le sont aussi. Il résulte alors de l'hypothèse de récurrence que $\dfrac{f'g-fg'}{g^2}$ est n fois dérivable sur I. Finalement, $\dfrac{f}{g}$ est $(n+1)$ fois dérivable sur I.

□

Nous introduisons enfin la notion de *classe* d'une fonction.

Définition B.108 *Soit $f\in\mathscr{F}(I,\mathbb{K})$.*

*i) Soit $n\in\mathbb{N}$. On dit que f est **de classe $\mathscr{C}^n$ sur** I si et seulement si f est n fois dérivable sur I et $f^{(n)}$ est continue sur I.*

*ii) On dit que f est **de classe $\mathscr{C}^\infty$ sur** I si et seulement si f est indéfiniment dérivable sur I.*

Pour $n\in\mathbb{N}\cup\{+\infty\}$, on note $\mathscr{C}^n(I,\mathbb{K})$ l'ensemble des applications de classe $\mathscr{C}^n$ de I dans $\mathbb{K}$.

Remarques.
— $f\in\mathscr{C}^0(I,\mathbb{K})$ si et seulement si f est continue sur I.
— Pour tout $(p,n)\in(\mathbb{N}\cup\{+\infty\})^2$ tel que $p\leq n$, on a $\mathscr{C}^n(I,\mathbb{K})\subset\mathscr{C}^p(I,\mathbb{K})$.

Extrema locaux d'une fonction réelle dérivable

Définitions B.109 *Soit $a\in I$ et $f\in\mathscr{F}(I,\mathbb{R})$. On dit que f admet*
- *un **maximum local en** a si et seulement si, au voisinage de a, $f(x)\leq f(a)$,*
- *un **minimum local en** a si et seulement si, au voisinage de a, $f(x)\geq f(a)$,*
- *un **maximum local strict en** a si et seulement si, au voisinage de a sauf en a, $f(x)<f(a)$,*
- *un **minimum local strict en** a si et seulement si, au voisinage de a sauf en a, $f(x)>f(a)$,*
- *un **extremum local en** a si et seulement si f admet un maximum local ou un minimum local en a,*
- *un **extremum local strict en** a si et seulement si f admet un maximum local strict ou un minimum local strict en a.*

Exemples.
- Toute application constante admet en tout point un maximum et un minimum local.
- L'application de $\mathbb{R}$ de $\mathbb{R}$ qui à x associe $|x|$ admet un minimum local strict en 0.

Proposition B.110 *Soit $f \in \mathscr{F}(I, \mathbb{R})$. Si f admet en un point intérieur a de I un extremum local et si f est dérivable en a, alors $f'(a) = 0$.*

DÉMONSTRATION. Supposons, pour fixer les idées, que f admet un maximum local en a. Puisque f est dérivable en a, on a

$$\begin{aligned} f'(a) &= \lim_{x\to a} \frac{f(x)-f(a)}{x-a} \\ &= \lim_{x\to a^+} \frac{f(x)-f(a)}{x-a} \geq 0 \\ &= \lim_{x\to a^-} \frac{f(x)-f(a)}{x-a} \leq 0, \end{aligned}$$

d'où $f'(a) = 0$. □

Remarques.
- La réciproque de cette proposition est fausse. Par exemple, l'application $x \mapsto x^3$ de $\mathbb{R}$ dans $\mathbb{R}$ à une dérivée nulle en 0 mais ne possède pas d'extremum en ce point.
- De fait, les extrema locaux d'une fonction définie sur un intervalle I seront recherchés aux points intérieurs de I où la dérivée de la fonction s'annule ou bien aux extrémités de I, où la fonction n'est pas dérivable.

Règle de L'Hôpital

Théorème B.111 (« règle de L'Hôpital[14] » [dLHô96]) *REPRENDRE Si f et g sont deux fonctions définies sur $[a, b]$, dérivables en a, s'annulant en a et telles que le quotient $\frac{f'(a)}{g'(a)}$ soit défini, alors $\lim_{x\to a^+} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}$.*

DÉMONSTRATION. A ECRIRE □

Théorème de Rolle

Le théorème des valeurs intermédiaires permet de démontrer le théorème suivant.

Théorème B.112 (« théorème de Rolle[15] ») *Soit $[a, b]$ un intervalle non vide de $\mathbb{R}$ et f une application de $[a, b]$ dans $\mathbb{R}$. Si f est continue sur $[a, b]$, dérivable sur $]a, b[$ et telle que $f(a) = f(b)$, alors il existe $c \in]a, b[$ tel que $f'(c) = 0$.*

DÉMONSTRATION. Puisque l'application f est continue sur le segment $[a, b]$, elle est bornée et atteint ses bornes (voir le théorème B.88). Notons $m = \inf_{x\in[a,b]} f(x)$ et $M = \sup_{x\in[a,b]} f(x)$. Si $M = m$, alors f est constante et $f'(x) = 0$ pour tout $x \in]a, b[$. Supposons $m < M$. Comme $f(a) = f(b)$, on a soit $M \neq f(a)$, soit $m \neq f(a)$. Ramenons-nous au cas $M \neq f(a)$. Il existe alors un point $c \in]a, b[$ tel que $f(c) = M$. Soit $x \in [a, b]$ tel que $f(x) \leq M = f(c)$. Si $x > c$, on a $\frac{f(x)-f(c)}{x-c} \leq 0$, et si $x < c$, on obtient $\frac{f(x)-f(c)}{x-c} \geq 0$. L'application f étant dérivable en c, nous obtenons, en passant à la limite, $f'(c) \leq 0$ et $f'(c) \geq 0$, d'où $f'(c) = 0$. □

Remarque. Le réel c n'est pas nécessairement unique.

