DAUPHINE
UNIVERSITÉ PARIS

DU MI2E deuxième année

Méthodes numériques

Introduction à l'analyse numérique et au calcul scientifique

Cours 2009/2010

Guillaume Legendre

Table des matières

Avant-propos

Ce document regroupe les notes d'un cours enseigné en deuxième année de licence de Mathématiques et Informatique appliquées à l'Économie et à l'Entreprise (MI2E) à l'université Paris-Dauphine. Cet enseignement se compose à la fois de cours magistraux et de séances de travaux dirigés et de travaux pratiques.

Son but est de présenter plusieurs méthodes numériques de base utilisées pour la résolution des systèmes linéaires, des équations non linéaires ou encore pour l'approximation des fonctions par interpolation polynomiale, ainsi que d'introduire aux étudiants les techniques d'analyse (théorique) de ces dernières, en abordant notamment les notions de *convergence*, de *précision* et de *stabilité*. 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 sont suivies d'une implémentation et d'applications réalisées par les étudiants avec les logiciels MATLAB® [1] et GNU OCTAVE [2].

Il est à noter que ce support de cours comporte des plusieurs passages qui ne seront 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. Enfin, les notes biographiques sont pour partie tirées de WIKIPEDIA [3].

Guillaume Legendre
Paris, décembre 2009.

Quelques références bibliographiques

En complément et pour approfondir les thèmes abordés dans ces pages, voici une sélection de plusieurs ouvrages de référence, plus ou moins accessibles selon la formation du lecteur, que l'on pourra consulter avec intérêt.

[AK02] G. Allaire and S. M. Kaber. *Algèbre linéaire numérique*. Mathématiques pour le deuxième cycle. Ellipses, 2002.

[Axe94] O. Axelsson. *Iterative solution methods*. Cambridge University Press, 1994.

[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.

[Gau97] W. Gautschi. *Numerical analysis : an introduction*. Birkhäuser, 1997.

[GV96] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, third edition, 1996.

[Hig96] N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 1996.

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

[PTVF07] 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.

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/`.

[QSS07] A. Quarteroni, R. Sacco, and F. Saleri. *Méthodes numériques. Algorithmes, analyse et applications*. Springer, 2007.

[SB02] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Springer-Verlag, third edition, 2002.

[SM03] E. Süli and D. F. Mayers. *An introduction to numerical analysis*. Cambridge University Press, 2003.

[Var00] R. S. Varga. *Matrix iterative analysis*, volume 27 of *Springer series in computational mathematics*. Springer-Verlag, second edition, 2000.

Chapitre 1

Généralités sur l'analyse numérique et le calcul scientifique

La mise en œuvre d'une méthode numérique sur une machine amène un certain nombre de difficulés supplémentaires, liées à une nécessaire représentation approchée des nombres.

1.1 Motivations

Un *algorithme* est un énoncé décrivant, à l'aide d'opérations élémentaires, toutes les étapes d'une démarche systématique permettant la résolution d'un problème spécifique. Un algorithme peut à son tour contenir des *sous-algorithmes* et doit pouvoir s'achever après un nombre *fini* d'opérations élémentaires afin de pouvoir être utilisé dans un programme informatique. La *mise en œuvre* d'un algorithme consiste en l'écriture de la série d'opérations élémentaires le composant dans un langage de programmation, ce que l'on appelle aussi fréquemment une *implementation*.

La *complexité* d'un algorithme est une mesure de son temps d'exécution. Calculer la complexité d'un algorithme fait donc partie de l'analyse de l'efficacité et du coût d'une méthode numérique.

pseudo-langage pour la description des algorithmes

on compte les soustractions comme des additions

1.2 Arithmétique en virgule flottante et erreurs d'arrondis

1.2.1 Représentation des nombres en machine

1.2.2 Erreurs d'arrondis

1.3 Stabilité et analyse d'erreur des méthodes numériques et conditionnement d'un problème

Première partie

Algèbre linéaire numérique

Préambule

algèbre linéaire numérique : deux/trois thèmes : résolution de systèmes linéaires (origine, grande taille, caractère particulier des matrices) et recherche de valeurs et vecteurs propres (motivations), problèmes aux moindres carrés

Dans les applications, beaucoup de matrices particulières : creuses, tridiagonales, symétriques, définies positives qui proviennent de formualtions de problèmes discrétisés par différentes méthodes différences finies, éléments finis, méthodes spectrales...

Resolution de systèmes linéaires : exemple de la discrétisation d'une équation différentielle par différences finies allaire 82, ciarlet

en statitisques ?

problème aux valeurs propres : vibrations d'un système mécanique, corde vibrante allaire 203-206

On est parfois seulement intéressé par le calcul d'une valeur propre particulière plutot que l'ensemble du spectre (motivations quarteroni 171-2)

Chapitre 2

Méthodes directes de résolution des systèmes linéaires

On considère la résolution du système linéaire

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

avec A une matrice d'ordre n à coefficients réels inversible et $\boldsymbol{b}$ un vecteur de $\mathbb{R}^n$, par des méthodes dites *directes*, c'est-à-dire fournissant, en l'absence d'erreurs d'arrondi, la solution *exacte* en un nombre *fini*[1] d'opérations élémentaires. On verra que 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 donné quelques éléments sur la résolution numérique des systèmes triangulaires, nous introduisons dans le détail la méthode d'élimination de Gauss. Ce procédé d'élimination est ensuite réinterprété en termes d'opérations matricielles, donnant lieu à une méthode de *factorisation* des matrices. Les propriétés de cette décomposition sont explorées et son application à des matrices particulières est ensuite étudiée. Le chapitre se conclut sur la présentation de quelques autres méthodes de factorisation.

2.1 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 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[3], ce qui s'avère bien plus coûteux que la résolution d'*un seul* système.

Considérons à présent un système linéaire dont la matrice A est inversible et triangulaire inférieure,

1. On oppose ici ce type de méthodes avec les méthodes dites *itératives*, qui nécessitent (en théorie) un nombre infini d'opérations pour obtenir 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 étudia tous les domaines des mathématiques et contribua à développer la plupart des branches des sciences.

3. Ces systèmes sont

$$A\boldsymbol{x}_i = \boldsymbol{e}_i,\ 1 \leq i \leq n,$$

où $\boldsymbol{e}_i$ désigne le $i^{\text{ième}}$ vecteur de la base canonique de $\mathbb{R}^n$.

c'est-à-dire de la forme

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

La matrice A étant inversible, ses termes diagonaux a_{ii}, $i = 1, \ldots, n$, sont tous non nuls [4] 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 subtitution"* en anglais), s'écrit

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_i &= \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\,x_j \right), \ i = 2, \ldots, n. \end{aligned} \tag{2.2}$$

L'algorithme mis en œuvre pour cette résolution effectue $\frac{n(n-1)}{2}$ additions et soustractions, $\frac{n(n-1)}{2}$ 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 $i^{\text{ième}}$, $2 \leq i \leq n$, composante du vecteur solution $\boldsymbol{x}$, on effectue un produit scalaire entre le vecteur constitué des $i-1$ premiers éléments de la $i^{\text{ième}}$ ligne de la matrice L 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 1).

Algorithme 1 Algorithme de la méthode de descente (version orientée ligne)

```
x_1 = b_1/a_11
pour i = 2 à n faire
  x_i = b_i
  pour j = 1 à i − 1 faire
    x_i = x_i − a_ij x_j
  fin pour
  x_i = x_i/a_ii
fin pour
```

On peut obtenir un algorithme *orienté colonne* implémentant la méthode en tirant parti du fait que la $i^{\text{ième}}$ 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 2, dans lequel la solution $\boldsymbol{x}$ calculée étant commodément stockée dans le tableau contenant initialement le second membre $\boldsymbol{b}$).

Exemple. 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 $x_2 = -\frac{1}{5}$. On a enfin

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

Algorithme 2 Algorithme de la méthode de descente (version orientée colonne)

pour $j = 1$ à $n-1$ **faire**
  $b_j = b_j / a_{jj}$
  pour $i = j+1$ à n **faire**
    $b_i = b_i - a_{ij}\, b_j$
  fin pour
fin pour
$b_n = b_n / a_{nn}$

Le choix d'une approche orientée ligne ou colonne dans l'écriture d'un même algorithme peut considérablement modifier ses performances et dépend 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 subtitution"* en anglais) suivante

$$\begin{aligned} x_n &= \frac{b_1}{a_{nn}} \\ x_i &= \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j \right), \; i = n-1, \ldots, 1, \end{aligned} \tag{2.3}$$

et dont le coût est aussi de n^2 opérations. Là encore, on peut produire des algorithmes orientés ligne ou colonne pour l'implémentation de la méthode.

Dans la pratique, il est utile de remarquer que seule la partie non nulle de la matrice nécessite d'être stockée[5] pour la résolution d'un système triangulaire, d'où une économie de mémoire conséquente dans le cas de grands systèmes.

2.2 Méthode d'élimination de Gauss

Une technique de choix pour ramener la résolution d'un système linéaire quelconque à celle d'un système triangulaire et la *méthode d'élimination de Gauss*. Celle-ci consiste en premier lieu à transformer, par des opérations simples sur les équations, ce 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[6] (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. Si A est inversible, la solution du système peut ensuite être obtenue par une méthode de remontée, mais le procédé d'élimination est en fait très général, la matrice pouvant être rectangulaire.

2.2.1 Élimination de Gauss 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 des lignes 2 à n du système en leur retranchant respectivement la première ligne multipliée par le coefficient $\frac{a_{i1}}{a_{11}}$, $= 2, \ldots, n$.

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

5. Les éléments de la matrice triangulaire sont généralement stockés dans un tableau à une seule entrée de dimension $\frac{n(n+1)}{2}$ en gérant la correspondance entre les indices i et j d'un élément de la matrice et l'indice $k(= k(i,j))$ de lélément le représentant dans le tableau. Par exemple, pour une matrice triangulaire inférieure stockée ligne par ligne, on vérifie facilement que $k(i,j) = j + \frac{i(i-1)}{2}$.

6. Il faut bien noter qu'on ne calcule en pratique jamais explicitement la matrice d'élimination M, mais seulement les produits MA et $M\boldsymbol{b}$.

En notant $A^{(2)}$ et $\boldsymbol{b}^{(2)}$ la matrice et le vecteur second membre résultant de ces opérations[7], on a alors

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

et le système $A^{(2)}\,\boldsymbol{x} = \boldsymbol{b}^{(2)}$ est équivalent au système de départ. En supposant le coefficient diagonal $a_{22}^{(2)}$ de $A^{(2)}$, on peut procéder à l'élimination de l'inconnue x_3 des lignes 3 à n de ce système, et ainsi de suite. On obtient, sous l'hypothèse $a_{kk}^{(k)} \neq 0$, $k = 1, \ldots, n-1$, une suite finie de matrices $A^{(k)}$, $k \geq 2$, de la forme

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \ldots & \ldots & \ldots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & & & a_{2n}^{(k)} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \ldots & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix}$$

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

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

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

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

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

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

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

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

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

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

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

7. On pose $A^{(1)} = A$ et $\boldsymbol{b}^{(1)} = \boldsymbol{b}$ pour être consistant.

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

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

Comme on l'a vu, la méthode de Gauss, dans sa forme sans échange, ne peut s'appliquer que si tous les pivots $a_{kk}^{(k)}$, $k = 1, \ldots, n-1$, sont non nuls, ce qui élimine 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 de pivot nul durant l'élimination, comme le montre l'exemple ci-dessous.

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

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

On a alors

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

et l'élimination s'interrompt à l'issue de la seconde étape, le pivot $a_{22}^{(2)}$ 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 ligne ou par colonne* et les matrices *symétriques définies positives* (voir à ce titre le théorème 2.9).

2.2.2 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 de Gauss sans échange pour mener l'élimination à son terme. Dans tout ce qui suit, les notations de la section 2.2.1 sont conservées.

À la première étape, au moins l'un des coefficients de la première colonne de la matrice $A^{(1)}(= A)$ est non nul, faute de quoi la matrice A ne serait pas inversible. On choisit [8] 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 $A^{(2)}$ et $\boldsymbol{b}^{(2)}$ la matrice et le second membre du système obtenu et l'on réitère ce procédé. À l'étape k, $2 \leq k \leq n-1$, la matrice $A^{(k)}$ est inversible [9], et donc l'un au moins des éléments $a_{ik}^{(k)}$, $k \leq i \leq n$, 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 $k^{\text{ième}}$ 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 $A^{(n)}$, dont le coefficient $a_{nn}^{(n)}$ 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*.

8. 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 2.2.4.

9. On a en effet que $\det(A^{(k)}) = \pm \det(A)$. On renvoie à la section 2.3.1 pour une justification de ce fait.

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

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

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

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

$$A^{(3)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & -2 & \frac{5}{3} \end{pmatrix} \text{ et } \boldsymbol{b}^{(3)} = \begin{pmatrix} 0 \\ 4 \\ -7 \\ \frac{10}{3} \end{pmatrix},$$

et

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

d'où la solution $\boldsymbol{x} = \begin{pmatrix} 1 & -1 & 0 & 2 \end{pmatrix}^T$. On note que l'on a procédé au cours de la deuxième étape à l'échange des deuxième et de la troisième lignes.

On pourra remarquer que si la matrice A est non inversible, alors tous les éléments $a_{ik}^{(k)}$, $k \leq i \leq n$, seront nuls pour au moins une valeur de k entre 1 et n. Si $k \neq n$, on n'a dans ce cas pas besoin de réaliser l'élimination de cette $k^{\text{ième}}$ (puisque cela est déjà fait) 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 $A^{(k+1)}$, $1 \leq k \leq n-1$, on effectue $(n-k+1)(n-k) = (n-k)^2+(n-k)$ additions, $(n-k+1)(n-k)$ multiplications et $n-k$ divisions, ce qui correspond à un total de $\frac{n(n^2-1)}{3}$ additions, $\frac{n(n^2-1)}{3}$ multiplications et $\frac{n(n-1)}{2}$ pour l'élimination complète. Pour la mise à jour du second membre à l'étape k, on a besoin de $n-k$ additions et multiplications, soit en tout $\frac{n(n-1)}{2}$ additions et multiplications. Enfin, il faut faire $\frac{n(n-1)}{2}$ additions et 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 de l'ordre de $\frac{n^3}{3}$ *additions,* $\frac{n^3}{3}$ *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.61) requiert, en utilisant un développement « brutal » par ligne ou colonne pour le calcul des déterminants, de l'ordre de $(n+1)!$ additions, $(n+2)!$ multiplications et n divisions. Ainsi, pour $n = 10$ par exemple, on obtient un compte d'environ 700 opérations pour la méthode d'élimination de Gauss contre près de 479000000 opérations pour la règle de Cramer !

2.2.3 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 délimination 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, et l'on peut alors résoudre le système associé comme expliqué dans la section A.4. La méthode d'élimination de Gauss constitue à ce titre un moyen simple de détermination du rang d'une matrice quelconque.

2.2.4 Erreurs d'arrondi et choix du pivot

Revenons à présent sur le choix du pivot à chaque étape de l'élimination. Si à la $k^{\text{ième}}$ l'élément $a_{kk}^{(k)}$ 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 des erreurs d'arrondi existant en pratique, cette manière de procéder est en général à proscrire, comme l'illustre l'exemple suivant

Exemple numérique tiré de [FM67]. Supposons que les calculs sont 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 est $x_1 = 1,0001$ et $x_2 = 0,9999$. En choisissant le nombre 10^{-4} comme pivot à la première étape de l'élimination de Gauss, on obtient le système triangulaire

$$\begin{pmatrix} 10^{-4} & 1 \\ 0 & -9990 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -9990 \end{pmatrix},$$

puisque les nombres $-10^4 + 1 = -9999$ et $-10^4 + 2 = -9998$ sont tous deux arrondis au même nombre -9990. La solution numérique calculée est alors :

$$x_1 = 0 \text{ et } x_2 = 1,$$

et est très éloignée 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 $-2 \times 10^{-4} + 2 = -0,9998$ sont arrondis au même nombre $0,999$. La solution calculée vaut

$$x_1 = 1 \text{ et } x_2 = 1,$$

ce qui est cette fois très satisfaisant.

En général, le changement de pivot n'a pas un effet aussi spectaculaire que dans l'exemple ci-dessus, 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 dans le cas où le pivot « naturel » est non nul, à choisir le plus grand pivot en valeur absolue. On peut pour cela suivre au début de la $k^{\text{ième}}$ étape, $1 \leq k \leq n-1$, de l'élimination

- soit une *stratégie de pivot partiel* : le pivot est un des éléments $a_{ik}^{(k)}$, $k \leq i \leq n$, de la $k^{\text{ième}}$ colonne situés sous la diagonale vérifiant
$$|a_{ik}^{(k)}| = \max_{k \leq p \leq n} |a_{pk}^{(k)}|,$$
- soit une *stratégie de pivot total* : le pivot est un des éléments de la sous-matrice $a_{ij}^{(k)}$, $k \leq i, j \leq n$, vérifiant
$$|a_{ij}^{(k)}| = \max_{k \leq p,q \leq n} |a_{pq}^{(k)}|.$$

Dans ce dernier cas, si le pivot n'est pas dans la $k^{\text{ième}}$ colonne, il faut procéder à un échange de colonnes en plus d'un éventuel échange de lignes.

Quelle que soit la stratégie adoptée, cette recherche de pivot doit également être prise en compte dans l'évaluation du coût global de la méthode d'élimination de Gauss.

2.2.5 Méthode d'élimination de Gauss–Jordan

La *méthode d'élimination de Gauss–Jordan*[10] est une variante de la méthode d'élimination de Gauss ramenant toute matrice sous forme *échelonnée réduite*. Dans le cas d'une matrice A inversible, cette méthode revient à chercher une matrice $\tilde{M}$ telle que la matrice $\tilde{M}A$ soit non pas triangulaire supérieure mais *diagonale*. 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.

Si elle est bien moins efficace[11] 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 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 $\begin{pmatrix} A & I_n \end{pmatrix}$. Au terme du processus d'élimination, le premier bloc contient la matrice identité et, si aucun échange de lignes n'a eu lieu, le second l'inverse de A.

Exemple. Soit $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$. La matrice augmentée est

$$\begin{pmatrix} A & I_n \end{pmatrix} = \left(\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right),$$

et l'on trouve successivement

$$k = 1, \quad \left(\begin{array}{ccc|ccc} 1 & -1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 3/2 & -1 & 1/2 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right),$$

$$k = 2, \quad \left(\begin{array}{ccc|ccc} 1 & 0 & -1/3 & 2/3 & 1/3 & 0 \\ 0 & 1 & -2/3 & 1/3 & 2/3 & 0 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array}\right),$$

$$k = 3, \quad \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3/4 & 1/2 & 1/4 \\ 0 & 1 & 0 & 1/2 & 1 & 1/2 \\ 0 & 0 & 1 & 1/4 & 1/2 & 3/4 \end{array}\right),$$

d'où $A^{-1} = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$.

2.3 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

10. Wilhelm Jordan (1^{er} mars 1842 - 17 avril 1899) était un géodésiste allemand. Il est connu parmi les mathématiciens pour le procédé d'élimination portant son nom et publié en 1888 dans son *Handbuch der Vermessungskunde*, Jordan améliorant la stabilité de l'algorithme d'élimination de Gauss pour l'appliquer à la résolution de problèmes aux moindres carrés en topographie.

11. Effectuons un compte des opérations effectuées pour la résolution d'un système de n équations à n inconnues. À chaque étape k, $1 \leq k \leq n$, il faut faire $(n-k+2)(n-1)$ additions, $(n-k+2)(n-1)$ multiplications et $(n-k+2)$ divisions pour mettre à jour la matrice et le second membre du système, mais le résolution du système (diagonal) final ne nécessite aucune opération supplémentaire. La résolution du système par la méthode d'élimination de Gauss–Jordan nécessite donc de l'ordre de n^3 opérations.

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 $U = A^{(n)}$ en reprenant la notation utilisée dans la section 2.2.1.

2.3.1 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 $A^{(k)}$, $1 \leq k \leq n-1$, 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

Soient $(m,n) \in (\mathbb{N}\backslash\{0,1\})^2$ et $A = (a_{ij})_{1\leq i\leq m,\ 1\leq j\leq m} \in M_{m,n}(\mathbb{K})$. On appelle *opérations élémentaires sur les lignes de* A les transformations suivantes :

- l'échange (entre elles) des $i^{\text{ième}}$ et $j^{\text{ième}}$ lignes de A,
- la multiplication de la $i^{\text{ième}}$ ligne de A par un scalaire $\lambda \in \mathbb{K}\backslash\{0\}$,
- le remplacement de la $i^{\text{ième}}$ ligne de A par la somme de cette même ligne avec la $j^{\text{ième}}$ ligne de A multipliée par un scalaire λ, où $\lambda \in \mathbb{K}\backslash\{0\}$.

Explicitons à présent les *opérations matricielles* correspondant à chacunes de ces opérations. Tout d'abord, échanger les $i^{\text{ième}}$ et $j^{\text{ième}}$ lignes, $(i,j) \in \{1\dots,n\}^2$, de la matrice A revient à la multiplier à gauche par la *matrice de permutation*

$$P_{ij} = \begin{pmatrix}
1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & \ddots & \ddots & & & & & & & & \vdots \\
\vdots & \ddots & 1 & \ddots & & & & & & & \vdots \\
\vdots & & \ddots & 0 & 0 & \cdots & 0 & 1 & & & \vdots \\
\vdots & & & 0 & 1 & \ddots & & 0 & & & \vdots \\
\vdots & & & \vdots & \ddots & \ddots & \ddots & \vdots & & & \vdots \\
\vdots & & & 0 & & \ddots & 1 & 0 & & & \vdots \\
\vdots & & & 1 & 0 & \dots & 0 & 0 & & & \vdots \\
\vdots & & & & & & & & 1 & & \vdots \\
\vdots & & & & & & & & & \ddots & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1
\end{pmatrix} = I_n + (E_{ij} + E_{ji} - E_{ii} - E_{jj}) \in M_n(\mathbb{K}).$$

Cette matrice est orthogonale, de déterminant valant -1.

La multiplication de la $i^{\text{ième}}$ ligne de la matrice A par un scalaire non nul λ s'effectue en multipliant la matrice par la *matrice de dilatation*

$$D_i(\lambda) = \begin{pmatrix}
1 & 0 & \dots & \dots & \dots & \dots & 0 \\
0 & \ddots & \ddots & & & & \vdots \\
\vdots & \ddots & 1 & \ddots & & & \vdots \\
\vdots & & \ddots & \lambda & \ddots & & \vdots \\
\vdots & & & \ddots & 1 & \ddots & \vdots \\
\vdots & & & & \ddots & \ddots & 0 \\
0 & \dots & \dots & \dots & \dots & 0 & 1
\end{pmatrix} = I_n + (\lambda - 1)E_{ii} \in M_n(\mathbb{K}).$$

Cette matrice est inversible et $D_i(\lambda)^{-1} = D_i\left(\frac{1}{\lambda}\right)$.

Enfin, le remplacement de la $i^{\text{ième}}$ ligne de A par la somme de la $i^{\text{ième}}$ ligne et de la $j^{\text{ième}}$, $i \neq j$ multipliée par un scalaire non nul λ est obtenu en multipliant à gauche la matrice A par la *matrice de transvection* (on suppose ici que $j < i$)

$$T_{ij}(\lambda) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \cdots & 0 & & \vdots \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ \vdots & & \lambda & \cdots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_n + \lambda\, E_{ij} \in M_n(\mathbb{K}).$$

Cette matrice a pour inverse $T_{ij}(-\lambda)$. On note que le produit de deux matrices de tranvection $T_{ij}(\lambda)$ et $T_{kl}(\mu)$, avec λ et μ deux scalaires non nuls et $(i,j) \neq (k,l)$, est commutatif et vaut

$$T_{ij}(\lambda)T_{kl}(\mu) = I_n + \lambda\, E_{ij} + \mu\, E_{kl}.$$

Ces trois types de matrices permettent de définir de manière analogue les opérations élémentaires sur les *colonnes* de A par des multiplications *à droite* de la matrice A (ce sont en effet des opérations élémentaires sur les lignes de la transposée de A).

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 alors le produit

$$M = E^{(n-1)} \dots E^{(2)} E^{(1)}$$

de $n-1$ *matrices d'élimination* définies par

$$E^{(k)} = \prod_{i=k+1}^{n} T_{ik}\left(-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+1\,k}^{(k)}}{a_{kk}^{(k)}} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & -\frac{a_{k+2\,k}^{(k)}}{a_{kk}^{(k)}} & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} & 0 & \dots & 0 & 1 \end{pmatrix}, \; 1 \leq k \leq n-1. \tag{2.4}$$

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

$$A = LU,$$

dans lequel $L = M^{-1}$ et $U = MA = A^{(n)}$ 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 (2.4) étant produit de matrices de transvection, il est facile de vérifier que son inverse vaut

$$(E^{(k)})^{-1} = \prod_{i=k+1}^{n} T_{ik}\left(\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & \frac{a_{k+1\,k}^{(k)}}{a_{kk}^{(k)}} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \frac{a_{k+2\,k}^{(k)}}{a_{kk}^{(k)}} & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} & 0 & \dots & 0 & 1 \end{pmatrix}, \ 1 \leq k \leq n-1,$$

et l'on a alors [12]

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1}\dots(E^{(n-1)})^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \frac{a_{21}^{(1)}}{a_{11}^{(1)}} & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \frac{a_{k+1\,k}^{(k)}}{a_{kk}^{(k)}} & 1 & \ddots & \vdots \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ \frac{a_{n1}^{(1)}}{a_{11}^{(1)}} & \dots & \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} & \dots & \frac{a_{n\,n-1}^{(n-1)}}{a_{n-1\,n-1}^{(n-1)}} & 1 \end{pmatrix}.$$

Si des échanges de lignes ont eu lieu lors de l'élimination, la matrice M s'écrit

$$M = E^{(n-1)}P^{(n-1)}\dots E^{(2)}P^{(2)}E^{(1)}P^{(1)},$$

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 $k^{\text{ième}}$ étape, soit la matrice identié si le pivot « naturel » est utilisé. En posant $P = P^{(n-1)}\dots P^{(2)}P^{(1)}$, on a cette fois-ci $L = PM^{-1}$ et $U = (MP^{-1})PA$, d'où

$$PA = LU.$$

Terminons cette section en montrant comment la méthode de factorisation LU fournit un procédé rapide de calcul du déterminant de la matrice A, qui n'est autre, au signe près, que le produit des pivots, puisque

$$\det(PA) = \det(LU) = \det(L)\det(U) = \det(U) = \left(\prod_{i=1}^{n} u_{ii}\right),$$

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.

12. La vérification est laissée en exercice.

2.3.2 Condition d'existence de la factorisation LU

Commençons par donner des conditions suffisantes assurant qu'il n'y aura pas d'échange de lignes durant l'élimination de Gauss, ce qui conduira bien à une telle *factorisation* pour la matrice A. 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).

Théorème 2.2 *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*

$$A_k = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}, \ 1 \leq k \leq n, \tag{2.5}$$

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 [13].

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

$$A_k = \begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^T & d \end{pmatrix},$$

avec $\boldsymbol{b}$ et $\boldsymbol{c}$ des vecteurs de $\mathbb{R}^k$ et d un nombre réel, et cherchons une factorisation de A_k de la forme

$$\begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^T & d \end{pmatrix} = \begin{pmatrix} L_{k-1} & \boldsymbol{0} \\ \boldsymbol{l}^T & 1 \end{pmatrix} \begin{pmatrix} U_{k-1} & \boldsymbol{u} \\ \boldsymbol{0}^T & \mu \end{pmatrix}$$

avec $\boldsymbol{l}$ et $\boldsymbol{u}$ des vecteurs de $\mathbb{R}^k$ et μ un nombre réel. En effectuant le produit de matrices et en identifiant par blocs avec A_k, on obtient

$$L_{k-1}U_{k-1} = A_{k-1}, \quad L_{k-1}\boldsymbol{u} = \boldsymbol{b}, \quad \boldsymbol{l}^T U_{k-1} = \boldsymbol{c}^T \ \text{ et } \ \boldsymbol{l}^T\boldsymbol{u} + \mu = d.$$

Si la première de ces égalités n'apporte aucune nouvelle information, les trois suivantes permettent de déterminer les vecteurs $\boldsymbol{l}$ et $\boldsymbol{u}$ et le scalaire 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 vecteurs $\boldsymbol{l}$ et $\boldsymbol{u}$ existent et sont uniques et $\mu = d - \boldsymbol{l}^T\boldsymbol{u}$. 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 2.4.1).

La factorisation LU est particulièrement avantageuse lorsque l'on doit résoudre plusieurs systèmes linéaires ayant tous A pour matrice, mais des seconds membres différents. En effet, il suffit de conserver les matrices L et U obtenues à l'issue de la factorisation pour ramener ensuite la résolution de chaque système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ à celle de deux systèmes triangulaires,

$$L\boldsymbol{y} = \boldsymbol{b}, \ \text{puis } U\boldsymbol{x} = \boldsymbol{y},$$

ce que l'on accomplit à chaque fois en $n(n-1)$ additions, $n(n-1)$ multiplications et $2n$ divisions.

13. 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 (cf. théorème 2.9).

Exemple d'application. 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 $\boldsymbol{x} = \frac{1}{3}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$.

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 traduction matricielle 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 ou total est appliquée à l'élimination de Gauss, on a plus précisément le résultat suivant.

Théorème 2.3 *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. total), une matrice triangulaire inférieure L, dont les éléments sont inférieur ou égaux à 1 en valeur absolue, et une matrice triangulaire supérieure U telles que*

$$PA = LU \quad (\textit{resp. } PAQ = LU).$$

Exemple. 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. En échangeant la deuxième et la troisième ligne, on arrive à

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

Les matrices d'élimination au 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 la matrice P est la matrice de permutation

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

d'où

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 7 & 0 & 1 \end{pmatrix}.$$

Dans le cas d'une factorisation de type $PA = LU$, la résolution du système linéaire (2.1) après factorisation s'effectue en appliquant tout d'abord la matrice de permutation P au vecteur $\boldsymbol{b}$ pour obtenir le second membre $P\boldsymbol{b}$ et en résolvant ensuite le système $L\boldsymbol{y} = P\boldsymbol{b}$ par une méthode de descente, puis le système $U\boldsymbol{x} = \boldsymbol{y}$ par une méthode de remontée.

Algorithme 3 Algorithme de factorisation LU version « kji »

pour $k = 1$ à $n - 1$ **faire**
  pour $i = k + 1$ à n **faire**
    $a_{ik} = a_{ik}/a_{kk}$
  fin pour
  pour $j = k + 1$ à n **faire**
    pour $i = k + 1$ à n **faire**
      $a_{ij} = a_{ij} - a_{ik}\, a_{kj}$
    fin pour
  fin pour
fin pour

2.3.3 Mise en œuvre et implémentation

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 3, écrit en pseudo-code, présente une première implémentation de la factorisation LU.

Cet algorithme contient trois boucles imbriquées, portant respectivement sur les indices k, j et i. Il peut être réecrit de plusieurs manières en modifiant (avec précaution) 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*, et *orienté colonne* dans le cas contraire. Dans LAPACK [ABB+99], une bibliothèque de programmes implémentant un grand nombre d'algorithmes pour la résolution numérique de problèmes d'algèbre linéaire, on dit que la version « kji », et donc 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 est d'effectuer le produit d'un scalaire par un vecteur puis d'additionnner le résultat avec un vecteur. La version « jki », également orientée colonne, de l'implémentation proposée dans l'algorithme 4 ci-après utilise des opérations *gaxpy* (acronyme pour *"generalized saxpy"*), l'opération de base étant cette fois-ci le produit d'une *matrice* par un vecteur puis une addition avec un vecteur.

Algorithme 4 Algorithme de factorisation LU version « jki »

pour $j = 1$ à n **faire**
  pour $k = 1$ à $j - 1$ **faire**
    pour $i = k + 1$ à n **faire**
      $a_{ij} = a_{ij} - a_{ik}\, a_{kj}$
    fin pour
  fin pour
  pour $i = j + 1$ à n **faire**
    $a_{ij} = a_{ij}/a_{jj}$
  fin pour
fin pour

Terminons sur un algorithme de factorisation LU dit *de forme compacte*, car nécessitant moins d'opérations intermédiaires que la méthode d'élimination de Gauss « classique » pour produire la factorisation [14]. Il s'agit de la *méthode de factorisation de Doolittle* [15] (la *méthode de factorisation de Crout* [Cro41] est obtenue de manière similaire, mais en choisissant que les éléments diagonaux de U, et non de L, sont tous égaux à 1). On l'obtient en remarquant que, si aucun échange de lignes n'est requis, la factorisation LU

14. Ceci était particulièrement avantageux avant l'avènement et la généralisation des machines à calculer.
15. 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.

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

$$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj}, \qquad j = k, \ldots, n,$$

$$l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rj} \right), \qquad i = k+1, \ldots, n,$$

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 implémentation de la méthode de Doolittle est proposée ci-dessous.

Algorithme 5 Algorithme de factorisation LU version « ijk »

```
pour i = 1 à n faire
  pour j = 2 à i faire
    a_{ij-1} = a_{ij-1} / a_{j-1j-1}
    pour k = 1 à j − 1 faire
      a_{ij} = a_{ij} − a_{ik} a_{kj}
    fin pour
  fin pour
  pour j = i + 1 à n faire
    pour k = 1 à i − 1 faire
      a_{ij} = a_{ij} − a_{ik} a_{kj}
    fin pour
  fin pour
fin pour
```

Bien évidemment, le choix de l'implémentation à employer préférentiellement dépend 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.

2.3.4 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 un cas général et/ou à une réduction considérable du coût des algorithmes utilisés, par exemple, pour leur factorisation. À ces quelques cas particuliers, il faut ajouter ceux des matrices symétriques et symétriques définies positives, abordés respectivement dans les sections 2.4.1 et 2.4.2.

Cas des matrices à diagonale strictement dominante

Certaines matrices produites par des méthodes de discrétisation des équations aux dérivées partielles (comme la méthode des éléments finis) possèdent la particularité d'être à diagonale dominante (voir la définition A.38). Le résultat suivant montre qu'une matrice à diagonale strictement dominante admet toujours une factorisation LU.

Théorème 2.4 *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*

$$|l_{ij}| \leq 1,\ 1 \leq i,j \leq n.$$

DÉMONSTRATION. Nous reprenons un argument provenant de [Wil61]. Supposons que A est une matrice à diagonale strictement dominante par colonnes. Posons $A^{(1)} = A$. On sait par hypothèse que

$$|a_{11}^{(1)}| > \sum_{j=2}^{n} |a_{j1}^{(1)}|,$$

et $a_{11}^{(1)}$ est donc non nul. L'application du procédé d'élimination sans échange donne

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{ij}^{(1)}}{a_{11}^{(1)}}\, a_{1j}^{(1)},\ 2 \leq i,j \leq n,$$

d'où, $\forall j \in \{2,\ldots,n\}$,

$$\begin{aligned}
\sum_{i=2}^{n} |a_{ij}^{(2)}| &\leq \sum_{i=2}^{n} \left(|a_{ij}^{(1)}| + \left| \frac{a_{ij}^{(1)}}{a_{11}^{(1)}} \right| |a_{1j}^{(1)}| \right) \\
&\leq \sum_{i=2}^{n} |a_{ij}^{(1)}| + |a_{1j}^{(1)}| \sum_{i=2}^{n} \left| \frac{a_{ij}^{(1)}}{a_{11}^{(1)}} \right| \\
&< \sum_{i=1}^{n} |a_{ij}^{(1)}|.
\end{aligned}$$

De plus, on a que

$$\begin{aligned}
|a_{ii}^{(2)}| &\geq |a_{ii}^{(1)}| - \left| \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \right| |a_{1i}^{(1)}| \\
&> \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}^{(1)}| - \left(1 - \sum_{\substack{j=2 \\ j\neq i}}^{n} \left| \frac{a_{j1}^{(1)}}{a_{11}^{(1)}} \right| \right) |a_{1i}^{(1)}| \\
&= \sum_{\substack{j=2 \\ j\neq i}}^{n} \left(|a_{ji}^{(1)}| + \left| \frac{a_{j1}^{(1)}}{a_{11}^{(1)}} \right| |a_{1i}^{(1)}| \right) \\
&\geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}^{(2)}|,
\end{aligned}$$

et $A^{(2)}$ est donc une matrice à diagonale strictement dominante par colonnes. Par un calcul analogue, on montre que si la matrice $A^{(k)}$, $2 \leq k \leq n-1$, 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^T est à diagonale strictement dominante par colonnes et admet donc une factorisation LU. On conclut alors en utilisant la proposition 2.8. □

Cas des matrices bandes

Les matrices bandes (voir la définition A.37) interviennent aussi très couramment dans la résolution de problèmes par des méthodes de différences finies ou d'éléments finis et il convient donc de tirer parti de la structure de ces matrices.

En particulier, le stockage d'une matrice bande A d'ordre n et de largeur de bande valant $p+q+1$ peut se faire dans un tableau de taille $(p+q+1)n$, les éléments de A étant stockés soit ligne par ligne, soit colonne par colonne. 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 $i^{\text{ième}}$ ligne de A, c'est-à-dire a_{ii-p}, est stocké dans le tableau à la $(p+q+1)(i-1)+1^{\text{ième}}$ 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.

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 largeur de bande (respectivement inférieure pour L et supérieure pour U) identique à celle de A. La zone mémoire allouée par le mode de stockage décrit ci-dessus afin de contenir une matrice bande est par conséquent de taille suffisante pour qu'on puisse y stocker sa factorisation.

Proposition 2.5 *La factorisation LU conserve de la structure bande des matrices.*

DÉMONSTRATION. Soit A matrice bande A d'ordre n et de largeur de bande valant $p+q+1$ admetttant une factorisation LU telle que

$$a_{ij}=\sum_{r=1}^{\min(i,j)} l_{ir}u_{rj},\ 1\leq i,j\leq n.$$

On raisonne alors 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 2.6 *Supposons l'entier n supérieur ou égal à 3. On dit qu'une matrice A de $M_n(\mathbb{R})$ est* ***tridiagonale*** *si*

$$a_{ij}=0 \text{ si } |i-j|>1,\ 1\leq i,j\leq n.$$

Supposons que la matrice tridiagonale réelle d'ordre n

$$A = \begin{pmatrix} a_1 & c_1 & 0 & \dots & 0 \\ b_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \dots & 0 & b_n & a_n \end{pmatrix}.$$

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, *i.e.*, $|a_1| > |c_1|$, $|c_i| > |b_1| + |c_i|$, $2 \leq i \leq n-1$, et $|a_n| > |c_n|$). Dans ce cas, les matrices L et U 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

$$v_i = c_i,\ i = 1, \dots, n-1,\ u_1 = a_1,\ l_j = \frac{a_j}{u_{j-1}},\ u_j = a_j - l_j\, c_{j-1},\ j = 2, \dots, n.$$

Cette méthode spécifique de factorisation LU d'une matrice tridiagonale est connue sous le nom d'*algorithme de Thomas*[16] et est un cas particulier de factorisation de Doolittle sans changement de pivot.

Si l'on souhaite résoudre le système linéaire $A\boldsymbol{x} = \boldsymbol{d}$, $\boldsymbol{d} \in \mathbb{R}^n$, dans lequel A est la matrice tridiagonale ci-dessus, on doit, en plus de la factorisation, résoudre les systèmes bidiagonaux $L\boldsymbol{y} = \boldsymbol{d}$ et $U\boldsymbol{x} = \boldsymbol{y}$. Les méthodes de descente et de remontée se résument dans ce cas au formules

$$y_1 = d_1,\ y_i = d_i - l_i\, y_{i-1},\ i = 2, \dots, n,$$

et

$$x_n = \frac{y_n}{u_n},\ x_j = \frac{1}{u_j}\left(y_j - c_j\, x_{j+1}\right),\ j = n-1, \dots, 1.$$

La factorisation d'une matrice d'ordre n par l'algorithme de Thomas requiert $n-1$ additions, multiplications et divisions, et les formules ci-dessus nécessitent à elles deux $2(n-1)$ additions et multiplications et n divisions. La résolution d'un système linéaire tridiagonal ne requiert donc que $8n-7$ opérations, ce qui constitue une réduction considérable par rapport au cas d'une matrice quelconque.

On peut réaliser l'implémentation d'une méthode de résolution d'un système linéaire tridiagonal d'ordre n, basée sur l'algorithme de Thomas, ne nécessitant que $3(n-1)$ additions, $3(n-1)$ multiplications et $2n$ divisions, soit une importante diminution du nombre d'opérations nécessaires par rapport à la méthode d'élimination de Gauss d'une matrice quelconque.

Phénomène de remplissage des matrices creuses

On parle de *matrice creuse* (*sparse matrix* en anglais) lorsque le nombre de ses coefficients non nuls est petit devant nombre total de coefficients (typiquement de l'ordre de n pour une matrice carrée d'ordre n, avec n grand). Un exemple simple de telles matrices est donné par les matrices triagonales. La structure des matrices creuses peut se décrire en terme de graphes et celles-ci sont pour cette raison très utilisées en

16. 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*).

optimisation combinatoire, et plus particulièrement en théorie des réseaux et en recherche opérationnelle (il semble d'ailleurs que l'appellation soit due à Markowitz [17]). Les matrices bandes produites par les méthodes courantes de résolution d'équations aux dérivées partielles (comme les méthodes de différences finies ou d'éléments finis) sont, en général, également creuses. On tire avantage de la structure d'une telle matrice en ne stockant que ses éléments non nuls, ce qui constitue un gain de place en mémoire non négligeable lorque l'on travaille avec des matrices de grande taille. Différents formats de stockage existent.

Un des inconvénients de la factorisation LU appliquées aux matrices creuses est qu'elle entraîne l'apparition d'un grand nombre de termes non nuls dans les matrices 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 un problème de structure de données, puisque le stockage utilisé pour la matrice à factoriser ne peut alors contenir sa factorisation.

On peut adapter le stockage en réalisant *a priori* une *factorisation symbolique* de la matrice, qui consiste à déterminer le nombre et la position des nouveaux coefficients créés au cours de la factorisation effective. Une renumérotation des inconnues et équations du système linéaire associé à la matrice creuse, en utilisant par exemple l'*algorithme de Cuthill–McKee*, permet aussi de limiter le remplissage en diminuant la largeur de bande de cette matrice (on pourra consulter l'article [GPS76] sur ce sujet).

2.4 Autres méthodes de factorisations

Nous présentons dans cette dernière section d'autres types de factorisation, adaptés à des matrices particulières. Il s'agit de la *factorisation* LDM^T d'une matrice carrée, qui devient la *factorisation* LDL^T lorsque cette matrice est symétrique, de la *factorisation de Cholesky* [18], pour une matrice *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.

2.4.1 Factorisation $\mathbf{LDM^T}$

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 2.7 *Sous les conditions du théorème 2.2, il existe une unique matrice triangulaire inférieure L, une unique matrice diagonale D et une unique matrice triangulaire supérieure M^T, les éléments diagonaux de L et M étant tous égaux à 1, telles que*

$$A = LDM^T.$$

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

$$A = LU = LDD^{-1}U.$$

17. Harry Max Markowitz (né le 24 août 1927) est un économiste américain, lauréat du « prix Nobel » d'économie 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.

18. André-Louis Cholesky (15 octobre 1875 - 31 août 1918) était un mathématicien et officier français. Il entra en 1895 à l'école polytechnique et effectua ensuite une carrière dans les services géographiques et topographiques de l'armée. On lui doit une méthode célèbre pour la résolution des systèmes d'équations linéaires dont la matrice est symétrique définie positive.

Il suffit alors de poser $M^T = D^{-1}U$ pour obtenir l'existence de la factorisation. Son unicité est une conséquence de l'unicité de la factorisation LU. □

L'intérêt de cette dernière factorisation prend son sens lorsque la matrice A est symétrique, puisque dans ce cas $M = L$. La factorisation résultante $A = LDL^T$ peut alors être calculée avec un coût environ deux fois moindre.

La factorisation LDM^T permet également de démontrer le résultat suivant.

Proposition 2.8 *Soit A une matrice carrée d'ordre n admettant une factorisation LU. Alors, sa transposée A^T admet une factorisation LU.*

DÉMONSTRATION. Puisque A admet une factorisation LU, elle admet aussi une factorisation LDM^T et l'on a

$$A^T = (LDM^T)^T = (M^T)^T D^T L^T = MDL^T.$$

La matrice A^T admet donc elle aussi une factorisation LDM^T et, par suite, une factorisation LU. □

2.4.2 Factorisation de Cholesky

Une matrice symétrique définie positive vérifie les conditions de la proposition 2.7 (en vertu du théorème A.51) et admet par conséquent une factorisation LDL^T, dont la matrice diagonale D est de plus à termes *strictement positifs*. Cette observation conduit à une factorisation ne faisant intervenir qu'une seule matrice triangulaire inférieure, appelée *factorisation de Cholesky*. Plus précisément, on a le résultat suivant.

Théorème 2.9 *Soit A une matrice symétrique définie positive d'ordre n. Alors, il existe une unique matrice triangulaire inférieure B dont les éléments diagonaux sont strictement positifs telle que*

$$A = BB^T.$$

DÉMONSTRATION. En notant A_k, $1 \leq k \leq n$, les n sous-matrices principales extraites de A (définies par (2.5)), on peut écrire, pour tout vecteur $\boldsymbol{w}$ de $\mathbb{R}^k$,

$$\boldsymbol{w}^T A_k \boldsymbol{w} = \boldsymbol{v}^T A \boldsymbol{v}, \text{ avec } v_i = w_i,\ 1 \leq i \leq k, \text{ et } v_i = 0,\ k+1 \leq i \leq n,$$

ce qui montre que les sous-matrices A_k sont symétriques définies positives. La matrice A vérifie donc les conditions du théorème 2.2 et admet 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 $BC = C^T B^T$, d'où $C(B^T)^{-1} = B^{-1}C^T = I_n$ (une matrice étant triangulaire supérieure, l'autre triangulaire inférieure et toutes deux à coefficients diagonaux égaux à 1) et donc $C = B^T$. 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

$$A = B_1 {B_1}^T = B_2 {B_2}^T,$$

d'où ${B_2}^{-1}B_1 = {B_2}^T(B_1^T)^{-1}$. Il existe donc une matrice diagonale D telle que ${B_2}^{-1}B_1 = D$ et, par conséquent, $B_1 = B_2 D$. Finalement, on a

$$A = B_2 {B_2}^T = B_2 D D^T {B_2}^T,$$

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 $A = BB^T$ 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 symétrique, il suffit que les relations ci-dessus soient vérifiées pour $j \leq i$ (par exemple), et l'on va donc construire les colonnes de B à partir des colonnes de A. On fixe donc j à 1 et on fait varier i de 1 à n :

$$\begin{array}{rclcrcl} a_{11} & = & (b_{11})^2, & \text{d'où} & b_{11} & = & \sqrt{a_{11}}, \\ a_{21} & = & b_{11}b_{21}, & \text{d'où} & b_{21} & = & \dfrac{a_{21}}{b_{11}}, \\ \vdots & & & & \vdots & & \\ a_{n1} & = & b_{11}b_{n1}, & \text{d'où} & b_{n1} & = & \dfrac{a_{n1}}{b_{11}}, \end{array}$$

pour déterminer la première colonne de B. La $j^{\text{ième}}$ colonne de B est obtenue en utilisant les relations

$$\begin{array}{rclcrcl} a_{jj} & = & (b_{j1})^2 + (b_{j2})^2 + \cdots + (b_{jj})^2, & \text{d'où} & b_{jj} & = & \sqrt{a_{jj} - \sum_{k=1}^{j-1}(b_{jk})^2}, \\ a_{j+1j} & = & b_{j1}b_{j+1\,1} + b_{j2}b_{j+1\,2} + \cdots + b_{jj}b_{j+1\,j}, & \text{d'où} & b_{j+1j} & = & \dfrac{a_{j+1j} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}}, \\ \vdots & & & & \vdots & & \\ a_{nj} & = & b_{j1}b_{n1} + b_{j2}b_{n2} + \cdots + b_{jj}b_{nj}, & \text{d'où} & b_{nj} & = & \dfrac{a_{nj} - \sum_{k=1}^{j-1} b_{jk}b_{nk}}{b_{jj}}, \end{array}$$

après avoir préalablement déterminé les $j-1$ premières colonnes, le théorème 2.9 assurant que les quantités sous les racines carrées sont strictement positives. Dans la pratique, on ne vérifie d'ailleurs pas que la matrice A est définie positive (simplement qu'elle est symétrique) avant commencer l'algorithme. En effet, si l'on trouve à l'étape k, $1 \leq k \leq n$, que $(b_{kk})^2 \leq 0$, 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 bien définie positive, car, pour toute matrice inversible B et tout vecteur $\boldsymbol{v}$ non nul, on a

$$\left(BB^T\boldsymbol{v}, \boldsymbol{v}\right) = \|B^T\boldsymbol{v}\|_2 > 0.$$

Il est à noter que le déterminant d'une matrice dont on connaît la factorisation de Cholesky est immédiat, puisque

$$\det(A) = \det(BB^T) = (\det(B))^2 = \left(\prod_{i=1}^{n} b_{ii}\right)^2.$$

Le nombre d'opérations élémentaires nécessaires pour effectuer la factorisation de Cholesky d'une matrice A symétrique définie positive d'ordre n par les formules ci-dessus est de $\frac{n^3-n}{6}$ additions, $\frac{n^3-n}{6}$ multiplications, $\frac{n(n-1)}{2}$ divisions et n extractions de racines carrées, soit un coût très 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 alors ajouter $n(n-1)$ additions, $n(n-1)$ multiplications et $2n$ divisions pour la résolution des systèmes triangulaires, soit au total de l'ordre de $\frac{n^3}{6}$ additions, $\frac{n^3}{6}$ multiplications, $\frac{n^2}{2}$ divisions et n extractions de racines carrées.

2.4.3 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 $Q^{-1} = Q^T$, 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 $R\boldsymbol{x} = Q^T\boldsymbol{b}$, 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, dont la preuve s'appuie sur le fameux *procédé d'orthonormalisation de Gram*[19]*–Schmidt*[20].

19. 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 la l'analyse numérique et de la théorie des nombres. Le procédé qui porte aujourd'hui son nom fut publié en 1883 dans un article intitulé " *On series expansions determined by the methods of least squares*".

20. Erhard Schmidt (13 janvier 1876 - 6 décembre 1959) était un mathématicien allemand. Il est considéré comme l'un

Théorème 2.10 *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 positifs, telles que*

$$A = QR.$$

Cette factorisation est unique.

DÉMONSTRATION. La matrice A étant inversible, ses colonnes, notées $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ forment une base de $\mathbb{R}^n$. On peut alors obtenir une base orthonormée $\{\boldsymbol{q}_i\}_{1 \leq i \leq n}$ de $\mathbb{R}^n$ à partir de la famille $\{\boldsymbol{a}_i\}_{1 \leq i \leq n}$ en appliquant le procédé d'orthonormalisation de Gram–Schmidt, *i.e.*

$$\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|_2},$$

$$\widetilde{\boldsymbol{q}}_{i+1} = \boldsymbol{a}_{i+1} - \sum_{k=1}^{i} (\boldsymbol{a}_{i+1}, \boldsymbol{q}_k)\, \boldsymbol{q}_k, \;\; \boldsymbol{q}_{i+1} = \frac{\widetilde{\boldsymbol{q}}_{i+1}}{\|\widetilde{\boldsymbol{q}}_{i+1}\|_2}, \;\; i = 1, \ldots, n-1.$$

On en déduit alors que

$$\boldsymbol{a}_i = \sum_{j=1}^{i} r_{ij}\, \boldsymbol{q}_j,$$

avec $r_{ii} = \|\boldsymbol{a}_i - \sum_{k=1}^{i-1} (\boldsymbol{a}_i, \boldsymbol{q}_k)\, \boldsymbol{q}_k\|_2 > 0$, $r_{ij} = \big(\boldsymbol{a}_i, \boldsymbol{q}_j\big)$ pour $1 \leq j \leq i-1$, et $r_{ij} = 0$ pour $i < j \leq n$, $1 \leq i \leq n$. 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 $\boldsymbol{q}_i$, $1 \leq i \leq n$, on vient d'établir que $A = QR$.
Pour montrer l'unicité de la factorisation, on suppose que

$$A = Q_1 R_1 = Q_2 R_2,$$

d'où

$$Q_2^T Q_1 = R_2 {R_1}^{-1}.$$

En posant $T = R_2 {R_1}^{-1}$, on a $TT^T = Q_2^T Q_1 (Q_2^T Q_1)^T = I_n$, 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 2.9). □

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 6 propose une implémentation de cette méthode pour le calcul de la factorisation QR d'une matrice inversible d'ordre n. Cette approche nécessite d'effectuer $(n-1)n^2$ additions, n^3 multiplications, n^2 divisions et n extractions de racines carrées pour le calcul de la matrice Q, soit de l'ordre de $2\,n^3$ opérations.

Algorithme 6 Algorithme du procédé d'orthonormalisation de Gram–Schmidt

pour $j = 1$ à n **faire**
  $\boldsymbol{v}_j = \boldsymbol{a}_j$
  pour $i = 1$ à $j - 1$ **faire**
    $r_{ij} = \boldsymbol{q}_i^* \boldsymbol{a}_j$
    $\boldsymbol{v}_j = \boldsymbol{v}_j - r_{ij}\, \boldsymbol{q}_i$
  fin pour
  $r_{jj} = \|\boldsymbol{v}_j\|_2$
  $\boldsymbol{q}_j = \boldsymbol{v}_j / r_{jj}$
fin pour

Sur machine cependant, la propagation des erreurs d'arrondis (plus particulièrement pour les problèmes de grande taille) fait que les vecteurs $\boldsymbol{q_i}$ calculés ne sont pas linéairement indépendants, ce qui empêche la matrice Q d'être exactement orthogonale. Ces instabilités numériques sont dues au fait que la procédure d'orthonormalisation produit des valeurs très petites, ce qui pose problème en arithmétique à virgule flottante. Il convient alors de recourir à une version plus stable de l'algorithme, appelée *procédé de Gram–Schmidt modifié* (voir algorithme 7).

Cette modification 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. Une différence majeure concerne alors le calcul des coefficients r_{ij}, puisque la méthode « originale »

des fondateurs de l'analyse fonctionnelle abstraite moderne.

Algorithme 7 Algorithme du procédé d'orthonormalisation de Gram–Schmidt modifié

pour $i = 1$ à n **faire**
  $\boldsymbol{v}_i = \boldsymbol{a}_i$
fin pour
pour $i = 1$ à n **faire**
  $r_{ii} = \|\boldsymbol{v}_i\|_2$
  $\boldsymbol{q}_i = \boldsymbol{v}_i / r_{ii}$
  pour $j = i + 1$ à n **faire**
    $r_{ij} = \boldsymbol{q}_i^* \boldsymbol{v}_j$
    $\boldsymbol{v}_j = \boldsymbol{v}_j - r_{ij}\, \boldsymbol{q}_i$
  fin pour
fin pour

fait intervenir une colonne $\boldsymbol{a}_j$ de la matrice à factoriser alors que sa variante utilise un vecteur déjà partiellement orthogonalisé. Pour cette raison, et malgré l'équivalence mathématique entre les deux versions du procédé, la seconde est préférable à la première lorsque les calculs sont effectués en arithmétique à virgule flottante. Celle-ci requiert $\frac{(2\,n+1)n(n-1)}{2}$ additions, n^3 multiplications, n^2 divisions et n extractions de racines carrées pour la factorisation d'une matrice inversible d'ordre n, soit encore de l'ordre de $2\,n^3$ opérations au total.

Indiquons à présent comment réaliser la factorisation QR de d'une matrice non inversible ou rectangulaire. Supposons pour commencer que la matrice A est d'ordre n et non inversible. L'ensemble $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n\}$ des colonnes de A forment alors une famille liée de vecteurs de $\mathbb{R}^n$ et il existe un entier k, $1 < k \leq n$, tel que la famille $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$ 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 $\mathrm{rg}(A)$ premières colonnes de $\tilde{A} = AP$ sont libres, les $n - \mathrm{rg}(A)$ colonnes restantes étant engendrées par les $\mathrm{rg}(A)$ premières (cette permutation peut d'ailleurs se faire au fur et à mesure du procédé d'orthornormalisation, en effectuant une permutation circulaire de la $k^{\text{ième}}$ à la $n^{\text{ième}}$ colonne dès que l'on trouve une norme nulle). On applique alors le procédé de Gram–Schmidt jusqu'à l'étape $\mathrm{rg}(A)$ pour construire une famille orthonormée $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{\mathrm{rg}(A)}\}$ que l'on complète ensuite par des vecteurs $\boldsymbol{q}_{\mathrm{rg}(A)+1}, \ldots, \boldsymbol{q}_n$ pour obtenir une base de orthonormée de $\mathbb{R}^n$. 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

$$\tilde{\boldsymbol{a}}_i = \begin{cases} \sum_{j=1}^{i} r_{ij}\, \boldsymbol{q}_j & \text{si } 1 \leq i \leq \mathrm{rg}(A), \\ \sum_{j=1}^{\mathrm{rg}(A)} r_{ij}\, \boldsymbol{q}_j & \text{si } \mathrm{rg}(A) + 1 \leq i \leq n, \end{cases}$$

avec $r_{ii} > 0$, $1 \leq i \leq \mathrm{rg}(A)$, et on note R la matrice carrée d'ordre n telle que

$$R = \begin{pmatrix} r_{11} & \cdots & \cdots & \cdots & r_{1n} \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & r_{\mathrm{rg}(A)\,\mathrm{rg}(A)} & \cdots & r_{\mathrm{rg}(A)\,n} \\ \vdots & & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

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 $\mathbb{R}^m$ et que $\mathrm{rg}(A) = m$, 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

$$R = \begin{pmatrix} r_{11} & \ldots & \ldots & \ldots & \ldots & r_{1n} \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \ldots & 0 & r_{mm} & \ldots & r_{mn} \end{pmatrix}.$$

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 $\mathrm{rg}(A) = n$ (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 $\mathbb{R}^m$. 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 modifiée, cette méthode reste relativement peu utilisée en pratique pour le calcul d'une factorisation QR, car on lui préfère la *méthode de Householder*[21] [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 2.11 *Soit $\boldsymbol{v}$ une vecteur non nul de $\mathbb{R}^n$. On appelle* ***matrice de Householder*** *associée au* ***vecteur de Householder $\boldsymbol{v}$****, et on note $H(\boldsymbol{v})$, la matrice définie par*

$$H(\boldsymbol{v}) = I_n - 2\,\frac{\boldsymbol{v}\boldsymbol{v}^T}{\boldsymbol{v}^T\boldsymbol{v}}. \tag{2.6}$$

On pose de plus $H(\mathbf{0}) = I_n$, ce qui permet de considérer la matrice identité comme une matrice de Householder.

21. 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.

Les matrices de Householder possèdent des propriétés intéressantes, que l'on résume dans le résultat suivant.

Lemme 2.12 *Soit $\boldsymbol{v}$ un vecteur non nul de $\mathbb{R}^n$ 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 $\mathbb{R}^n$ et $\boldsymbol{e}$ est un vecteur unitaire tels que $\boldsymbol{x} \neq \pm\|\boldsymbol{x}\|_2\,\boldsymbol{e}$, on a*

$$H\left(\boldsymbol{x} \pm \|\boldsymbol{x}\|_2\,\boldsymbol{e}\right)\boldsymbol{x} = \mp\|\boldsymbol{x}\|_2\,\boldsymbol{e}.$$

DÉMONSTRATION. Il est facile de voir que $H(\boldsymbol{v}) = H(\boldsymbol{v})^T$. Par ailleurs, on vérifie que

$$H(\boldsymbol{v})^2 = I_n - 4\,\frac{\boldsymbol{v}\boldsymbol{v}^T}{\|\boldsymbol{v}\|_2{}^2} + 4\,\frac{\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{v}\boldsymbol{v}^T}{\|\boldsymbol{v}\|_2{}^4} = I_n - 4\,\frac{\boldsymbol{v}\boldsymbol{v}^T}{\|\boldsymbol{v}\|_2{}^2} + 4\,\frac{\|\boldsymbol{v}\|_2{}^2\,\boldsymbol{v}\boldsymbol{v}^T}{\|\boldsymbol{v}\|_2{}^4} = I_n.$$

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 $\mathbb{R}^n$. On a

$$\begin{aligned} H\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)\boldsymbol{x} &= \boldsymbol{x} - 2\,\frac{\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)^T}{\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)^T\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)} \\ &= \boldsymbol{x} - 2\,\frac{\left(\|\boldsymbol{x}\|_2{}^2 + \|\boldsymbol{x}\|_2\,v_1\right)\left(\boldsymbol{x} + \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)}{2\,\|\boldsymbol{x}\|_2{}^2 + 2\,\|\boldsymbol{x}\|_2\,v_1} \\ &= -\|\boldsymbol{x}\|_2\,\boldsymbol{e}_1. \end{aligned}$$

On obtient par un calcul similaire que $H\left(\boldsymbol{x} - \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1\right)\boldsymbol{x} = \|\boldsymbol{x}\|_2\,\boldsymbol{e}_1$. □

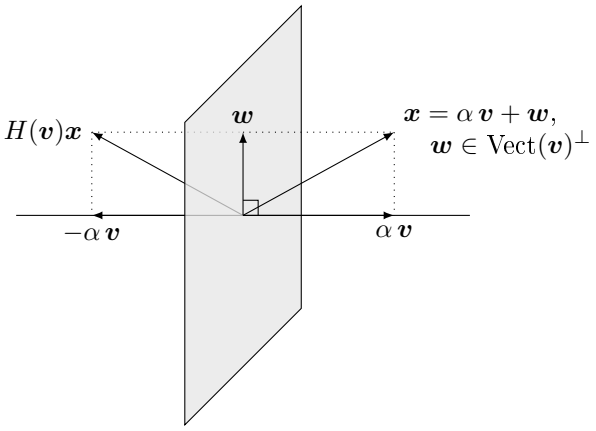

FIGURE 2.1 – Transformation d'un vecteur $\boldsymbol{x}$ de l'espace par la matrice de Householder $H(\boldsymbol{v})$.

La matrice de Householder $H(\boldsymbol{v})$ est la matrice de la *symétrie orthogonale par rapport à l'hyperplan orthogonal à $\boldsymbol{v}$* (voir figure 2.1). Les matrices de Householder peuvent par conséquent être utilisées pour annuler certaines composantes d'un vecteur $\boldsymbol{x}$ de $\mathbb{R}^n$ donné, comme le montre l'exemple suivant.

Exemple. Soit $\boldsymbol{x} = \begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix}^T$ et choisissons $\boldsymbol{e} = \boldsymbol{e}_3$. 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. 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 $k^{\text{ième}}$ 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 $k^{\text{ième}}$ colonne de $A^{(k)}$.

Soit $\tilde{\boldsymbol{a}}^{(k)}$ le vecteur de $\mathbb{R}^{n-k+1}$ contenant les éléments $a_{ik}^{(k)}$, $k \leq i \leq n$, de $A^{(k)}$. Si $\sum_{i=k+1}^{n} |a_{ik}^{(k)}| = 0$, alors $A^{(k)}$ est déjà de la « forme » de $A^{(k+1)}$ et on pose $H^{(k)} = I_n$. Si $\sum_{i=k+1}^{n} |a_{ik}^{(k)}| > 0$, alors il existe, en vertu du lemme 2.12, un vecteur $\tilde{\boldsymbol{v}}^{(k)}$ de $\mathbb{R}^{n-k+1}$, donné par

$$\tilde{\boldsymbol{v}}^{(k)} = \tilde{\boldsymbol{a}}^{(k)} \pm \|\tilde{\boldsymbol{a}}^{(k)}\|_2 \, \tilde{\boldsymbol{e}}_1^{(n-k+1)}, \tag{2.7}$$

où $\tilde{\boldsymbol{e}}_1^{(n-k+1)}$ désigne le premier vecteur de la base canonique de $\mathbb{R}^{n-k+1}$, 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

$$H^{(k)} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H(\tilde{\boldsymbol{v}}^{(k)}) \end{pmatrix} = H(\boldsymbol{v}^{(k)}), \text{ avec } \boldsymbol{v}^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\boldsymbol{v}}^{(k)} \end{pmatrix} \in \mathbb{R}^n. \tag{2.8}$$

On réitère ces opérations jusqu'à obtenir la matrice triangulaire supérieure

$$A^{(n)} = H^{(n-1)} \dots H^{(1)} A^{(1)},$$

et alors $A^{(n)} = R$ et $Q = \left(H^{(n-1)} \dots H^{(1)}\right)^T = H^{(1)} \dots H^{(n-1)}$. 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 (2.7) lors de la construction du vecteur de Householder à la $k^{\text{ième}}$ 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

$$A^{(n+1)} = \begin{pmatrix} a_{11}^{(n+1)} & \dots & a_{1n}^{(n+1)} \\ 0 & \ddots & \vdots \\ \vdots & \ddots & a_{nn}^{(n+1)} \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}.$$

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.55. 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 ceci etait 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 de Householder $H(\boldsymbol{v})$ avec $\boldsymbol{v}$ un vecteur de $\mathbb{R}^m$. En utilisant (2.6), on obtient que

$$H(v)M = M - \frac{2}{\|\boldsymbol{v}\|_2^2}\,\boldsymbol{v}(M^T\boldsymbol{v})^T.$$

Ainsi, le produit $H(v)M$ se ramène *grosso modo* à un produit scalaire (le coefficient $\beta = \frac{2}{\|\boldsymbol{v}\|_2^2}$), un produit matrice-vecteur (le produit $\boldsymbol{w} = M^T\boldsymbol{v}$), un produit vecteur-vecteur (la matrice $\boldsymbol{v}(\beta\,\boldsymbol{w})^T$) suivi de la différence de deux matrices et nécessite au total $2m^2-1$ additions, $2m(m+1)$ multiplications et une division. Ce résultat est à comparer aux $2m-1$ additions, $m(m+2)$ multiplications et une division requises pour la contruction de $H(\boldsymbol{v})$, ajoutées aux $m^2(m-1)$ additions et m^3 multiplications nécessaires au produit de deux matrices quelconques. Par des considérations analogues, on a

$$MH(v) = M - \frac{2}{\|\boldsymbol{v}\|_2^2}\,(M\boldsymbol{v})\boldsymbol{v}^T$$

Une conséquence de cette remarque est que l'on n'a pas, *a priori*, pas à stocker, ni même à calculer la matrice Q lors de la résolution d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ par la méthode 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 $\frac{1}{3}\,(n+1)(2n^2+7n+3)$ additions, $\frac{2}{3}(n+1)(n+2)(n+3)$ 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 (2.8)), $1 \leq k \leq n-1$, 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.

Pour aller plus loin

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 »* (*Jiŭzhāng Suànshù*), compilé entre le deuxième et le premier siècle avant J.-C..

La sensibilité aux erreurs d'arrondis de la méthode d'élimination de Gauss, sujet que nous n'avons pas précisément abordé, a été étudiée par Wilkinson dans [Wil61]. Cet article constitue l'une des premières contributions majeures à l'analyse d'erreur *a priori* rétrograde des méthodes directes.

Une d'autre façon de parvenir la factorisation QR d'une matrice est d'utiliser des *matrices de rotation de Givens*[22] pour annuler les coefficients sous-diagonaux de la matrice à factoriser [Giv58], en parcourant

22. James Wallace Givens, Jr. (14 décembre 1910 - 5 mars 1993) était un mathématicien américain et un pionnier de l'informatique et du calcul scientifique. Il reste connu pour les matrices de rotation portant son nom.

ligne par ligne ou colonne par colonne. Ces matrices orthogonales apparaissent aussi dans la *méthode de Jacobi*[23] pour le calcul des valeurs propres d'une matrice symétrique (voir le chapitre 4).

Références du chapitre

[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*. Society for Industrial and Applied Mathematics, third edition, 1999.

[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 :1235–1240, 1941.

[FM67] G. E. Forsythe and C. B. Moler. *Computer solution of linear systems*. Series in automatic computation. Prentice-Hall, 1967.

[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.

[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.

[Hou58] A. S. Householder. Unitary triangularization of a nonsymmetric matrix. *J. ACM*, 5(4) :339–342, 1958.

[Wil61] J. H. Wilkinson. Error analysis of direct methods of matrix inversion. *J. ACM*, 8(3) :281–330, 1961.

23. 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, la théorie des déterminants. Un très grand nombre de résultats d'algèbre et d'analyse portent ou utilisent son nom.

Chapitre 3

Méthodes itératives de résolution des systèmes linéaires

L'idée des méthodes itératives de résolution des 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 la solution du système (2.1). Dans ce chapitre, on va présenter des méthodes itératives parmi les plus simples à mettre en œuvre, à savoir les méthodes de *Jacobi*, de *Gauss–Seidel*[1] et leurs variantes. Dans ces méthodes, qualifées de *méthodes itératives linéaires stationnaires du premier ordre*, la suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ est obtenue, à partir d'un vecteur initial arbitraire $\boldsymbol{x}^{(0)}$, par une relation de récurrence de la forme

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c},\ \forall k \in \mathbb{N}, \tag{3.2}$$

où la matrice carrée B, appelée *matrice d'itération* de la méthode, et le vecteur $\boldsymbol{c}$ dépendent de la matrice A et du second membre $\boldsymbol{b}$ du système à résoudre.

Pour une matrice pleine, le coût de calcul de ces méthodes 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 $\frac{2}{3}\,n^3$ opérations. Ainsi, une méthode itérative ne sera compétitive que si elle converge en un nombre d'itérations indépendant de, ou bien croissant de manière sous-linéaire avec, l'entier n. Cependant, les méthodes directes peuvent s'avérer particulièrement coûteuses pour les grandes matrices creuses (comme celles issues de la discrétisation d'équations différentielles ou aux dérivées partielles[2]) et les méthodes itératives sont souvent associées à la résolution de ce type de systèmes linéaires.

Avant d'aborder leur description, on va donner quelques résultats généraux de convergence et de stabilité, ainsi que des principes de comparaison (en terme de *« vitesse » de convergence*), d'une classe de méthodes itératives de la forme (3.2). Des résultats plus précis pour les méthodes présentées, mais s'appuyant sur des cas particuliers, comme celui de systèmes dont la matrice A est *symétrique définie positive*, sont établis en fin de chapitre.

3.1 Généralités

Dans cette section, nous abordons quelques aspects généraux des méthodes itératives de résolution de systèmes linéaires de la forme (3.2). Dans toute la suite, nous nous plaçons dans le cas de matrices et de vecteurs complexes, mais les résultats sont bien sûr valables dans le cas réel.

1. 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.

2. Il existe néanmoins des solveurs éfficaces basés sur des méthodes directes pour ces cas particuliers (voir par exemple [DER86]).

Commençons par une définition naturelle.

Définition 3.1 *On dit que la méthode itérative est* ***convergente*** *si l'on a* (3.1) *pour tout initialisation* $\boldsymbol{x}^{(0)}$ *dans* $\mathbb{C}^n$.

Nous introduisons ensuite une condition qu'une méthode itérative de la forme (3.2) doit nécessairement satisfaire pour qu'elle puisse converger vers la solution de (2.1).

Définition 3.2 *Une méthode itérative de la forme* (3.2) *est dite* ***consistante*** *avec* (2.1) *si* B *et* $\boldsymbol{c}$ *sont tels que l'on a* $\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}$, *le vecteur* $\boldsymbol{x}$ *étant la solution de* (2.1), *ou, de manière équivalente,* $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$.

Définitions 3.3 *On appelle* ***erreur*** *(resp.* ***résidu****) à l'itération* k, $k \in \mathbb{N}$, *de la méthode itérative le vecteur* $\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ *(resp.* $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$*), où* $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ *est la solution de* (2.1).

On déduit de ces définitions qu'une méthode itérative consistante de la forme (3.2) converge si et seulement si $\lim\limits_{k\to+\infty} \boldsymbol{e}^{(k)} = \boldsymbol{0}$ (soit encore si $\lim\limits_{k\to+\infty} \boldsymbol{r}^{(k)} = \lim\limits_{k\to+\infty} A\boldsymbol{e}^{(k)} = \boldsymbol{0}$).

La seule propriété de consistance ne suffisant pas à assurer que la méthode considérée converge, nous donnons dans le résultat suivant un critère fondamental de convergence.

Théorème 3.4 *Si une méthode de la forme* (3.2) *est consistante, celle-ci est convergente si et seulement si* $\rho(B) < 1$.

Démonstration. La méthode étant supposée consistante, l'erreur à l'itération $k+1$, $k \in \mathbb{N}$, vérifie la relation de récurrence

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x} = B\boldsymbol{x}^{(k)} - \boldsymbol{c} - (B\boldsymbol{x} - \boldsymbol{c}) = B\big(\boldsymbol{x}^{(k)} - \boldsymbol{x}\big) = B\boldsymbol{e}^{(k)}.$$

On déduit alors le résultat du théorème A.58. □

En pratique, le rayon spectral d'une matrice est difficile à calculer, mais on a déduit du théorème A.56 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 telle que $\|B\| < 1$.

Une autre question à laquelle on se trouve confronté lorsque l'on est en présence de deux méthodes itératives convergentes est de savoir laquelle des deux converge le plus rapidement. Une réponse est fournie par le résultat suivant : la méthode la plus « rapide » est celle dont la matrice a le plus petit rayon spectral.

Théorème 3.5 *Soit* $\|\cdot\|$ *une norme vectorielle quelconque. On considère deux méthodes itératives consistantes avec* (2.1),

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c} \text{ et } \tilde{\boldsymbol{x}}^{(k+1)} = \tilde{B}\tilde{\boldsymbol{x}}^{(k)} + \tilde{\boldsymbol{c}},\ k \geq 0,$$

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*

$$k \geq N \Rightarrow \sup_{\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| = 1} \left(\frac{\|\tilde{\boldsymbol{x}}^{(k)} - \boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|} \right)^{1/k} \geq \frac{\rho(\tilde{B})}{\rho(B) + \varepsilon},$$

où $\boldsymbol{x}$ *désigne la solution de* (2.1).

Démonstration. D'après le théorème A.59, étant donné $\varepsilon > 0$, il existe un entier N, dépendant de ε, tel que

$$k \geq N \Rightarrow \sup_{\|\boldsymbol{e}^{(0)}\| = 1} \|B^k \boldsymbol{e}^{(0)}\|^{1/k} \leq (\rho(B) + \varepsilon).$$

Par ailleurs, pour tout entier $k \geq N$, 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.56 et en notant $\|\cdot\|$ la norme matricielle subordonnée à la norme vectorielle considérée. Ceci achève de démontrer l'assertion. □

Parlons à présent de l'utilisation d'une méthode itérative pour le calcul d'une solution *approchée* de (2.1). En pratique, il conviendrait de mettre fin aux calculs à la première itération pour laquelle l'erreur est « suffisament petite », c'est-à-dire 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 vectorielle 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, les résidus $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ étant très faciles 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 doit choisir δ telle que $\delta \leq \frac{\varepsilon}{\|A^{-1}\|}$. Ce critère peut par conséquent 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 $\mathrm{cond}(A)$ désigne le conditionnement de la matrice A relativement à la norme subordonnée $\|\cdot\|$ considérée.

Un autre critère parfois utilisé dans la pratique est basé sur l'*incrément* $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$, $k \in \mathbb{N}$. L'erreur d'une méthode itérative de la forme (3.2) vérifiant la relation de récurrence $\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)}$, $\forall k \in \mathbb{N}$, on obtient, par utilisation de l'inégalité triangulaire,

$$\|\boldsymbol{e}^{(k+1)}\| \leq \|B\|\,\|\boldsymbol{e}^{(k)}\| \leq \|B\|\left(\|\boldsymbol{e}^{(k+1)}\| + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|\right),\ \forall k \in \mathbb{N},$$

d'où

$$\|\boldsymbol{e}^{(k+1)}\| \leq \frac{\|B\|}{1 - \|B\|}\,\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|,\ \forall k \in \mathbb{N}.$$

Les méthodes de Jacobi et de Gauss–Seidel que nous allons présenter font partie de la famille de méthodes itératives de la forme

$$M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b},\ k \geq 0, \tag{3.3}$$

basées sur la « décomposition » (*"splitting"* en anglais) de la matrice A,

$$A = M - N, \tag{3.4}$$

avec M une matrice inversible. Pour que la formule ci-dessus soit utilisable en pratique, il faut par ailleurs que la matrice M soit *« facilement »* inversible, c'est-à-dire que l'on doit pouvoir résoudre simplement et à faible coût un système linéaire ayant M pour matrice. On verra en effet que, pour les méthodes précitées, M est une matrice respectivement diagonale et triangulaire inférieure.

En vertu du théorème 3.4, la convergence d'une méthode consistante définie par (3.3) dépend de la valeur du rayon spectral de sa matrice d'itération, $M^{-1}N$. Plusieurs résultats de convergence, propres aux méthodes de Jacobi et de Gauss–Seidel (ainsi que leurs variantes relaxées), sont donnés dans la section 3.5. Le résultat ci-dessous garantit la convergence d'une méthode itérative associée à une décomposition $A = M - N$ *quelconque* d'une matrice A hermitienne[3] définie positive.

3. Comme on l'a déjà mentionné, tous les résultats énoncés le sont dans le cas complexe, mais restent vrais dans le cas réel en remplaçant le mot « hermitien » par « symétrique ».

Théorème 3.6 *Soit A une matrice hermitienne définie positive, que l'on décompose sous la forme* (3.4) *avec M une matrice inversible. Si la matrice hermitienne M^*+N est définie positive, alors $\rho(M^{-1}N) < 1$.*

DÉMONSTRATION. La matrice A (que l'on suppose d'ordre n) étant hermitienne, la matrice $M^* + N$ est effectivement hermitienne puisque

$$M^* + N = M^* + M - A = M + M^* - A^* = M + N^*.$$

La matrice A étant par ailleurs définie positive, l'application $\|\cdot\|$ de $\mathbb{C}^n$ dans $\mathbb{R}$ définie par

$$\|\boldsymbol{v}\| = (\boldsymbol{v}^* A \boldsymbol{v})^{1/2},$$

définit une norme vectorielle, et on note également $\|\cdot\|$ la norme matricielle qui lui est subordonnée.