14. Guillaume François Antoine de L'Hôpital (1661 - 2 février 1704) était un mathématicien français. Son nom est associé à une règle permettant le calcul d'une limite de quotient de forme indéterminée.

15. Michel Rolle (21 avril 1652 - 8 novembre 1719) était un mathématicien français. S'il inventa la notation $\sqrt[n]{x}$ pour désigner la racine n^{e} d'un réel x, il reste principalement connu pour avoir établi en 1691, dans le cas particulier des polynômes réels à une variable, une première version du théorème portant aujourd'hui son nom.

Théorème des accroissements finis

Le théorème de Rolle permet à son tour de prouver le résultat suivant, appelé le *théorème des accroissements finis*.

Théorème B.113 (« théorème des accroissements finis ») *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une application de $[a,b]$ dans $\mathbb{R}$. Si f est continue sur $[a,b]$ et dérivable sur $]a,b[$, alors il existe $c \in]a,b[$ tel que*

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$

DÉMONSTRATION. Considérons la fonction $\varphi : [a,b] \to \mathbb{R}$ définie par

$$\varphi(x) = f(x) - \frac{f(b)-f(a)}{b-a}(x-a).$$

Il est clair que φ est continue sur $[a,b]$, dérivable sur $]a,b[$ et que $\varphi(a) = \varphi(b)$. En appliquant le théorème de Rolle à φ, on obtient qu'il existe $c \in]a,b[$ tel que $\varphi'(c) = 0$, c'est-à-dire tel que

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$

□

Remarque. Là encore, le réel c n'est pas forcément unique.

On déduit directement l'*inégalité des accroissements finis* du théorème B.113. Celle-ci est plus générale que le théorème du même nom, dans la mesure où elle s'applique à d'autres fonctions que les fonctions d'une variable réelle à valeurs dans $\mathbb{R}$, comme par exemple les fonctions de $\mathbb{R}$ dans $\mathbb{C}$ ou de $\mathbb{R}^n$ $(n \in \mathbb{N}^*)$ dans $\mathbb{R}$.

Théorème B.114 (« inégalité des accroissements finis ») *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$. Si f est une fonction continue sur $[a,b]$, dérivable sur $]a,b[$ et qu'il existe un réel $M > 0$ tel que*

$$\forall x \in]a,b[,\ \left|f'(x)\right| \leq M,$$

alors on a

$$|f(b)-f(a)| \leq M\,|b-a|\,.$$

Sens de variation d'une fonction dérivable

Les résultats précédents permettent d'établir un lien entre le sens de variation d'une fonction et le signe de sa dérivée. Nous avons la

Proposition B.115 *Soit $(a,b) \in \overline{\mathbb{R}}^2$, tel que $a < b$, et f une fonction dérivable sur $]a,b[$. Alors*

i) f est croissante si et seulement si $(\forall x \in]a,b[,\ f'(x) \geq 0)$,

ii) f est décroissante si et seulement si $(\forall x \in]a,b[,\ f'(x) \leq 0)$,

iii) f est constante si et seulement si $(\forall x \in]a,b[,\ f'(x) = 0)$.

DÉMONSTRATION.

i) Supposons f croissante. Soit $x_0 \in]a,b[$, pour tout $x \in]a,b[$ tel que $x > x_0$, on a

$$\frac{f(x)-f(x_0)}{x-x_0} \geq 0.$$

En passant à la limite $x \to x_0$, on déduit que $f'(x_0) \geq 0$. Réciproquement, supposons que, pour tout $x \in]a,b[$, $f'(x) \geq 0$. Soit x_1 et x_2 deux éléments de $]a,b[$ tels que $x_1 < x_2$. En appliquant le théorème des accroissements finis à f sur $[x_1,x_2]$, on voit qu'il existe $c \in [x_1,x_2]$ tel que

$$f(x_2)-f(x_1) = f'(c)(x_2-x_1) \geq 0,$$

et f est par conséquent croissante sur]a,b[.

ii) L'application f est décroissante si et seulement si $-f$ est croissante ; ii) résulte donc de i).

iii) L'application f est constante si et seulement si elle est à la fois croissante et décroissante ; iii) résulte donc de i) et de ii).

□

Formules de Taylor

Le théorème suivant constitue une généralisation du théorème des accroissements finis.

Théorème B.116 (« formule de Taylor[16]–Lagrange ») *Soit n un entier naturel, $[a,b]$ un intervalle non vide de $\mathbb{R}$ et $f : [a,b] \to \mathbb{R}$ une fonction de classe $\mathscr{C}^n$ sur $[a,b]$. On suppose de plus que $f^{(n)}$ est dérivable sur $]a,b[$. Alors, il existe $c \in]a,b[$ tel que*

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(b-a)^{n+1},$$

soit encore

$$f(b) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(b-a)^k + \frac{f^{(n+1)}(c)}{(n+1)!}(b-a)^{n+1}.$$

DÉMONSTRATION. Soit A le réel tel que

$$\frac{(b-a)^{n+1}}{(n+1)!}A = f(b) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(b-a)^k.$$

Il s'agit de montrer que $A = f^{(n+1)}(c)$, avec $c \in]a,b[$. On définit pour cela la fonction $\varphi : [a,b] \to \mathbb{R}$ comme suit

$$\varphi(x) = f(b) - \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}(b-x)^k - \frac{(b-x)^{n+1}}{(n+1)!}A.$$

Cette fonction est continue sur $[a,b]$, dérivable sur $]a,b[$ et vérifie par ailleurs $\varphi(a) = \varphi(b) = 0$. D'après le théorème de Rolle, il existe donc $c \in]a,b[$ tel que $\varphi'(c) = 0$. Or, pour tout $x \in]a,b[$, on a

$$\begin{aligned}\varphi'(x) &= \sum_{k=1}^{n} \frac{f^{(k)}(x)}{(k-1)!}(b-x)^{k-1} - \sum_{k=0}^{n} \frac{f^{(k+1)}(x)}{k!}(b-x)^k + \frac{(b-x)^n}{n!}A \\ &= \frac{(b-x)^n}{n!}\left(-f^{(n+1)}(x) + A\right).\end{aligned}$$

Par conséquent, on déduit de $\varphi'(c) = 0$ que $A = f^{(n+1)}(c)$. □

Remarques.

— Le terme $\dfrac{f^{(n+1)}(c)}{(n+1)!}(b-a)^{n+1}$ est appelé *reste de Lagrange*.

— Dans le cas particulier $n = 0$, on retrouve l'égalité du théorème des accroissements finis.

Théorème B.117 (« formule de Taylor–Young[17] ») *Soit n un entier naturel, I un intervalle ouvert non vide de $\mathbb{R}$, a un point de I et $f : I \to \mathbb{R}$ une fonction admettant une dérivée n^e au point a. Alors, il existe une fonction ϵ à valeurs réelles, définie sur I et vérifiant $\lim\limits_{x \to a} \epsilon(x) = 0$, telle que, pour tout x appartenant à I,*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + (x-a)^n \epsilon(x).$$

DÉMONSTRATION. La formule se démontre par récurrence sur l'entier n, en considérant l'assertion équivalente : *pour toute fonction $f : I \to \mathbb{R}$, n fois dérivable au point a, on a*

$$\lim_{\substack{x \to a \\ x \neq a}} \frac{1}{(x-a)^n}\left(f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k\right) = 0.$$

Pour $n = 0$, le résultat est immédiat. Pour $n = 1$, l'assertion découle de la dérivabilité de la fonction f au point a.

16. Brook Taylor (18 août 1685 - 30 novembre 1731) était un mathématicien, artiste peintre et musicien anglais. Il inventa le calcul aux différences finies et découvrit l'intégration par parties.

17. William Henry Young (20 octobre 1863 - 7 juillet 1942) était un mathématicien anglais. Ses études portèrent principalement sur la théorie de la mesure et de l'intégration, les séries de Fourier et le calcul différentiel des fonctions de plusieurs variables.

Soit à présent $n \geq 2$ et f une fonction n fois dérivable en a. On suppose l'assertion vérifiée jusqu'au rang $n-1$. La dérivée f', définie dans un voisinage ouvert du point a, est une fonction $n-1$ fois dérivable en a et, par hypothèse de récurrence, pour tout $\varepsilon > 0$, il existe un réel $\eta_\varepsilon > 0$ tel que

$$\left| f'(x) - \sum_{k=0}^{n-1} \frac{f^{(k+1)}(a)}{k!} (x-a)^k \right| \leq \varepsilon \, |x-a|^{n-1}, \ \forall x \in I \cap]a-\eta_\varepsilon, a+\eta_\varepsilon[.$$

On définit alors, pour tout $t \in I \cap]a-\eta_\varepsilon, a+\eta_\varepsilon[$, la fonction dérivable

$$g(t) = f(t) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (t-a)^k,$$

telle que $g(a) = 0$. Il résulte de la majoration ci-dessus et de l'inégalité des accroissements finis que

$$|g(x) - g(a)| \leq \varepsilon \, |x-a|^{n-1} \, |x-a|, \ \forall x \in I \cap]a-\eta_\epsilon, a+\eta_\epsilon[,$$

soit encore

$$\frac{1}{|x-a|^n} \left| f(x) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k \right| \leq \varepsilon, \ \forall x \in I \cap]a-\eta_\epsilon, a+\eta_\epsilon[,$$

ce qui implique l'assertion au rang n. □

B.4 Intégrales *

Cette section est consacrée à la notion d'*intégrabilité au sens de Riemann* des fonctions, qui permet d'aborder le calcul numérique des intégrales par des formules de quadrature au chapitre 7. Dans toute cette section, on désigne par $[a, b]$ un intervalle borné et non vide de $\mathbb{R}$.

B.4.1 Intégrabilité au sens de Riemann *

INTRO ?

Définition B.118 (subdivision d'un intervalle) *Soit n un entier naturel strictement plus grand que 1. On appelle **subdivision de** $[a, b]$ toute famille de points $\sigma = \{x_i\}_{i=0,\ldots,n}$ telle que*

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Si σ et τ sont deux subdivisions d'un même intervalle, on dit que τ est *plus fine* que σ si $\sigma \subset \tau$. D'autre part, le *pas* d'une subdivision σ est le réel défini par $h(\sigma) = \max_{i \in \{1,\ldots,n\}} |x_i - x_{i-1}|$.

Définition B.119 (fonction en escalier) *Une application réelle f définie sur $[a, b]$ est dite **en escalier** sur $[a, b]$ s'il existe une subdivision $\sigma = \{x_i\}_{i=0,\ldots,n}$ de $[a, b]$ telle que f soit constante sur chaque intervalle $]x_{i-1} - x_i[$, $i = 1, \ldots, n$.*

On remarque qu'une fonction en escalier sur un intervalle ne prend qu'un nombre fini de valeurs. Elle est donc bornée et ne possède qu'un nombre fini de points de discontinuité. L'ensemble des fonctions en escalier sur un intervalle $[a, b]$, que l'on note $\mathscr{E}([a, b])$, est un sous-espace vectoriel des fonctions réelles définies sur $[a, b]$.

On dit qu'une subdivision $\sigma = \{x_i\}_{i=0,\ldots,n}$ d'un intervalle $[a, b]$ est *adaptée* à une fonction en escalier sur le même intervalle si cette dernière est constante sur chaque intervalle $]x_{i-1} - x_i[$, $i = 1, \ldots, n$.