On va maintenant établir que $\|M^{-1}N\| < 1$. Par définition, on a

$$\|M^{-1}N\| = \|I_n - M^{-1}A\| = \sup_{\|\boldsymbol{v}\|=1} \|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|.$$

D'autre part, pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$ 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 fontion de $\mathbb{C}^n$ dans $\mathbb{R}$ qui à $\boldsymbol{v}$ associe $\|\boldsymbol{v}-M^{-1}A\boldsymbol{v}\|$ étant continue sur le compact $\{\boldsymbol{v} \in \mathbb{C}^n \mid \|\boldsymbol{v}\| = 1\}$, elle y atteint sa borne supérieure, ce qui achève la démonstration. □

Les méthodes itératives de la forme (3.3) étant destinées à être utilisées sur des machines dont les calculs sont entachés d'erreurs d'arrondis, il convient de s'assurer que leur convergence ne s'en trouve pas détruite ou encore qu'elles ne convergent pas vers des vecteurs qui ne sont pas la solution de 2.1. Le résultat de stabilité suivant montre qu'il n'en est rien.

Théorème 3.7 *Soit A une matrice inversible d'ordre n, décomposée sous la forme* (3.4)*, avec M une matrice inversible et $\rho(M^{-1}N) < 1$, $\boldsymbol{b}$ un vecteur de $\mathbb{C}^n$ 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 le vecteur $\boldsymbol{x}^{(k+1)}$, $k \in \mathbb{N}$, est donné par*

$$\boldsymbol{x}^{(k+1)} = M^{-1}N\boldsymbol{x}^{(k)} + M^{-1}\boldsymbol{b} + \boldsymbol{\epsilon}^{(k)} \tag{3.5}$$

avec $\boldsymbol{\epsilon}^{(k)} \in \mathbb{C}^n$, et qu'il existe une norme vectoriel $\|\cdot\|$ et une constante positive ϵ telle que, pour tout entier naturel k,

$$\|\epsilon^{(k)}\| \leq \epsilon.$$

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 (3.5), l'erreur à l'étape $k+1$ vérifie la relation de récurrence

$$\boldsymbol{e}^{(k+1)} = M^{-1}N\boldsymbol{e}^{(k)} + \boldsymbol{\epsilon}^{(k)},\ \forall k \geq 0,$$

dont on déduit que

$$\boldsymbol{e}^{(k)} = (M^{-1}N)^k\boldsymbol{e}^{(0)} + \sum_{i=0}^{k-1}(M^{-1}N)^i\boldsymbol{\epsilon}^{(k-i-1)},\ \forall k \geq 0.$$

Puisque $\rho(M^{-1}N) < 1$, il existe, par application du théorème (A.56), 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 $\mathbb{C}^n$ étant équivalentes, il existe une constante C, strictement plus grande que 1 et ne dépendant que de $M^{-1}N$, telle que

$$C^{-1}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_s \leq C\,\|\boldsymbol{v}\|,\ \forall \boldsymbol{v} \in \mathbb{C}^n.$$

Par majoration, il vient alors

$$\|\boldsymbol{e}^{(k)}\|_s \leq \|M^{-1}N\|_s{}^k \|\boldsymbol{e}^{(0)}\|_s + C\,\epsilon \sum_{i=0}^{k-1} \|M^{-1}N\|_s{}^i \leq \|M^{-1}N\|_s{}^k \|\boldsymbol{e}^{(0)}\|_s + \frac{C\,\epsilon}{1+\|M^{-1}N\|_s},\ \forall k \geq 0,$$

d'où on tire le résultat en posant $K = \dfrac{C^2}{1+\|M^{-1}N\|_s}$. □

3.2 Méthodes de Jacobi et de sur-relaxation

Observons que, si les coefficients diagonaux de la matrice A sont non nuls, il est possible d'isoler la $i^{\text{ième}}$ inconnue dans la $i^{\text{ième}}$ équation de (2.1), $1 \leq i \leq n$ et l'on obtient alors le système linéaire équivalent

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j \right),\ i = 1, \ldots, n.$$

Le méthode de Jacobi 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

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} \right),\ i = 1, \ldots, n,\ k \in \mathbb{N}, \tag{3.6}$$

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 dont les coefficients sont les coefficients diagonaux de A, $d_{ij} = a_{ij}\,\delta_{ij}$, E est la matrice triangulaire inférieure de coefficients $e_{ij} = -a_{ij}$ si $i > j$ et 0 autrement, et F est la matrice triangulaire supérieure telle que $f_{ij} = -a_{ij}$ si $i < j$ et 0 autrement, avec $1 \leq i, j \leq n$. On a ainsi $A = D - (E + F)$ et la matrice d'itération de la méthode est donnée par

$$B_J = D^{-1}(E + F).$$

On note que la matrice diagonale D doit être inversible. Cette condition n'est cependant pas très restrictive dans la mesure où l'ordre des équations et des inconnues peut être modifié.

Une généralisation de la méthode de Jacobi est la *méthode de sur-relaxation de Jacobi* (*Jacobi over-relaxation* (*JOR*) en anglais), dans laquelle un paramètre de relaxation est introduit. Les relations de récurrence deviennent

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{\substack{i=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} \right) + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n,\ k \in \mathbb{N},$$

ce qui correspond à la matrice d'itération suivante

$$B_J(\omega) = \omega\, B_J + (1-\omega)\, I_n. \tag{3.7}$$

Cette méthode est consistante pour toute valeur de ω non nulle et coïncide avec la méthode de Jacobi pour $\omega = 1$. L'idée de relaxer la méthode repose sur le fait que, si l'efficacié de la méthode se mesure par le rayon spectral de la matrice d'itération, alors, puisque $\rho(B_J(\omega))$ est une fonction continue de ω, on peut trouver une valeur de ω pour laquelle ce rayon spectral est le plus petit possible et qui donne donc une méthode itérative plus efficace que la méthode de Jacobi. Ce type de raisonnement s'applique également à la méthode de Gauss–Seidel (voir la section suivante).

L'étude des méthodes de relaxation pour un type de matrices donné consiste en général à déterminer, s'ils existent, un intervalle I de $\mathbb{R}$ ne contenant pas l'origine tel que, pour tout ω choisi dans I, la méthode converge, et un paramètre de relaxation optimal $\omega_0 \in I$ tel que (dans le cas de la méthode de sur-relaxation)

$$\rho(B_J(\omega_0)) = \inf_{\omega \in I} \rho(B_J(\omega)).$$

3.3 Méthodes de Gauss–Seidel et de sur-relaxation successive

Remarquons à présent que, lors du calcul du vecteur $\boldsymbol{x}^{(k+1)}$ par les formules de récurrence (3.6), les premières $i-1^{\text{ièmes}}$ composantes de $\boldsymbol{x}^{(k+1)}$ sont connues lors de la détermination de $i^{\text{ième}}$, $2 \leq i \leq n$. La méthode de Gauss–Seidel utilise ce fait en se servant des composantes du vecteur $\boldsymbol{x}^{(k+1)}$ déjà obtenues pour le calcul des suivantes. On a alors

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right),\ i = 1, \ldots, n,\ k \in \mathbb{N}, \tag{3.8}$$

ce qui revient à poser $M = D - E$ et $N = F$ dans la décomposition (3.4), d'où la matrice d'itération associée

$$B_{GS} = (D - E)^{-1} F.$$

Pour que la méthode soit bien définie, il faut que la matrice D soit inversible, mais, là encore, cette condition n'est pas très restrictive en pratique.

On peut également introduire dans cette méthode un paramètre de relaxation ω. On parle alors de *méthode de sur-relaxation successive* (*successive over-relaxation* (*SOR*) en anglais), définie par

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) + (1 - \omega)\, x_i^{(k)},\ i = 1, \ldots, n,\ k \in \mathbb{N},$$

et dont la matrice d'itération est

$$B_{GS}(\omega) = (I_n - \omega\, D^{-1} E)^{-1} \left((1 - \omega)\, I_n + \omega\, D^{-1} F \right).$$

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*, de *sous-relaxation* si $\omega < 1$. Il s'avère que la valeur du paramètre optimal est, en géneral, plus grande que 1, d'où le nom de la méthode.

3.4 Remarques sur l'implémentation des méthodes itératives

Parlons à présent de l'implémentation des méthodes de Jacobi et de Gauss–Seidel, et de leurs variantes, en utilisant un test d'arrêt basé sur le résidu. Dans ce cas, il convient tout d'abord de remarquer que les méthodes itératives de la forme (3.3) peuvent également s'écrire

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + M^{-1} \boldsymbol{r}^{(k)},\ k \geq 0, \tag{3.9}$$

où le vecteur $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ est le résidu à l'étape k. C'est sur cette dernière écriture que reposeront les algorithmes proposés pour les différentes méthodes.

Pour l'initialisation de la méthode itérative, on choisit généralement, sauf si l'on possède *a priori* des 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 :

- calcul du résidu,
- résolution du système linéaire ayant M pour matrice et le résidu comme second membre,
- mise à jour de l'approximation de la solution,

jusqu'à ce que la norme du résidu soit plus petite qu'une tolérance prescrite. Dans la pratique, il est aussi nécessaire de limiter le nombre d'itérations, afin d'éliminer les problèmes liés à la non-convergence d'une méthode.

Le nombre d'opérations élémentaires requises à chaque itération pour un système linéaire d'ordre n se décompose en $2n$ additions et n^2 multiplications pour le calcul du résidu, n divisions (pour la méthode de Jacobi) ou $\frac{n(n-1)}{2}$ additions, $\frac{n(n-1)}{2}$ multiplications et n divisions (pour la méthode de Gauss–Seidel) 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 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 de l'ordre de $\frac{1}{2}\,n^2$ additions et $\frac{3}{2}\,n^2$ multiplications s'avère donc très favorable par rapport à celui des méthodes directes du chapitre 2 si le nombre d'itérations à effectuer reste petit devant n.

Terminons en remarquant que, dans la méthode de Jacobi (ou JOR), chaque composante de l'approximation de la solution peut être calculée indépendamment des autres. Cette méthode est donc facilement parallélisable. Au contraire, pour la méthode de Gauss–Seidel (ou SOR), ce calcul ne peut se faire que séquentiellement, mais sans qu'on ait toutefois besoin de stocker l'approximation de la solution à l'étape précédente, d'où un gain de mémoire.

3.5 Convergence des méthodes de Jacobi et Gauss–Seidel

Avant de considérer la résolution de systèmes linéaires dont les matrices possèdent des propriétés particulières, commençons par un résultat général pour la méthode de sur-relaxation successive.

Théorème 3.8 ***(condition nécessaire de convergence pour la méthode SOR)*** *Le rayon spectral de la matrice de la méthode de sur-relaxation successive vérifie toujours l'inégalité*

$$\rho(B_{GS}(\omega)) \geq |\omega - 1|, \ \forall \omega > 0.$$

Cette méthode ne peut donc converger que si $\omega \in]0, 2[$.

Démonstration. On remarque que le déterminant de $B_{GS}(\omega)$ vaut

$$\det(B_{GS}(\omega)) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = (1-\omega)^n,$$

compte tenu des structures des matrices, respectivement diagonale et triangulaires, D, E et F. En notant λ_i, $1 \leq i \leq n$, les valeurs propres de cette matrice, on en déduit alors que

$$\rho(B_{GS}(\omega))^n \geq \prod_{i=1}^{n} |\lambda_i| = |\det(B_{GS}(\omega))| = |1-\omega|^n.$$

□

Nous indiquons également le résultat suivant concernant la méthode de sur-relaxation de Jacobi.

Théorème 3.9 *Si la méthode de Jacobi converge, alors la méthode de sur-relaxation de Jacobi converge pour* $0 < \omega \leq 1$.

Démonstration. D'après (3.7), les valeur propres de la matrice $B_J(\omega)$ sont

$$\eta_k = \omega\,\lambda_k + 1 - \omega, \ k = 1, \ldots, n,$$

où les nombres λ_k sont les valeurs propres de la matrice B_J. En posant $\lambda_k = r_k\, e^{\mathrm{i}\theta_k}$, on a alors

$$|\eta_k| = \omega^2 {r_k}^2 + 2\omega r_k \cos(\theta_k)(1-\omega) + (1-\omega)^2 \leq (\omega r_k + 1 - \omega)^2, \ k = 1, \ldots, n,$$

qui est strictement inférieur à 1 si $0 < \omega \leq 1$. □

3.5.1 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. Dans le contexte des méthodes itératives, on est en mesure d'établir des résultats de convergence *a priori* pour de telles matrices.

Théorème 3.10 *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 $i = 1, \ldots, n$. 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.

On considère à présent l'erreur à l'itération $k+1$, $k \in \mathbb{N}$, de la méthode de Gauss–Seidel qui vérifie

$$e_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}}\, e_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}}\, e_j^{(k)},\ 1 \leq i \leq n.$$

On va établir que

$$\|e^{(k+1)}\|_\infty \leq r\, \|e^{(k)}\|_\infty,\ \forall k \in \mathbb{N},$$

en raisonnant par récurrence sur l'indice i, $1 \leq i \leq n$, des composantes du vecteur. Pour $i = 1$, on a

$$e_1^{(k+1)} = -\sum_{j=2}^{n} \frac{a_{1j}}{a_{11}}\, e_j^{(k)},\ \text{d'où } |e_1^{(k+1)}| \leq r\, \|e^{(k)}\|_\infty.$$

Supposons que $|e_j^{(k+1)}| \leq r\, \|e^{(k)}\|_\infty$ pour $j = 1, \ldots, i-1$. On a alors

$$|e_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k+1)}| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |e_j^{(k)}| \leq \|\boldsymbol{e}^{(k)}\|_\infty \left(r \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \right) < \|\boldsymbol{e}^{(k)}\|_\infty \sum_{\substack{i=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|,$$

d'où $|e_i^{(k+1)}| \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty$, ce qui achève la preuve par récurrence. On a par conséquent

$$\|\boldsymbol{e}^{(k)}\|_\infty \leq r\, \|\boldsymbol{e}^{(k)}\|_\infty \leq \cdots \leq r^k\, \|\boldsymbol{e}^{(0)}\|_\infty,$$

et, par suite,

$$\lim_{k\to+\infty} \|\boldsymbol{e}^{(k)}\|_\infty = 0,$$

ce qui prouve la convergence de la méthode de Gauss–Seidel. □

3.5.2 Cas des matrices hermitiennes définies positives

Dans le cas de matrices hermitiennes définies positives, on peut établir que la condition nécessaire de convergence de la méthode de sur-relaxation successive du théorème 3.8 est suffisante.

Théorème 3.11 ***(condition suffisante de convergence de la méthode SOR)*** *Si la matrice A est hermitienne définie positive, alors la méthode de sur-relaxation successive converge si $\omega \in]0, 2[$.*

DÉMONSTRATION. La matrice A étant hermitienne, on a $D - E - F = D^* - E^* - F^*$, 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

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega} D + F = \frac{2-\omega}{\omega}.$$

La matrice D est elle aussi définie positive. En effet, en notant A_k, $1 \leq k \leq n$, les sous-matrices principales de A, on a $\sigma(D) = \cup_{k=1}^{n} \sigma(A_k)$, chacune de ces sous-matrices étant définie positive (c'est une conséquence du théorème A.51). La matrice $M^* + N$ est donc définie positive si et seulement si $0 < \omega < 2$ et il suffit pour conclure d'appliquer le théorème 3.6. □

Donnons également un résultat du même type pour la méthode de sur-relaxation de Jacobi.

Théorème 3.12 ***(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 $\omega \in \left]0, \frac{2}{\rho(D^{-1}A)}\right[$.*

DÉMONSTRATION. Puisque la matrice A est hermitienne, on peut utiliser le théorème 3.6 à condition que la matrice hermitienne $\frac{2}{\omega} D - A$ soit définie positive. Ses valeurs propres étant données par $\frac{2}{\omega} d_{ii} - \lambda_i$, où les λ_i sont les valeurs propres de la matrice A, $i = 1, \ldots, n$, ceci implique

$$0 < \omega < \frac{2\, d_{ii}}{\lambda_i},\ i = 1, \ldots, n,$$

d'où le résultat. □

3.5.3 Cas des matrices tridiagonales

On peut comparer les méthodes de Jacobi, de Gauss–Seidel et de sur-relaxation successive dans le cas particulier des matrices tridiagonales.

Théorème 3.13 *Si A est une matrice tridiagonale, alors les rayons spectraux des matrices d'itération des méthodes de Jacobi et 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 démontrer ce résultat, on a besoin d'un lemme technique.

Lemme 3.14 *Pour tout scalaire non nul μ, on définit la matrice tridiagonale $A(\mu)$ d'ordre n par*

$$A(\mu) = \begin{pmatrix} a_1 & \mu^{-1}c_1 & 0 & \dots & 0 \\ \mu\, b_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \mu^{-1}c_{n-1} \\ 0 & \dots & 0 & \mu\, b_n & a_n \end{pmatrix}. \tag{3.10}$$

Le déterminant de cette matrice ne dépend pas de μ. En particulier, on a $\det(A(\mu)) = \det(A(1))$.

Démonstration. Les matrices $A(\mu)$ et $A(1)$ sont semblables, car si l'on introduit la matrice diagonale d'ordre n inversible (μ étant non nul)

$$Q(\mu) = \begin{pmatrix} \mu & & & \\ & \mu^2 & & \\ & & \ddots & \\ & & & \mu^n \end{pmatrix},$$

on a $A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$, d'où le résultat. □

Démonstration du théorème 3.13. 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

$$p_{B_J}(\lambda) = \det(B_J - \lambda\, I_n) = \det(-D^{-1})\det(\lambda\, D - E - F).$$

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

$$p_{B_{GS}}(\lambda) = \det(B_{GS} - \lambda\, I_n) = \det((E-D)^{-1})\det(\lambda\, D - \lambda\, E - F).$$

Compte tenu de la structure tridiagonale de A, la matrice $A(\mu) = \lambda^2 D - \mu\lambda^2 E - \mu^{-1}F$ est bien de la forme (3.10) et l'application du lemme 3.14 avec le choix $\mu = \lambda^{-1}$ montre que

$$\det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda\, E - \lambda\, F) = \lambda^n \det(\lambda\, D - E - F),$$

d'où

$$p_{B_{GS}}(\lambda^2) = \frac{\det(-D)}{\det(E-D)}\,\lambda^n p_J(\lambda) = \lambda^n p_J(\lambda).$$

De cette dernière relation, on déduit que, pour tout λ non nul,

$$\lambda^2 \in \sigma(B_{GS}) \Leftrightarrow \pm\lambda \in \sigma(B_J),$$

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 opposées non nulles de matrice B_J.

Si la matrice tridiagonale est de plus hermitienne définie positive, le théorème 3.11 assure que la méthode de sur-relaxation successive converge pour $0 < \omega < 2$. La méthode de Gauss-Seidel (qui correspond au choix $\omega = 1$ dans cette dernière méthode) est donc elle aussi convergente, ainsi que la méthode de Jacobi en vertu du théorème 3.13. 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 3.15 *Si A est une matrice tridiagonale hermitienne définie positive, alors, la méthode de sur-relaxation successive converge pour $0 < \omega < 2$ et il existe un unique paramètre optimal,*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \rho(B_J)^2}},$$

minimisant le rayon spectral de la matrice d'itération de cette méthode.

Démonstration. La matrice A étant hermitienne définie positive, on sait par le théorème 3.11 que la méthode de sur-relaxation successive est convergente si et seulement si $0 < \omega < 2$. Il nous reste donc à déterminer la valeur du paramètre optimal ω_0.

Pour cela, on commence par définir, pour tout scalaire μ non nul, la matrice

$$A(\mu) = \frac{\lambda^2 + \omega - 1}{\omega} D - \mu\lambda^2 E - \frac{1}{\mu} F.$$

Par une application du lemme 3.14, on obtient 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

$$p_{B_{GS}(\omega)}(\lambda^2) = \det\left(\left(E - \frac{D}{\omega}\right)^{-1}\right) \det\left(\frac{\lambda^2 + \omega - 1}{\omega} D - \lambda^2 E - F\right),$$

il vient

$$p_{B_{GS}(\omega)}(\lambda^2) = \frac{\det\left(E - \frac{D}{\omega}\right)}{\det(-D)} \lambda^n p_{B_J}\left(\frac{\lambda^2 + \omega - 1}{\lambda\omega}\right).$$

On déduit que, pour tout λ non nul,

$$\lambda^2 \in \sigma(B_{GS}(\omega)) \Leftrightarrow \pm\frac{\lambda^2 + \omega - 1}{\lambda\omega} \in \sigma(B_J).$$

Ainsi, pour toute valeur propre α de la matrice B_J, le nombre $-\alpha$ est aussi une valeur propre et les carrés $\eta_\pm(\alpha, \omega)$ des deux racines

$$\lambda_\pm(\alpha, \omega) = \frac{\alpha\omega \pm \sqrt{\alpha^2\omega^2 - 4(\omega - 1)}}{2}$$

de l'équation du second degré en λ

$$\frac{\lambda^2 + \omega - 1}{\lambda\omega} = \alpha,$$

sont des valeurs propres de la matrice $B_{GS}(\omega)$. Par conséquent, on a la caractérisation suivante

$$\rho(B_{GS}(\omega)) = \max_{\alpha \in \sigma(B_J)} \max\left\{|\eta_+(\alpha, \omega)|, |\eta_-(\alpha, \omega)|\right\}.$$

On va maintenant montrer que les valeurs propres de la matrice B_J sont réelles. On a

$$B_J \boldsymbol{v} = \alpha\, \boldsymbol{v} \Leftrightarrow (E + F)\boldsymbol{v} = \alpha D \boldsymbol{v} \Leftrightarrow A\boldsymbol{v} = (1 - \alpha)\boldsymbol{v} \Rightarrow (A\boldsymbol{v}, \boldsymbol{v}) = (1 - \alpha)(D\boldsymbol{v}, \boldsymbol{v})$$

et donc $(1 - \alpha) \in \mathbb{R}_+$, puisque les matrices A et D sont définies positives. Pour déterminer le rayon spectral $\rho(B_{GS}(\omega))$, il suffit donc d'étudier la fonction

$$\begin{array}{rccl} M\ :\ [0,1[\times]0,2[& \mapsto & \mathbb{R} \\ (\alpha, \omega) & \to & \max\left\{|\eta_+(\alpha, \omega)|, |\eta_-(\alpha, \omega)|\right\}, \end{array}$$

puisque $\eta_+(-\alpha,\omega) = \eta_-(\alpha,\omega)$, car $\eta_\pm(\alpha,\omega) = \frac{1}{2}(\alpha^2\omega^2 - 2(\omega-1)) \pm \frac{\alpha\omega}{2}(\alpha^2\omega^2 - 4(\omega-1))^{1/2}$, et $|\alpha| < 1$ d'une part et que la méthode ne peut converger si $\omega \notin]0,2[$ d'autre part. Pour $\alpha = 0$, on vérifie que

$$M(0,\omega) = |\omega - 1|.$$

Pour $0 < \alpha < 1$, le trinôme $\omega \to \alpha^2\omega^2 - 4(\omega-1)$ possède deux racines réelles $\omega_\pm(\alpha)$ vérifiant

$$1 < \omega_+(\alpha) = \frac{2}{1+\sqrt{1-\alpha^2}} < 2 < \omega_-(\alpha) = \frac{2}{1-\sqrt{1-\alpha^2}}.$$

Si $\frac{2}{1+\sqrt{1-\alpha^2}} < \omega < 2$, alors les nombres complexes $\eta_+(\alpha,\omega)$ et $\eta_-(\alpha,\omega)$ sont conjugués et un calcul simple montre que

$$M(\alpha,\omega) = |\eta_+(\alpha,\omega)| = |\eta_-(\alpha,\omega)| = \omega - 1.$$

Si $0 < \omega < \frac{2}{1+\sqrt{1-\alpha^2}}$, alors on voit facilement que

$$M(\alpha,\omega) = \eta_+(\alpha,\omega) = \lambda_+(\alpha,\omega)^2.$$

On a ainsi, pour $0 < \alpha < 1$ et $0 < \omega < \frac{2}{1+\sqrt{1-\alpha^2}}$, on a

$$\frac{\partial M}{\partial \alpha}(\alpha,\omega) = 2\,\lambda_+(\alpha,\omega)\,\frac{\partial \lambda_+}{\partial \alpha}(\alpha,\omega) = \lambda_+(\alpha,\omega)\left(\omega + \frac{\alpha\omega^2}{\sqrt{\alpha^2\omega^2 - 4(\omega-1)}}\right) > 0,$$

et donc, à ω fixé,

$$\max_{\alpha\in\sigma(B_J)} |\eta_+(\alpha,\omega)| = |\eta_+(\rho(B_J),\omega)|.$$

On va enfin pouvoir minimiser le rayon spectral $\rho(B_{GS}(\omega))$ par rapport à ω. Pour $0 < \omega < \frac{2}{1+\sqrt{1-\rho(B_j)^2}}$, il vient

$$\begin{aligned}\frac{\partial}{\partial \alpha}|\eta_+(\rho(B_J),\omega)| = 2\,\lambda_+(\rho(B_J),\omega)\,\frac{\partial \lambda_+}{\partial \omega}(\rho(B_J),\omega) &= \lambda_+(\rho(B_J),\omega)\left(\rho(B_J) + \frac{\rho(B_J)\omega - 2}{2\sqrt{\rho(B_J)^2\omega^2 - 4(\omega-1)}}\right)\\ &= 2\,\lambda_+(\rho(B_J),\omega)\frac{\rho(B_J)\lambda_+(\rho(B_J),\omega) - 1}{\sqrt{\rho(B_J)^2\omega^2 - 4(\omega-1)}}.\end{aligned}$$

Sachant que $0 < \rho(B_J) < 1$, on trouve que le minimum de $|\eta_+(\rho(B_J),\omega)|$ 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 $]0,2[$, le minimum de $\rho(B_{GS}(\omega))$ est atteint en

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(B_j)^2}},$$

et que l'on a alors $\rho(B_{GS}(\omega_0)) = \omega_0 - 1$ (voir la figure 3.1). □

Pour aller plus loin

La généralisation de la relation de récurrence (3.9), par l'introduction d'un paramètre de relaxation ou d'accélération α, conduit à la large classe des *méthodes de Richardson*[4] *stationnaires*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\, M^{-1}\boldsymbol{r}^{(k)},\ k \geq 0. \tag{3.11}$$

Si le paramètre α dépend de l'itération, c'est-à-dire

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)} M^{-1}\boldsymbol{r}^{(k)},\ k \geq 0,$$

4. Lewis Fry Richardson (11 octobre 1881 - 30 septembre 1953) était un mathématicien, météorologiste et psychologue anglais. 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.

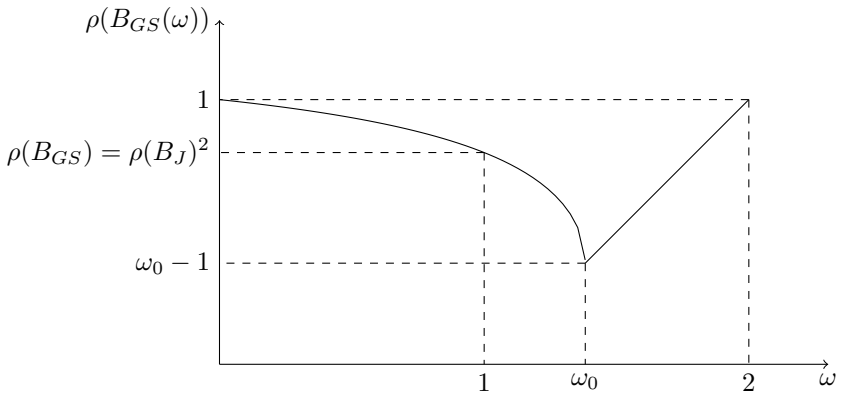

FIGURE 3.1 – Valeur du rayon spectral de la matrice d'itération $B_{GS}(\omega)$ en fonction du paramètre de relaxation ω dans le cas d'une matrice A tridiagonale hermitienne définie positive.

on parle de méthode de Richardson *instationnaire*. Dans ce cadre, les méthodes de Jacobi et de Gauss–Seidel (resp. JOR et SOR) peuvent être vues comme des méthodes de Richardson avec $\alpha = 1$ (resp. $\alpha = \omega$) et respectivement $M = D$ et $M = D - E$. Bien évidemment, de nombreux autres choix ont été proposés pour le *préconditionneur* (la matrice M^{-1}) et le paramètre d'accélération de la méthode. Nous renvoyons à la littérature spécialisée, et notamment au livre de Saad [Saa03], pour plus de détails.

D'un point de vue pratique, les méthodes itératives présentées dans ce chapitre ont été supplantées par la *méthode du gradient conjugué* [HS52] et ses généralisations. Celle-ci fait partie des méthodes dites *à direction de descente*, dont le point de départ est la minimisation de la fonction

$$J(\boldsymbol{x}) = \frac{1}{2}\,\boldsymbol{x}^* A\boldsymbol{x} - \boldsymbol{x}^*\boldsymbol{b},\ \forall \boldsymbol{x} \in \mathbb{C}^n,$$

avec A une matrice d'ordre n hermitienne définie positive et $\boldsymbol{b}$ un vecteur de $\mathbb{C}^n$. Dans ce cas, 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

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{p}^{(k)},\ k \geq 0,$$

où le vecteur $\boldsymbol{p}^{(k)}$ et le scalaire $\alpha^{(k)}$ sont respectivement la *direction de descente* et le *pas 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 et d'un pas de descente indépendant de l'itération conduit à une méthode de Richardson stationnaire (il suffit en effet de choisir $M = I_n$ dans (3.11)) appelée *méthode du gradient à pas fixe*. La *méthode du gradient à pas optimal* est obtenue en déterminant le pas de descente $\alpha^{(k)}$, $k \geq 0$, à chaque étape (c'est une méthode de Richardson instationnaire) de manière à minimiser la norme de l'erreur $\|\boldsymbol{e}^{(k+1)}\|$, avec $\|\cdot\|$ une norme vectorielle adaptée. 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) 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*[5], une propriété fondamentale étant que le vecteur $\boldsymbol{x}^{(k)}$, $k \geq 0$, minimise la fonction J sur l'espace affine $\boldsymbol{x}^{(0)} + \mathcal{K}_k$, avec $\mathcal{K}_k = \mathrm{Vect}\{\boldsymbol{r}^{(0)}, A\boldsymbol{r}^{(0)}, \ldots, A^{k-1}\boldsymbol{r}^{(0)}\}$ 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 A ne permet pas de définir un produit scalaire (hermitien) sur $\mathbb{C}^n$, ce point intervenant de

5. Alexeï Nikolaïevitch Krylov (Алексе́й Никола́евич Крылов, 15 août 1863 - 26 octobre 1945) était un ingénieur naval, mathématicien et mémorialiste russe. Il est célèbre pour ses travaux en mathématiques appliquées, notamment un article paru en 1931, consacré aux problèmes aux valeurs propres et introduisant ce que l'on appelle aujourd'hui les *sous-espaces de Krylov*.

manière critique dans les propriétés de la fonction J. Cependant, le cadre des méthodes de Krylov est propice à la construction de méthodes itératives consistant à minimiser la norme $\|\cdot\|_2$ du résidu. Parmi les nombreuses méthodes existantes, citons la *méthode du gradient biconjugué* (*biconjugate gradient method* (*BiCG*) en anglais) [Fle76], la *méthode orthomin* [Vin76] ou la méthode du résidu minimal généralisée (*generalized minimal residual method* (*GMRES*) en anglais) [SS86].

Références du chapitre

[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 *Numerical analysis - proceedings of the Dundee conference on numerical analysis, 1975*, pages 73–89. Springer, 1976.

[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.

[Saa03] Y. Saad. *Iterative methods for sparse linear systems.* SIAM, second edition, 2003.

[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.

[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.

Chapitre 4

Calcul de valeurs et de vecteurs propres

Nous abordons dans ce chapitre le problème du calcul de valeurs propres (et, éventuellement, de vecteur propres d'une matrice d'ordre n. 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, on pourrait naïvement penser qu'il suffit de factoriser ce dernier pour les obetnir. On sait cependant (par le théorème d'Abel [1]–Ruffini [2]) qu'il n'est pas toujours possible d'exprimer les racines d'un polynôme de degré supérieur ou égal à 5 à partir des coefficients du polynôme et d'opérations élémentaires (addition, soustraction, multiplication, division et extraction de racines). Par conséquant, 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 donc recours à des méthodes itératives.

Parmi ces méthodes, il convient distinguer les méthodes 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 vecteurs propres et dont nous étudierons les propriétés de convergence. Dans le cas de la détermination du spectre d'une matrice réelle symétrique A, nous présentons ensuite une technique de construction d'une suite de matrices, orthogonalement semblables à A, convergeant vers une matrice diagonale dont les coefficients sont les valeurs propres de A, la *méthode de Jacobi*.

4.1 Localisation des valeurs propres

Certaines méthodes de calcul des valeurs propres permettant d'approcher une valeur propre bien spécifique, il peut être utile d'avoir une idée de la localisation des valeurs propres dans le plan complexe. Dans ce domaine, une première estimation est donnée par le théorème A.56, dont on déduit que, pour toute matrice carrée A et pour toute norme matricielle consistante $\|\cdot\|$, on a

$$|\lambda| \leq \|A\|, \ \forall \lambda \in \sigma(A).$$

Cette inégalité, bien que souvent grossière, montre que toutes les valeurs propres de A sont contenues dans un disque de rayon $\|A\|$ et centrée en l'origine du plan complexe. Une autre estimation de localisation des valeurs propres *a priori*, plus précise mais néanmoins très simple, est fournie par le théorème 4.2.

1. 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.

2. Paolo Ruffini (22 septembre 1765 - 10 mai 1822) était un médecin et mathématicien italien. Son nom est lié 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.

Définition 4.1 *(« **disques de Gershgorin**[3] ») Soit A une matrice de $M_n(\mathbb{C})$. Les **disques de Gershgorin** D_i, $i = 1, \ldots, n$, sont les régions du plan complexe définies par*

$$D_i = \{z \in \mathbb{C} \,|\, |z - a_{ii}| \leq R_i\}\,, \text{ avec } R_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|. \tag{4.1}$$

Théorème 4.2 *(« **théorème des disques de Gershgorin** ») Si A est une matrice d'ordre n, alors*

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} D_i,$$

où les D_i sont les disques de Gershgorin définis par (4.1).

Démonstration. Supposons que $\lambda \in \mathbb{C}$ soit une valeur propre de A. Il existe alors un vecteur non nul $\boldsymbol{v}$ de $\mathbb{C}^n$ tel que $A\boldsymbol{v} = \lambda \boldsymbol{v}$, c'est-à-dire

$$\sum_{j=1}^{n} a_{ij} v_j = \lambda\, v_i,\ i = 1, \ldots, n.$$

Soit v_k, avec $k \in \{1, \ldots, n\}$, 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, d'où le résultat. □

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 4.1). La transposée A^T de A possèdant le même spectre que A, on obtient de manière immédiate une première amélioration du résultat.

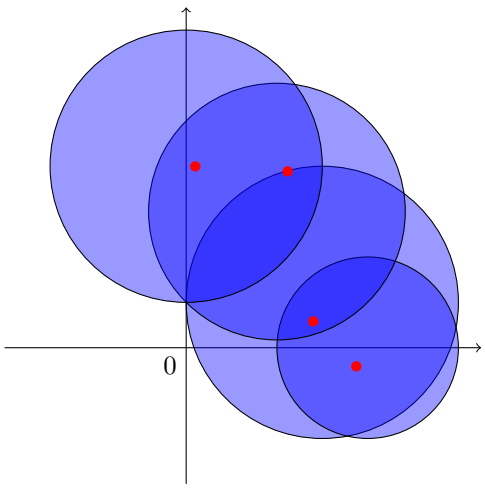

Figure 4.1 – Représentation dans le plan complexe des valeurs propres (en rouge) et des disques de Gershgorin (en bleu) de la matrice complexe .

3. Semyon Aranovitch Gershgorin (Семён Аранович Гершгорин, 24 août 1901 - 30 mai 1933) était un mathématicien biélorusse (soviétique) qui travailla en algèbre et en théorie des fonctions d'une variable complexe. Dans son article *Über die Abgrenzung der Eigenwerte einer Matrix* publié en 1931, il donna des estimations permettant de localiser dans le plan complexe les valeurs propres d'une matrice carrée.

Proposition 4.3 *Si A est une matrice d'ordre n, alors*

$$\sigma(A) \subseteq \left(\bigcup_{i=1}^{n} D_i\right) \cap \left(\bigcup_{j=1}^{n} D'_j\right),$$

où les ensembles D'_j, $j = 1, \ldots, n$, sont tels que

$$D'_j = \{z \in \mathbb{C} \,|\, |z - a_{jj}| \leq C_j\}, \text{ avec } C_j = \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|,$$

et les D_i sont définis par (4.1)

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 composantes connexes.

Théorème 4.4 ***(« second théorème de Gershgorin »)*** *Soit A est une matrice d'ordre n, avec $n \geq 2$. On suppose qu'il existe un entier p compris entre A et $n-1$ tel que l'on puisse diviser la réunion des disques de Gershgorin en deux sous-ensembles disjoints de p et $n-p$ disques. Alors, le premier premier sous-ensemble contient exactement p valeurs propres, chacune étant comptée avec sa multiplicité algébrique, les valeurs propres restantes étant dans le second sous-ensemble.*

Démonstration. wikipedia/traduire We shall use a so-called homotopy (or continuation) argument. For $0 \leq \varepsilon \leq 1$, we consider the matrix $B(\varepsilon) = (b_{ij}(\varepsilon)) \in M_n(\mathbb{C})$, where

$$b_{ij}(\varepsilon) = \Big\{ a_{ii} \text{ if } i = j, \varepsilon\, a_{ij} \text{ if } i \neq j.$$

Then, $B(1) = A$, and $B(0)$ is the diagonal matrix whose diagonal elements coincide with those of A. Each of the eigenvalues of $B(0)$ is therefore the centre of one of the Gerschgorin discs of A; thus exactly p of the eigenvalues of $B(0)$ lie in the union of the discs in $D(p)$. Now, the eigenvalues of $B(\varepsilon)$ are the zeros of its characteristic polynomial, which is a polynomial whose coefficients are continuous functions of ε; hence the zeros of this polynomial are also continuous functions of ε. Thus as ε increases from 0 to 1 the eigenvalues of $B(\varepsilon)$ move along continuous paths in the complex plane, and at the same time the radii of the Gerschgorin discs increase from 0 to the radii of the Gerschgorin discs of A. Since p of the eigenvalues lie in the union of the discs in $D(p)$ when $\varepsilon = 0$, and these discs are disjoint from all of the discs in $D(q)$, these p eigenvalues must still lie in the union of the discs in $D(p)$ when $\varepsilon = 1$, and the theorem is proved. □

remarque sur stabilité et conditionnement

4.2 Méthode de la puissance

La méthode de la puissance fournit une très bonne approximation des valeurs propres extrémales d'une matrice et de vecteurs propres associés. Dans la suite, on note λ_1 et λ_n les valeurs propres d'une matrice A d'ordre ayant ayant respectivement le plus petit et le plus grand module.