Définition B.120 (intégrale d'une fonction en escalier) *Soit f une fonction en escalier sur $[a, b]$ et $\sigma = \{x_i\}_{i=0,\ldots,n}$ une subdivision de $[a, b]$ adaptée à f. On appelle **intégrale de f sur l'intervalle** $[a, b]$ le réel*

$$\int_a^b f(x) \, \mathrm{d}x = \sum_{i=1}^{n} c_i \, (x_i - x_{i-1}), \tag{B.2}$$

où le réel c_i, $1 \leq i \leq n$, désigne la valeur prise par f sur l'intervalle $]x_{i-1}, x_i[$.

L'intégrale d'une fonction en escalier est bien définie, comme le montre le résultat suivant.

Proposition B.121 *La valeur de l'intégrale* (B.2) *est indépendante de la subdivision adaptée choisie.*

DÉMONSTRATION. □

Proposition B.122 (propriétés de l'intégrale des fonctions en escalier) *Soit λ un réel et f et g deux fonctions de $\mathscr{E}([a,b])$. On a les propriétés suivantes.*

1. $\displaystyle\int_a^b (\lambda f(x)+g(x))\,\mathrm{d}x = \lambda\int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$
2. *Si f est positive sur $[a,b]$, alors* $\displaystyle\int_a^b f(x)\,\mathrm{d}x \geq 0.$
3. $\displaystyle\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)|\,\mathrm{d}x.$

DÉMONSTRATION. A ECRIRE □

En déduire propriété de monotonie

Il s'agit maintenant de donner un sens à l'intégrale d'une fonction lorsque celle-ci n'est pas en escalier. Pour cela, on va définir, pour toute fonction bornée sur l'intervalle $[a,b]$ son *intégrale supérieure* et son *intégrale inférieure*.

definition des intégrales

Définition B.123 (« intégrabilité au sens de Riemann ») *On dit qu'une fonction f définie et bornée sur $[a,b]$ est **intégrale au sens de Riemann sur** $[a,b]$ si son intégrale inférieure $I^-(f)$ et son intégrale supérieure $I^+(f)$ sont égales. L'intégrale de f est alors cette valeur commune,*

$$\int_a^b f(x)\,\mathrm{d}x = I^-(f) = I^+(f).$$

On a la caractérisation suivante.

Proposition B.124 *On dit qu'une fonction f définie et bornée sur $[a,b]$ est intégrable au sens de Riemann sur $[a,b]$ si et seulement si, pour tout $\varepsilon > 0$, on peut trouver des fonction en escalier g_ε^- et g_ε^+ telles que $g_\varepsilon^- \leq f \leq g_\varepsilon^+$ et*

$$\int_a^b \left(g_\varepsilon^+(x) - g_\varepsilon^-(x)\right)\mathrm{d}x \leq \varepsilon.$$

On notera que l'on peut définir d'une autre façon les intégrales supérieure et inférieure d'une fonction bornée en utilisant des fonctions en escalier particulières.

Définition B.125 (« sommes de Darboux ») *Soit $[a,b]$ un intervalle borné et non vide de $\mathbb{R}$, f une fonction bornée sur $[a,b]$ et σ une subdivision d'ordre n de $[a,b]$. On appelle **somme de Darboux inférieure** (resp. **supérieure**) de f relativement à σ la quantité*

$$s(f,\sigma) = \sum_{i=1}^{n}(x_i - x_{i-1}) \inf_{x\in[x_{i-1},x_i]} f(x) \text{ (resp. } S(f,\sigma) = \sum_{i=1}^{n}(x_i - x_{i-1}) \sup_{x\in[x_{i-1},x_i]} f(x)).$$

Proposition B.126 *REPRENDRE Les ensembles $\{s(f,\sigma)\}_{\sigma\in S_{a,b}}$ et $\{S(f,\sigma)\}_{\sigma\in S_{a,b}}$ admettent respectivement une borne supérieure et une borne inférieure.*

DÉMONSTRATION. Si σ et τ sont deux subdivisions de $[a,b]$ telles que $\sigma \subset \tau$, alors

$$s(f,\sigma) \leq s(f,\tau) \text{ et } S(f,\sigma) \geq S(f,\tau).$$

Si σ et τ sont deux subdivisions quelconques de $[a,b]$, on a

$$s(f,\sigma) \leq S(f,\tau)$$

□

On en déduit une autre définition de l'intégrale au sens de Riemann : une fonction f est dite intégrable au sens de Riemann sur le segment $[a,b]$ si elle est définie et bornée sur $[a,b]$ et si

$$\sup_{\sigma\in S_{a,b}} s(f,\sigma) = \inf_{\sigma\in S_{a,b}} S(f,\sigma).$$

Cette valeur commune est l'intégrale de f sur l'intervalle $[a,b]$, notée $\displaystyle\int_a^b f(x)\,\mathrm{d}x$.

REPRENDRE L'idée dans la définition des fonctions intégrables est que l'encadrement entre les sommes de Darboux inférieures et les sommes de Darboux supérieures peut être rendu aussi précis que l'on veut, déterminant ainsi un réel unique. Il est commode d'utiliser le critère d'intégrabilité suivant.

On obtient alors la version suivante du critère de la proposition ref.