4.2.1 Approximation de la valeur propre de plus grand module

Soit A une matrice de $M_n(\mathbb{C}$ diagonalisable et soit V une matrice des vecteurs propres $\boldsymbol{v}_j$, $j = 1, \ldots, n$, normalisés (c'est-à-dire de normes euclidiennes égales à 1) associés. On suppose que les valeurs propres de A sont ordonnées de la manière suivante

$$|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_n|,$$

et supposons que λ_n soit de multiplicité algébrique égale à 1 et que la dernière des inégalités ci-dessus est stricte. Sous ces hypothèses, λ_n est appelée *valeur propre dominante* de A.

Étant donné un vecteur initial arbitraire $\boldsymbol{q}^{(0)}$ de $\mathbb{C}^n$ normalisé, on considère pour $k = 1, 2, \ldots$ la méthode itérative suivante

$$\begin{aligned}\boldsymbol{z}^{(k)} &= A\boldsymbol{q}^{(k-1)},\\ \boldsymbol{q}^{(k)} &= \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|_2},\\ \nu^{(k)} &= (\boldsymbol{q}^{(k)})^* A\boldsymbol{q}^{(k)},\end{aligned}$$

appelée méthode de la puissance.

Analysons ses propriétés de convergence. Par récurrence sur k, on peut vérifier que

$$\boldsymbol{q}^{(k)} = \frac{A^k\boldsymbol{q}^{(0)}}{\|A^k\boldsymbol{q}^{(0)}\|_2},\ k \geq 1.$$

Cette relation rend explicite le rôle joué par les puissance de la matrice A. Ayant supposé cette dernière diagonalisable, il existe une base de vecteurs propres de A dans laquelle on peut décomposer le vecteur $\boldsymbol{q}^{(0)}$:

$$\boldsymbol{q}^{(0)} = \sum_{i=1}^{n} \alpha_i\, \boldsymbol{q}_i.$$

Comme $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, on a

$$A^k\boldsymbol{q}^{(0)} = \alpha_n\lambda_n^k \left(\boldsymbol{v}_n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n}\left(\frac{\lambda_i}{\lambda_n}\right)^k \boldsymbol{v}_i\right),\ k \geq 1.$$

Puisque $|\frac{\lambda_i}{\lambda_n}| < 1$, la composante le long de $\boldsymbol{v}_1$ du vecteur $A\boldsymbol{q}^{(0)}$ (et donc celle de $\boldsymbol{q}^{(k)}$) augmente quand k augmente en module, tandis que les composantes suivant les autres directions diminuent. On obtient alors, en utilisant les deux dernières relations,

$$\boldsymbol{q}^{(k)} = \frac{\alpha_n\lambda_n^k(\boldsymbol{v}_n + \boldsymbol{y}^{(k)}}{\|\alpha_n\lambda_n^k(\boldsymbol{v}_n + \boldsymbol{y}^{(k)}\|_2},$$

où $\boldsymbol{y}^{(k)}$ désigne un vecteur tendant vers 0 quand k tend vers l'infini. Le vecteur $\boldsymbol{q}^{(k)}$ s'aligne donc avec un vecteur propre associé à la valeur propre dominante quand k tend vers l'infini. On a de plus l'estimation d'erreur suivante à l'étape k.

Théorème 4.5 *Soit A une matrice diagonalisable d'ordre n dont les valeurs propres satisfont*

$$|\lambda_1| \leq |\lambda_2| \leq \cdots < |\lambda_n|.$$

En supposant $\alpha_n \neq 0$ dans ..., il existe une constante $C > 0$ telle que

$$\|\widetilde{\boldsymbol{q}}^{(k)} - \boldsymbol{v}_n\|_2 \leq C\,|\frac{\lambda_{n-1}}{\lambda_n}|^k,\ k \geq 1,$$

où $\widetilde{\boldsymbol{q}}^{(k)} = \boldsymbol{v}_n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n}\left(\frac{\lambda_i}{\lambda_n}\right)^k \boldsymbol{v}_i$.

Démonstration. quarteroni 173 □

Dans le cas réel symetrique, la convergence est plus rapide (Allaire 217, GVL 406-407)
Remarques sur test d'arret
deflation

4.2.2 Approximation de la valeur propre de plus petit module : la méthode de la puissance inverse

shift

4.3 Méthode de Jacobi pour les matrices symétriques

La méthode de Jacobi se sert de la structure particulière des matrices symétriques pour construire une suite de matrice convergeant vers la *forme de Schur* (voir le théoème A.29) diagonale, orthogonalement semblable, de la matrice symétrique. Elle utilise pour cela les *matrices de Givens*.

4.3.1 Matrices de rotation de Givens

Les matrices de rotation de Givens sont des matrices orthogonales qui permettent, tout comme les matrices de Householder présentées dans la section 2.4.3, 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 comme

$$
\begin{aligned}
G(p,q,\theta) &= \begin{pmatrix}
1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & 1 & & & & & & & & & \vdots \\
\vdots & & \ddots & & & & & & & & \vdots \\
\vdots & & & \cos(\theta) & & & & \sin(\theta) & & & \vdots \\
\vdots & & & & 1 & & & & & & \vdots \\
\vdots & & & & & \ddots & & & & & \vdots \\
\vdots & & & & & & 1 & & & & \vdots \\
\vdots & & & -\sin(\theta) & & & & \cos(\theta) & & & \vdots \\
\vdots & & & & & & & & 1 & & \vdots \\
\vdots & & & & & & & & & \ddots & 0 \\
0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1
\end{pmatrix} \\
&= I_n - (1-\cos(\theta))(E_{pp}+E_{qq}) + \sin(\theta)(E_{pq}-E_{qp}).
\end{aligned}
$$

Cette matrice repésente la rotation d'angle θ (dans le sens trigonométrique) dans le plan des $p^{\text{ième}}$ et $q^{\text{ième}}$ vecteur de la base canonique de $\mathbb{R}^n$.

completer Ciarlet 114-113

4.3.2 Méthode de Jacobi « classique »

quarteroni 198-199

4.3.3 Méthode de Jacobi cyclique

matrices symétriques tridiagonales, suites de Sturm GL 8.4

Pour aller plus loin

expliquer le principe de la méthode QR pour le calcul de toutes les valeurs propres d'une matrice carrée quelconque

Deuxième partie

Traitement numérique des fonctions

Chapitre 5

Résolution des équations non linéaires

Nous nous intéressons dans ce chapitre à l'approximation des zéros (ou racines dans le cas d'un polynôme) d'une fonction réelle d'une variable réelle, c'est-à-dire, étant donnés un intervalle $I \subseteq \mathbb{R}$ et une application f de I dans $\mathbb{R}$, la résolution approchée du problème : *trouver* $\xi \in \mathbb{R}$ *(ou plus généralement* $\mathbb{C}$*) tel que*

$$f(\xi) = 0.$$

Ce problème intervient notamment dans l'étude générale de fonctions d'une variable réelle, qu'elle soit motivée ou non par des applications [1], pour lesquelles des solutions exactes de ce type d'équation ne sont pas 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 l'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 la donnée 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 des méthodes itératives, en introduisant notamment la notion d'ordre de convergence, nous présentons plusieurs méthodes parmi les plus connues et les plus utilisées : tout d'abord des méthodes dites *d'encadrement* comme les méthodes de dichotomie et de la fausse position, puis les méthodes de la corde, de Newton [3]–Raphson [4], qui sont toutes deux des

1. Essayons néanmoins de donner deux exemples, l'un issu de la physique, l'autre de l'économie.
Supposons tout d'abord que l'on cherche à déterminer le volume V occupé par n molécules d'un gaz de van der Waals de température T et de pression p. L'équation d'état (c'est-à-dire l'équation liant les variables d'état que sont n, p, T et V) d'un tel gaz s'écrit

$$\left(p + a\left(\frac{n}{V}\right)^2\right)(V - nb) = nk_BT,$$

où les coefficients a (pression de cohésion) et b (covolume) dépendent de la nature du gaz considéré et k_B désigne la constante de Boltzmann. On est donc amené à résoudre une équation non linéaire d'inconnue V et de fonction $f(V) = \left(p + a\left(\frac{n}{V}\right)^2\right)(V - nb) - nk_BT$.
Admettons maintenant 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. Le 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 $f(R) = M - V\frac{1+R}{R}\left((1+R)^n - 1\right) = 0$.

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. 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 la création du calcul infinitésimal.

4. Joseph Raphson (v. 1648 - v. 1715) était un mathématicien anglais. Son travail le plus notable est son ouvrage *Analysis*

méthodes de point fixe, 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 brièvement abordées en fin de chapitre.

5.1 Généralités

5.1.1 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 un des critères discriminants lors du choix d'une méthode), il nous faut introduire quelques définitions.

Définition 5.1 ***(ordre d'une suite convergente)*** *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels convergeant vers une limite ξ. On dit que cette suite* ***convergente d'ordre*** *$r \geq 1$, s'il existe deux constantes $0 < C_1 \leq C_2 < +\infty$ telles que*

$$C_1 \leq \frac{|x^{(k+1)} - \xi|}{|x^{(k)} - \xi|^r} \leq C_2, \ \forall k \geq k_0, \tag{5.1}$$

où k_0 appartient à $\mathbb{N}$.

Par extension, une méthode itérative produisant une suite convergente vérifiant les relations (5.1) sera également dite *d'ordre r*. On notera que, dans plusieurs ouvrages, on trouve l'ordre d'une suite défini uniquement par le fait qu'il existe une constante $C \geq 0$ telle que, pour tout $k \geq k_0 \geq 0$, $|x^{(k+1)} - \xi| \leq C|x^{(k)} - \xi|^r$. Il faut cependant observer 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 (5.1), faute de quoi la suite ne pourrait converger.

La définition 5.1 est très générale et n'exige pas que la suite $\left(\frac{|x^{(k+1)}-\xi|}{|x^{(k)}-\xi|^r}\right)_{k\in\mathbb{N}}$ admette une limite quand k tend vers l'infini. Lorsque c'est le cas, on a coutume de se servir de la définition suivante.

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*

$$\lim_{k\to+\infty} \frac{|x^{(k+1)} - \xi|}{|x^{(k)} - \xi|^r} = \mu. \tag{5.2}$$

Dans le cas particulier où $r = 1$, on dit que la suite ***converge linéairement*** *si*

$$\lim_{k\to+\infty} \frac{|x^{(k+1)} - \xi|}{|x^{(k)} - \xi|} = \mu, \ \text{ avec } \mu \in]0, 1[,$$

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 2 est dite *quadratique*, celle d'ordre 3 *cubique*.
Si cette 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 permettre de fournir l'ordre d'une suite dont la « vitesse de convergence » est variable, ce qui se traduit par le fait que la limite (5.2) 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*** *r* ***au moins*** *vers une limite ξ s'il existe une suite $(\varepsilon^{(k)})_{k\in\mathbb{N}}$ vérifiant*

$$|x^{(k)} - \xi| \leq \varepsilon^{(k)}, \ \forall k \in \mathbb{N}, \tag{5.3}$$

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.

et un réel $\nu > 0$ *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 (5.3), mais ceci est souvent impossible à obtenir en pratique.

Finissons en indiquant que les notions d'ordre et de constante asymptotique d'erreur ne sont pas purement théoriques et sont en relation avec le nombre de chiffres exacts obtenus dans l'approximation de ξ. Posons en effet $\delta^{(k)} = -\log_{10}(|x^{(k)} - \xi|)$; $\delta^{(k)}$ est alors le nombre de chiffres significatifs décimaux exacts de $x^{(k)}$. Pour k suffisament 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.

5.1.2 Critères d'arrêt

En cas de convergence, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode itérative tend vers le zéro ξ quand k tend vers l'infini. Pour l'utilisation pratique d'une telle méthode, il faut introduire un *critère d'arrêt* (comme cétait déjà le cas pour les méthodes itératives de résolution de systèmes linéaires au chapitre 3) pour 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 (imposer un nombre maximum d'itérations constituant une troisième possibilié) : l'un basé sur l'incrément et l'autre sur le résidu.

Soit $\varepsilon > 0$ la tolérance fixée pour le calcul approché de ξ. Dans le cas d'un *contrôle de l'incrément*, les itérations s'achèvent dès que

$$|x^{(k+1)} - x^{(k)}| < \varepsilon. \tag{5.4}$$

Si l'on choisit de *contrôler le résidu*, on met fin aux itérations dès que

$$|f(x^{(k)})| < \varepsilon. \tag{5.5}$$

Selon les cas, chacun de ces critères peut s'avérer soit trop restrictif, soit trop optimiste.

COMPLETER par conditionnement de l'equation et analyse d'erreur + dessins (quarteroni p 212 + 336-7)

5.2 Méthodes d'encadrement

Cette première classe de méthodes repose sur la propriété fondamentale suivante, relative à l'existence de zéros d'une appliquation d'une variable réelle à valeurs réelles.

Théorème 5.4 ***(existence d'un zéro d'une fonction continue)*** *Soit un intervalle non vide* $[a, b]$ *de* $\mathbb{R}$ *et* f *une application continue de* $[a, b]$ *dans* $\mathbb{R}$ *vérifiant* $f(a)f(b) < 0$. *Alors il existe* $\xi \in]a, b[$ *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 théorème B.1 en annexe). □

5.2.1 Méthode de dichotomie

La *méthode de dichotomie* (ou *méthode de la bissection*) suppose que la fonction f est continue un intervalle $[a, b]$, n'admet qu'un seul zéro $\xi \in]a, b[$ et vérifie $f(a)f(b) < 0$.

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 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 $f(x^{(k)})f(b^{(k)}) < 0$.

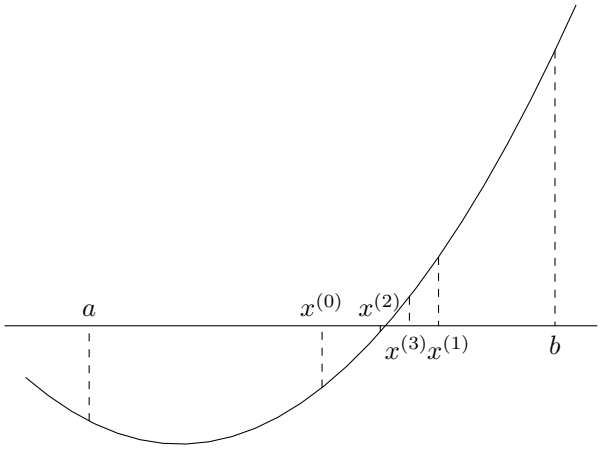

FIGURE 5.1 – Construction des premiers itérés de la méthode de dichotomie.

La figure 5.1 illustre la mise en œuvre de la méthode. Concernant la convergence de cette méthode, on a le résultat suivant, dont la preuve est laissée en exercice.

Proposition 5.5 *Soit f une fonction continue sur un intervalle $[a, b]$, vérifiant $f(a)f(b) < 0$, et soit $\xi \in]a, b[$ l'unique solution de l'équation $f(x) = 0$. 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*

$$|x^{(k)} - \xi| \leq \frac{b-a}{2^{k+1}}, \ \forall k \in \mathbb{N}. \tag{5.6}$$

Il ressort de cette proposition que la méthode de dichotomie converge de manière certaine : c'est une méthode *globalement convergente*. L'estimation d'erreur (5.6) 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

$$\frac{b-a}{2^{k+1}} \leq \varepsilon \Leftrightarrow k \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)} - 1. \tag{5.7}$$

Ainsi, pour améliorer la précision de l'approximation du zéro d'un ordre de grandeur, c'est-à-dire trouver $k > j$ tel que $|x^{(k)} - \xi| = \frac{1}{10}|x^{(j)} - \xi|$, il faut effectuer $k - j = \frac{\ln(10)}{\ln(2)} \simeq 3.32$ itérations. La convergence de cet algorithme est donc *lente*. Enfin, la méthode de dichotomie ne garantit pas une réduction monotone

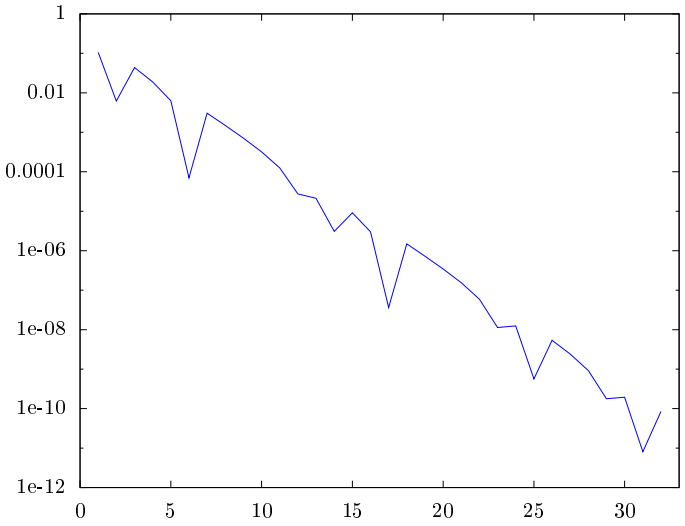

FIGURE 5.2 – Historique de la convergence, c'est-à-dire le tracé de l'erreur $|x^{(k)} - \xi|$ en fonction k, de la méthode de dichotomie pour l'approximation de la racine $\xi = 0,9061798459...$ du polynôme de Legendre[5] de degré 5, $P_5(x) = \frac{x}{8}\,(63x^4 - 70\,x^2 + 15)$, dont les racines se 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 $30,89735...$ de l'estimation (5.7). On observe que l'erreur a un comportement oscillant, mais diminue néanmoins en moyenne.

de l'erreur absolue d'une itération à l'autre, comme on le constate sur la figure 5.2. Ce n'est donc pas une méthode d'ordre un au sens de la définition 5.1.

On gardera donc à l'esprit que la méthode de dichotomie est une méthode robuste permettant d'obtenir une approximation raisonnable du zéro ξ pouvant servir à l'initialisation d'une méthode dont la convergence est plus rapide mais seulement *locale*, comme la méthode de Newton–Raphson (voir la section 5.3.4).

Exemple. On utilise la méthode de dichotomie pour approcher la racine du polynôme $f(x) = x^3 + 2x^2 - 3x - 1$ contenue dans l'intervalle $[1,2]$ (cette fonction est en effet continue et on a $f(1) = -1$ et $f(2) = 9$), avec une précision égale à 10^{-4}. Le tableau suivant 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.

5. 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.

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

5.2.2 Méthode de la fausse position

La *méthode de la fausse position*, dite encore méthode *regula falsi*, est une méthode d'encadrement combinant les possibilités de la méthode de dichotomie avec celles de la méthode de la sécante, que nous introduisons dans la section 5.3.5. L'idée est d'utiliser l'information fournie par les valeurs de la fonction f aux extrémités de l'intervalle d'encadrement pour remédier à la lente vitesse de convergence de la méthode de dichotomie (cette dernière ne tenant compte que du signe de la fonction). Sous des hypothèses raisonnables de régularité sur f, on peut en effet montrer que la convergence de cette méthode est *linéaire*.

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

$$x^{(k)} = a^{(k)} - \frac{a^{(k)} - b^{(k)}}{f(a^{(k)}) - f(b^{(k)})} f(a^{(k)}) = b^{(k)} - \frac{b^{(k)} - a^{(k)}}{f(b^{(k)}) - f(a^{(k)})} f(b^{(k)}) = \frac{f(a^{(k)})b^{(k)} - f(b^{(k)})a^{(k)}}{f(a^{(k)}) - f(b^{(k)})}. \quad (5.8)$$

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 $x^{(k)}$ ainsi 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. Par ailleurs, si f est une fonction linéaire, on voit que le zéro est obtenu après une itération plutôt qu'une infinité.

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, à la différence de la méthode de dichotomie, vers zéro, comme l'illustre l'exemple ci-dessous.

Exemple. On reprend l'exemple précédent en utilisant cette fois la méthode de la fausse position. Le tableau présente ...

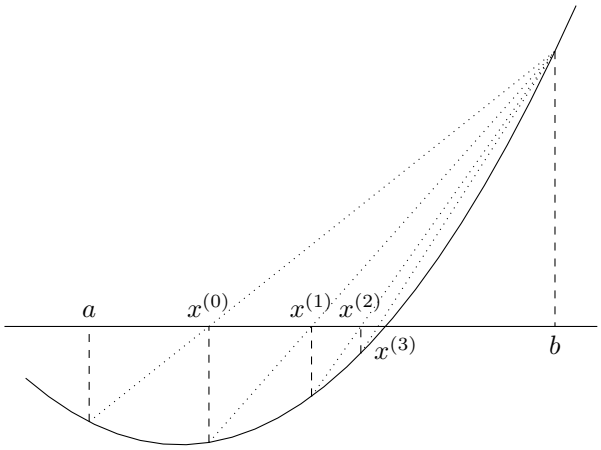

FIGURE 5.3 – Construction des premiers itérés de la méthode de la fausse position.

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

On observe que la borne de droite de l'intervalle d'encadrement initial est conservée tout au long du calcul.

De fait, compte tenu des hypothèses sur f, on peut voir que la méthode conduit inévitablement à partir d'un certain rang à l'une des configurations présentées à la Figure 5.4, pour chacune desquelles l'une des deux bornes de l'intervalle d'encadrement n'est plus jamais modifiée tandis que l'autre converge de manière monotone vers le zéro de la fonction. La méthode se comporte alors comme une *méthode de point fixe* (comparer à ce titre (5.9) avec (5.10)).

Sous des hypothèses de régularité légèrement restrictives [6] sur f, on peut établir le résultat de convergence suivant pour la méthode de la fausse position.

Théorème 5.6 *Soit f une fonction de classe $\mathscr{C}^2$ sur un intervalle $[a,b]$, vérifiant $f(a)f(b)<0$, et soit $\xi \in]a,b[$ l'unique solution de l'équation $f(x)=0$. Alors, la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode de la fausse position converge linéairement vers ξ.*

DÉMONSTRATION. Si f est une fonction affine, la méthode converge en une étape. Sinon, l'une des configurations illustrées à la figure 5.4 est obligatoirement atteinte par la méthode 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.
Supposons à présent que $f'(x)>0$ (f croissante) et $f''(x)>0$ (f convexe) sur l'intervalle $[a,b]$ (c'est la première configuration décrite plus haut). On remplace alors à l'étape $k+1$, $k\geq 0$, l'intervalle $[x^{(k)},b]$ par $[x^{(k+1)},b]$, où la borne $x^{(k+1)}$ est donnée (en choisissant, de manière un peu abusive, $x^{(0)}$ égal à a) par la formule

$$x^{(k+1)} = x^{(k)} - \frac{b-x^{(k)}}{f(b)-f(x^{(k)})} f(x^{(k)}),\ k\geq 0. \tag{5.9}$$

6. L'hypothèse « *f dérivable* » est en effet suffisante pour établir le résultat, mais demanderait une preuve plus élaborée.

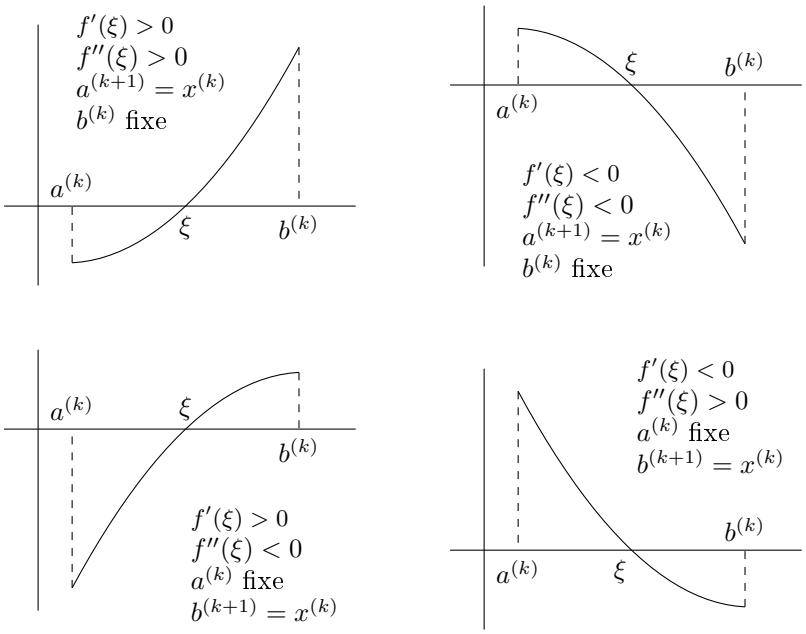

FIGURE 5.4 – Différentes configurations atteintes par la méthode de la fausse position à partir d'un certain rang.

Étudions à présent la suite $(x^{(k)})_{k\in\mathbb{N}}$ en posant $g(x) = x - \frac{b-x}{f(b)-f(x)}f(x)$, d'où $x^{(k+1)} = g(x^{(k)})$, $\forall k \in \mathbb{N}$. La fonction g est de manière évidente de classe $\mathscr{C}^1$ sur $[a,b[$ et continue en b, avec $g(b) = b - \frac{f(b)}{f'(b)}$. On a par ailleurs

$$g'(x) = 1 - \frac{b-x}{f(b)-f(x)}f'(x) + \frac{f(b)-f(x)-(b-x)f'(x)}{(f(b)-f(x))^2}f(x) = \frac{f(b)-f(x)-(b-x)f'(x)}{(f(b)-f(x))^2}f(b),\ \forall x \in [a,b[,$$

dont on déduit la continuité de g' en b, avec $g'(b) = \frac{f(b)f''(b)}{2f'(b)^2}$. L'application g est donc de classe $\mathscr{C}^1$ sur $[a,b]$.
La fonction f étant supposée convexe sur $[a,b]$, on a $f(b)-f(x)-(b-x)f'(x) \geq 0$, $\forall x \in [a,b]$, ainsi que $f(b) > 0$, puisque f est croissante et $f(a)f(b) < 0$. Par conséquent, g est croissante sur $[a,b]$ et alors $g([a,b]) \subset [g(a), g(b)]$. Enfin, on utilise la croissance de f et le fait que $f(a) < 0$ et $f(b) > 0$ pour montrer que $g(a) = a - \frac{b-a}{f(b)-f(a)}f(a) \geq a$ et $g(b) = b - \frac{f(b)}{f'(b)} \leq b$.
La suite $(x^{(k)})_{k\in\mathbb{N}}$ est donc croissante et majorée par b; elle converge vers une limite $\ell \in [a,b]$, qui vérifie, par continuité de g, $g(\ell) = \ell$. Puisque $x^{(0)} = a < \xi$, on a, par récurrence, $x^{(k)} \leq \xi$, $\forall k \in \mathbb{N}$, et donc $\ell \leq \xi$, d'où $\ell \in]a,b[$ et, par suite, $f(\ell) = 0$ donc $\ell = \xi$, par unicité de ξ.
Pour prouver que la convergence est linéaire, on doit montrer que

$$0 < \lim_{k\to+\infty} \frac{x^{(k+1)} - \xi}{x^{(k)} - \xi} < 1.$$

Or, le théorème des accroissements finis (voir théorème B.3 en annexe) et la continuité de la fonction g' impliquent que

$$\lim_{k\to+\infty} \frac{x^{(k+1)} - \xi}{x^{(k)} - \xi} = g'(\xi) = 1 - \frac{b-\xi}{f(b)-f(\xi)}f'(\xi),$$

et, f étant strictement convexe, il est alors facile de voir que la pente de la droite passant par les points $(\xi, 0)$ et $(b, f(b))$ est strictement plus grande que celle de la tangente à la courbe représentative de f au point ξ, d'où la conclusion.

La même technique de démonstration s'adapte pour traiter les trois cas (pour lesquels les signes de $f'(x)$ et $f''(x)$ sont constants sur $[a,b]$) restants, ce qui achève la preuve. □

On notera que le critère d'arrêt des itérations de la méthode doit nécessairement être basé sur la valeur du résidu $f(x^{(k)})$, puisque la longueur de l'intervalle d'encadrement du zéro de f ne tend pas nécessairement vers zéro.

5.3 Méthodes de point fixe

Les méthodes d'approximation de zéros introduites dans la suite 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 que la suite $(x^{(k)})_{k\in\mathbb{N}}$ produite par l'un des algorithmes présentés prend ses valeurs dans un intervalle fixé *a priori*.

REPRENDRE!!!

D'autre part, comme nous l'avons déjà vu avec la méthode de la fausse position, prendre en compte les informations données par les valeurs de la fonction f et même, dans le cas où celle-ci est différentiable, celles de sa dérivée aux points $x^{(k)}$, $k \in \mathbb{N}$, peut conduire à des propriétés de convergence améliorées. On verra que les méthodes présentées exploitent ce « principe » sous différentes formes.

Les sections 5.3.3 et 5.3.4 sont respectivement consacrées aux méthodes de la corde et de Newton–Raphson, qui sont ensuite analysées à la section 5.3 dans le cadre général des méthodes de point fixe. La méthode de la sécante est introduite et analysée dans la section 5.3.5.

5.3.1 Principe

La famille de méthodes que nous allons maintenant introduire utilise le fait que le problème $f(x) = 0$ peut toujours ramener au problème équivalent $x - g(x) = 0$, pour lequel on a le résultat suivant.

Théorème 5.7 ***(« théorème du point fixe de Brouwer**[7] **»)*** *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 $f(x) = x - g(x)$. On a alors $f(a) = a - g(a) \leq 0$ et $f(b) = b - g(b) \geq 0$, puisque $g(x) \in [a, b]$ pour tout $x \in [a, b]$. Par conséquent, f est une fonction continue sur $[a, b]$, 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 $0 = f(\xi) = \xi - g(\xi)$. □

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 cela 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. Réécrivons à présent le problème en posant $g(x) = \ln(2x + 1)$. Cette dernière fonction est continue et croissante sur l'intervalle $[1, 2]$, à valeurs dans lui-même. Elle satisfait donc les conditions 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 donné $x^{(0)}$ (appartenant à $[a, b]$), on pose

$$x^{(k+1)} = g(x^{(k)}),\ k \geq 0. \tag{5.10}$$

On dit que la relation (5.10) est une *itération de point fixe*. 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 que

$$\xi = \lim_{k\to+\infty} x^{(k+1)} = \lim_{k\to+\infty} g(x^{(k)}) = g\left(\lim_{k\to+\infty} x^{(k)}\right) = g(\xi),$$

la deuxième égalité provenant de la définition (5.10) de la suite récurrente et la troisième étant une conséquence de la continuité de g.

7. Luitzen Egbertus Jan Brouwer (27 février 1881 - 2 décembre 1966) était mathématicien et philosophe néerlandais. Ses apports concernèrent principalement la topologie et la logique formelle.

5.3.2 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 de convergence vers un zéro de la fonction f. 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*

$$|g(x) - g(y)| \leq K|x - y|, \ \forall x \in [a,b], \ \forall y \in [a,b]. \tag{5.11}$$

On notera que la *constante de Lipschitz*[8] de 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*[9] (également attribué à Picard [10]), dont l'énoncé général vaut pour toute application contractante définie sur un *espace métrique complet*.

Théorème 5.9 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application contractante de $[a,b]$ dans lui-même. 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 deux estimations suivantes :*

$$|x^{(k)} - \xi| \leq K^k |x^{(0)} - \xi|, \ \forall k \geq 0, \tag{5.12}$$

$$|x^{(k)} - \xi| \leq \frac{K}{1-K} |x^{(k)} - x^{(k-1)}|, \ \forall k \geq 1. \tag{5.13}$$

Démonstration. On commence par montrer que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est une suite de Cauchy. En effet, pour tout entier k non nul, on a

$$|x^{(k+1)} - x^{(k)}| = |g(x^{(k)}) - g(x^{(k-1)})| \leq K|x^{(k)} - x^{(k-1)}|,$$

par hypothèse, et on obtient par récurrence que

$$|x^{(k+1)} - x^{(k)}| \leq K^k |x^{(1)} - x^{(0)}|, \ \forall k \in \mathbb{N}.$$

On en déduit, par une application répétée de l'inégalité triangulaire, que, $\forall k \in \mathbb{N}$, $\forall p > 2$,

$$\begin{aligned} |x^{(k+p)} - x^{(k)}| &\leq |x^{(k+p)} - x^{(k+p-1)}| + |x^{(k+p-1)} - x^{(k+p-2)}| + \cdots + |x^{(k+1)} - x^{(k)}| \\ &\leq \left(K^{p-1} + K^{p-2} + \cdots + 1\right) |x^{(k+1)} - x^{(k)}| \\ &\leq \frac{1 - K^p}{1 - K} K^k |x^{(1)} - x^{(0)}|, \end{aligned}$$

le dernier membre tendant vers zéro lorsque 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 [11], 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 récurrence sur k en écrivant que

$$|x^{(k)} - \xi| = |g(x^{(k-1)}) - g(\xi)| \leq |x^{(k-1)} - \xi|, \ \forall k \geq 1,$$

8. 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.

9. 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.

10. 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.

11. C'est par hypothèse une application K-lipschitzienne.

et la seconde est obtenue en utilisant que

$$|x^{(k+p)} - x^{(k)}| \leq \frac{1-K^p}{1-K}|x^{(k+1)} - x^{(k)}| \leq \frac{1-K^p}{1-K}\, K|x^{(k)} - x^{(k-1)}|,\ \forall k \geq 1,\ \forall p \geq 1,$$

et en faisant tendre 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 intiale $x^{(0)}$ dans l'intervalle $[a, b]$: c'est donc un nouvel exemple de convergence *globale*. Par ailleurs, l'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 est aussi particulièrement utile d'un point de vue applicatif, car elle fournit à chaque étape un majorant de la distance à la limite (sans pour autant la connaître) en fonction d'une quantité connue. Il est 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*

$$|x^{(k)} - \xi| \leq \varepsilon,\ \forall k \geq k_0(\varepsilon).$$

On a alors la majoration

$$k_0(\varepsilon) \leq \left\lfloor \frac{\ln(\varepsilon) + \ln(1-K) - \ln(|x^{(1)} - x^{(0)}|)}{\ln(K)} \right\rfloor + 1,$$

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

$$|x^{(0)} - \xi| \leq |x^{(0)} - x^{(1)}| + |x^{(1)} - \xi| \leq |x^{(0)} - x^{(1)}| + K|x^{(0)} - \xi|,$$

d'où

$$|x^{(0)} - \xi| \leq \frac{K}{1-K}|x^{(0)} - x^{(1)}|.$$

En substituant cette expression dans (5.13), on obtient que

$$|x^{(k)} - \xi| \leq \frac{K^k}{1-K}|x^{(0)} - x^{(1)}|,$$

et on aura en particulier $|x^{(k)} - \xi| \leq \varepsilon$ si k est tel que

$$\frac{K^k}{1-K}|x^{(0)} - x^{(1)}| \leq \varepsilon.$$

En prenant le logarithme népérien de chacun des membres de cette dernière inégalité, on arrive à

$$k \geq \frac{\ln(\varepsilon) + \ln(1-K) - \ln(|x^{(1)} - x^{(0)}|)}{\ln(K)},$$

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 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*

$$|g'(x)| \leq K < 1,\ \forall x \in [a, b].$$

Alors, g est une application contractante sur $[a, b]$.

DÉMONSTRATION. D'après le théorème des accroissements finis (voir théorème B.3 en annexe), 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

$$|g(x) - g(y)| = |g'(c)||x - y|,$$

d'où le résultat. □

On est alors en mesure 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.3 en annexe), il existe, pour tout $k \geq 0$, 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. La quantité $|g'(\xi)|$ est appelée, par comparaison avec la constante C apparaissant dans (5.1), *facteur de convergence asymptotique* de la méthode.

Encore une fois, il est souvent difficile en pratique de déterminer *a priori* un intervalle $[a,b]$ sur lequel les hypothèses de la proposition 5.11. Il est néanmoins possible de se contenter d'hypothèses plus faibles, au prix d'un résultat de convergence seulement *locale*.

Théorème 5.13 *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 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 suffisament proche de ξ.*

DÉMONSTRATION. Par hypothèses sur la fonction g, il existe un réel $h >$ tel que 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 $0 < \delta \leq h$, tel que $|g'(x)| \leq L$, avec $L < 1$, pour tout x appartenant à I_δ. Pour cela, il suffit de poser $L = \frac{1}{2}\left(1 + |g'(\xi)|\right)$ et d'utiliser la continuité de g' pour choisir $\delta \leq h$ de manière à ce que

$$|g'(x) - g'(\xi)| \leq \frac{1}{2}\left(1 - |g'(\xi)|\right), \ \forall x \in I_\delta.$$

On en déduit alors que

$$|g'(x)| \leq |g'(x) - g'(\xi)| + |g'(\xi)| \leq \frac{1}{2}\left(1 - |g'(\xi)|\right) + |g'(\xi)| = L, \ \forall x \in I_\delta.$$

Supposons à présent que, pour un entier k donné, le terme $x^{(k)}$ de la suite définie par la relation de récurrence (5.10) appartient à I_δ. On a alors, en vertu du théorème des accroissements finis (voir théorème B.3 en annexe),

$$x^{(k+1)} - \xi = g(x^{(k)}) - \xi = g(x^{(k)}) - g(\xi) = g'(\eta^{(k)})(x^{(k)} - \xi),$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ, d'où

$$|x^{(k+1)} - \xi| \leq L|x^{(k)} - \xi|,$$

et $x^{(k+1)}$ appartient donc lui aussi à I_δ. On montre alors par récurrence que, si $x^{(0)}$ appartient à I_δ, alors $x^{(k)}$ également, $\forall k \geq 0$, et que

$$|x^{(k)} - \xi| \leq L^k |x^{(0)} - \xi|,$$

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 $|g'(x^{(k)})| > 1$, on obtient $|x^{(k+1)} - \xi| > |x^{(k)} - \xi|$ 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 théorème 5.13 conduisent à l'introduction des définitions 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 $x^{(0)}$ suffisament 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**.*

conditions suffisantes + exemples
methodes d'ordre superieur gautschi 235, quarteroni 225

5.3.3 Méthode de relaxation ou de la corde

Nous avons vu dans la section 5.3.1 que l'on pouvait obtenir de diverses manières une fonction g dont les points fixes sont les zéros de la fonction f. Beaucoup de méthodes parmi les plus courantes s'appuient néanmoins sur le choix de la forme suivante

$$g(x) = x + h(x)f(x), \tag{5.15}$$

avec h une fonction satisfaisant $0 < |h(x)| < +\infty$ sur le domaine de définition (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 h est alors celui conduisant à la *méthode de relaxation*, qui consiste en la construction d'une suite $(x^{(k)})_{k\in\mathbb{N}}$ satisfaisant la relation de récurrence

$$x^{(k+1)} = x^{(k)} - \lambda\, f(x^{(k)}),\ \forall k \geq 0, \tag{5.16}$$

avec λ un réel fixé, la valeur de $x^{(0)}$ étant donnée.

En supposant f différentiable dans un voisinage de son zéro ξ, rien ne garantit que la méthode converge si $f'(\xi) = 0$ mais on voit qu'on peut facilement assurer la convergence locale de cette méthode si ξ est un zéro simple et λ est tel que $0 < \lambda f'(\xi) < 2$. Ceci est rigoureusement établi dans le théorème suivant.