Proposition B.127 (« critère de Darboux ») *Pour qu'une fonction réelle f définie et bornée sur $[a,b]$ soit intégrable au sens de Riemann sur $[a,b]$, il faut et il suffit que, pour tout réel $\varepsilon>0$, il existe une subdivision σ de $[a,b]$ telle que $S(f,\sigma)-s(f,\sigma)<\varepsilon$.*

DÉMONSTRATION. REPRENDRE Supposons f intégrable et donnons nous $\varepsilon>0$. D'après la définition de borne supérieure, il existe une subdivision Y de $[a,b]$ telle que $s(f,Y)>\int_a^b f(x)\,\mathrm{d}x-\varepsilon/2$. De même, il existe une subdivision Z telle que $S(f,Z)<\int_a^b f(x)\,\mathrm{d}x+\varepsilon/2$. En posant $X=Y\cup Z$, on obtient $S(f,X)-s(f,X)\leq S(f,Z)-s(f,Y)<\varepsilon$.

Réciproquement, supposons le critère vérifié. Pour tout $\varepsilon>0$, on peut donc trouver une subdivision X telle que $S(f,X)-s(f,X)<\varepsilon$, et donc comme $s(f,X)\leq s(f)\leq S(f)\leq S(f,X)$ on a $S(f)-s(f)<\varepsilon$. Comme ceci doit avoir lieu pour tout $\varepsilon>0$, c'est que $s(f)=S(f)$ et donc que f est intégrable sur $[a,b]$. □

sommes de Riemann ?

Terminons en donnant quelques propriétés de l'intégrale de Riemann.

Proposition B.128 *On a les propriétés suivantes.*

1. *L'ensemble des fonction intégrables au sens de Riemann sur $[a,b]$ est un espace vectoriel.*
2. *L'intégrale au sens de Riemann est une forme linéaire positive.*

DÉMONSTRATION. □

B.4.2 Classes de fonctions intégrables *

Caractériser les fonctions intégrables au sens de Riemann n'est pas chose aisée, c'est d'ailleurs dans le cadre d'une théorie de l'intégration plus « aboutie », due à Lebesgue, que l'on peut le faire. Nous nous contenterons ici de d'indiquer deux exemples élémentaires de classes de fonctions intégrables.

Proposition B.129 *Les fonctions monotones sur $[a,b]$ sont intégrables au sens de Riemann.*

DÉMONSTRATION. REPRENDRE On va traiter le cas de f croissante. Tout d'abord, f est bornée sur $[a,b]$ puisque pour tout $x\in[a,b]$, on a $f(a)\leq f(x)\leq f(b)$. Soit $sigma_n$ une subdivision régulière. Puisque f est croissante, la borne supérieure (respectivement inférieure) de f sur $[a_{i-1},a_i]$ est $f(a_i)$ (resp. $f(a_{i-1})$). On a donc

$$s(f,\sigma_n)=\sum_{i=1}^n \frac{b-a}{n} f(a_{i-1}),\ S(f,\sigma_n)=\sum_{i=1}^n \frac{b-a}{n} f(a_i).$$

Ceci donne $S(f,\sigma_n)-s(f,\sigma_n)=(f(b)-f(a))(b-a)/n$. Si on se donne $\varepsilon>0$, alors en choisissant l'entier $n>\varepsilon/(f(b)-f(a))(b-a)$, on obtient $S(f,\sigma_n)-s(f,\sigma_n)<\varepsilon$. Le critère d'intégrabilité est bien vérifié. □

Proposition B.130 *Les fonctions continues sur $[a,b]$ sont intégrables au sens de Riemann.*

DÉMONSTRATION. REPRENDRE On sait déjà que si f est continue sur $[a,b]$, elle est bornée sur $[a,b]$. On sait aussi par le théorème de Heine que f est uniformément continue. Donc, quand on se donne $\varepsilon>0$, il existe $\eta>0$ tel que, pour tous x,y de $[a,b]$, si $|x-y|<\eta$ alors $|f(x)-f(y)|<\varepsilon/(b-a)$. Choisissons un entier $n>(b-a)/\eta$. Sur chaque segment $[a_{i-1},a_i]$ découpé par la subdivision σ_n, la fonction f atteint sa borne inférieure m_i et sa borne supérieure M_i : on a $m_i=f(x_i)$ et

$M_i = f(y_i)$, et comme x_i et y_i appartiennent tous les deux à l'intervalle $[a_{i-1}, a_i]$ de longueur $(b-a)/n < \eta$, on doit avoir $M_i - m_i < \varepsilon/(b-a)$. Donc

$$S(f,\sigma_n) - s(f,\sigma_n) = \sum_{i=1}^{n} (M_i - m_i)\frac{b-a}{n} < \sum_{i=1}^{n} \frac{\varepsilon}{n} = \varepsilon,$$

et le critère d'intégrabilité est vérifié. □

A DEPLACER/SUPPRIMER Pour qu'une fonction soit intégrable au sens de Riemann sur $\mathbb{R}$, il est nécessaire qu'elle soit bornée et *à support compact* [18].

B.4.3 Théorème fondamental de l'analyse et intégration par parties **

formule de Chasles [19]

Théorème B.131 *[« théorème fondamental de l'analyse »] Soit f une fonction réelle définie et continue sur $[a,b]$. Alors, la fonction F, définie sur $[a,b]$ par*

$$\forall x \in [a,b],\ F(x) = \int_a^x f(t)\,\mathrm{d}t,$$

est dérivable sur $[a,b]$ et sa dérivée est f.

Si la fonction f est seulement intégrable au sens de Riemann sur $[a,b]$ et continue en un point x_0 de $[a,b]$, alors la fonction F est dérivable en x_0 et $F'(x_0) = f(x_0)$.

DÉMONSTRATION. a ecrire □

Definition primitive

Proposition B.132 (« formule d'intégration par parties ») *Soient f et g deux fonctions de classe $\mathscr{C}^1$ sur $[a,b]$. On alors*

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = -\int_a^b f'(x)g(x)\,\mathrm{d}t + f(b)g(b) - f(a)g(a).$$

DÉMONSTRATION. a ecrire □

B.4.4 Formules de la moyenne

Les résultats qui suivent fournissent d'autres exemples de conséquence du théorème des valeurs intermédiaires.