Théorème 5.15 *Soit f une fonction réelle de classe $\mathscr{C}^1$ dans un voisinage du réel ξ tel que $f(\xi) = 0$. Supposons que $f'(\xi) \neq 0$. Alors il existe un ensemble de réels λ 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 suffisament proche de ξ.*

Démonstration. Supposons que $f'(\xi) > 0$, la preuve étant identique, aux 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 $f'(x) \geq \frac{1}{2} f'(\xi)$ dans l'intervalle $I_\delta = [\xi - \delta, \xi + \delta]$. Posons alors $M = \max_{x\in I_\delta} f'(x)$. On a alors

$$1 - \lambda M \leq 1 - \lambda f'(x) \leq 1 - \frac{\lambda}{2} f'(\xi),\ \forall x \in I_\delta.$$

On choisit alors λ de façon à ce que $\lambda M - 1 = 1 - \frac{\lambda}{2} f'(\xi)$, c'est-à-dire

$$\lambda = \frac{4}{2M + f'(\xi)}.$$

En posant $g(x) = x - \lambda f(x)$, on obtient que

$$g'(x) \leq \frac{2M - f'(\xi)}{2M + f'(\xi)} < 1,\ \forall x \in I_\delta,$$

et la convergence se déduit alors du théorème 5.12. □

D'un point de vue géométrique, le point $x^{(k+1)}$ dans (5.16) est, à chaque itération, l'abscisse du point d'intersection entre la droite de pente $1/\lambda$ passant par le point $(x^{(k)}, f(x^{(k)}))$ et l'axe des abscisses (voir figure ref). Elle 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. Connaissant un intervalle d'encadrement $[a, b]$ de ξ, on a coutume de définir la méthode de la corde par

$$x^{(k+1)} = x^{(k)} - \frac{b-a}{f(b)-f(a)} f(x^{(k)}),\ \forall k \geq 0, \tag{5.17}$$

avec $x^{(0)}$ donné dans $[a, b]$. Sous les hypothèses du théorème 5.15, 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.

5.3.4 Méthode de Newton–Raphson

En supposant la fonction f est de classe $\mathscr{C}^1$ et que ξ est un zéro simple, la *méthode de Newton–Raphson* fait le choix

$$h(x) = \frac{1}{f'(x)}$$

dans (5.15). La relation de récurrence définissant cette méthode est alors

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})},\ \forall k \geq 0, \tag{5.18}$$

l'initialisation $x^{(0)}$ étant donnée.

Cette méthode peut être interprétée comme une *linéarisation de l'équation $f(x) = 0$ au point $x = x^{(k)}$*. 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 résoud 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'abcisse 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 figure 5.5).

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.

Théorème 5.16 *Soit f une fonction réelle de classe $\mathscr{C}^2$ dans un voisinage du réel ξ tel que $f(\xi) = 0$. Supposons que $f'(\xi) \neq 0$. 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 suffisament proche de ξ.*

Démonstration. a ecrire □

Théorème 5.17 *resultat de convergence globale (hyp. signes derivées) suli 35*

Démonstration. a ecrire □

exemple de mise en échec pour $x^3 - 2x + 2$ 0 et 1

***méthode de Newton modifiée* pour zeros d'ordre superieur** quarteroni 227

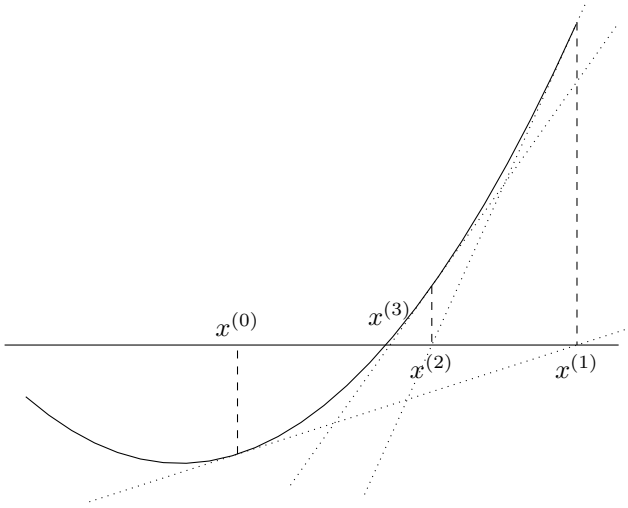

FIGURE 5.5 – Construction des premiers itérés de la méthode de Newton–Raphson.

5.3.5 Méthode de la sécante

La *méthode de la sécante* 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 méthode de Newton dans laquelle la donnée de la dérivée $f'(x^{(k)})$ serait remplacée 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[12], à calculer ou encore coûteuse à évaluer.

À partir de la donnée de deux valeurs initiales $x^{(-1)}$ et $x^{(0)}$, telles que $x^{(-1)} \neq x^{(0)}$, on utilise la relation de récurrence

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}),\ \forall k \geq 0, \tag{5.19}$$

pour obtenir les approximations successives du zéro recherché. **remarque sur le nom avec dessin 5.6**

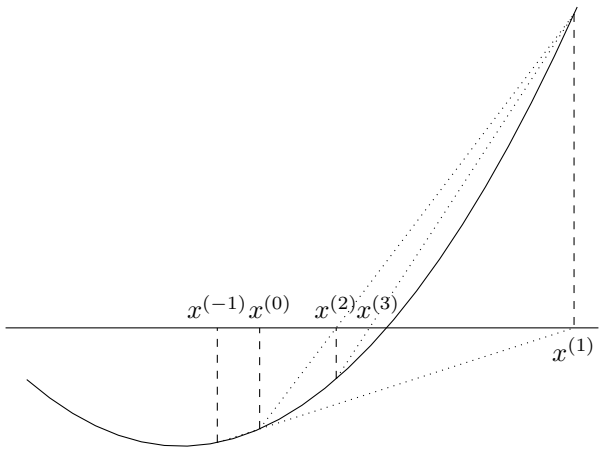

FIGURE 5.6 – Construction des premiers itérés de la méthode de la sécante.

12. 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 x est un paramètre de la donnée initiale du problème associé.

Bien que l'on doive disposer de deux estimations de ξ avant de pouvoir utiliser la relation de récurrence (5.19), 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, dont la relation (5.18) demande de connaître les valeurs de $f(x^{(k)})$ et de $f'(x^{(k)})$. Cependant, à la différence de la méthode de la fausse position, rien n'assure qu'au moins un zéro de f se trouve entre $x^{(k-1)}$ et $x^{(k)}$, pour tout $k \in \mathbb{N}$. Enfin, comparée à la méthode de la corde, elle nécessite le calcul de « mise à jour » du quotient apparaissant dans (5.19). Le bénéfice tiré de cet effort supplémentaire est bien une vitesse de convergence *superlinéaire*, mais cette convergence n'est plus que *locale*, comme le montre le résultat suivant [13].

Théorème 5.18 *Supposons que f est 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* (5.19) *converge vers ξ avec un ordre (au moins) égal à $\frac{1}{2}(1+\sqrt{5}) = 1,6180339887\ldots$*

Démonstration. Nous allons tout d'abord prouver la convergence locale de la méthode. À cette fin, introduisons, pour $\delta > 0$, l'ensemble $I_\delta = \{x \in \mathbb{R} \,|\, |x-\xi| \leq \delta\}$ et supposons que f est classe $\mathscr{C}^2$ dans ce voisinage de ξ. Pour δ suffisament petit, définissons

$$M(\delta) = \max_{\substack{s \in I_\delta \\ t \in I_\delta}} \left| \frac{f''(s)}{2f'(t)} \right|,$$

et supposons que δ est tel que [14]

$$\delta M(\delta) < 1. \tag{5.20}$$

Le nombre ξ est l'unique zéro de f contenu dans I_δ. En effet, en appliquant la formule de Taylor–Lagrange (voir théorème B.5 en annexe) à l'ordre deux à f au point ξ, on trouve que

$$f(x) = f(\xi) + (x-\xi)f'(\xi) + \frac{1}{2}(x-\xi)^2 f''(c),$$

avec c compris entre x et ξ. Si $x \in I_\delta$, on a également $c \in I_\delta$ et on obtient

$$f(x) = (x-\xi)f'(\xi)\left(1 + (x-\xi)\frac{f''(c)}{2f'(\xi)}\right).$$

Si $x \in I_\delta$ et $x \neq \xi$, les trois facteurs dans le membre de droite sont tous différents de zéro (le dernier parce que $\left|(x-\xi)\frac{f''(c)}{2f'(\xi)}\right| < \delta M(\delta) < 1$) et la fonction f ne s'annule qu'en ξ sur l'intervalle I_δ.

Montrons à présent que, quelles que soient les initialisations $x^{(-1)}$ et $x^{(0)}$, avec $x^{(-1)} \neq x^{(0)}$, dans I_δ, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode de la sécante converge vers ξ en prouvant que, pour tout $k \geq 0$, $x^{(k)}$ appartient à I_δ et que deux itérés successifs $x^{(k)}$ et $x^{(k-1)}$ sont distincts, sauf si $f(x^{(k)}) = 0$ pour k donné, auquel cas la méthode aura convergé en un nombre fini d'itérations.
On raisonne par récurrence, le résultat étant vrai par hypothèse pour $k = 0$. Supposons que $x^{(k)}$ et $x^{(k-1)}$ appartiennent à I_δ, avec $x^{(k)} \neq x^{(k-1)}$, pour $k \geq 1$. Utilisons (5.19) 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 (5.19) et en se servant que $f(\xi) = 0$, il vient

$$x^{(k+1)} - \xi = x^{(k)} - \xi - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}) = (x^{(k)} - \xi)\frac{f[x^{(k-1)}, x^{(k)}] - f[x^{(k)}, \xi]}{f[x^{(k-1)}, x^{(k)}]},$$

où l'on a noté, en employant la notation des *différences divisées* (dont on anticipe l'introduction dans le chapitre 6),

$$f[x, y] = \frac{f(x) - f(y)}{x - y}.$$

Par la relation de récurrence pour les différences divisées (6.9), la dernière égalité se réécrit alors

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)(x^{(k-1)} - \xi)\frac{f[x^{(k-1)}, x^{(k)}, \xi]}{f[x^{(k-1)}, x^{(k)}]}.$$

13. Notons qu'on ne peut utiliser les techniques introduites pour les méthodes de point fixe pour établir un résultat de convergence, la relation (5.19) ne pouvant s'écrire sous la forme (5.10) voulue.

14. Notons que $\lim_{\delta \to 0} M(\delta) = \left|\frac{f''(\xi)}{2f'(\xi)}\right| < +\infty$, on peut donc bien satisfaire la condition (5.20) pour δ assez petit.

Par application du théorème des accroissements finis (voir théorème B.3 an annexe), 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

$$f[x^{(k-1)}, x^{(k)}] = f'(\zeta^{(k)}) \text{ et } f[x^{(k-1)}, x^{(k)}, \xi] = \frac{1}{2} f''(\eta^{(k)}).$$

On en déduit alors que

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)(x^{(k-1)} - \xi)\frac{f''(\eta^{(k)})}{2\, f'(\zeta^{(k)})}, \tag{5.21}$$

d'où

$$|x^{(k+1)} - \xi| \leq \delta^2 \left| \frac{f''(\eta^{(k)})}{2\, f'(\zeta^{(k)})} \right| \leq \delta(\delta M(\delta)) < \delta,$$

et $x^{(k+1)}$ appartient à I_δ. Par ailleurs, il est clair d'après la relation (5.19) que $x^{(k+1)}$ est différent de $x^{(k)}$, excepté si $f(x^{(k)})$ est nulle.
En revenant à (5.21), il vient alors que

$$|x^{(k+1)} - \xi| \leq \delta M(\delta) |x^{(k)} - \xi|, \ \forall k \geq 0,$$

et donc

$$|x^{(k+1)} - \xi| \leq (\delta M(\delta))^{k+1} |x^{(0)} - \xi|, \ \forall k \geq 0,$$

ce qui permet de prouver que la méthode converge.

Il reste à vérifier 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 (5.21) que

$$|x^{(k+1)} - \xi| \leq M(\delta) |x^{(k)} - \xi|\, |x^{(k-1)} - \xi|, \ \forall k \geq 0.$$

En posant $E^{(k)} = M(\delta)|x^{(k)} - \xi|$, $\forall k \geq 0$, on obtient, après multiplication de l'inégalité ci-dessus par $M(\delta)$, la relation

$$E^{(k+1)} \leq E^{(k)}\, E^{(k-1)}, \ \forall k \geq 0.$$

Soit $E = \max(E^{(-1)}, {E^{(0)}}^{1/r})$. On va établir par récurrence que

$$E^{(k)} \leq E^{r^{k+1}}, \ \forall k \geq 0.$$

Cette inégalité est en effet trivialement vérifiée pour $k = 0$. En la supposant vraie jusqu'au rang k, $k \geq 1$, 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

$$|x^{(k)} - \xi| \leq \varepsilon^{(k)}, \text{ avec } \varepsilon^{(k)} = \frac{1}{M(\delta)} E^{r^{k+1}}, \ \forall k \geq 0,$$

avec $E < 1$ par hypothèses sur δ, $x^{(-1)}$ et $x^{(0)}$. Il reste à remarquer que

$$\frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^r} = M(\delta)^{r-1} \frac{E^{r^{k+2}}}{{E^{r^{k+1}}}^r} = M(\delta)^{r-1}, \ \forall k \geq 0,$$

et à utiliser la définition 5.3 pour conclure. □

5.4 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 polynôme p_n de degré $n \geq 0$:

$$p_n(x) = \sum_{i=0}^{n} a_i\, x^i, \tag{5.22}$$

les coefficients a_i, $i = 0, \ldots, n$, étant des nombres réels donnés.

S'il est trivial de résoudre les équations algébriques du premier degré [15] et que la forme des solutions des équations du second degré [16] est bien connue, il existe aussi des expressions analytiques pour les solutions des équations de degré trois et quatre, publiées par Cardano [17] en 1545 dans son *Artis Magnæ, Sive de Regulis Algebraicis Liber Unus* (les formules étant respectivement dues à del Ferro [18] et Tartaglia [19] pour le troisième et à Ferrari [20] pour le quatrième). Par contre, 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.

5.4.1 Évaluation des polynômes et de leurs dérivées

Nous allons à présent décrire la *méthode de Horner* [21], qui permet l'évaluation efficace d'un polynôme et de sa dérivée en un point donné. Celle-ci repose sur le fait que tout polynôme $p_n \in \mathbb{P}_n$ peut s'écrire sous la forme

$$p_n(x) = a_0 + x\,(a_1 + x\,(a_2 + \cdots + x\,(a_{n-1} + a_n\,x)\ldots))\,. \tag{5.23}$$

Si les formes (5.22) et (5.23) sont algébriquement équivalentes, la première nécessite n additions et $2n-1$ multiplications alors que la seconde ne requiert que n additions et n multiplications.

L'algorithme pour évaluer le polynôme 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 $b_0 = p_n(z)$.

Application. Évaluons le polynôme $7\,x^4 + 5\,x^3 - 2\,x + 8$ 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 $8,5625$.

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, $i = n, n-1, \ldots, 0$, du polynôme 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.

Application. Pour l'exemple d'application précédent, on trouve le tableau suivant [22]

	7	5	−2	0	8
0	7	8,5	2,25	1,125	8,5625

.

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

$$p_n(x) = (x - z)\,q_{n-1}(x) + r_0,$$

15. Ce sont les équations du type $a\,x + b = 0$, avec $a \neq 0$, dont la solution est donnée par $x = -\frac{b}{a}$.

16. Ce sont les équations de la forme $a\,x^2 + b\,x + c = 0$, avec $a \neq 0$, dont les solutions sont données par $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

17. 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.

18. 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.

19. 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.

20. Lodovico Ferrari (2 février 1522 - 5 octobre 1565) était un mathmaticien italien. Élève de Cardano, il est à l'origine de la méthode de résolution des équations algébriques du quatrième degré.

21. William George Horner (1786 - 22 septembre 1837) était un mathématicien anglais. 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.

22. 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.

où le quotient $q_{n-1} \in \mathbb{P}_{n-1}$ est un polynôme dépendant de z par l'intermédiaire de ses coefficients, puisque

$$q_{n-1}(x) = \sum_{i=1}^{n} b_i \, x^{i-1},$$

et 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. Effectuons la division euclidienne du polynôme $4\,x^3 - 7\,x^2 + 3\,x - 5$ par $x - 2$. En construisant un tableau comme précédemment, soit

$$\begin{array}{c|cccc} & 4 & -7 & 3 & -5 \\ \hline 0 & 4 & 1 & 5 & 5 \end{array},$$

on obtient $4\,x^3 - 7\,x^2 + 3\,x - 5 = (x-2)(4\,x^2 + x + 5) + 5$.

Appliquons de nouveau la méthode pour effectuer la division du quotient q_{n-1} par $(x - z)$. On trouve

$$q_{n-1}(x) = (x - z)\, q_{n-2}(x) + r_1,$$

avec $q_{n-2} \in \mathbb{P}_{n-2}$ et r_1 une constante, avec

$$q_{n-2}(x) = \sum_{i=2}^{n} b_i \, x^{i-2} \text{ et } r_1 = c_1,$$

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

$$p_n(x) = (x - z)^2\, q_{n-2}(x) + r_1\,(x - z) + r_0,$$

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 du polynôme p_n au point z. On arrive en effet à

$$p_n(x) = r_n\,(x - z)^n + \cdots + r_1\,(x - \xi) + r_0, \tag{5.24}$$

après $n + 1$ itérations de la méthode que l'on peut résumer dans un tableau synthétique comme on l'a déjà fait

$$\begin{array}{c|cccccc} & a_n & a_{n-1} & \ldots & a_2 & a_1 & a_0 \\ \hline 0 & b_n & b_{n-1} & \ldots & b_2 & b_1 & r_0 \\ 0 & c_n & c_{n-1} & \ldots & c_2 & r_1 & \\ \cdot & \cdot & \cdot & \ldots & r_2 & & \\ \vdots & \vdots & \vdots & \iddots & & & \\ \cdot & \cdot & r_{n-1} & & & & \\ 0 & r_n & & & & & \end{array} \tag{5.25}$$

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 (5.24), on montre alors que

$$r_j = \frac{1}{j!}\, p_n^{(j)}(z), \ j = 0, \ldots, n,$$

où $p_n^{(j)}$ désigne la $j^{\text{ième}}$ dérivée du polynôme p_n.

Le calcul de l'ensemble du tableau (5.25) demande $\frac{1}{2}(n^2 + n)$ additions et autant de multiplications.

5.4.2 Méthode de Newton–Horner

Compte tenu des remarques de la section précédente, on voit que la méthode de Newton peut judicieusement être adaptée pour la recherche des racines d'un polynôme, en exploitant la méthode de Horner étant pour calculer le quotient apparaissant dans ...

$$z^{(k+1)} = z^{(k)} - \frac{p(z^{(k)})}{p'(z^{(k)})} = ... \tag{5.26}$$

seulement besoin des deux premières colonnes du tableau synthetique

si q_{n-1} est le polynôme associé à p_n, il vient en dérivant par rapport à x

$$p'_n(x) = q_{n-1}(x;z) + (x-z)\, q'_{n-1}(x;z),$$

d'où $p'_n(z) = q_{n-1}(z;z)$. Grâce à cette identité, la méthode de Newton–Horner pour l'approximation d'une racine r_j prend la forme suivante : étant donné une estimation initiale $r_j^{(0)}$ de la racine, calculer

$$r_j^{(k+1)} = r_j^{(k)} - \frac{p_n(r_j^{(k)})}{p'_n(r_j^{(k)})} = r_j^{(k)} - \frac{p_n(r_j^{(k)})}{q_{n-1}(r_j^{(k)}; r_j^{(k)})},\ k \geq 0.$$

Pour un polynôme de degré n, le coût de chaque itération de l'algorithme est égal à $4n$. ? ? ? ? Si la racine est complexe, il est nécessaire de travailler en arithmétique complexe et de prendre la donnée initiale dans $\mathbb{C}$.

voir Gautschi 237

5.4.3 Méthode de Muller

ON OUBLIE!!!! En parler cependant... méthode introduite en 1956 par D. E. Muller [Mul56]

capable de calculer des zéros complexes d'une fonction f même en partant d'une donnée initiale réelle, sa convergence est presque quadratique.

5.4.4 Déflation

AJOUTER manque de stabilite a cause des additions (raffinement necessaire lors de la deflation)

Une fois une approximation d'une racine du polynôme obtenue, on effectue une division de celui-ci par $(x - ...)$ et on applique de nouveau la méthode de recherche de zéros au polynôme quotient pour l'approximation d'une autre racine. Ce procédé itératif, permettant l'approximation successive de *toutes* les racines d'un polynôme, est appelé *déflation*. Associé à la méthode de Newton–Horner, il exploite pleinement la méthode de Horner.

REPRENDRE et COMPLETER

On peut alors, à chaque étape, améliorer la précision en utilisant l'approximation $\tilde{r}_j$ obtenue d'une racine comme donnée initiale de la méthode de Newton–Horner (par exemple) appliquée au polynôme original p_n, c'est la *phase de raffinement* de la méthode.

le processus de déflation est affecté d'erreurs d'arrondi. Pour améliorer sa stabilité, on peut commencer par approcher la racine r_1 de module minimum (qui est la plus sensible au mauvais conditionnement du problème), puis continuer avec les suivante jusqu'à celle de plus grand module.

quateroni 228, Stoer 306

Pour aller plus loin

variantes de la méthode de la fausse position

Pour un résumé du développement historique de la méthode de Newton–Raphson, on pourra se consulter l'article de Ypma [Ypm95]. Cette méthode est l'une des plus célébrées et utilisées des mathématiques appliquées. Elle se généralise à la résolution de toute équation non linéaire, qu'elle porte sur une variable complexe ou de $\mathbb{R}^d$, $d \geq 1$, mais aussi de systèmes d'équations non linéaires ou d'équations fonctionnelles

dans les espaces de Banach (la dérivée étant alors entendue au sens de la dérivée de Fréchet [23]). Elle est un élément éssentiel de la démonstration du fameux *théorème de Nash*[24]*–Moser*[25], un résultat d'inversion locale formulé dans une classe particulière d'espaces de Fréchet. On peut l'utiliser pour traiter le 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, 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 et abondante. Nous renvoyons le lecteur intéressé à l'ouvrage [BGLS06] en guise d'introduction.

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 façon 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

$$z^{(k+1)} = z^{(k)} - \frac{p(z^{(k)})}{p'(z^{(k)})},\ k \geq 0,$$

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*, plus particulièrement l'ensemble de Julia [26] associé à la fonction méromorphe $z \mapsto z - \frac{p(z)}{p'(z)}$, et sa représentation donne lieu, selon le polynôme considéré, à des images particulièrement surprenantes (voir la figure 5.7 ci-dessous).

Il peut s'avérer intéressant, notamment pour obtenir des estimations, de savoir combien de racines réelles d'un polynôme sont contenues dans un intervalle donné. On peut pour cela utiliser les *suites de Sturm*[27]. On trouvera plus de détails dans [IK94].

Pour un aperçu historique et une présentation d'algorithmes récents concernant la résolution des équations algébriques, on pourra consulter l'article de Pan [Pan97].

Références du chapitre

[BGLS06] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization, theoretical and practical aspects.* Universitext. Springer, second edition, 2006.

[Mul56] D. E. Muller. A method for solving algebraic equations using an automatic computer. *Math. Tab. Aids Comp.*, 10(56) :208–215, 1956.

23. Maurice René Fréchet (2 septembre 1878 - 4 juin 1973) était un mathématicien français. Très prolifique, il fit d'importantes contributions en topologie, en probabilités et en statistique.

24. John Forbes Nash, Jr. (né le 13 juin 1928) est un mathématicien et économiste américain. Il s'est principalement intéressé à la théorie des jeux, la géométrie différentielle et aux équations aux dérivées partielles. Il a partagé le « prix Nobel » d'économie en 1994 avec Reinhard Selten et John Harsanyi pour leurs travaux en théorie des jeux.

25. 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.

26. 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.*

27. Jacques Charles François Sturm (29 septembre 1803 - 15 décembre 1855) était un mathématicien français d'origine allemande.

[Pan97] V. Y. Pan. Solving a polynomial equation : some history and recent progress. *SIAM Rev.*, 39(2) :187–220, 1997.

[Ypm95] T. J. Ypma. Historical development of the Newton–Raphson method. *SIAM Rev.*, 37(4) :531–551, 1995.

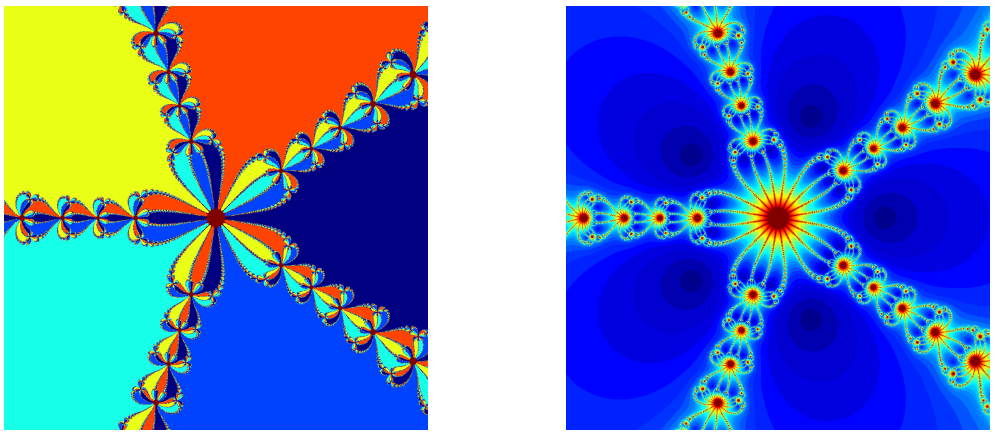

FIGURE 5.7 – Illustration de l'utilisation de la méthode de Newton pour la recherche des racines complexes de l'équation $z^5 - 1 = 0$. À gauche, on a représenté les bassins de convergence de la méthode : chaque point $z^{(0)}$ (choisi ici tel que $|\mathrm{Re}(z^{(0)})| \leq 2$ et $|\mathrm{Im}(z^{(0)})| \leq 2$) servant d'initialisation est coloré en fonction de la racine atteinte en cas de convergence (une sixième couleur étant attibuée 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.

Chapitre 6

Interpolation polynomiale

Soit n un entier positif. Étant donné une famille de $n+1$ points $(x_i, y_i)_{i=0,\ldots,n}$ distincts du plan, l'*interpolation* est une technique consistant à construire une courbe d'un type donné passant par les points (x_i, y_i). Les quantités y_i, $i = 0, \ldots, n$, peuvent en effet représenter les valeurs aux *nœuds* x_i, $i = 0, \ldots, n$, d'une fonction f connues analytiquement, et l'on cherche alors à remplacer f par une fonction plus simple à manipuler en vue d'un calcul numérique faisant intervenir des dérivées et/ou des intégrales, ou bien encore des données expérimentales, auquel cas on vise à obtenir une représentation ou même une loi empirique pour celles-ci lorsque leur nombre est important.

Dans un problème d'*interpolation polynomiale de Lagrange*[1], on cherche en particulier à déterminer un polynôme de degré n dont le graphe passe par ces $n+1$ points, c'est-à-dire à trouver $\Pi_n \in \mathbb{P}_n$ vérifiant $\Pi_n(x_i) = y_i$ pour $i = 0, \ldots, n$. On dit alors que le polynôme Π_n *interpole* les quantités $\{y_i\}_{i=0,\ldots,n}$ aux nœuds $\{x_i\}_{i=0,\ldots,n}$. Le choix de polynômes n'est pas le seul possible : l'*interpolation trigonométrique* utilise des polynômes trigonométriques et est largement utilisée pour la mise en œuvre de l'analyse de Fourier[2]. Cependant, la régularité, la facilité de calcul d'une valeur en un point (grâce à la méthode de Horner) et les nombreuses autres propriétés des polynômes en font une classe de fonctions particulièrement intéressante d'un point de vue pratique. L'interpolation polynomiale est pour cette raison un outil de premier plan pour l'approximation numérique des fonctions.

Dans ce chapitre, on traite majoritairement de l'interpolation de Lagrange, qui constitue la base théorique principale de l'interpolation polynomiale. Après en avoir donné les principes et les propriétés, nous considérons les aspects pratiques du calcul du *polynôme d'interpolation de Lagrange* ainsi que l'étude de l'*erreur d'interpolation*, qui est l'erreur commise lorsque l'on substitue à une fonction donnée son polynôme d'interpolation. Quelques exemples d'*interpolation par morceaux* concluent cette (brève) présentation.

On suppose une fois pour toutes que $\{(x_i, y_i)\}_{i=0,\ldots,n}$, $n \geq 0$, est une famille de $n+1$ points dont les abcisses x_i sont toutes deux à deux distinctes. Afin d'alléger la rédaction, on appellera souvent dans ce chapitre (la section consacrée à l'interpolation par morceaux faisant toutefois exception) polynôme d'interpolation le polynôme de Lagrange associé aux points $\{(x_i, y_i)\}_{i=0,\ldots,n}$.

6.1 Polynôme d'interpolation de Lagrange

Le *polynôme d'interpolation*, encore appelé *polynôme interpolant, de Lagrange associé aux points* $\{(x_i, y_i)\}_{i=0,\ldots,n}$ est défini comme étant la solution du problème d'interpolation polynomiale mentionné en introduction. Commençons par montrer que ce problème est *bien posé*, c'est-à-dire qu'il admet une unique solution.

1. Joseph Louis Lagrange (Giuseppe Lodovico Lagrangia en italien, 25 janvier 1736 - 10 avril 1813) était un mathématicien et astronome franco-italien. Fondateur du calcul des variations avec Euler, il a également produit d'importantes contributions tant en analyse qu'en géométrie, en théorie des groupes et en mécanique.

2. 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.

Théorème 6.1 *Soit n un entier positif. 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 $\Pi_n \in \mathbb{P}_n$ tel que $\Pi_n(x_i) = y_i$ pour $i = 0, \ldots, n$.*

DÉMONSTRATION. Le polynôme Π_n recherché étant de degré n, on peut poser

$$\Pi_n(x) = \sum_{j=0}^{n} a_j \, x^j, \ \forall x \in \mathbb{R},$$

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 :

$$a_0 + a_1 \, x_i + \cdots + a_n {x_i}^n = y_i, \ i = 0, \ldots, n. \tag{6.1}$$

Ce système possède une unique solution si et seulement si la matrice carrée qui lui est associée est inversible. Or, il se trouve que le déterminant de cette dernière est un *déterminant de Vandermonde*[3] dont on peut montrer (preuve est laissée en exercice) qu'il 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. □

Remarque. Une autre façon de prouver l'unicité du polynôme d'interpolation est la suivante. Supposons 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'annule alors en $n+1$ points distincts, elle est donc nulle d'après le théorème fondamental de l'algèbre.

Pour construire le polynôme d'interpolation Π_n, il suffit donc de résoudre le système (6.1). Il a cependant été démontré (voir [Gau75]) que le nombre de conditionnement des matrices de Vandermonde peut être grand, ce qui conduit à des erreurs importantes lors de la résolution numérique par des méthodes directes, cette résolution s'avérant également coûteuse (de l'ordre de $O(n^3)$ opérations arithmétiques) lorsque le nombre de nœuds d'interpolation est important. Plusieurs auteurs ont proposé des méthodes rapides (de l'ordre de $O(n^3)$ opérations arithmétiques) et numériquement stables pour la résolution des systèmes de Vandermonde, mais celles-ci s'appuient sur la *forme de Newton* du polynôme d'interpolation (voir la section 6.1.2). On pourra consulter la bibliographie en fin de chapitre pour des références.

Une autre possibilité consiste à écrire le polynôme d'interpolation non pas dans la base canonique mais dans une base adaptée, pour laquelle la matrice du système linéaire associé au problème est diagonale : la base des *polynômes de Lagrange*.

6.1.1 Forme de Lagrange du polynôme d'interpolation

Commençons par introduire les polynômes de Lagrange et leurs propriétés.

Définition 6.2 *On appelle **polynômes de Lagrange associés aux nœuds** $\{x_i\}_{i=0,\ldots,n}$, $n \ge 1$, les $n+1$ polynômes $l_i \in \mathbb{P}_n$, $i = 0, \ldots, n$, définis par*

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{6.2}$$

Bien que communément employée pour ne pas alourdir les écritures, la notation l_i, $i = 0, \ldots, n$, utilisée pour les polyômes de Lagrange ne fait pas explicitement apparaître leur degré. La valeur de l'entier n est en général claire compte tenu du contexte, mais il faudra cependant bien garder cette remarque à l'esprit, puisque l'on peut être amené à augmenter n (voir la section 6.1.2) et même à le faire tendre

3. 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.

tendre vers l'infini (voir la section 6.1.4). Ajoutons que, si l'on a exigé que n soit supérieur ou égal à 1 dans la définition, le cas trivial $n = 0$ peut en fait être inclus dans tout ce qui va suivre en posant $l_0 \equiv 1$ si $n = 0$.

Proposition 6.3 *Les polynômes de Lagrange $\{l_i\}_{i=0,\ldots,n}$, $n \geq 0$, sont tous de degré n, vérifient $l_i(x_k) = \delta_{ik}$, $i, k = 0, \ldots, n$, et forment une base de $\mathbb{P}_n$.*

Démonstration. Le résultat est évident si $n = 0$. Si $n \geq 1$, les deux premières propriétés découlent directement de la définition (6.2) des polynômes de Lagrange. On déduit ensuite de la deuxième propriété que, si le polynôme $\sum_{i=0}^{n} \lambda_i \, l_i$, $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, est identiquement nul, alors on a

$$0 = \sum_{i=0}^{n} \lambda_i \, l_i(x_j) = \lambda_j, \ \forall j \in \{1, \ldots, n\}.$$

La famille $\{l_i\}_{i=0,\ldots,n}$ 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, $-0,5$, 0, $0,5$ et 1.

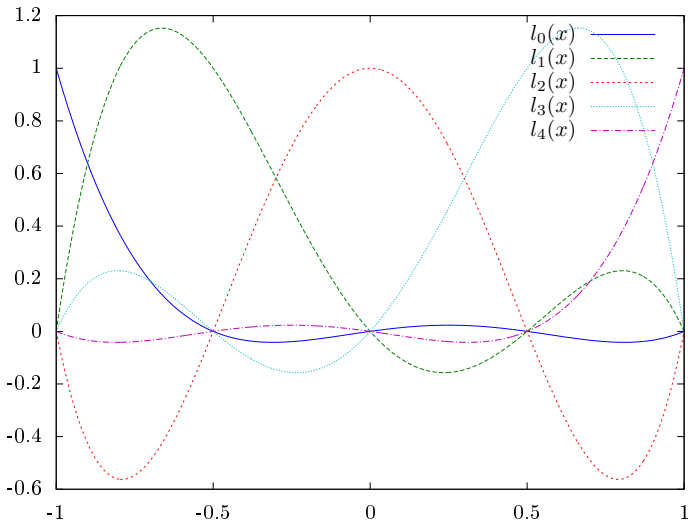

Figure 6.1 – Graphes des polynômes de Lagrange $l_i(x)$, $i = 0, \ldots, 5$ associés à des nœuds équidistribués sur l'intervalle $[-1, 1]$.

On déduit de la proposition 6.3 le résultat suivant.

Théorème 6.4 *Soit n un entier positif. Étant donné $n + 1$ points distincts $x_0, \ldots, x_n$ et $n + 1$ valeurs $y_0, \ldots, y_n$, le polynôme d'interpolation $\Pi_n \in \mathbb{P}_n$ tel que $\Pi_n(x_i) = y_i$, $i = 0, \ldots, n$, est donné par la* ***formule d'interpolation de Lagrange***

$$\Pi_n(x) = \sum_{i=0}^{n} y_i \, l_i(x). \tag{6.3}$$

Démonstration. Pour établir (6.3), on utilise que les polynômes $\{l_i\}_{i=0,\ldots,n}$ 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

$$y_j = \Pi_n(x_j) = \sum_{i=0}^{n} \mu_i \, l_i(x_j) = \mu_j, \ \forall j \in \{1, \ldots, n\}.$$

□

6.1.2 Forme de Newton du polynôme d'interpolation

L'utilisation de la formule d'interpolation de Lagrange (6.3) s'avère peu commode d'un point de vue pratique. D'une part, l'expression (6.2) montre qu'on peut difficilement déduire un polynôme de Lagrange d'un autre. D'autre part, imaginons que l'on connaisse le polynôme d'interpolation Π_{n-1} associé aux n paires (x_i, y_i), $i = 0, \ldots, n-1$, et que, étant donné un couple supplémentaire (x_n, y_n), on souhaite calculer le polynôme Π_n. L'écriture (6.3) de Π_n, faisant intervenir la base des polynômes de Lagrange (6.2) associés aux nœuds $\{x_i\}_{i=0,\ldots,n}$, conduit au calcul de *tous* les polynômes (de degré n) l_i, $i = 0, \ldots, n$. La *forme de Newton* du polynôme d'interpolation offre une alternative, bien moins onéreuse en coût de calcul, à la détermination de Π_n, en écrivant ce dernier comme la somme de 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 dépend des nœuds $\{x_i\}_{i=0,\ldots,n-1}$ et d'un seul coefficient inconnu).

Nous allons à présent expliciter la forme de Newton du polynôme d'interpolation Π_n. Posons pour cela

$$\Pi_n(x) = \Pi_{n-1}(x) + q_n(x), \tag{6.4}$$

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

$$\omega_n(x) = \prod_{i=0}^{n-1}(x - x_i) \tag{6.5}$$

le *polynôme de Newton de degré n associé aux nœuds* $\{x_i\}_{i=0,\ldots,n-1}$ et déterminons le coefficient a_n. Comme $\Pi_n(x_n) = y_n$, on déduit de (6.4) 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 $n^{ième}$ *différence divisée de Newton* et se note généralement

$$a_n = y[x_0, x_1, \ldots, x_n],\ n \geq 1.$$

On a par conséquent

$$\Pi_n(x) = \Pi_{n-1}(x) + y[x_0, x_1, \ldots, x_n]\,\omega_n(x). \tag{6.6}$$

En posant $y[x_0] = y_0$ et $\omega_0 \equiv 1$, on obtient, à partir de (6.6) et en raisonnant par récurrence sur le degré n, que

$$\Pi_n(x) = \sum_{i=0}^{n} y[x_0, \ldots, x_i]\,\omega_i(x), \tag{6.7}$$

qui est, en vertu de l'unicité du polynôme d'interpolation, le même polynôme que celui définit par (6.3). La forme (6.7) 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[4] de $\mathbb{P}_n$ formée par la famille de polynômes de Newton $\{\omega_i\}_{i=0,\ldots,n}$.

Une propriété des différences divisées

On peut vérifier, à titre d'exercice, que la formule (6.3) se réécrit, en fonction du polynôme de Newton de degré $n+1$, de la manière suivante

$$\Pi_n(x) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)}\, y_i. \tag{6.8}$$

4. 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).