Théorème B.133 (« première formule de la moyenne ») *Soit f une fonction réelle définie et continue sur $[a,b]$. Alors, il existe un réel c strictement compris entre a et b vérifiant*

$$\frac{1}{b-a}\int_a^b f(t)\,\mathrm{d}t = f(c).$$

DÉMONSTRATION. La fonction f étant continue sur l'intervalle $[a,b]$, on pose $m = \inf_{x\in[a,b]} f(x)$ et $M = \sup_{x\in[a,b]} f(x)$ et on a alors

$$m(b-a) \le \int_a^b f(t)\,\mathrm{d}t \le M(b-a).$$

La conclusion s'obtient grâce au théorème des valeurs intermédiaires. □

18. On rappelle que l'on définit le *support* d'une fonction réelle d'une variable réelle par $\mathrm{supp}(f) = \overline{\{x \in \mathbb{R} \mid f(x) \neq 0\}}$. Dire qu'une fonction est à support compact signifie alors qu'elle est nulle en dehors d'un ensemble borné.

19. Michel Chasles (15 novembre 1793 - 18 décembre 1880) était un mathématicien et historien des mathématiques français. Auteur d'ouvrages de référence en géométrie, il est aussi connu pour sa solution au problème d'énumération de coniques tangentes à cinq coniques données contenues dans un plan ou un théorème de géodésie physique montrant que toute fonction harmonique peut se représenter par un potentiel de simple couche sur l'une quelconque de ses surfaces équipotentielles.

Théorème B.134 (« première formule de la moyenne généralisée ») *Soit f une fonction réelle définie et continue sur $[a,b]$ et g une fonction réelle définie, continue et positive sur $[a,b]$. Alors, il existe un réel c strictement compris entre a et b vérifiant*

$$\int_a^b f(t)g(t)\,\mathrm{d}t = f(c)\int_a^b g(t)\,\mathrm{d}t.$$

DÉMONSTRATION. La fonction f étant continue sur l'intervalle $[a,b]$, on pose $m = \inf_{x\in[a,b]} f(x)$ et $M = \sup_{x\in[a,b]} f(x)$. Par positivité de la fonction g, on obtient

$$m\,g(x) \le f(x)\,g(x) \le M\,g(x),\ \forall x \in [a,b].$$

En intégrant ces inégalités entre a et b, il vient

$$m\int_a^b g(t)\,\mathrm{d}t \le \int_a^b f(t)\,g(t)\,\mathrm{d}t \le M\int_a^b g(t)\,\mathrm{d}t.$$

Si l'intégrale de g entre a et b est nulle, le résultat est trivialement vérifié. Sinon, on a

$$m \le \frac{\int_a^b f(t)\,g(t)\,\mathrm{d}t}{\int_a^b g(t)\,\mathrm{d}t} \le M,$$

et on conclut grâce au théorème des valeurs intermédiaires. □

On note que, dans ce dernier théorème, on peut simplement demander à ce que la fonction g soit intégrable au sens de Riemann, plutôt que continue, sur $[a,b]$.

Théorème B.135 (« formule de la moyenne discrète ») *Soit f une fonction réelle définie et continue sur $[a,b]$, x_j, $j=0,\dots,n$, $n+1$ points de $[a,b]$ et δ_j, $j=0,\dots,n$, $n+1$ constantes toutes de même signe. Alors, il existe un réel c compris entre a et b vérifiant*

$$\sum_{j=0}^{n}\delta_j f(x_j) = f(c)\sum_{i=0}^{n}\delta_j.$$

DÉMONSTRATION. La fonction f étant continue sur l'intervalle $[a,b]$, on pose $m = \inf_{x\in[a,b]} f(x)$ et $M = \sup_{x\in[a,b]} f(x)$ et l'on note $\underline{x}$ et $\overline{x}$ les points de $[a,b]$ vérifiant $f(\underline{x}) = m$ et $f(\overline{x}) = M$. On a alors

$$m\sum_{j=0}^{n}\delta_j \le \sum_{j=0}^{n}\delta_j f(x_j) \le M\sum_{j=0}^{n}\delta_j.$$

On considère à présent, pour tout point x de $[a,b]$, la fonction continue $F(x) = f(x)\sum_{j=0}^{n}\delta_j$. D'après les inégalités ci-dessus, on a

$$F(\underline{x}) \le \sum_{j=0}^{n}\delta_j f(x_j) \le F(\overline{x}),$$

et l'on déduit du théorème des valeurs intermédiaires qu'il existe un point c, strictement compris entre $\underline{x}$ et $\overline{x}$, tel que $F(c) = \sum_{j=0}^{n}\delta_j f(x_j)$, ce qui achève la preuve. □

Références

[Bol17] B. BOLZANO. *Rein analytischer Beweis des Lehrsatzes, daß zwischen je zwey Werthen, die ein entgegengeseßtes Resultat gewähren, wenigstens eine reelle Wurzel der Gleichung liege*. Gottlieb Haase, 1817.

[Ces90] E. CESÀRO. Sur la multiplication des séries. *Bull. Sci. Math. (2)*, 14(1) :114-120, 1890.

[dLHô96] G. F. A. de L'HÔPITAL. *Analyse des infiniment petits, pour l'intelligence des lignes courbes*. Imprimerie Royale, 1696.