En utilisant alors (6.7) pour identifier $y[x_0,\ldots,x_n]$ avec le coefficient lui correspondant dans (6.8), on obtient la forme explicite suivante pour cette différence divisée :

$$y[x_0,\ldots,x_n] = \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)}.$$

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 simple manipulation algébrique, on obtient la formule de récurrence

$$y[x_0,\ldots,x_n] = \frac{y[x_1,\ldots,x_n] - y[x_0,\ldots,x_{n-1}]}{x_n - x_0}, \tag{6.9}$$

de laquelle ces quantités tirent leur nom et qui fournit un procédé pour leur calcul effectif. Ce dernier consiste en la construction du tableau suivant

$$\begin{array}{c|ccccc} x_0 & y[x_0] & & & & \\ x_1 & y[x_1] & y[x_0,x_1] & & & \\ x_2 & y[x_2] & y[x_1,x_2] & y[x_0,x_1,x_2] & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ x_n & y[x_n] & y[x_{n-1},x_n] & y[x_{n-2},x_{n-1},x_n] & \cdots & y[x_0,\ldots,x_n] \end{array} \tag{6.10}$$

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 (6.7) du polynôme d'interpolation de Lagrange sont les $n+1$ coefficients diagonaux du tableau (6.10). Leur évaluation requiert $n(n+1)$ additions et $\frac{1}{2}n(n+1)$ divisions. Si l'on dispose d'une valeur $y_{n+1}(= y[x_{n+1}])$ associée à un nouveau nœud x_{n+1}, on n'a qu'à calculer une ligne supplémentaire $(y[x_n,x_{n+1}] \ \cdots \ y[x_0,\ldots,x_{n+1}])$ pour construire le polynôme $\Pi_{n+1}(x)$ à partir de $\Pi_n(x)$ en lui ajoutant $y[x_0,\ldots,x_{n+1}]\,\omega_{n+1}(x)$, ce qui nécessite $2(n+1)$ additions et $n+1$ divisions.

Application. 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 (6.7) et de la méthode de calcul des différences divisées basée sur la formule (6.9). 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

$$\Pi_3(x) = 4 - \frac{5}{2}(x+1) + \frac{5}{2}(x+1)(x-1) - (x+1)(x-1)(x-2).$$

6.1.3 Algorithme de Neville

Si l'on ne cherche pas à contruire le polynôme d'interpolation de Lagrange mais simplement à connaître sa valeur en un point donné, 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 6.5 *Soient x_{i_k}, $k=0,\ldots,n$, $n+1$ nœuds distincts et y_{i_k}, $k=0,\ldots,n$, $n+1$ valeurs. On note $\Pi_{x_{i_0},x_{i_1},\ldots,x_{i_n}}$ le polynôme d'interpolation 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_k}, $k=0,\dots,n$, $n+3$ nœuds distincts et y_i, y_j, y_{i_k}, $k=0,\dots,n$, $n+3$ valeurs, on a

$$\Pi_{x_{i_0},\dots,x_{i_n},x_i,x_j}(x) = \frac{(x-x_j)\,\Pi_{x_{i_0},\dots,x_{i_n},x_i}(x) - (x-x_i)\,\Pi_{x_{i_0},\dots,x_{i_n},x_j}(x)}{x_i - x_j},\ \forall x \in \mathbb{R}. \tag{6.11}$$

DÉMONSTRATION. Notons $q(x)$ le membre de droite de (6.11). Les polynômes $\Pi_{x_{i_0},\dots,x_{i_n},x_i}$ et $\Pi_{x_{i_0},\dots,x_{i_n},x_i}$ étant tous deux de degré $n+1$, le polynôme q est de degré inférieur ou égal à $n+2$. On vérifie ensuite que

$$q(x_{i_k}) = \frac{(x_{i_k}-x_j)\,\Pi_{x_{i_0},\dots,x_{i_n},x_i}(x_{i_k}) - (x_{i_k}-x_i)\,\Pi_{x_{i_0},\dots,x_{i_n},x_j}(x_{i_k})}{x_i - x_j} = y_{i_k},\ k=0,\dots,n,$$

et

$$q(x_i) = \frac{(x_i-x_j)\,\Pi_{x_{i_0},\dots,x_{i_n},x_i}(x_i)}{x_i-x_j} = y_i,\ q(x_j) = -\frac{(x_j-x_i)\,\Pi_{x_{i_0},\dots,x_{i_n},x_j}(x_j)}{x_i-x_j} = y_j.$$

On en déduit que $q = \Pi_{x_{i_0},\dots,x_{i_n},x_i,x_j}$ par unicité du polynôme d'interpolation. □

Dans la classe de méthodes utilisant l'identité (6.11), l'une des plus connues est l'*algorithme de Neville*, qui consiste à calculer de proche en proche les valeurs au point x considéré des polynômes d'interpolation, de degré croissant, associés à des sous-ensembles des points $\{(x_i,y_i)\}_{i=0,\dots,n}$. À la manière de (6.10) pour les différences divisées, cette construction peut s'organiser dans un tableau synthétique :

$$\begin{array}{c|cccccc} x_0 & \Pi_{x_0}(x)=y_0 & & & & & \\ x_1 & \Pi_{x_1}(x)=y_1 & \Pi_{x_0,x_1}(x) & & & & \\ x_2 & \Pi_{x_2}(x)=y_2 & \Pi_{x_1,x_2}(x) & \Pi_{x_0,x_1,x_2}(x) & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \\ x_n & \Pi_{x_n}(x)=y_n & \Pi_{x_{n-1},x_n}(x) & \Pi_{x_{n-2},x_{n-1},x_n}(x) & \cdots & \Pi_{x_0,\dots,x_n}(x) \end{array} \tag{6.12}$$

Le point x 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), en appliquant (6.11). Par exemple, la valeur $\Pi_{x_0,x_1,x_2}(x)$ est donnée par

$$\Pi_{x_0,x_1,x_2}(x) = \frac{(x-x_2)\Pi_{x_0,x_1}(x) - (x-x_0)\Pi_{x_1,x_2}(x)}{x_0-x_2}.$$

Application A ecrire

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*[5], utilisant (6.11) mais avec des polynômes d'interpolation intermédiaires différents, ce qui conduit au tableau suivant :

$$\begin{array}{c|cccccc} x_0 & \Pi_{x_0}(x)=y_0 & & & & & \\ x_1 & \Pi_{x_1}(x)=y_1 & \Pi_{x_0,x_1}(x) & & & & \\ x_2 & \Pi_{x_2}(x)=y_2 & \Pi_{x_0,x_2}(x) & \Pi_{x_0,x_1,x_2}(x) & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \\ x_n & \Pi_{x_n}(x)=y_n & \Pi_{x_0,x_n}(x) & \Pi_{x_0,x_1,x_n}(x) & \cdots & \Pi_{x_0,\dots,x_n}(x) \end{array} \tag{6.13}$$

6.1.4 Interpolation polynomiale d'une fonction

L'intérêt de remplacer une fonction f quelconque par un polynôme l'approchant aussi précisément que voulu est évident d'un point de vue numérique et informatique, puisqu'il est très aisé de stocker et de manipuler (additionner, multiplier, dériver, intégrer...) un polynôme dans une machine. Pour ce faire, il est naturel d'utiliser un polynôme d'interpolation prenant la valeur $y_i = f(x_i)$ au nœud x_i, $i=0,\dots,n$.

5. 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.

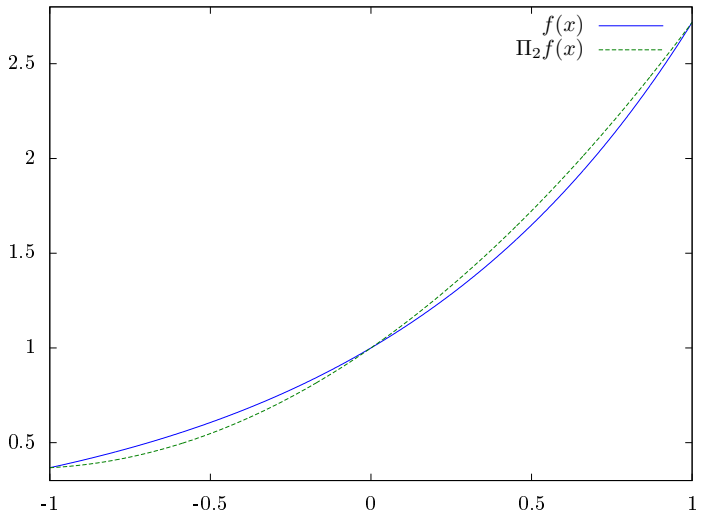

FIGURE 6.2 – Graphes de la fonction $f(x) = e^x$ et de son polynôme d'interpolation de Lagrange de degré deux à nœuds équirépartis sur l'intervalle $[-1, 1]$.

Polynôme d'interpolation de Lagrange d'une fonction

Définition 6.6 *Soient $n+1$ nœuds distincts x_i, $i = 0, \ldots, n$, et f une fonction donnée. 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. Construisons le polynôme d'interpolation de Lagrange de degré deux de la fonction $x \mapsto e^x$ sur l'intervalle $[-1, 1]$, avec comme nœuds d'interpolation les points $x_0 = -1$, $x_1 = 0$ et $x_2 = 1$. Nous avons :

$$l_0(x) = \frac{1}{2}\,x(x-1),\ l_1(x) = 1 - x^2 \text{ et } l_2(x) = \frac{1}{2}\,x(x+1),$$

d'où

$$\Pi_2 f(x) = \frac{1}{2}\,x(x-1)\,e^{-1} + (1-x^2)\,e^0 + \frac{1}{2}\,x(x+1)\,e^1,$$

ce que l'on peut encore écrire $\Pi_2 f(x) = 1 + \sinh(1)x + (\cosh(1) - 1)x^2$.

Du point de vue de l'approximation polynomiale, le polynôme d'interpolation de Lagrange de la fonction f aux nœuds x_i, $i = 0, \ldots, n$, comme le polynôme de degré n minimisant l'*erreur d'approximation*, basée sur une semi-norme, suivante

$$\|f - p_n\| = \sum_{i=0}^{n} |f(x_i) - p_n(x_i)|,\ \forall p_n \in \mathbb{P}_n.$$

Bien que les valeurs de f et de son polynôme d'interpolation soient les mêmes 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 suffisament régulière, on peut établir le résultat suivant, qui donne une estimation de cette différence.

Théorème 6.7 *Soient $n+1$ nœuds distincts x_i, $i = 0, \ldots, n$, contenus dans un intervalle $[a, b]$ non vide de $\mathbb{R}$ et f une fonction supposée de classe $\mathscr{C}^{n+1}$ sur $[a, b]$. L'erreur d'interpolation en tout point x de $[a, b]$ est alors donnée par*

$$f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(c)}{(n+1)!}\,\omega_{n+1}(x), \tag{6.14}$$

où $c \in]a,b[$ et ω_{n+1} est le polynôme de Newton de degré $n+1$ associé à la famille $\{x_i\}_{i=0,\dots,n}$. On a de plus

$$|f(x) - \Pi_n f(x)| \leq \frac{M_{n+1}}{(n+1)!} |\omega_{n+1}(x)|, \tag{6.15}$$

avec $M_{n+1} = \max_{x\in[a,b]} |f^{(n+1)}(x)|$.

DÉMONSTRATION. Si le point x coïncide avec l'un des nœuds d'interpolation, les deux membres de (6.14) sont nuls et l'égalité est trivialement vérifiée. Supposons à présent que x est un point fixé de $[a,b]$, différent de x_i pour $i = 0,\dots,n$, et introduisons la fonction auxiliaire

$$\varphi(t) = f(t) - \Pi_n f(t) - \frac{f(x) - \Pi_n f(x)}{\omega_{n+1}(x)} \omega_{n+1}(t), \ \forall t \in [a,b].$$

Celle-ci est de classe $\mathscr{C}^{n+1}$ sur $[a,b]$ (en vertu des hypothèses sur la fonction f) et s'annule en $n+2$ points (puisque $\varphi(x) = \varphi(x_0) = \varphi(x_1) = \cdots = \varphi(x_n) = 0$). D'après le théorème de Rolle (voir théorème B.2 en annexe), la fonction φ' possède au moins $n+1$ zéros distincts dans l'intervalle $]a,b[$ et, par récurrence, $\varphi^{(j)}$, $0 \leq j \leq n+1$, admet au moins $n+2-j$ zéros distincts. Par conséquent, il existe c appartenant à $]a,b[$ tel que $\varphi^{(n+1)}(c) = 0$, ce qui s'écrit encore

$$f^{(n+1)}(c) - \frac{f(x) - \Pi_n f(x)}{\omega_{n+1}(x)} (n+1)! = 0,$$

d'où (6.14).
Pour prouver (6.15), il suffit de remarquer que $f^{(n+1)}$, et donc $|f^{(n+1)}|$, est une fonction continue sur $[a,b]$. Ainsi, l'application $x \mapsto |f^{(n+1)}(x)|$ est bornée sur $[a,b]$ et atteint son maximum sur cet intervalle ; l'inégalité (6.15) se déduit donc de (6.14). □

La forme de Newton du polynôme d'interpolation permet d'obtenir une autre expression que (6.14) pour l'erreur d'interpolation. Soit en effet $\Pi_n f$ le polynôme d'interpolation de f aux nœuds $x_0,\dots,x_n$ et soit t un nœud arbitraire distinct des précédents. Si l'on désigne par $\Pi_{n+1} f$ le polynôme interpolant f aux nœuds $x_0,\dots,x_n$ et t, on a, en utilisant (6.7),

$$\Pi_{n+1} f(x) = \Pi_n f(x) + f[x_0,\dots,x_n,t](x-x_0)\dots(x-x_n).$$

Puisque $\Pi_{n+1} f(t) = f(t)$, on obtient

$$f(t) = \Pi_n f(t) + f[x_0,\dots,x_n,t](t-x_0)\dots(t-x_n),$$

d'où, en prenant $t = x$ et compte tenu de la définition (6.5) des polynômes de Newton,

$$f(x) - \Pi_n f(x) = f[x_0,\dots,x_n,x]\,\omega_{n+1}(x). \tag{6.16}$$

Cette nouvelle formule 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 6.7, en posant $x = x_{n+1} \in [a,b]$ et en comparant (6.16) avec (6.14), il vient

$$f[x_0,\dots,x_{n+1}] = \frac{f^{(n+1)}(c)}{(n+1)!},$$

avec $c \in]a,b[$. L'intérêt de cette dernière identité vient du fait que la forme de Newton de $\Pi_n f$ peut alors être vue comme un développement de Taylor de f en x_0 (à condition $|x_n - x_0|$ ne soit pas trop grand) tronqué à l'ordre n. A VOIR !

Convergence des polynômes d'interpolation et phénomène de Runge [6]

Nous nous intéressons dans cette section à la question de la convergence uniforme du polynôme d'interpolation d'une fonction f vers cette dernière lorsque le nombre de nœuds n tend vers l'infini. Comme ce polynôme dépend de la distribution des nœuds d'interpolation, il est nécessaire de formuler

6. Carl David Tolmé Runge (30 août 1856 - 3 janvier 1927) était un mathématicien et physicien allemand. Il a co-développé la méthode de Runge-Kutta, qui est l'une des plus utilisées pour la résolution numérique d'équations différentielles.

ce problème de manière un peu plus spécifique. Nous supposerons ici que l'on fait le choix, très courant, 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

$$x_i = a + \frac{i(b-a)}{n},\ i = 0, \ldots, n,\ \forall n \in \mathbb{N}^*.$$

Au regard de l'estimation (6.15), 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 lié au comportement de M_{n+1} lorsque n augmente. En effet, si

$$\lim_{n\to+\infty} \frac{M_{n+1}}{(n+1)!} \max_{x\in[a,b]} |\omega_{n+1}(x)| = 0,$$

on en déduit immédiatement que

$$\lim_{n\to+\infty} \max_{x\in[a,b]} |f(x) - \Pi_n f(x)| = 0,$$

c'est-à-dire que la suite des polynômes d'interpolation de la fonction f associés à des nœuds équirépartis sur l'intervalle $[a,b]$ converge vers f quand n tend vers l'infini, uniformément sur $[a,b]$.

Malheureusement, il existe des fonctions pour lesquelles la quantité $M_{n+1} \max_{x\in[a,b]} |\omega_{n+1}(x)|$ tend vers l'infini *plus rapidement* que $(n+1)!$ lorsque n tend vers l'infini. Un exemple célèbre d'un tel cas « pathologique » est celui dû à Runge, dans lequel on considère le polynôme d'interpolation avec nœuds équirépartis de la fonction

$$f(x) = \frac{1}{1+x^2}$$

sur l'intervalle $[-5,5]$. Les valeurs du maximum de la valeur absolue de l'erreur d'interpolation pour f en fonction du degré d'interpolation sont présentées dans la table 6.1 pour des valeurs paires de n allant de 2 à 24. On observe une croissance exponentielle de l'erreur avec n. La figure 6.3 représente 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 permet de mettre en évidence le phénomène de divergence de l'interpolation au voisinage des extrémités de l'intervalle.

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 – 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 « manque » de régularité de la fonction f, qui est de classe $\mathscr{C}^\infty$ sur $\mathbb{R}$ et dont toutes les dérivées sont bornées sur $[-5,5]$, mais est lié au fait que la série de Taylor de la fonction d'une variable complexe $z \mapsto f(z)$ n'est convergente que dans le disque ouvert de centre $z = 0$ et de rayon égal à 1, la fonction possédant deux pôles sur l'axe imaginaire en $z = \pm i$.

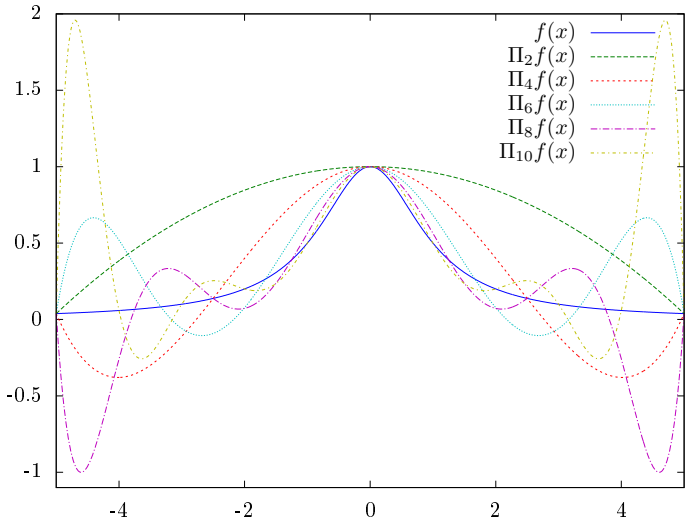

FIGURE 6.3 – Graphes de la fonction de Runge $f(x) = \dfrac{1}{1+x^2}$ et de cinq de ses polynômes d'interpolation à nœuds équirépartis sur l'intervalle $[-5, 5]$.

Notons qu'un choix convenable des nœuds d'interpolation permet d'établir un résultat de convergence uniforme du polynôme d'interpolation $\Pi_n f$ vers toute fonction f continue. C'est le cas notamment avec les *points de Tchebychev*[7] (voir le tableau 6.2 et la figure 6.4).

degré n	$\max_{x\in[-5,5]} \lvert f(x) - \Pi_n f(x)\rvert$
2	0,6006
4	0,2017
6	0,15602
8	0,17083
10	0,10915
12	0,06921
14	0,0466
16	0,03261
18	0,02249
20	0,01533
22	0,01036
24	0,00695

TABLE 6.2 – Erreur d'interpolation de Lagrange utilisant les points de Tchebychev 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 $[-5, 5]$.

6.2 Interpolation polynomiale par morceaux

Jusqu'à présent, nous n'avons attaqué le problème de l'approximation sur un intervalle $[a, b]$ d'une fonction f par l'interpolation de Lagrange qu'en un sens *global*, c'est-à-dire en cherchant à n'utiliser qu'une seule expression analytique de l'interpolant (un seul polynôme) sur $[a, b]$. Pour obtenir une approximation

7. Pafnouti Lvovitch Tchebychev (Пафну́тий Льво́вич Чебышёв, 4 mai 1821 - 26 novembre 1894) était un mathématicien russe. Il est connu pour ses travaux dans le domaine des probabilités et des statistiques.

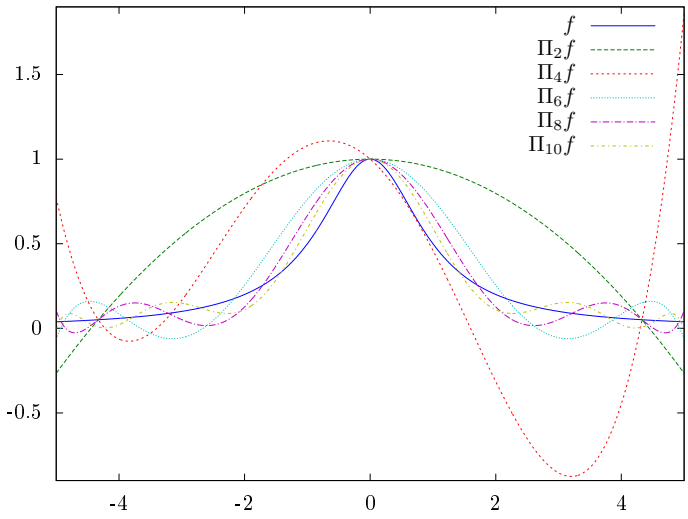

FIGURE 6.4 – Graphes de la fonction de Runge $f(x) = \dfrac{1}{1+x^2}$ et de cinq de ses polynômes d'interpolation aux points de Tchebychev sur l'intervalle $[-5, 5]$.

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 6.1.4 montre que la convergence uniforme de $\Pi_n f$ 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 on emploie une interpolation polynomiale de bas degré. On parle alors d'*interpolation polynomiale par morceaux*. L'idée naturelle suivie ici est que toute fonction peut être approchée de manière arbitrairement précise par des polynômes de degré un (ou même zéro) *sur des intervalles suffisamment petits*.

Dans toute cette section, on désigne par $[a, b]$ un intervalle non vide de $\mathbb{R}$ et par f une application de $[a, b]$ dans $\mathbb{R}$. On considère également $n+1$ nœuds x_j, $j = 0, \ldots, n$, tels que $a = x_0 < x_1 < \cdots < x_n = b$, réalisant une partition de $[a, b]$ en n sous-intervalles. 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*.

6.2.1 Interpolation de Lagrange par morceaux

On peut par exemple utiliser l'interpolation de Lagrange de bas degré (y compris avec des nœuds équirépartis) sur chaque sous-intervalle. DEVELOPPER

Soit une partition τ_h de $[a, b]$ en N sous-intervalles $I_j = [x_j, x_{j+1}]$ de longueur h_j, avec $h = \max\limits_{0 \le j \le N-1} h_j$, tels que $[a, b] = \bigcup_{j=0}^{N-1} I_j$ + interpolation de Lagrange sur chaque intervalle I_j en $n+1$ nœuds équirépartis $\{x_j^{(i)}\}_{0 \le i \le n}$ avec n petit. Pour $n \ge 1$ et τ_h donnée, on introduit

$$X_h^n = \left\{ v \in \mathscr{C}^0([a, b]),\ v_{|_{I_j}} \in \mathbb{P}_n(I_j),\ \forall I_j \in \tau_h \right\}$$

qui est l'espace des fonctions continues sur $[a, b]$ dont la restriction à chaque sous-intervalle I_j est polynomiale de degré inférieur ou égal à n. Pour toute fonction f continue sur $[a, b]$, le polynôme d'interpolation par morceaux $\Pi_h^n f$ coïncide sur chaque I_j avec l'interpolant de $f_{|_{I_j}}$ aux $n+1$ nœuds $x_j^{(i)}$ $0 \le i \le n$. Par conséquent, si f est de classe $\mathscr{C}^{n+1}$ sur $[a, b]$, on obtient en utilisant (6.15) dans chaque sous-intervalle

$$\|f - \Pi_h^n f\|_\infty \le C\, h^{n+1} \|f^{(n+1)}\|_\infty,$$

où $\|f\|_\infty = \max\limits_{x \in [a,b]} |f(x)|$.

6.2.2 Splines d'interpolation

Suli chap. 11 p. 303, quarteroni page 277

Définition 6.8 *Soit $k \in \mathbb{N}^*$. La fonction s_k sur l'intervalle $[a,b]$ est une* ***spline de degré*** *k relative aux nœuds $\{x_j\}_{j=0,\ldots,n}$ si elle vérifie*

$$s_{k|_{[x_j,x_{j+1}]}} \in \mathbb{P}_k,\ j = 0,\ldots,n-1, \tag{6.17}$$

$$s_k \in \mathscr{C}^{k-1}([a,b]). \tag{6.18}$$

Il ressort de cette définition que tout polynôme de degré k est une spline, mais une spline est en général, et même quasiment toujours en pratique, constituée de polynômes différents sur chaque sous-intervalle $[x_j, x_{j+1}]$, $j = 0,\ldots,n$, et la dérivée $k^{\text{ième}}$ de la spline s_k peut donc présenter une discontinuité en chacun des *nœuds internes* $x_1,\ldots,x_n$. Un nœud en lequel se produit une telle discontinuité est appelé un *nœud actif*.

On constate aussi que les conditions (6.17) et (6.18) ne suffisent pas pour caractériser une spline d'interpolation de degré k. En effet, la condition (6.17) signifie que la restriction $s_{k,j} = s_{k|_{[x_j,x_{j+1}]}}$ de la spline s_k au sous-intervalle $[x_j, x_{j+1}]$ peut s'écrire

$$s_{k,j}(x) = \sum_{i=0}^{k} s_{ij}(x - x_j)^i,\ \forall x \in [x_j, x_{j+1}],\ \forall j \in \{0,\ldots,n-1\},$$

et il faut donc déterminer les $(k+1)n$ coefficients s_{ij}, $i = 0,\ldots,k$, $j = 0,\ldots,n-1$. La condition (6.18) se traduit par

$$s_{k,j}^{(m)}(x_j) = s_{k,j}^{(m)}(x_j),\ j = 1,\ldots,n-1,\ m = 0,\ldots,k-1,$$

ce qui revient à obtenir $(n-1)k$ équations sur ces coefficients. Par ailleurs, une spline d'interpolation de la fonction f doit aussi vérifier

$$s_k(x_j) = f(x_j),\ j = 0,\ldots,n,$$

et il ne reste donc que $(k+1)n - (n-1)k - (n+1) = k-1$ contraintes à imposer. Le choix (arbitraire) de ces dernières définit alors le type de splines d'interpolation utilisé :

1. soit $s_k^{(m)}(a) = s_k^{(m)}(b)$, $m = 0,\ldots,k-1$, et l'on parle de splines *périodiques*,
2. soit, pour $k = 2l-1$, $l \geq 2$, $s_k^{(l+j)}(a) = s_k^{(l+j)}(b) = 0$, $j = 0,\ldots,l-2$, et les splines sont dites *naturelles*.

Parmi les deux cas traités, on retiendra principalement celui des *splines d'interpolation cubiques*, qui sont les splines de plus petit degré permettant d'obtenir une approximation de classe $\mathscr{C}^2$ de la fonction interpolée.

Splines d'interpolation linéaires

contruction + base (fonctions chapeau)

Splines d'interpolation cubiques

Les *splines d'interpolation cubiques* sont les splines de plus petit degré permettant une approximation de classe $\mathscr{C}^2$. contruction + bases (quarteroni + evocation B-splines ?) + resultat de minimisation (rapport au nom de spline voir suli page 311)

Pour aller plus loin

REPRENDRE Lorsque la fonction f différentiable, on peut généraliser l'interpolation de Lagrange pour prendre en compte, en plus des valeurs nodales de f, les valeurs de ses dérivées en tout ou partie des nœuds. On parle alors d'*interpolation de Hermite*[8]

8. 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.

Le lecteur intéressé par le phénomène de Runge pourra consulter l'article de J. F. Epperson [Epp87], dans lequel l'auteur aborde de manière simple ce problème de divergence sous l'angle de l'analyse complexe.

L'article [BP70] présente une méthode de résolution efficace des systèmes de Vandermonde, géréralisée dans [Hig88] à des systèmes plus généraux.

références pour les splines [de 01]

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* (voir par exemple [Cia78]), qui, tout comme les méthodes spectrales, est très utilisée pour la résolution numérique des équations aux dérivées partielles.

Références du chapitre

[BP70] Å. Bjork and V. Pereyra. Solution of Vandermonde systems of equations. *Math. Comp.*, 24(112) :893–903, 1970.

[Cia78] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 4 of *Studies in mathematics and its applications*. North-Holland Publishing Company, 1978.

[de 01] C. de Boor. *A practical guide to splines*, volume 27 of *Applied mathematical sciences*. Springer Verlag, revised edition, 2001.

[Epp87] J. F. Epperson. On the Runge example. *Amer. Math. Monthly*, 94(4) :329–341, 1987.

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

[Hig88] N. J. Higham. Fast solution of Vandermonde-like systems involving orthogonal polynomials. *IMA J. Numer. Anal.*, 8(4) :473–486, 1988.

[SB02] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Springer-Verlag, third edition, 2002.

[Tre00] L. N. Trefethen. *Spectral methods in* MATLAB. SIAM, 2000.

Chapitre 7

Intégration numérique

Le calcul d'une intégrale définie de la forme

$$I(f) = \int_a^b f(x)\,\mathrm{d}x,$$

où f est une fonction continue sur l'intervalle borné $[a, b]$ à valeurs dans $\mathbb{R}$, est un problème classique intervenant dans de nombreux domaines, qu'ils soient scientifiques ou non. Cette évaluation peut cependant s'avérer difficile en pratique, même lorsque l'on dispose d'une expression analytique de l'intégrale, voire impossible (c'est le cas par exemple lorsque la fonction f est la solution d'une équation différentielle qu'on ne sait pas explicitement résoudre ou bien lorsqu'on ne connaît pas de primitive de f, même en ayant recours à des techniques de changement de variable ou d'intégration par parties). Dans ce chapitre, nous introduisons des *formules de quadrature*, qui consistent à approcher la valeur de l'intégrale par une somme pondérée finie de valeurs de la fonction f en des points choisis ; en d'autres mots, ces formules fournissent une approximation de $I(f)$ par la quantité

$$I_n(f) = \sum_{i=0}^{n} \alpha_i\, f(x_i), \tag{7.1}$$

avec $n \geq 0$, les coefficients $\{\alpha_i\}_{i=0,\ldots,n}$ étant réels et dépendant de l'entier n et les points $\{x_i\}_{i=0,\ldots,n}$ appartenant à $[a, b]$. Nous limitons notre exposé aux formules de Newton–Cotes [1], qui sont un cas particulier de *formules de quadrature interpolatoires*.

7.1 Quelques généralités sur les formules de quadrature

Dans l'expression (7.1), les points x_i et les coefficients α_i, $i = 0, \ldots, n$, sont respectivement appelés *nœuds* et *poids* de la formule de quadrature.

Comme pour les problèmes d'interpolation étudiés au chapitre précédent, la précision d'une formule de quadrature pour une fonction $f \in \mathscr{C}^0([a, b])$ donnée se mesure notamment en évaluant l'*erreur de quadrature*

$$E_n(f) = I(f) - I_n(f).$$

Pour toute formule de quadrature, on définit par ailleurs son *degré d'exactitude* comme le plus grand entier $r \geq 0$ pour lequel

$$I(f) = I_n(f),\ \forall f \in \mathbb{P}_m,\ \forall m \in \{0, \ldots, r\}.$$

Enfin, une *formule de quadrature interpolatoire* est obtenue en remplacant la fonction f dans l'intégrale par son polynôme d'interpolation (de Lagrange ou de Hermite selon les cas). On a le résultat suivant.

1. 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.

Théorème 7.1 *Soit n un entier positif. Toute formule de quadrature interpolatoire à $n+1$ nœuds a un degré d'exactitude au moins égal à n, et réciproquement.*

DÉMONSTRATION. Montrons tout d'abord l'assertion. Si la fonction f appartient à $\mathbb{P}_n$, alors $\Pi_n f = f$, où $\Pi_n f$ désigne le polynôme d'interpolation.... Par définition, on a alors

$$I_n(f) = \int_a^b \Pi_n f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x = I(f).$$

reciproque dans IK66 □

7.2 Formules de Newton–Cotes

Les formules de quadrature de Newton–Cotes sont basées sur l'interpolation de Lagrange à nœuds équirépartis dans l'intervalle $[a,b]$. Pour n un entier positif fixé, notons $x_i = x_0 + ih$, $i = 0,\ldots,n$, les nœuds de quadrature. Il existe deux types de formules de Newton–Cotes :

- les formules *fermées*, 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 formules bien connues du *trapèze* $(n=1)$ et de *Simpson*[2] $(n=2)$ sont des cas particuliers,
- les formules *ouvertes*, pour lesquelles $x_0 = a+h$, $x_n = b-h$ et $h = \frac{b-a}{n+2}$ $(n \geq 0)$, auxquelles appartient la formule du point milieu $(n=0)$.

Soit f un fonction continue sur l'intervalle $[a,b]$. Une fois fixé l'ensemble des nœuds $\{x_i\}_{i=0,\ldots,n}$, la formule est obtenue en construisant le polynôme d'interpolation de Lagrange $\Pi_n f$ de f, puis en posant

$$I_n(f) = \int_a^b \Pi_n f(x)\,\mathrm{d}x.$$

On a alors

$$I_n(f) = \int_a^b \left(\sum_{i=0}^n f(x_i)\,l_i(x)\right)\mathrm{d}x = \sum_{i=0}^n f(x_i)\int_a^b l_i(x)\,\mathrm{d}x,$$

et, en identifiant, on trouve que les poids sont donnés par les intégrales de polynômes de Lagrange sur l'intervalle $[a,b]$

$$\alpha_i = \int_a^b l_i(x)\,\mathrm{d}x,\ i = 0,\ldots,n.$$

Une propriété intéressante est que ces poids ne dépendent explicitement que de n et h et pas de l'intervalle d'intégration $[a,b]$.

Preuve dans Quarteroni page 299

quelques tables

Présentons maintenant plus en détails quelques cas particuliers des formules de quadrature de Newton–Cotes.

Formule du point milieu. Cette formule (aussi appelée *formule du rectangle*) est obtenue en remplaçant la fonction f par la valeur qu'elle prend au milieu de l'intervalle $[a,b]$ (voir la figure 7.1), d'où

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right). \tag{7.2}$$

Le poids de quadrature vaut donc $\alpha_0 = b-a$ et le nœud est $x_0 = \dfrac{a+b}{2}$.

En supposant la fonction f de classe $\mathscr{C}^2$ sur $[a,b]$, on peut utiliser le théorème 7.2 pour montrer que l'erreur de quadrature de cette formule vaut

$$E_0(f) = -\frac{f''(c)}{24}(b-a)^3,\ c \in]a,b[.$$

Son degré d'exactitude est par conséquent égal à 1.

2. 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.

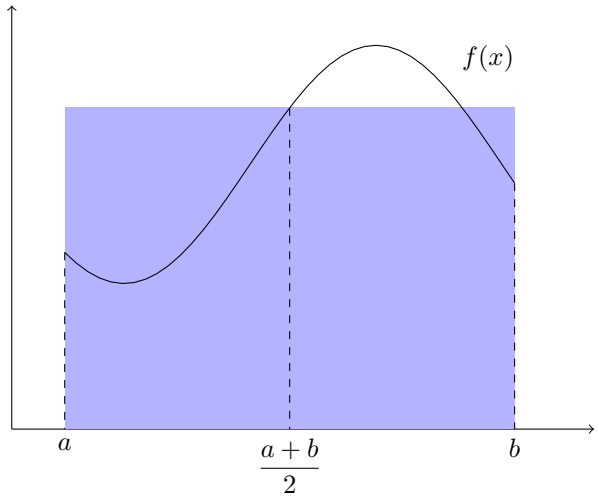

FIGURE 7.1 – Formule du point milieu. La valeur approchée de l'intégrale $I(f)$ correspond à l'aire colorée en bleu.

Formule du trapèze. On obtient cette formule en remplaçant la fonction f par son polynôme d'interpolation de Lagrange de degré un aux points a et b (voir la figure 7.2). On alors

$$I_2(f) = \frac{b-a}{2}\left[f(a) + f(b)\right]. \tag{7.3}$$

Les poids de quadrature valent $\alpha_0 = \alpha_1 = \dfrac{b-a}{2}$, tandis que les nœuds sont $x_0 = a$ et $x_1 = b$.

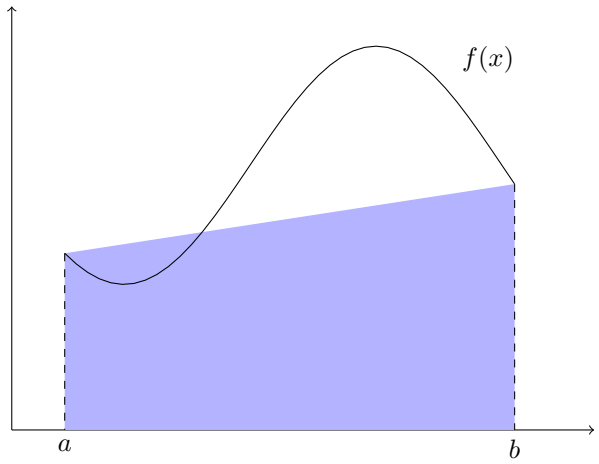

FIGURE 7.2 – Formule du trapèze. La valeur approchée de l'intégrale $I(f)$ correspond à l'aire colorée en bleu.

En supposant f de classe $\mathscr{C}^2$ sur $[a,b]$, on obtient la valeur suivante pour l'erreur de quadrature

$$E_1(f) = -\frac{f''(c)}{12}(b-a)^3,\ c \in]a,b[.$$

et l'on en déduit que cette formule à un degré d'exactitude égal à 1.

Formule de Simpson. Cette dernière formule est obtenue en substituant 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

$$I_1(f) = \frac{b-a}{6}\left[f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{7.4}$$

Les poids de quadrature sont donnés par $\alpha_0 = \alpha_2 = \dfrac{b-a}{6}$ et $\alpha_1 = 2\dfrac{b-a}{3}$.

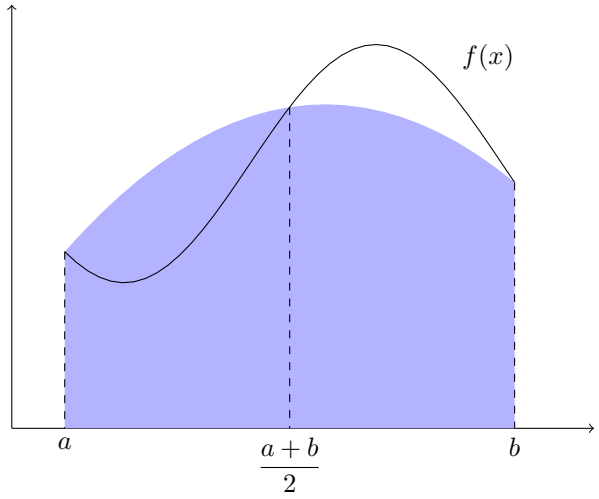

FIGURE 7.3 – Formule de Simpson. La valeur approchée de l'intégrale $I(f)$ correspond à l'aire colorée en bleu.

On montre, grâce au théorème 7.2, que, si la fonction f est de classe $\mathscr{C}^4$ sur l'intervalle $[a, b]$, l'erreur de quadrature peut s'écrire

$$E_2(f) = -\frac{f^{(4)}(c)}{2880}\,(b-a)^5,\ c \in]a, b[.$$

Cette formule a donc un degré d'exactitude égal à 3.

7.3 Estimations d'erreur

Démontrons maintenant le résultat utilisé pour l'obtention des estimations des erreurs de quadrature des formules du point milieu, du trapèze et de Simpson fournies plus haut.

Théorème 7.2 *pour les formules fermées*

$$E_n(f) = \begin{cases} \dfrac{M_n}{(n+2)!}\,f^{(n+2)}(c),\ M_n = \displaystyle\int_a^b t\,\omega_{n+1}(t)\,\mathrm{d}t < 0, & \textit{si } n \textit{ est pair,} \\ \dfrac{M_n}{(n+1)!}\,f^{(n+1)}(c),\ M_n = \displaystyle\int_a^b \omega_n(t)\,\mathrm{d}t < 0, & \textit{si } n \textit{ est impair,} \end{cases}$$

avec $a < c < b$.

pour les formules ouvertes

$$E_n(f) = \begin{cases} \dfrac{M'_n}{(n+2)!}\,f^{(n+2)}(c),\ M'_n = \displaystyle\int_a^b t\,\omega_{n+1}(t)\,\mathrm{d}t > 0, & \textit{si } n \textit{ est pair,} \\ \dfrac{M'_n}{(n+1)!}\,f^{(n+1)}(c),\ M'_n = \displaystyle\int_a^b \omega_n(t)\,\mathrm{d}t > 0, & \textit{si } n \textit{ est impair,} \end{cases}$$

avec $a < c < b$. *Le degré d'exactitude des formules de Newton–Cotes est donc égal à* $n+1$ *lorsque* n *est pair et* n *lorsque* n *est impair.*

DÉMONSTRATION. a ecrire IK 308-314 □

7.4 Formules de quadrature composites

pendant de l'interpolation par morceaux
Quarteroni page 304

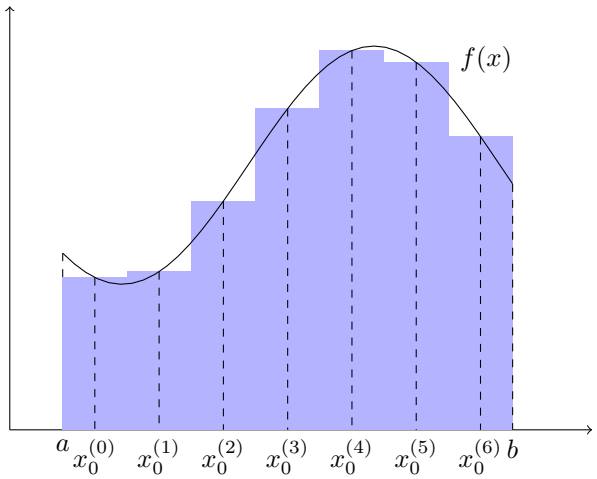

FIGURE 7.4 – Formule du point milieu composite à sept sous-intervalles sur $[a, b]$. La valeur approchée de l'intégrale $I(f)$ correspond à l'aire colorée en bleu.

Pour aller plus loin

formules de quadrature de Gauss (lien avec interpolation de Hermite et polynômes orthogonaux), Si on cherche a utiliser les extrémité de l'intervalle : une, formules de Radau, les deux, Lobatto
Clenshaw–Curtis
méthode de Romberg

References dans Suli p. 230 pdf

[DR84] P. J. Davis and P. Rabinowitz. *Methods of numerical integration*. Computer sciences and applied mathematics. Academic Press, second edition, 1984.

Annexe A

Rappels et compléments d'algèbre linéaire et d'analyse matricielle

On rappelle dans cette annexe un certain nombre de résultats relatifs aux espaces vectoriels de dimension *finie* et aux matrices, qui sont utilisés dans l'ensemble du cours. La plupart des notions abordées sont supposées déjà connues du lecteur, à l'exception peut-être des normes matricielles.

Dans toute la suite, on désigne par $\mathbb{K}$ le corps des *scalaires*, avec $\mathbb{K} = \mathbb{R}$, le corps des nombres réels, ou bien $\mathbb{K} = \mathbb{C}$, le corps des nombres complexes.

A.1 Espaces vectoriels

Nous commençons par rappeler la notion d'*espace vectoriel*.

Définition A.1 *Un* ***espace vectoriel sur*** $\mathbb{K}$ *est un ensemble non vide E sur lequel est définie une loi interne notée $+$, appelée* ***addition****, et une loi externe notée*[1] $\cdot$*, appelée* ***multiplication par un scalaire****, possédant les propriétés suivantes :*

1. $(E, +)$ *est un groupe commutatif (ou abélien),*
2. $\forall (\lambda, \mu) \in \mathbb{K}^2$ *et* $\forall \boldsymbol{v} \in E$, $(\lambda + \mu)\, x = \lambda\, \boldsymbol{v} + \mu\, \boldsymbol{v}$,
3. $\forall (\lambda, \mu) \in \mathbb{K}^2$ *et* $\forall \boldsymbol{v} \in E$, $\lambda(\mu\, \boldsymbol{v}) = (\lambda\mu)\, \boldsymbol{v}$,
4. $\forall \lambda \in \mathbb{K}$ *et* $\forall (\boldsymbol{v}, \boldsymbol{w}) \in E^2$, $\lambda(\boldsymbol{v} + \boldsymbol{w}) = \lambda\, \boldsymbol{v} + \lambda\, \boldsymbol{v}$,
5. $\forall \boldsymbol{v} \in E$, $1_{\mathbb{K}}\, \boldsymbol{v} = \boldsymbol{v}$,

le scalaire $1_{\mathbb{K}}$ étant l'élément unitaire du corps $\mathbb{K}$. Les éléments de l'espace vectoriel E sont appelés ***vecteurs****.*

Définition A.2 *On dit qu'une partie non vide F d'un espace vectoriel E est un* ***sous-espace vectoriel*** *de E si et seulement si*

$$\forall (\boldsymbol{v}, \boldsymbol{w}) \in F^2,\ \forall (\lambda, \mu) \in \mathbb{K}^2,\ \lambda\, \boldsymbol{v} + \mu\, \boldsymbol{w} \in F.$$

En particulier, l'ensemble des combinaisons linéaires d'une famille $\{\boldsymbol{v}_i\}_{i=1,\dots,p}$ de p vecteurs de E est un sous-espace vectoriel de E, appelé *sous-espace engendré* par la famille de vecteurs. On le note

$$\mathrm{Vect}\{\boldsymbol{v}_1, \dots,\ \boldsymbol{x}_p\} = \{\boldsymbol{v} = \lambda_1\, \boldsymbol{v}_1 + \cdots + \lambda_p\, \boldsymbol{v}_p,\ \text{avec } \lambda_i \in \mathbb{K},\ i = 1, \dots, p\}\,.$$

La famille $\{\boldsymbol{v}_i\}_{i=1,\dots,p}$ est alors appelée *famille génératrice* de ce sous-espace.

Définition A.3 *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****.*

1. Dans la pratique, on omet souvent d'écrire la symbole « $\cdot$ ». C'est ce que nous faisons ici.

Dans toute la suite, nous ne considérons que des espaces vectoriels de dimension finie.

Définitions A.4 *Une famille de vecteurs $\{\boldsymbol{v}_i\}_{i=1,\dots,p}$ d'un espace vectoriel E est dite* ***libre*** *si les vecteurs $\boldsymbol{v}_1,\dots,\boldsymbol{v}_p$ sont* ***linéairement indépendants****, c'est-à-dire si la relation*

$$\lambda_1\,\boldsymbol{v}_1+\cdots+\lambda_p\,\boldsymbol{v}_p=\boldsymbol{0},$$

où $\boldsymbol{0}$ 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****.*

On appelle *base* de l'espace vectoriel E toute famille libre et génératrice de E. Si la famille $\{\boldsymbol{e}_i\}_{i=1,\dots,n}$ est une base de E, tout vecteur de E admet une décomposition unique de la forme

$$\boldsymbol{v}=\sum_{i=1}^{n}v_i\,\boldsymbol{e}_i,\ \forall\boldsymbol{v}\in E,$$

les scalaires v_i, $i=1,\dots,n$, étant appelés les *composantes* du vecteur $\boldsymbol{v}$ dans la base $\{\boldsymbol{e}_i\}_{i=1,\dots,n}$. On a de plus les résultats suivants.

Théorème A.5 *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 $\mathscr{G}=\{\boldsymbol{g}_1,\dots,\boldsymbol{g}_n\}$ est une famille génératrice de E et si $\mathscr{F}=\{\boldsymbol{f}_1,\dots,\boldsymbol{f}_n,\boldsymbol{f}_{n+1}\}$ 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 $\{a_{ij}\}_{i=1,\dots,n+1,\ j=1,\dots,n}$ de scalaires telle que

$$\begin{array}{ccccccccc}
\boldsymbol{f}_1 & = & a_{11}\,\boldsymbol{g}_1 & + & \dots & + & a_{1n-1}\,\boldsymbol{g}_{n-1} & + & a_{1n}\,\boldsymbol{g}_n,\\
\boldsymbol{f}_2 & = & a_{21}\,\boldsymbol{g}_1 & + & \dots & + & a_{2n-1}\,\boldsymbol{g}_{n-1} & + & a_{2n}\,\boldsymbol{g}_n,\\
\vdots & & \vdots & & & & \vdots & & \vdots\\
\boldsymbol{f}_n & = & a_{n1}\,\boldsymbol{g}_1 & + & \dots & + & a_{nn-1}\,\boldsymbol{g}_{n-1} & + & a_{nn}\,\boldsymbol{g}_n,\\
\boldsymbol{f}_{n+1} & = & a_{n+11}\,\boldsymbol{g}_1 & + & \dots & + & a_{n+1n-1}\,\boldsymbol{g}_{n-1} & + & a_{n+1n}\,\boldsymbol{g}_n.
\end{array}$$

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 $\mathrm{Vect}\{\boldsymbol{g}_i\}_{i=1,\dots,n-1}$; de l'hypothèse de récurrence, on déduit que la famille $\{\boldsymbol{f}_i\}_{i=1,\dots,n}$ 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 $\mathrm{Vect}\{\boldsymbol{g}_i\}_{i=1,\dots,n-1}$. Par hypothèse de récurrence, la famille $\{\boldsymbol{h}_1,\dots,\boldsymbol{h}_n\}$ est liée : il existe des scalaires $\lambda_1,\dots,\lambda_n$ 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. □

Corollaire A.6 *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 $\mathscr{B}'$. □

Définition A.7 *Le cardinal d'une base quelconque d'un espace vectoriel E de dimension finie s'appelle la* ***dimension de*** *E et se note* $\dim E$.

A.2 Matrices

Soit m et n deux entiers strictement positifs. On appelle *matrice* à m lignes et n colonnes à coefficients dans $\mathbb{K}$ un ensemble A de mn scalaires a_{ij} de $\mathbb{K}$, $i=1,\dots,m$, $j=1,\dots,n$, présentés dans le tableau rectangulaire suivant

$$A=\begin{pmatrix}
a_{11} & a_{12} & \dots & a_{1n}\\
a_{21} & a_{22} & \dots & a_{2n}\\
\vdots & \vdots & & \vdots\\
a_{m1} & a_{m2} & \dots & a_{mn}
\end{pmatrix}.$$

Les scalaires a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, sont appelés *coefficients*, ou *éléments*, de la matrice A, le premier indice i étant celui de la ligne de l'élément et le second j étant celui de la colonne. Ainsi, l'ensemble des coefficients $a_{i1}, \ldots, a_{in}$ est la $i^{\text{ième}}$ *ligne* de la matrice et l'ensemble $a_{1j}, \ldots, a_{mj}$ est la $j^{\text{ième}}$ *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 $M_{m,n}(\mathbb{K})$ l'ensemble des matrices à m lignes et n 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 et on note $M_n(\mathbb{K})$ l'ensemble correspondant. Lorsque $m \neq n$, on parle de matrice *rectangulaire*.

On appelle *diagonale* d'une matrice A d'ordre n l'ensemble des coefficents a_{ii}, $i = 1, \ldots, n$. 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 $A \in M_{m,n}(\mathbb{R})$, on note $A^T \in M_{n,m}(\mathbb{R})$ la *matrice transposée*[2] de A telle que

$$(A^T)_{ij} = (A)_{ji},\ 1 \leq i \leq n,\ 1 \leq j \leq m.$$

On a alors $(A^T)^T = A$. De même, étant donné $A \in M_{m,n}(\mathbb{C})$, on note $A^* \in M_{n,m}(\mathbb{C})$ la *matrice adjointe* de A telle que

$$(A^*)_{ij} = \overline{(A)_{ji}},\ 1 \leq i \leq n,\ 1 \leq j \leq m,$$

le scalaire $\overline{z}$ désignant le nombre complexe conjugué du nombre z, et on $(A^*)^* = A$.

On appelle *vecteur ligne* (resp. *vecteur colonne*) 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 $\boldsymbol{v}$ dans la base $\{\boldsymbol{e}_i\}_{i=1\ldots,n}$ par

$$\boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

et que le *vecteur tranposé* $\boldsymbol{v}^T$ (resp. *vecteur adjoint* $\boldsymbol{v}^*$) de $\boldsymbol{v}$ sera alors représenté par le vecteur ligne suivant

$$\boldsymbol{v}^T = \begin{pmatrix} v_1 & v_2 & \ldots & v_n \end{pmatrix} \text{ (resp. } \boldsymbol{v}^* = \begin{pmatrix} \overline{v_1} & \overline{v_2} & \ldots & \overline{v_n} \end{pmatrix}).$$

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*.

Définition A.8 ***(sous-matrice)*** *Soit A une matrice de $M_{m,n}(\mathbb{K})$. Soient $1 \leq i_1 < \cdots < i_p \leq m$ et $1 \leq j_1 < \cdots < j_q \leq n$ 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*** *de A.*

Il est aussi très courant d'associer à une matrice une décomposition en sous-matrices.

Définition A.9 ***(décomposition par blocs d'une matrice)*** *Une matrice A de $M_{m,n}(\mathbb{K})$ est dite* ***décomposée par blocs*** *si elle s'écrit*

$$A = \begin{pmatrix} A_{11} & A_{12} & \ldots & A_{1N} \\ A_{21} & A_{22} & \ldots & A_{2N} \\ \vdots & \vdots & & \vdots \\ A_{M1} & A_{M2} & \ldots & A_{MN} \end{pmatrix},$$

où les ***blocs*** *A_{IJ}, $1 \leq I \leq M$, $1 \leq J \leq N$, sont des sous-matrices de A.*

L'intérêt de telles décompositions par blocs réside dans le fait que certaines opérations définies sur les matrices restent formellement les mêmes, les coefficients de la matrice étant remplacés par ses sous-matrices.

2. On peut aussi définir la matrice transposée d'une matrice complexe, mais cette notion n'a en général que peu d'intérêt dans ce cas.

A.2.1 Opérations sur les matrices

Nous rappelons à présent quelques opérations essentielles définies sur les matrices.

Définition A.10 ***(égalité de matrices)*** *Soit A et B deux matrices de $M_{m,n}(\mathbb{K})$. On dit que A est* ***égale*** *à B si $a_{ij} = b_{ij}$ pour $i = 1, \ldots, m$, $j = 1, \ldots, n$.*

Définition A.11 ***(somme de deux matrices)*** *Soit A et B deux matrices de $M_{m,n}(\mathbb{K})$. On appelle* ***somme*** *des matrices A et B la matrice C de $M_{m,n}(\mathbb{K})$ dont les coefficients sont $c_{ij} = a_{ij} + b_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.*

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

$$(A + B)^T = A^T + B^T \text{ et } (A + B)^* = A^* + B^*, \ \forall A, B \in M_{m,n}(\mathbb{K}).$$

Définition A.12 ***(multiplication d'une matrice par un scalaire)*** *Soit A une matrice de $M_{m,n}(\mathbb{K})$ et λ un scalaire. Le résultat de la* ***multiplication de la matrice*** *A* ***par le scalaire*** *λ est la matrice C de $M_{m,n}(\mathbb{K})$ dont les coefficients sont $c_{ij} = \lambda\, a_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.*

On a

$$(\alpha\, A)^T = \alpha\, A^T \text{ et } (\alpha\, A)^* = \overline{\alpha}\, A^*, \ \forall \alpha \in \mathbb{K}, \ \forall A \in M_{m,n}(\mathbb{K}).$$

Muni des deux dernières opérations, l'ensemble $M_{m,n}(\mathbb{K})$ est un espace vectoriel sur $\mathbb{K}$ (la vérification est laissée en exercice). On appelle alors *base canonique de $M_{m,n}(\mathbb{K})$* l'ensemble des mn matrices E_{kl}, $k = 1, \ldots, m$, $l = 1, \ldots, n$, de $M_{m,n}(\mathbb{K})$ dont les éléments sont définis par

$$(E_{kl})_{ij} = \begin{cases} 0 & \text{si } i \neq k \text{ ou } j \neq l \\ 1 & \text{si } i = k \text{ et } j = l \end{cases}, \ 1 \leq i \leq m, \ 1 \leq j \leq n.$$

Définition A.13 ***(produit de deux matrices)*** *Soit A une matrice de $M_{m,p}(\mathbb{K})$ et B une matrice de $M_{p,n}(\mathbb{K})$. Le* ***produit*** *des matrices A et B est la matrice C de $M_{m,n}(\mathbb{K})$ dont les coefficients sont donnés par $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{jk}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.*

Le produit de matrices est associatif et distributif par rapport à la somme de matrices, mais il n'est pas commutatif en général. Dans le cas de matrices carrées, on dit que deux matrices A et B *commutent* 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*, définie par

$$I_n = (\delta_{ij})_{1 \leq i,j \leq n},$$

avec δ_{ij} le *symbole de Kronecker*[3],

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

Cette matrice est, par définition, la seule matrice d'ordre n telle que $AI_n = I_n A = A$ pour toute matrice A d'ordre n. Muni de la multiplication par un scalaire, de la somme et du produit de matrice l'ensemble $M_n(\mathbb{K})$ est une algèbre (la vérification est laissée en exercice).

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 $A^0 = I_n$. On rapelle enfin que l'on a

$$(AB)^T = B^T A^T \text{ et } (AB)^* = B^* A^*, \ \forall A \in M_{m,p}(\mathbb{K}), \ \forall B \in M_{p,n}(\mathbb{K}).$$

Terminons en indiquant que toutes ces opérations peuvent s'étendre au cas de matrices décomposées par blocs, pourvu que la taille de chacun des blocs soit telle que les opérations soient bien définies. On a par exemple le résultat suivant.

3. 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.

Lemme A.14 ***(produit de matrices décomposées par blocs)*** *Soient A et B deux matrices de tailles compatibles pour effectuer le produit AB. Si A admet une décomposition en blocs $(A_{IK})_{1\leq I\leq M,\ 1\leq K\leq N}$ de formats respectifs (r_I, s_K) et B admet une décomposition compatible en blocs $(B_{KJ})_{1\leq K\leq N,\ 1\leq J\leq P}$ de formats respectifs (s_K, t_J), alors le produit $C = AB$ peut aussi s'écrire comme une matrice par blocs $(C_{IJ})_{1\leq I\leq M,\ 1\leq J\leq P}$, de formats respectifs (r_I, t_J) et donnés par*

$$C_{IJ} = \sum_{K=1}^{N} A_{IK} B_{KJ},\ 1 \leq I \leq M,\ 1 \leq J \leq P.$$

A.2.2 Liens entre applications linéaires et matrices

Dans cette 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. Pour cela, quelques rappels sont nécessaires.

Définition A.15 ***(application linéaire)*** *Soient E et F deux espaces vectoriels sur le même corps $\mathbb{K}$ et f une application de E dans F. On dit que f est une* ***application linéaire*** *si*

$$f(\lambda\,\boldsymbol{v} + \boldsymbol{w}) = \lambda\, f(\boldsymbol{v}) + f'\boldsymbol{w}),\ \forall(\boldsymbol{v}, \boldsymbol{w}) \in E^2,\ \forall\lambda \in \mathbb{K}.$$

L'ensemble des applications linéaires de E dans F est noté $\mathscr{L}(E, F)$.

Définitions A.16 *Soit f une application de $\mathscr{L}(E, F)$. On appelle* ***noyau*** *(kernel en anglais) de f, et l'on note* $\operatorname{Ker} f$*, l'ensemble*

$$\operatorname{Ker} f = \{\boldsymbol{x} \in E \,|\, f(\boldsymbol{x}) = \boldsymbol{0}\}\,.$$

On dit que f est ***injective*** *si* $\operatorname{Ker} f = \{\boldsymbol{0}\}$.
On appelle ***image*** *de f, et l'on note* $\operatorname{Im} f$*, l'ensemble*

$$\operatorname{Im} f = \{\boldsymbol{y} \in F \,|\, \exists \boldsymbol{x} \in E,\ \boldsymbol{y} = f\boldsymbol{x})\}\,,$$

et le ***rang*** *de f est la dimension de* $\operatorname{Im} f$*. L'application f est dite* ***surjective*** *si* $\operatorname{Im} f = F$.
Enfin, on dit que f est ***bijective****, ou que c'est un* ***isomorphisme****, si elle est injective et bijective.*

Le résultat suivant permet de relier les dimensions du noyau et de l'image d'une application linéaire.

Théorème A.17 ***(« théorème du rang »)*** *Soit E et F deux espaces vectoriels sur $\mathbb{K}$ de dimension finie. Pour toute application f de $\mathscr{L}(E, F)$, on a*

$$\dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(E).$$

Démonstration. Notons $n = \dim E$. Le sous-espace-vectoriel $\operatorname{Ker} f$ de E admet au moins une base $\{\boldsymbol{e}_i\}_{i=1,\ldots,p}$ que l'on peut compléter en une base $\{\boldsymbol{e}_i\}_{i=1,\ldots,n}$ de E. Nous allons montrer que $\{f(\boldsymbol{e}_{p+1}), \ldots, f(\boldsymbol{e}_n)\}$ est une base de $\operatorname{Im} f$. Les vecteurs $f(\boldsymbol{e}_i)$, $p+1 \leq i \leq n$, sont à l'évidence des éléments de $\operatorname{Im} f$. Soit l'ensemble $\{\lambda_{p+1}, \ldots, \lambda_n\} \in \mathbb{K}^{n-p}$ tel que

$$\sum_{i=p+1}^{n} \lambda_i\, f(\boldsymbol{e}_i) = \boldsymbol{0}.$$

On a alors $f\left(\sum_{i=p+1}^{n} \lambda_i\, \boldsymbol{e}_i\right) = \boldsymbol{0}$, et donc $\sum_{i=p+1}^{n} \lambda_i\, \boldsymbol{e}_i \in \operatorname{Ker} f$.
Il existe donc un ensemble $\{\mu_1, \ldots, \mu_p\} \in \mathbb{K}^p$ tel 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 $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p\}$ est libre, on en déduit que $\lambda_{p+1} = \cdots = \lambda_n = 0$, ce qui montre que $\{f(\boldsymbol{e}_{p+1}), \ldots, f(\boldsymbol{e}_n)\}$ est libre.
Soit maintenant $\boldsymbol{y} \in \operatorname{Im} f$. Il existe $\boldsymbol{x} \in E$ tel que $\boldsymbol{y} = f(\boldsymbol{x})$. Comme $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ engendre E, il existe $\{\alpha_1, \ldots, \alpha_n\} \in \mathbb{K}^n$ tel que $\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i\, \boldsymbol{e}_i$. On a alors

$$\boldsymbol{y} = f(\boldsymbol{x}) = f\left(\sum_{i=1}^{n} \alpha_i\, \boldsymbol{e}_i\right) = \sum_{i=1}^{n} \alpha_i\, f(\boldsymbol{e}_i) = \sum_{i=p+1}^{n} \alpha_i\, f(\boldsymbol{e}_i),$$

puisque les vecteurs $\boldsymbol{e}_i$, $1 \leq i \leq p$, appartiennent au noyau de f. La famille $\{f(\boldsymbol{e}_{p+1}), \ldots, f(\boldsymbol{e}_n)\}$ engendre donc $\mathrm{Im}\, f$ et c'est une base de ce sous-espace de F. On conclut alors

$$\dim(\mathrm{Im}\, f) = n - p = \dim E - \dim(\mathrm{Ker}\, f).$$

□

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 $\{\boldsymbol{e}_i\}_{i=1,\ldots,n}$ une base de E et $\{\boldsymbol{f}_i\}_{i=1,\ldots,m}$ une base de F. Pour toute application linéaire f de E dans F, on peut écrire que

$$f(\boldsymbol{e}_j) = \sum_{i=1}^{m} a_{ij}\, \boldsymbol{f}_i,\ 1 \leq j \leq n, \tag{A.1}$$

ce qui conduit à la définition suivante.

Définition A.18 *On appelle* ***représentation matricielle*** *de l'application linéaire f de $\mathscr{L}(E,F)$, relativement aux bases $\{\boldsymbol{e}_i\}_{i=1,\ldots,n}$ et $\{\boldsymbol{f}_i\}_{i=1,\ldots,m}$, la matrice A de $M_{m,n}(\mathbb{K})$ ayant pour coefficients les scalaires a_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, définis de manière unique par les relations* (A.1).

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 $M_{m,n}(\mathbb{K})$. Cet isomorphisme n'est cependant pas intrinsèque, puisque la représentation matricielle dépend des bases choisies pour E et 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.19 ***(noyau, image et rang d'une matrice)*** *Soit A une matrice de $M_{m,n}(\mathbb{K})$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Le* ***noyau*** *de A est le sous-espace vectoriel de $\mathbb{K}^n$ défini par*

$$\mathrm{Ker}\,(A) = \{\boldsymbol{x} \in \mathbb{K}^n \mid A\boldsymbol{x} = \boldsymbol{0}\}.$$

*L'**image** de A est le sous-espace vectoriel de $\mathbb{K}^m$ défini par*

$$\mathrm{Im}\,(A) = \{\boldsymbol{y} \in \mathbb{K}^n \mid \exists \boldsymbol{x} \in \mathbb{K}^n \text{ tel que } A\boldsymbol{x} = \boldsymbol{y}\},$$

*et le **rang** de A est la dimension de cette image,*

$$\mathrm{rg}(A) = \dim(\mathrm{Im}\,(A)).$$

En vertu du théorème du rang, on a, pour toute matrice de $M_{m,n}(\mathbb{K})$, la relation

$$\dim(\mathrm{Ker}\,(A)) + \mathrm{rg}(A) = n.$$

Enfin, une matrice A de $M_{m,n}(\mathbb{K})$ est dite *de rang maximum* si $\mathrm{rg}(A) = \min(m,n)$.

A.2.3 Inverse d'une matrice

Définitions A.20 *Soit A une matrice d'ordre n. On dit que A est* ***inversible*** *(ou* ***régulière****) s'il existe une (unique) matrice, notée A^{-1}, telle que $AA^{-1} = A^{-1}A = I_n$ (A^{-1} est appelée la* ***matrice inverse*** *de A). Une matrice non inversible est dite* ***singulière****.*

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 $\mathrm{rg}(A) = n$.

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 :

$$(AB)^{-1} = B^{-1}A^{-1},\ (A^T)^{-1} = (A^{-1})^T,\ (A^*)^{-1} = (A^{-1})^* \text{ et } (\alpha A)^{-1} = \frac{1}{\alpha}\, A^{-1},\ \forall \alpha \in \mathbb{K}^*.$$

A.2.4 Trace et déterminant d'une matrice

Nous rappellons dans cette section les notions de *trace* et de *déterminant* d'une matrice carrée.

Définition A.21 ***(trace d'une matrice)*** *La* ***trace*** *d'une matrice A d'ordre n est la somme de ses coefficients diagonaux :*

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii},$$

On montre facilement les relations

$$\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B),\ \mathrm{tr}(AB) = \mathrm{tr}(BA),\ \mathrm{tr}(\alpha A) = \alpha\,\mathrm{tr}(A),\ \forall \alpha \in \mathbb{K},\ \forall A, B \in M_n(\mathbb{K}),$$

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.22 ***(déterminant d'une matrice)*** *On appelle* ***déterminant*** *d'une matrice A d'ordre n le scalaire défini par la formule de Leibniz*[4]

$$\det(A) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{\sigma(i)i},$$

où $\varepsilon(\sigma)$ désigne la signature d'une permutation[5] *σ de $\mathfrak{S}_n$.*

Par propriété des permutations, on a $\det(A^T) = \det(A)$ 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 $\det(A) = \det(A^T)$) de A par un scalaire α multiplie le déterminant par ce scalaire. On a notamment

$$\det(\alpha A) = \alpha^n \det(A),\ \forall \alpha \in \mathbb{K},\ \forall A \in M_n(\mathbb{K}).$$

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 les déterminant en algèbre linéaire.

On rapelle enfin que le déterminant est un *morphisme de 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,

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

4. 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 le calcul intégral et différentiel indépendamment de Newton et introduisit les notations en usage aujourd'hui.

5. Rappelons qu'une *permutation* d'un ensemble est une bijection de cet ensemble dans lui-même. 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

$$\varepsilon(\sigma) = \prod_{1 \leq i < j \leq n} \frac{\sigma(i) - \sigma(j)}{i - j}.$$

Définition A.23 ***(déterminant extrait)*** *Soit A une matrice de $M_{m,n}(\mathbb{K})$ et q un entier strictement positif. On appelle* ***déterminant extrait d'ordre*** *q le déterminant de n'importe quelle matrice d'ordre q obtenue à partir de A en éliminant $m-q$ lignes et $n-q$ colonnes.*

La démonstration du résultat suivant est immédiate.

Proposition A.24 *Le rang d'une matrice A de $M_{m,n}(\mathbb{K})$ est donné par l'ordre maximum des déterminants extraits non nuls de A.*

Définitions A.25 ***(mineur, cofacteur)*** *Soit A une matrice d'ordre n. On appelle* ***mineur*** *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 $i^{ième}$ et de la $j^{ième}$ colonne de A. On appelle* ***cofacteur*** *associé à ce même élément le scalaire*

$$\mathrm{Cof}_{ij}(A) = (-1)^{i+j} \begin{vmatrix} a_{11} & \dots & a_{1j-1} & a_{1j+1} & \dots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-11} & \dots & a_{i-1j-1} & a_{i-1j+1} & \dots & a_{i-1n} \\ a_{i+11} & \dots & a_{i+1j-1} & a_{i+1j+1} & \dots & a_{i+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj-1} & a_{nj+1} & \dots & a_{nn} \end{vmatrix}.$$

Enfin, on appelle ***matrice des cofacteurs****, ou* ***comatrice****, de A est la matrice d'ordre n constituée de l'ensemble des cofacteurs de A,*

$$\widetilde{A} = (\mathrm{Cof}_{ij}(A))_{1 \leq i,j \leq n}.$$

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,

$$\det(A + \alpha\, E_{ij}) = \det(A) + \alpha \begin{vmatrix} a_{11} & \dots & a_{1j-1} & 0 & a_{1j+1} \dots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ a_{i-11} & \dots & a_{i-1j-1} & 0 & a_{i-1j+1} \dots & a_{i-1n} \\ a_{i1} & \dots & a_{ij-1} & 1 & a_{ij+1} \dots & a_{in} \\ a_{i+11} & \dots & a_{i+1j-1} & 0 & a_{i+1j+1} \dots & a_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & a_{nj-1} & 0 & a_{nj+1} \dots & a_{nn} \end{vmatrix} = \det(A) + \alpha\, \mathrm{Cof}_{ij}(A).$$

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.26 ***(« formule de Laplace[6] »)*** *Soit A une matrice d'ordre n. On a*

$$\det(A) = \sum_{k=1}^{n} a_{ik}\, \mathrm{Cof}_{ik}(A) = \sum_{k=1}^{n} a_{kj}\, \mathrm{Cof}_{kj}(A),\ \forall (i,j) \in \{1,\dots,n\}^2.$$

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 $j^{ième}$ 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 $a_{ij}\, \mathrm{Cof}_{ij}$, $i = 1,\dots,n$, d'où le résultat. □

Proposition A.27 *Soit A une matrice d'ordre n. On a*

$$A\widetilde{A}^T = \widetilde{A}^T A = \det(A)\, I_n.$$

6. 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.

Démonstration. Considérons la matrice, de déterminant nul, obtenue en remplaçant la $j^{\text{ième}}$ colonne de A, $j \in \{1, \ldots, n\}$, par une colonnes de zéros et ajoutons lui les n matrices $a_{ik}\, E_{ik}$, $i = 1, \ldots, n$, avec $k \in \{1, \ldots, n\}$ et $k \neq j$. La matrice résultante est également de déterminant nul, puisque deux de ses colonnes sont identiques. Ceci signifie que

$$\sum_{i=1}^{n} a_{ik}\, \mathrm{Cof}_{ij}(A) = 0,\ \forall (j,k) \in \{1, \ldots, n\}^2,\ j \neq k.$$

En ajoutant le cas $k = j$, on trouve

$$\sum_{i=1}^{n} a_{ik}\, \mathrm{Cof}_{ij}(A) = \det(A)\, \delta_{jk}\ \forall (j,k) \in \{1, \ldots, n\}^2,$$

ce qu'on traduit matriciellement par $A\widetilde{A}^T = \det(A)\, I_n$. La seconde formule découle du fait que $\det(A^T) = \det(A)$. □

Lorsque la matrice A est inversible, on a obtenu une formule pour l'inverse de A,

$$A^{-1} = \frac{1}{\det(A)}\, \widetilde{A}^T,$$

qui ne nécessite que des calculs de déterminants.

A.2.5 Valeurs et vecteurs propres

Les *valeurs propres* d'une matrice A d'ordre n sont les n racines λ_i, $i = 1, \ldots, n$, réelles ou complexes, distinctes ou confondues, du *polynôme caractéristique*

$$\lambda \in \mathbb{C} \to \det(A - \lambda\, I_n)$$

associé à A. Le *spectre* d'une matrice A, noté $\sigma(A)$, est l'ensemble des valeurs propres de A. On rappelle les propriétés suivantes :

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i,\ \det(A) = \prod_{i=1}^{n} \lambda_i.$$

Par conséquent, la matrice A est singulière si au moins une de ses valeurs propres est nulle.

Enfin, le *rayon spectral* d'une matrice A est le nombre défini par

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|.$$

$\det(A^T - \lambda\, I_n) = \det((A - \lambda\, I_n)^T) = \det(A - \lambda\, I_n)$ d'où $\sigma(A^T) = \sigma(A)$.

À toute valeur propre λ d'une matrice A est associé au moins un vecteur non nul $\boldsymbol{v}$ tel que

$$A\boldsymbol{v} = \lambda\, \boldsymbol{v}$$

et appelé *vecteur propre* de la matrice A correspondant à la valeur propre λ. Le sous-espace vectoriel constitué de la réunion de l'ensemble des vecteurs propres associés à une valeur propre λ et du vecteur nul est appelé *sous-espace propre correspondant à la valeur propre* λ. Il coïncide par définition avec $\mathrm{Ker}\,(A - \lambda\, I_n)$ et sa dimension est $n - \mathrm{rg}(A - \lambda\, I_n)$. On appelle cette dernière *multiplicité géométrique* de λ et elle ne peut jamais être supérieure à la *multiplicité algébrique* de λ, définie comme 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*.

A.2.6 Matrices semblables

On a la définition suivante.

Définition A.28 ***(matrices semblables)*** *On dit que deux matrices A et B d'ordre n son* ***semblables*** *s'il existe une matrice d'ordre n inversible P 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). On voit que deux matrices sont semblables si et seulement si elles représentent le même endomorphisme dans deux bases éventuellement 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 mêmes spectre). Ces applications sont appelées *invariants de similitude*.

L'exploitation la notion de matrices semblables permet entre autres 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 [7].

Théorème A.29 ***(« décomposition de Schur*** [8] ***»)*** *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.*

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. □

Parmi les différents résultats qu'implique la décomposition de Schur, 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. Ceci 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 4.3 du chapitre 4).

Ajoutons qu'il ne faut pas confondre la notion de matrices semblables avec celle de *matrices équivalentes* [9]. En revanche, si deux matrices sont semblables, alors elles sont équivalentes.

7. Dans la démonstration de ce résultat, on fait appel à plusieurs notions qui sont abordées (ultérieurement...) dans la section A.3.

8. Issai Schur (Исáй Шур, 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 plusieurs concepts et résultats mathématiques variés.

9. 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 $B = PAQ$.

Enfin, une matrice A d'ordre n est dite *diagonalisable* si elle est semblable à une matrice *diagonale* (voir la définition A.33). On note que, dans ce cas, les éléments diagonaux de la matrice $P^{-1}AP$ sont les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$ de la matrice A, et que la $j^{\text{ième}}$ colonne de la matrice P, $1 \leq j \leq n$, est formée des composantes (relativement à la même base que pour la matrice A) d'un vecteur propre associé à λ_j. Ainsi, une matrice est diagonalisable si et seulement s'il existe une base de vecteurs propres.

A.2.7 Quelques matrices particulières

Matrices symétriques et hermitiennes

Définitions A.30 *Soit $A \in M_n(\mathbb{R})$. On dit que la matrice A est **symétrique** si $A = A^T$, **antisymétrique** si $A = -A^T$ et **orthogonale** si $A^T A = AA^T = I_n$.*

Définitions A.31 *On dit que la matrice A est **hermitienne** si $A = A^*$, **unitaire** si $A^*A = AA^* = I_n$ et **normale** 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.20 qu'une matrice orthogonale est telle que $A^{-1} = A^T$ et qu'une matrice unitaire est telle que $A^{-1} = A^*$.

Les matrices symétriques et hermitiennes vérifient un résultat de diagonalisation tout à fait remarquable, que nous citons ici sans démonstration.

Théorème A.32 ***(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 $P^{-1}AP$ soit une matrice diagonale. Les éléments diagonaux de cette matrice sont les valeurs propres de A, qui sont réelles.*

Matrices diagonales

Les matrices *diagonales* interviennent à de nombreuses reprise en algèbre linéaire. Elles vérifient des propriétés qui rendent leur manipulation particulièrement aisée d'un point de vue calculatoire.

Définition A.33 ***(matrice diagonale)*** *Une matrice A d'ordre n est dite **diagonale** si on a $a_{ij} = 0$ pour les couples d'indices $(i, j) \in \{1, \ldots, n\}^2$ tels que $i \neq j$.*

La démonstration du lemme suivant est laissée au lecteur.

Lemme A.34 *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 diagoanux. Une matrice diagonale A est donc inversible si et seulement si tous ses éléments diagonaux sont non nuls et, dans ce cas, son inverse est une matrice diagonale dont les élements diagonaux sont les inverses des élements diagonaux correspondants de A.*

Matrices triangulaires

Les matrices *triangulaires* forment une classe de matrices intervenant très couramment en algèbre linéaire numérique.

Définition A.35 ***(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 $(i, j) \in \{1, \ldots, n\}^2$ tels que $i > j$ (resp. $i < j$).*

Une matrice à la fois triangulaire supérieure est 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.36 *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 sont 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.*

Matrices bandes

Une *matrice bande* est une matrice carrée dont les coefficients non nuls sont localisés dans une « bande » autour de la diagonale principale. Plus précisement, on a la définition suivante.