Annexe C

Treize articles majeurs de l'analyse numérique

En 1993, L. N. Trefethen publia une liste de treize articles scientifiques utilisés pour un enseignement intitulé *Classic Papers in Numerical Analysis* et prodigué à l'université de Cornell (aux États-Unis d'Amérique). Si cette liste ne peut être exhaustive, elle constitue néanmoins un ensemble de lectures incontournables pour quiconque est intéressé par les méthodes numériques et leur analyse. On a reproduit ci-après le message originel dans son intégralité, suivi des références bibliographiques complètes des articles en question.

```
From: Nick Trefethen
Date: Thu, 6 May 93 10:36:16 -0400
Subject: Classic Papers in Numerical Analysis

                  "CLASSIC PAPERS IN NUMERICAL ANALYSIS"

NA-Netters may be interested to hear of my experiences this spring teaching a
seminar with the above title to a dozen Cornell graduate students (three of
whom were actually post-docs or faculty).  Comp. Sci. 722 met once a week for
two hours, and in the course of the semester we read thirteen papers:

   1. Cooley & Tukey (1965)              the Fast Fourier Transform
   2. Courant, Friedrichs & Lewy (1928)  finite difference methods for PDE
   3. Householder (1958)                 QR factorization of matrices
   4. Curtiss & Hirschfelder (1952)      stiffness of ODEs; BD formulas
   5. de Boor (1972)                     calculations with B-splines
   6. Courant (1943)                     finite element methods for PDE
   7. Golub & Kahan (1965)               the singular value decomposition
   8. Brandt (1977)                      multigrid algorithms
   9. Hestenes & Stiefel (1952)          the conjugate gradient iteration
  10. Fletcher & Powell (1963)           optimization via quasi-Newton updates
  11. Wanner, Hairer & Norsett (1978)    order stars and applications to ODE
  12. Karmarkar (1984)                   interior pt. methods for linear prog.
  13. Greengard & Rokhlin (1987)         multipole methods for particles

Most weeks, one or two related readings were also assigned, typically from a
recent textbook or survey article.  For example, along with the Fletcher &
Powell paper we read an extract from the 1983 text by Dennis & Schnabel.

Our weekly meetings followed a regular format.  First, this week's Historian
distributed a handout containing information he/she had obtained about the
historical context of the paper, including biographical information about the
author(s) and a plot of citations as a function of time.  Next, the
Mathematician gave a presentation of some of the central ideas of the paper.
```

Third and fourth, two Experimentalists reported the results of Matlab, C, or Fortran experiments conducted to illustrate some of the properties of the algorithm under discussion. Finally, the Professor added a few remarks.

To me and at least some of the students, this course provided a satisfying vision of the broad scope of numerical analysis and a sense of excitement at what a diversity of beautiful and powerful ideas have been invented in this field. The thirteen papers were selected partly for their variety; they touch upon nearly all the main problems of numerical computation. We found that although they vary greatly in style, most are quite readable. Indeed it was a pleasure, week after week, to be in the hands of the masters. These authors are for the most part extraordinary people, including some about whom most numerical analysts know little (such as Hirschfelder, one of the leading American chemists of this century).

We were struck by how young many of the authors were when they wrote these papers (average age: 34), and by how short an influential paper can be (Householder: 3.3 pages, Cooley & Tukey: 4.4). Our readings also uncovered a few surprises. For example, Curtiss and Hirschfelder inexplicably define stiffness in terms of exponentially diverging trajectories, not converging ones; nevertheless they invent the right cure for the problem in the shape of backward differentiation formulas. For another example, did you know that the classic SVD paper by Golub & Kahan makes no mention of the QR algorithm?

Our thirteen papers fall into three categories:

```
          Finite algorithms for finite problems: papers 1,3,5
      Infinite algorithms for infinite problems: papers 2,4,6,7,10,11
        Infinite algorithms for finite problems: papers 8,9,12,13
```

(An infinite algorithm is one that depends on an iteration or discretization parameter; an infinite problem is one for which all exact algorithms must be infinite.) The third category is particularly interesting. Evidently four of the most exciting modern developments in numerical analysis -- multigrid iterations, conjugate gradient iterations, interior point methods, and multipole methods -- have in common that they depend on the approximate computation of quantities that might in principle be computed exactly.

Most readers of this note will have thought of other classic authors and papers that should have been on the list. We agree! We are saving up ideas for the next run of CS 722 in a couple of years.

Nick Trefethen
Dept. of Computer Science
Cornell University

Références

[Bra77] A. BRANDT. Multi-level adaptive solutions to boundary-value problems. *Math. Comput.*, 31(138):333–390, 1977. DOI: 10.1090/S0025-5718-1977-0431719-X.

[CFL67] R. COURANT, K. FRIEDRICHS, and H. LEWY. On the partial difference equations of mathematical physics. *IBM J.*, 11(2):215–234, 1967. DOI: 10.1147/rd.112.0215.

[CH52] C. F. CURTISS and J. O. HIRSCHFELDER. Integration of stiff equations. *Proc. Nat. Acad. Sci. U.S.A.*, 38(3):235–243, 1952. DOI: 10.1073/pnas.38.3.235.

[Cou43] R. COURANT. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49(1):1–23, 1943. DOI: 10.1090/S0002-9904-1943-07818-4.

[CT65] J. W. COOLEY and J. W. TUKEY. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19(90):297–301, 1965. DOI: 10.1090/S0025-5718-1965-0178586-1.

[dBoo72] C. de BOOR. On calculating with *B*-splines. *IMA J. Appl. Math.*, 6(1):50–62, 1972. DOI: 10.1016/0021-9045(72)90080-9.

[FP63] R. FLETCHER and M. J. D. POWELL. A rapidly convergent descent method for minimization. *Comput. J.*, 6(2):163–168, 1963. DOI: 10.1093/comjnl/6.2.163.

[GK65] G. GOLUB and W. KAHAN. Calculating the singular values and pseudo-inverse of a matrix. *J. SIAM Ser. B Numer. Anal.*, 2(2):205–224, 1965. DOI: 10.1137/0702016.

[GR87] L. GREENGARD and V. ROKHLIN. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987. DOI: 10.1016/0021-9991(87)90140-9.

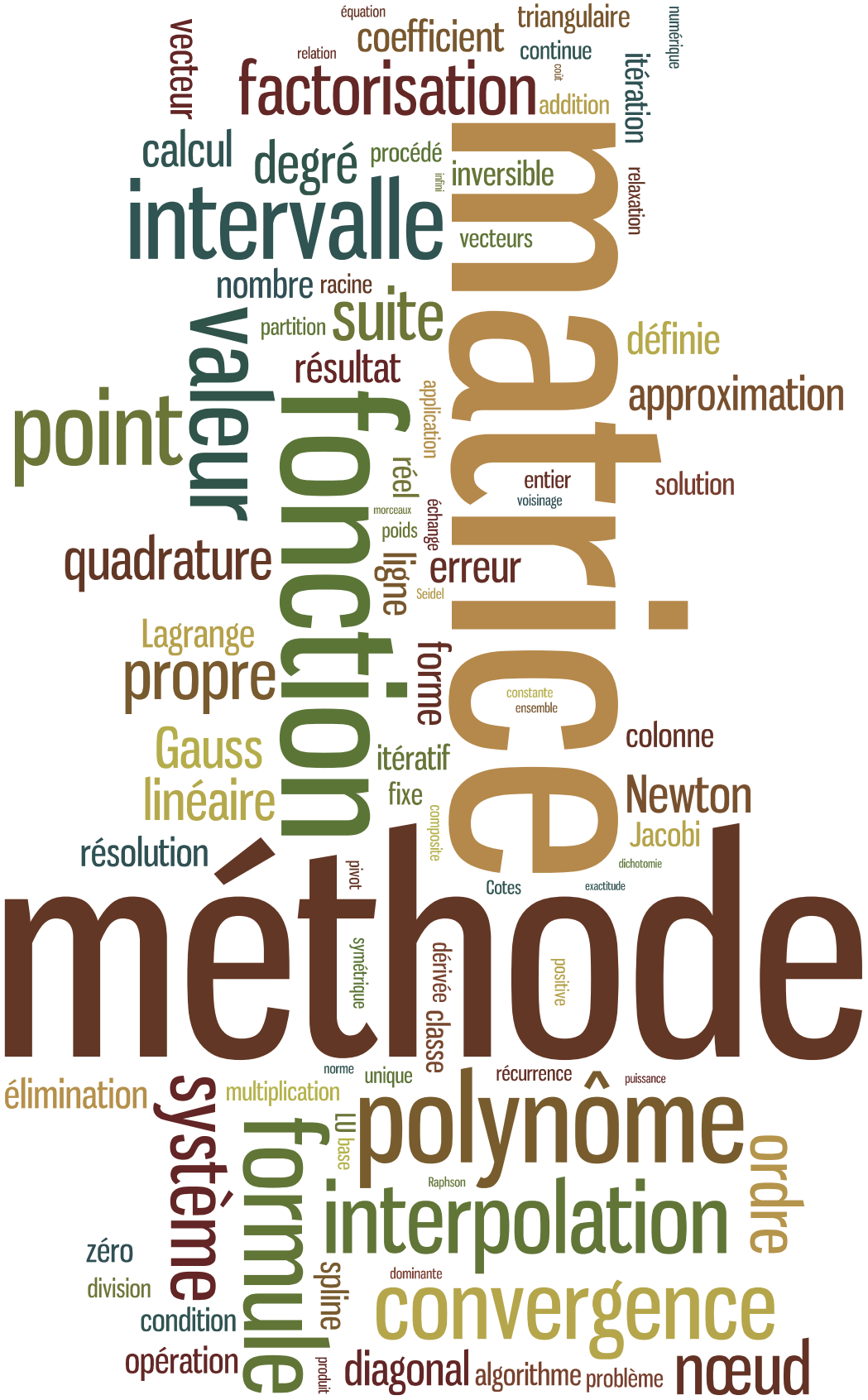
[Hou58] A. S. HOUSEHOLDER. Unitary triangularization of a nonsymmetric matrix. *J. Assoc. Comput. Mach.*, 5(4):339–342, 1958. DOI: 10.1145/320941.320947.

[HS52] M. R. HESTENES and E. STIEFEL. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49(6):409–436, 1952. DOI: 10.6028/jres.049.044.

[Kar84] N. KARMARKAR. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984. DOI: 10.1007/BF02579150.

[WHN78] G. WANNER, E. HAIRER, and S. P. NØRSETT. Order stars and stability theorems. *BIT*, 18(4):475–489, 1978. DOI: 10.1007/BF01932026.

Index


méthode
matrice
fonction
polynôme
factorisation
intervalle
point
valeur
suite
formule
système
interpolation
convergence
nœud
approximation
définie
solution
quadrature
erreur
propre
Gauss
linéaire
Newton
Jacobi
résolution
élimination
multiplication
ordre
zéro
division
condition
opération
diagonal
algorithme
problème
dérivée
classe
forme
ligne
résultat
partition
nombre
racine
calcul
degré
procédé
inversible
vecteurs
vecteur
coefficient
triangulaire
continue
addition
itération
relaxation
numérique
équation
relation
coût
infini
application
réel
entier
voisinage
échange
morceaux
poids
Seidel
Lagrange
constante
ensemble
colonne
itératif
fixe
composite
dichotomie
Cotes
exactitude
pivot
symétrique
positive
norme
unique
récurrence
puissance
LU
base
Raphson
spline
dominante
produit