Définition A.37 *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 $(i,j)\in\{1,\ldots,n\}^2$ tels que $i-j>p$ ou $j-i>q$. La **largeur de bande** 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.*

Matrices à diagonale dominante

Les matrices *à diagonale dominante* possèdent des propriétés remarquables pour les différentes méthodes de résolution de systèmes linéaires présentées aux chapitres 2 et 3.

Définition A.38 *On dit qu'une matrice A d'ordre n est **à diagonale dominante par lignes** (respectivement **par colonnes**) si*

$$|a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \ \ (\textit{resp. } |a_{ii}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ji}|),\ 1\leq i\leq n.$$

*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 [10].

Théorème A.39 *Soit A une matrice d'ordre n à diagonale strictement dominante (par lignes ou par colonnes). Alors, A 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

$$\sum_{j=1}^{n} a_{ij}\,x_j = 0,\ 1\leq i\leq n.$$

Le vecteur $\boldsymbol{x}$ étant non nul, il existe un indice i_0 dans $\{1,\ldots,n\}$ tel que $0\neq|x_{i_0}|=\max\limits_{1\leq i\leq n}|x_i|$ et l'on a alors

$$-a_{i_0i_0}x_{i_0} = \sum_{\substack{j=1\\ j\neq i_0}}^{n} a_{i_0j}x_j,$$

d'où

$$|a_{i_0i_0}| \leq \sum_{\substack{j=1\\ j\neq i_0}}^{n} |a_{i_0j}|\frac{|x_j|}{|x_{i_0}|} \leq \sum_{\substack{j=1\\ j\neq i_0}}^{n} |a_{i_0j}|,$$

10. Ce théoème semble avoir été redécouvert de nombreuses fois de manière totalement indépendante (voir la liste de références dans [Tau49]).

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^T, qui est une matrice à diagonale strictement dominante par lignes, est inversible et on utilise que $\det(A^T) = \det(A)$. □

A.3 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.3.1 Définitions

Nous rappelons dans cette section plusieurs définitions et propriétés à caractère général relatives aux normes et aux produits scalaires sur un espace vectoriel.

Définition A.40 ***(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*

1. $\|\boldsymbol{v}\| \geq 0$, $\forall \boldsymbol{v} \in E$, *et* $\|\boldsymbol{v}\| = 0$ *si et seulement si* $\boldsymbol{v} = \boldsymbol{0}$,
2. $\|\alpha\, \boldsymbol{v}\| = |\alpha|\, \|\boldsymbol{v}\|$, $\forall \alpha \in \mathbb{K}$, $\forall \boldsymbol{v} \in E$,
3. *elle vérifie l'****inégalité triangulaire****, c'est-à-dire*

$$\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|, \ \forall \boldsymbol{u}, \boldsymbol{v} \in E.$$

On appelle *espace vectoriel normé* un espace vectoriel muni d'une norme. C'est un cas particulier d'espace métrique dans lequel la distance entre deux éléments est donné par

$$d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|, \ \forall \boldsymbol{u}, \boldsymbol{v} \in E.$$

Définition A.41 ***(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*** *s'il existe deux constantes positives c et C telles que*

$$c\, \|\boldsymbol{v}\|_* \leq \|\boldsymbol{v}\|_{**} \leq C\, \|\boldsymbol{v}\|_*, \ \forall \boldsymbol{v} \in E.$$

Définition A.42 ***(produit scalaire)*** *Un* ***produit scalaire*** *(resp.* ***produit scalaire hermitien****) sur un espace vectoriel E sur $\mathbb{R}$ (resp. $\mathbb{C}$) est une application $(\cdot\,,\cdot)$ de $E \times E$ dans $\mathbb{R}$ (resp. $\mathbb{C}$) possédant les propriétés suivantes :*

1. *elle est* ***bilinéaire*** *(resp.* ***sesquilinéaire****), c'est-à-dire linéaire par rapport à la première variable*

$$(\alpha\, \boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w}) = \alpha\, (\boldsymbol{u}, \boldsymbol{w}) + (\boldsymbol{u}, \boldsymbol{v}), \ \forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in E, \ \forall \alpha \in \mathbb{R} \ \textit{(resp. } \mathbb{C}\textit{)},$$

   et linéaire (resp. antilinéaire) par rapport à la seconde

$$(\boldsymbol{u}, \alpha\, \boldsymbol{v} + \boldsymbol{w}) = \alpha\, (\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{u}, \boldsymbol{w}) \ \textit{(resp. } (\boldsymbol{u}, \alpha\, \boldsymbol{v} + \boldsymbol{w}) = \overline{\alpha}\, (\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{u}, \boldsymbol{w})\textit{)},$$
$$\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in E, \ \forall \alpha \in \mathbb{R} \ \textit{(resp. } \mathbb{C}\textit{)},$$

2. *elle est* ***symétrique*** *(resp.* ***à symétrie hermitienne****), c'est-à-dire*

$$(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{v}, \boldsymbol{u}) \ \textit{(resp. } (\boldsymbol{u}, \boldsymbol{v}) = \overline{(\boldsymbol{v}, \boldsymbol{u})}\textit{)}, \ \forall \boldsymbol{u}, \boldsymbol{v} \in E,$$

3. *elle est* ***définie positive***[11]*, c'est-à-dire*

$$(\boldsymbol{v}, \boldsymbol{v}) \geq 0, \ \forall \boldsymbol{v} \in E, \ \textit{et } (\boldsymbol{v}, \boldsymbol{v}) = 0 \ \textit{si et seulement si } \boldsymbol{v} = \boldsymbol{0}.$$

11. On dit aussi qu'elle est *non dégénérée positive*.

Définition A.43 ***(espace euclidien)*** *On appelle* ***espace euclidien*** *tout espace vectoriel sur $\mathbb{R}$ de dimension finie muni d'un produit scalaire.*

Lemme A.44 ***(« inégalité de Cauchy*** [12] ***–Schwarz*** [13] ***»)*** *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni du produit scalaire $(\cdot\,,\cdot)$. On a*

$$|(\boldsymbol{u},\boldsymbol{v})| \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\|,\ \forall \boldsymbol{u},\boldsymbol{v} \in E,$$

où l'on a noté $\|\boldsymbol{v}\| = \sqrt{(\boldsymbol{v},\boldsymbol{v})}$, $\forall \boldsymbol{v} \in E$, avec égalité si et seulement si $\boldsymbol{u}$ et $\boldsymbol{v}$ sont linéairement dépendants.

Démonstration. Soit $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de E. On va démontrer le résultat dans le cas réel. Dans le cas complexe, on se ramène au cas réel en multipliant $\boldsymbol{u}$ par un scalaire de la forme $e^{\mathrm{i}\theta}$, avec θ réel, de manière à ce le produit $(e^{\mathrm{i}\theta}\boldsymbol{u},\boldsymbol{v})$ est réel. On considère l'application qui à tout réel t associe $\|\boldsymbol{u} - t\boldsymbol{v}\|$. On a, par propriétés du produit scalaire,

$$0 \leq \|\boldsymbol{u} - t\,\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + 2t\,(\boldsymbol{u},\boldsymbol{v}) + t^2\|\boldsymbol{v})\|^2,\ \forall t \in \mathbb{R}.$$

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 annoncé. En outre, on a égalité lorsque le discriminant est nul, ce qui signifie que le polynôme possède une racine réelle λ d'où $\boldsymbol{u} + \lambda\,\boldsymbol{v} = \boldsymbol{0}$. □

À tout produit scalaire, on peut associer une norme particulière comme le montre le théorème suivant.

Théorème A.45 *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ et $(\cdot\,,\cdot)$ un produit scalaire sur E. L'application $\|\cdot\|$, définie par*

$$\|\boldsymbol{v}\| = \sqrt{(\boldsymbol{v},\boldsymbol{v})},\ \forall \boldsymbol{v} \in E,$$

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.40. 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

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + (\boldsymbol{u},\boldsymbol{v}) + (\boldsymbol{v},\boldsymbol{u}) + \|\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + (\boldsymbol{u},\boldsymbol{v}) + \overline{(\boldsymbol{u},\boldsymbol{v})} + \|\boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + 2\,\mathrm{Re}((\boldsymbol{u},\boldsymbol{v})) + \|\boldsymbol{v}\|^2.$$

Par utilisation de l'inégalité de Cauchy–Schwarz, on obtient alors

$$\|\boldsymbol{u}+\boldsymbol{v}\|^2 \leq \|\boldsymbol{u}\|^2 + 2\,\|\boldsymbol{u}\|\,\|\boldsymbol{v}\| + \|\boldsymbol{v}\|^2 = (\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)^2\,.$$

□

Définition A.46 *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$ muni d'un produit scalaire $(\cdot\,,\cdot)$. On dit que deux vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ de E sont* ***orthogonaux****, ce que l'on note $\boldsymbol{u} \perp \boldsymbol{v}$, si $(\boldsymbol{u},\boldsymbol{v}) = 0$. Par extension, un vecteur $\boldsymbol{v}$ de E est* ***orthogonal à une partie*** *G* ***de*** *E, ce que l'on note $\boldsymbol{v} \perp G$, si le vecteur $\boldsymbol{v}$ est orthogonal à tout vecteur de G. Enfin, un ensemble de vecteurs $\{\boldsymbol{u}_i\}_{i=1,\ldots,m}$, $2 \leq m \leq n$, de E est dit* ***orthonormal*** *s'il vérifie*

$$(\boldsymbol{u}_i,\boldsymbol{u}_j) = \delta_{ij},\ 1 \leq i,j \leq m.$$

12. 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.

13. 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.

A.3.2 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

$$(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{v}^T\boldsymbol{u} = \boldsymbol{u}^T\boldsymbol{v} = \sum_{i=1}^{n} u_i v_i \text{ si } \mathbb{K} = \mathbb{R},$$
$$(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{v}^*\boldsymbol{u} = \overline{\boldsymbol{u}^*\boldsymbol{v}} = \sum_{i=1}^{n} u_i \overline{v_i} \text{ si } \mathbb{K} = \mathbb{C},$$

est appelée *produit scalaire canonique* (et *produit scalaire euclidien* lorsque $\mathbb{K} = \mathbb{R}$). On note La norme induite par ce produit scalaire, appelée *norme euclidienne* dans le cas réel, est alors

$$\|\boldsymbol{v}\|_2 = \sqrt{(\boldsymbol{v},\boldsymbol{v})} = \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2}.$$

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 tout matrice orthogonale (resp. unitaire) U,

$$(U\boldsymbol{u}, U\boldsymbol{v}) = (U^T U\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v}) \text{ (resp. } (U\boldsymbol{u}, U\boldsymbol{v}) = (U^* U\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u},\boldsymbol{v})),\ \forall \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n \text{ (resp. } \mathbb{C}^n).$$

D'autres normes couramment utilisées sont les normes

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|,$$

et

$$\|\boldsymbol{v}\|_\infty = \max_{1\le i\le n} |v_i|.$$

Plus généralement, on a le résultat suivant.

Théorème A.47 *Soit E un espace vectoriel sur $\mathbb{R}$ ou $\mathbb{C}$, de dimension finie n. Pour tout nombre réel $p \ge 1$, l'application $\|\cdot\|_p$ définie par*

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p},\ \forall \boldsymbol{v} \in E,$$

est une norme.

Démonstration. Pour $p = 1$, la preuve est immédiate et on va donc considérer que p est strictement plus grand que 1. 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 E. D'après l'inégalité ci-dessus, on a

$$\frac{|u_i v_i|}{\|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q} \le \frac{1}{p}\frac{|u_i|^p}{\|\boldsymbol{u}\|_p} + \frac{1}{q}\frac{|v_i|^q}{\|\boldsymbol{v}\|_q},\ 1 \le i \le \dim(E),$$

d'où, par sommation,

$$\sum_{i=1}^{n} |u_i v_i| \leq \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q.$$

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

$$(|u_i| + |v_i|)^p = |u_i| \, (|u_i| + |v_i|)^{p-1} + |v_i| \, (|u_i| + |v_i|)^{p-1}, \ 1 \leq i \leq n,$$

d'où, après sommation et utilisation de l'inégalité précédemment établie,

$$\sum_{i=1}^{n} (|u_i| + |v_i|)^p \leq (\|\boldsymbol{u}\|_p + \|\boldsymbol{v}\|_p) \left(\sum_{i=1}^{n} (|u_i| + |v_i|)^{(p-1)q} \right)^{1/q}.$$

L'inégalité triangulaire découle alors de la relation $(p-1)q = p$. □

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 $\mathbb{C}^n$, on a par exemple

$$\|\boldsymbol{v}\|_2 \leq \|\boldsymbol{v}\|_1 \leq \sqrt{n} \, \|\boldsymbol{v}\|_2 \text{ et } \|\boldsymbol{v}\|_\infty \leq \|\boldsymbol{v}\|_1 \leq n \, \|\boldsymbol{v}\|_\infty.$$

Nous pouvons maintenant introduire la notion de matrice *symétrique définie positive*, dont les propriétés sont intéressantes pour les méthodes de résolution de systèmes linéaires étudiées dans les chapitres 2 et 3.

Définition A.48 ***(matrice définie positive)*** *Une matrice d'ordre n est dite* ***définie positive*** *sur $\mathbb{C}^n$ si $(A\boldsymbol{x}, \boldsymbol{x}) \geq 0$, $\forall \boldsymbol{x} \in \mathbb{C}^n$, avec $(A\boldsymbol{x}, \boldsymbol{x}) = 0$ si et seulement si $\boldsymbol{x} = \boldsymbol{0}$.*

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 $\frac{1}{2}(A + A^T)$, est définie positive sur $\mathbb{R}^n$. Plus généralement, on a le résultat suivant montre qu'une matrice 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.49 *Soit A une matrice de $M_n(\mathbb{C})$ (resp. $\mathbb{R}$). 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

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \overline{a_{ij} v_j} v_i = \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{a_{ij} v_j v_i} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_j \overline{v_i} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ji} v_i \overline{v_j},$$

ce qui implique

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (\overline{a_{ij}} - a_{ji}) v_i \overline{v_j} = 0, \ \forall \boldsymbol{v} \in \mathbb{C}^n.$$

Par des choix appropriés du vecteur $\boldsymbol{v}$, on en déduit que $\overline{a_{ij}} = a_{ji}$, pour tous i, j dans $\{1, \ldots, n\}$. □

Deux propriétés principales des matrices définies positives sont résumées ci-dessous.

Théorème A.50 *Une matrice est définie positive sur $\mathbb{C}^n$ si et seulement si elle est hermitienne et ses valeurs propres sont strictement positives. En particulier, un matrice définie positive est inversible.*

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, $i = 1, \ldots, n$, 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 $i = 1, \ldots, n$, 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. □

Le résultat classique suivant fournit une caractérisation simple des matrices symétriques (ou hermitienne) définies positives.

Théorème A.51 *(« critère de Sylvester[14] ») Une matrice symétrique ou hermitienne d'ordre n est définie positive si et seulement tous ses mineurs principaux sont strictement positifs, c'est-à-dire si toutes les sous-matrices principales*

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}, \ 1 \leq k \leq n,$$

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 $(\cdot\,,\cdot)_{\mathbb{R}^n}$ désigne le produit scalaire euclidien sur $\mathbb{R}^n$.

Pour $n = 1$, la matrice A est un nombre réel, $A = (a_{11})$, et $(Ax, x)_{\mathbb{R}} = a_{11}\, x^2$ 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 $n-1$, $n \geq 2$, et prouvons-le pour celles d'ordre n. Soit A une telle matrice. On note respectivement λ_i et $\boldsymbol{v}_i$, $1 \leq i \leq n$ les valeurs et vecteurs propres de A, l'ensemble $\{\boldsymbol{v}_i\}_{1\leq i\leq n}$ formant par ailleurs une base orthonormée de $\mathbb{R}^n$.

Observons que

$$\left(A \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix} \right)_{\mathbb{R}^n} = \left(A_{n-1} \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \right)_{\mathbb{R}^{n-1}}$$

Puisque $(A\boldsymbol{x}, \boldsymbol{x})_{\mathbb{R}^n} > 0$ pour tout vecteur $\boldsymbol{x}$ non nul de $\mathbb{R}^n$, ceci est donc en particulier vrai pour tous les vecteurs de la forme

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}.$$

Par conséquent, la matrice A_{n-1} est définie positive et tous ses mineurs principaux, qui ne sont autres que les $n-1$ mineurs principaux 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 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 $(A(\alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j), \alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j)_{\mathbb{R}^n} > 0$. Mais, on a par ailleurs

$$(A(\alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j), \alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j)_{\mathbb{R}^n} = \alpha^2\lambda_i + \beta^2\lambda_j < 0,$$

d'où une contradiction. □

A.3.3 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 souvent à une norme définie sur un espace de matrices de satisfaire une propriété de *sous-multiplicativité* qui la rend intéressante en pratique[15]. On parle alors 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.

14. James Joseph Sylvester (3 septembre 1814 - 13 mars 1897) était un mathématicien et géomètre anglais. Il travailla sur les formes algébriques, particulièrement 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*.

15. Sur $M_n(\mathbb{K})$, une telle norme est en effet une *norme d'algèbre*.

Définition A.52 ***(normes consistantes)*** *On dit que trois normes, toutes notées $\|\cdot\|$ et respectivement définies sur $\mathbb{C}^m$, $M_{m,n}(\mathbb{C})$ et $\mathbb{C}^n$, sont* ***consistantes*** *si*

$$\|A\boldsymbol{v}\| \leq \|A\|\,\|\boldsymbol{v}\|,\ \forall A \in \mathbb{C}^{m\times n},\ \forall \boldsymbol{v} \in \mathbb{C}^n.$$

Définition A.53 ***(norme matricielle)*** *Une* ***norme matricielle*** *sur $M_n(\mathbb{C})$ est une application de $M_n(\mathbb{C})$ dans $\mathbb{R}$ vérifiant les propriétés d'une norme (voir la définition A.40) ainsi que la propriété de* ***sous-multiplicativité*** *suivante :*

$$\|AB\| \leq \|A\|\,\|B\|,\ \forall A, B \in M_n(\mathbb{C}). \tag{A.2}$$

Toutes les normes sur $M_n(\mathbb{C})$ ne sont pas des normes matricielles comme le montre l'exemple suivant, tiré de [GV96].

Exemple. La norme $\|\cdot\|_\Delta$, définie sur $M_n(\mathbb{C})$ par

$$\|A\|_\Delta = \max_{1\leq i,j\leq n} |a_{ij}|,\ \forall A \in M_n(\mathbb{C}),$$

ne satisfait pas la propriété de sous-multiplicativité (A.2), puisque pour

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

on a $2 = \|A^2\|_\Delta > \|A\|_\Delta^2 = 1$.

On remarquera aussi qu'il existe toujours une norme vectorielle avec laquelle une norme matricielle donnée est consistante. En effet, étant donnée une norme matricielle $\|\cdot\|$ et un vecteur non nul quelconque $\boldsymbol{u}$ dans $\mathbb{C}^n$, il suffit de définir la norme vectorielle par

$$\|\boldsymbol{v}\| = \|\boldsymbol{v}\boldsymbol{u}^*\|,\ \forall \boldsymbol{v} \in \mathbb{C}^n.$$

Ainsi, il n'est pas nécessaire de préciser explicitement la norme vectorielle avec laquelle la norme matricielle est consistante.

Exemple. L'application définie par

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} = \sqrt{\mathrm{tr}(AA^*)},\ \forall A \in M_n(\mathbb{C}),$$

est une norme matricielle (la démonstration est laissée en exercice), appelée *norme de Frobenius*[16], consistante avec la norme vectorielle euclidienne $\|\cdot\|_2$, car on a

$$\|A\boldsymbol{v}\|_2^2 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}\, x_j \right|^2 \leq \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}|^2 \sum_{j=1}^{n} |x_j|^2 \right) = \|A\|_F^2 \|\boldsymbol{v}\|_2^2.$$

On remarque que l'on a $\|I_n\|_F = \sqrt{n}$.

Proposition A.54 ***(norme matricielle subordonnée)*** *Étant donné une norme vectorielle $\|\cdot\|$ sur $\mathbb{C}^n$, l'application $\|\cdot\|$ de $M_n(\mathbb{C})$ dans $\mathbb{R}$ définie par*

$$\|A\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \|\boldsymbol{v}\|\leq 1}} \|A\boldsymbol{v}\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \|\boldsymbol{v}\|= 1}} \|A\boldsymbol{v}\|, \tag{A.3}$$

est une norme matricielle.

16. 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.

DÉMONSTRATION. On remarque tout d'abord que la quantité $\|A\|$ est bien définie pour tout matrice d'ordre n : ceci découle de la continuité de l'application de $\mathbb{C}^n$ dans $\mathbb{R}$ qui à un vecteur $\boldsymbol{v}$ associe $\|A\boldsymbol{v}\|$ sur la sphère unité, qui est compacte puisqu'on est en dimension finie. La vérification des propriétés satisfaites par une norme matricielle est alors immédiate. □

On déduit de la définition (A.3) que $\|I_n\| = 1$ pour toute norme matricielle subordonnée $\|\cdot\|$. Un bon exemple de norme matricielle n'étant pas subordonnée à une norme vectorielle est la norme de Frobenius, pour laquelle on a déjà vu que $\|I_n\|_F = \sqrt{n}$.

La proposition suivante donne le calcul des normes subordonnées aux normes vectorielles $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$.

Proposition A.55 *Soit A une matrice carrée d'ordre n. On a*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \tag{A.4}$$

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \|A^*\|_2, \tag{A.5}$$

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{A.6}$$

Par ailleurs, la norme $\|\cdot\|_2$ est invariante par transformation unitaire et si A est normale, alors $\|A\|_2 = \rho(A)$.

DÉMONSTRATION. Pour tout vecteur $\boldsymbol{v}$ de $\mathbb{C}^n$, on a

$$\|A\boldsymbol{v}\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} v_j \right| \leq \sum_{j=1}^{n} |v_j| \sum_{i=1}^{n} |a_{ij}| \leq \left(\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \right) \|\boldsymbol{v}\|_1.$$

Pour montrer (A.4), on construit un vecteur (qui dépendra de la matrice A) tel que l'on ait égalité dans l'inégalité ci-dessus. Il suffit pour cela de considérer pour cela le vecteur $\boldsymbol{u}$ de composantes

$$u_i = 0 \text{ pour } i \neq j_0,\ u_{j_0} = 1,$$

où j_0 est un indice vérifiant

$$\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ij_0}|.$$

De la même manière, on prouve (A.6) en écrivant

$$\|A\boldsymbol{v}\|_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^{n} a_{ij} v_j \right| \leq \left(\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \right) \|\boldsymbol{v}\|_\infty,$$

et en choisissant le vecteur $\boldsymbol{u}$ tel que

$$u_j = \frac{\overline{a_{i_0 j}}}{|a_{i_0 j}|} \text{ si } a_{i_0 j} \neq 0,\ u_j = 1 \text{ sinon},$$

avec i_0 un indice satisfaisant

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i_0 j}|.$$

On prouve à présent (A.5). La matrice A^*A étant hermitienne, il existe (voir le théorème A.32) 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

$$\|A\|_2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \sqrt{\frac{(A^*A\boldsymbol{v},\boldsymbol{v})}{(\boldsymbol{v},\boldsymbol{v})}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\frac{(U^*A^*AU\boldsymbol{w},\boldsymbol{w})}{(\boldsymbol{w},\boldsymbol{w})}} = \sup_{\substack{\boldsymbol{w}\in\mathbb{C}^n \\ \boldsymbol{w}\neq\boldsymbol{0}}} \sqrt{\sum_{i=1}^{n} \mu_i \frac{|w_i|^2}{\sum_{j=1}^{n} |w_i|^2}} = \sqrt{\max_{1\leq i\leq n} \mu_i}.$$

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 $\|A\|_2 = \|A^*\|_2 = \rho(AA^*)$.

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

$$\|UA\|_2^2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|UA\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{(U^*UA\boldsymbol{v},\boldsymbol{v})}{\|\boldsymbol{v}\|_2^2} = \|A\|_2.$$

Le changement de variable $\boldsymbol{u} = U\boldsymbol{v}$ vérifiant $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2$, on a par ailleurs

$$\|AU\|_2^2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|AU\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2^2}{\|U^{-1}\boldsymbol{u}\|_2^2} = \sup_{\substack{\boldsymbol{u}\in\mathbb{C}^n \\ \boldsymbol{u}\neq\boldsymbol{0}}} \frac{\|A\boldsymbol{u}\|_2^2}{\|\boldsymbol{u}\|_2^2} = \|A\|_2.$$

Enfin, si A est une matrice normale, alors elle diagonalisable dans une base orthonormée de vecteurs propres (voir le théorème A.32 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. On observe tout d'abord que $\|A\|_1 = \|A^*\|_\infty$, et l'on a $\|A\|_1 = \|A\|_\infty$ et $\|A\|_2 = \rho(A)$ si A est une matrice hermitienne (donc normale). Si U est une matrice unitaire (donc normale), on a alors $\|U\|_2 = \rho(I_n) = 1$. La norme $\|A\|_2$ n'est autre que la plus grande valeur singulière [17] de la matrice A et son calcul pratique est donc beaucoup plus difficile et coûteux que celui de $\|A\|_1$ ou $\|A\|_\infty$

Il est également clair à l'examen de la démonstration ci-dessus que les expressions trouvées pour $\|A\|_1$, $\|A\|_2$ et $\|A\|_\infty$ sont encore valables pour des matrices rectangulaires. Dans ce cas cependant, ces applications ne sont plus des normes matricielles mais de simples normes sur un espace vectoriel de matrices donné, puisque le produit de telles matrices n'a en général pas de sens.

Enfin, si l'on a montré qu'il existait des normes matricielles et des matrices A vérifiant l'égalité $\|A\| = \rho(A)$, il faut insister sur le fait que le rayon spectral n'est pas une norme (par exemple, toute matrice triangulaire non nulle dont les coefficients diagonaux sont nuls a un rayon spectral égal à zéro). 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.58).

Théorème A.56 *Soit A une matrice carrée d'ordre n et $\|\cdot\|$ une norme matricielle. Alors, on a*

$$\rho(A) \leq \|A\|.$$

D'autre part, étant donné une matrice A et un nombre strictement positif ε, il existe au moins une norme matricielle subordonnée telle que

$$\|A\| \leq \rho(A) + \varepsilon.$$

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 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 se trouve démontrée.

17. On appelle *valeurs singulières* d'une matrice carrée A les racines carrées positives de la matrice carrée hermitienne A^*A (ou A^TA si la matrice A est réelle).

Soit maintenant A une matrice d'ordre n. 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 suffisament 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

$$\sum_{j=i+1}^{n} \delta^{j-i}\,|t_{ij}| \leq \varepsilon,\ 1 \leq i \leq n-1.$$

On a alors

$$\|(UD_\delta)^{-1}A(UD_\delta)\|_\infty = \max_{1\leq i\leq n} \sum_{j=i}^{n} \delta^{j-i}|t_{ij}| \leq \rho(A) + \varepsilon.$$

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 $\|(UD_\delta)^{-1}\cdot\|_\infty$. □

Théorème A.57 *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*

$$\frac{1}{1+\|A\|} \leq \|(I_n - A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

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 definition 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

$$1 = \|I_n\| \leq \|I_n - A\|\,\|(I_n - A)^{-1}\| \leq (1 + \|A\|)\,\|(I_n - A)^{-1}\|,$$

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

$$\|(I_n - A)^{-1}\| \leq 1 + \|A\|\,\|(I_n - A)^{-1}\|,$$

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 utilise alors le théorème A.56 pour conclure. □

Le résultat qui suit donne des conditions nécessaires et suffisantes pour que la suite formée des puissances successives d'une matrice carrée donnée converge vers la matrice nulle. Il fournit un critère fondamental de convergence pour les *méthodes itératives de résolution des systèmes linéaires* introduites dans le chapitre 3.

Théorème A.58 *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 correspondante. 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

$$A\boldsymbol{v} = \lambda\,\boldsymbol{v} \text{ et } |\lambda| \leq 1.$$

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.56. Il reste à montrer que (iv) implique (i). Il suffit pour cela d'utiliser l'inégalité

$$\|A^k\| \leq \|A\|^k,\ \forall k \in \mathbb{N},$$

vérifiée par la norme subordonnée de l'énoncé. □

On peut maintenant prouver le résultat suivant, qui précise un peu plus le lien existant entre la norme matricielle et le rayon spectral d'une matrice.

Théorème A.59 *Soit A une matrice carrée et $\|\cdot\|$ une norme matricielle. Alors, on a*

$$\lim_{k\to+\infty} \|A^k\|^{1/k} = \rho(A).$$

Démonstration. Puisque $\rho(A) \leq \|A\|$ d'après le théorème A.56 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

$$A_\varepsilon = \frac{A}{\rho(A)+\varepsilon}$$

vérifie $\rho(A_\varepsilon) < 1$ et on déduit du théorème A.58 que $\lim_{k\to+\infty} A_\varepsilon^k = 0$. Par conséquent, il existe un entier l, dépendant de ε, tel que

$$k \geq l \Rightarrow \|A_\varepsilon^k\| = \frac{\|A^k\|}{(\rho(A)+\varepsilon)^k} \leq 1.$$

Ceci implique que

$$k \geq l \Rightarrow \|A_\varepsilon^k\|^{1/k} \leq \rho(A)+\varepsilon,$$

et démontre donc l'égalité cherchée. □

A.4 Systèmes linéaires

Soit m et n deux entiers strictement positifs. Résoudre un *système linéaire de m équations à n inconnues et à coefficents dans un corps* $\mathbb{K}$ consiste à trouver la ou les solutions, s'il en existe, de l'équation algébrique

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

où A est une matrice de $M_{m,n}(\mathbb{K})$, appelée *matrice du système*, $\boldsymbol{b}$ est un vecteur de $\mathbb{K}^m$, appelé *second membre du système*, et $\boldsymbol{x}$ est un vecteur de $\mathbb{K}^n$, appelé *inconnue du système*. On dit que le vecteur $\boldsymbol{x}$ est *solution du système* ci-dessus si ces composantes vérifient les m équations

$$\sum_{j=1}^{n} a_{ij}\,x_k = b_i,\ i = 1,\ldots,m,$$

du système. 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 rappellons des résultats sur l'existence et l'unicité éventuelle des solutions de systèmes linéaires et leur détermination.

A.4.1 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.60 *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 un 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.61 ***(« règle de Cramer[18] »)*** *On suppose les vecteurs $\boldsymbol{a}_j$, $j = 1, \ldots, n$, de $\mathbb{K}^n$ désignent les colonnes d'une matrice inversible A de $M_n(\mathbb{K})$. Les composante de la solution du système $A\boldsymbol{x} = \boldsymbol{b}$ sont données par*

$$x_i = \frac{\det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n)}{\det(A)}, \ i = 1, \ldots, n.$$

Démonstration. Le déterminant étant une forme multilinéaire alternée, on a

$$\det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \lambda\,\boldsymbol{a}_i + \mu\,\boldsymbol{a}_j, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n) = \lambda \det(A), \ \forall i, j \in \{1, \ldots, n\}, \ i \neq j, \ \forall \lambda, \mu \in \mathbb{K}.$$

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 $\mathbb{K}^n$ 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

$$\det\left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \sum\nolimits_{j=1}^{n} x_j\,\boldsymbol{a}_j, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n\right) = \det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n) = x_i \det(A), \ i = 1, \ldots, n.$$

d'où la formule. □

On appelle *système de Cramer* tout système d'équations linéaires dont la matrice est inversible.

Aussi séduisante qu'elle soit, la règle de Cramer s'avère parfaitement inefficace en pratique. Le problème provient de l'évaluation des déterminants intervenant dans les formules, qui nécessite bien trop d'opérations si l'on applique une méthode récursive de calcul du déterminant.

A.4.2 Systèmes linéaires sur- ou sous-dimensionnés

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 $m < n$, 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 $m > n$, 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.62 *Il existe une solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ si et seulement si le second membre $\boldsymbol{b}$ appartient à l'image de A. La solution est unique si est seuelement 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 (théorème A.17).

Lemme A.63 *Si $m < n$, alors $\dim\ker(A) \geq n - m \geq 1$, et s'il existe une solution au système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, il en existe une infinité.*

18. Gabriel Cramer (31 juillet 1704 - 4 janvier 1752) était un mathématicien suisse. Le travail par lequel il est le mieux connu est son traité *Introduction à l'analyse des lignes courbes algébriques* publié en 1750.

A.4.3 Systèmes échelonnés

Nous abordons maintenant le cas de systèmes linéaires dont les matrices sont *échelonnées*. 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 2.2 du chapitre 2) est de ramener un système linéaire quelconque à un système échelonné équivalent (c'est-à-dire ayant le même ensemble de solutions), plus simple à résoudre.

Définition A.64 ***(matrice échelonnée)*** *Une matrice A de $M_{m,n}(\mathbb{K})$ est dite* ***échelonnée*** *ou* ***en échelons*** *s'il existe un entier r, $1 \leq r \leq \min(m,n)$ et une suite d'entiers $1 \leq j_1 < j_2 < \ldots < j_r \leq n$ 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 $r < i \leq m$ et $1 \leq j \leq n$, c'est-à-dire que toutes les lignes après les r premières sont nulles.*

Exemple. 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 échelonnée dont les pivots sont 1, 2 et 3.

On déduit immédiatement de la définition précédente que le rang d'une matrice échelonnée est égal au nombre r de pivots. Dans un système linéaire *échelonné*, c'est-à-dire associé à une matrice échelonnée, de m équations à n inconnues, 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 échelonné $A\boldsymbol{x} = \boldsymbol{b}$ de m équations à n inconnues et de rang r. Commençons par discuter de la compatibilité de ce système. Tout d'abord, si $r = m$, le système linéaire est compatible et ses équations sont linéairement indépendantes. Sinon, c'est-à-dire si $r < m$, les $m - r$ dernières lignes de la matrice A sont nulles et le système linéaire n'est donc compatible que si les $m - r$ dernières composantes du vecteur $\boldsymbol{b}$ sont également nulles, ce qui revient à vérifier $m - r$ *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 2.1 du chapitre 2).
- Si $r = n < m$, la solution existe, puisque le système est supposé satisfaire les $m - r$ 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 $r < n \leq m$ 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 m 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 m lignes et $n - r$ colonnes, constituée des colonnes de A ne contenant pas de pivot, et $\boldsymbol{x}_S$ est un vecteur de $\mathbb{K}^{n-r}$ 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 $\mathbb{K}^n$ (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 $n-r$ systèmes linéaires $A_P \boldsymbol{x}_{P_k} = \boldsymbol{b} - A_S \boldsymbol{e}_k^{(n-r)}$, $1 \leq k \leq n-r$, où $\boldsymbol{e}_k^{(n-r)}$ désigne le $k^{\text{ième}}$ vecteur de la base canonique de $\mathbb{K}^{n-r}$ (ceci revient à fixer la valeur de la $k^{\text{ième}}$ inconnue secondaire à 1 et celles des autres à zéro), le vecteur de base $\boldsymbol{x}_k$ correspondant étant

$$\boldsymbol{s}_k = \begin{pmatrix} \boldsymbol{x}_{P_k} \\ \boldsymbol{e}_k^{(n-r)} \end{pmatrix}.$$

La solution générale du système est alors de la forme

$$\boldsymbol{x} = \boldsymbol{s}_0 + \sum_{k=1}^{n-r} c_k \, \boldsymbol{s}_k,$$

avec les c_k, $1 \leq k \leq n-r$, des scalaires.

A.4.4 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 déteriorer notablement la précision de la solution obtenue. Afin de quantifier la sensibilité de la solution d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ vis-à-vis des perturbations des données A et $\boldsymbol{b}$, on utilise la notion de *conditionnement* d'une matrice inversible, introduite par Turing[19] dans [Tur48]. Celle-ci n'est qu'un cas particulier de la notion générale introduite dans la section 1.3.

Définition A.65 ***(conditionnement d'une matrice)*** *Soit $\|\cdot\|$ une norme matricielle subordonnée. Pour toute matrice inversible A d'ordre n, on appelle* ***conditionnement*** *de A relativement à la norme matricielle $\|\cdot\|$ le nombre*

$$\mathrm{cond}(A) = \|A\| \, \|A^{-1}\|.$$

La valeur du conditionnement d'une matrice dépendant en général de la norme subordonnée 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 $\mathrm{cond}(A) \geq 1$ puisque $1 = \|I_n\| = \|AA^{-1}\| \geq \|A\| \, \|A^{-1}\|$. D'autres propriétés évidentes du conditionnement sont rassemblées dans le résultat suivant.

Théorème A.66 *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.

19. Alan Mathison Turing (23 juin 1912 - 7 juin 1954) était un mathématicien et informaticien anglais, spécialiste de la logique et de la cryptanalyse. Il fut l'auteur de l'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.

(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) :*

$$UU^* = I_n \Rightarrow \mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^*AU).$$

DÉMONSTRATION. Les propriétés (1) découlement de la définition du conditionnement et des propriétés de la norme.
On a établi dans la proposition A.55 que $\|A\|_2 = \sqrt{\rho(A^*A)}$ et donc, d'après la définition des valeurs singulières de A, on a $\|A\|_2 = \mu_n$. 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 la propriété (2). La propriété (3) résulte de l'égalité $\|A\|_2 = \rho(A)$ vérifiée par les matrices normales (voir encore la proposition A.55). Si A est une matrice orthogonale ou unitaire, légalité $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(I_n)} = 1$ entraîne la propriété (4). Enfin, la propriété (5) est une conséquence de l'invariance par transformation unitaire de la norme $\|\cdot\|_2$ (voir une nouvelle fois la proposition A.55). □

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.67 *Soit A une matrice inversible d'ordre n et $\boldsymbol{b}$ un vecteur non nul de taille correspondante. 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*

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}.$$

Si $\boldsymbol{x}$ et $\boldsymbol{x}+\delta\boldsymbol{x}$ sont les solutions respectives 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

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}+\delta\boldsymbol{x}\|} \leq \mathrm{cond}(A)\,\frac{\|\delta A\|}{\|A\|}.$$

De plus, ces deux inégalités sont optimales, c'est-à-dire que l'on peut trouver une 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 $\|\boldsymbol{b}\| \leq \|A\|\,\|\delta\boldsymbol{x}\|$, on en déduit la première inégalité. Pour la seconde inégalité, on tire de légalité $A\delta\boldsymbol{x} + \delta A(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{0}$ la majoration $\|\delta\boldsymbol{x}\| \leq \|A^{-1}\|\,\|\delta A\|\,\|\boldsymbol{x}+\delta\boldsymbol{x}\|$, dont on déduit le résultat.
Le fait que les inégalités sont optimales découle du fait que, pout tout norme matricielle subordonnée et toute matrice A d'ordre n, il existe un vecteur $\boldsymbol{y}$ non nul tel que $\|A\boldsymbol{y}\| = \|A\|\,\|\boldsymbol{y}\|$ (voir la démonstration de la proposition A.54). □

Bien qu'optimales, les inégalités de ce dernier résultat sont, en général, pessimistes. Elles conduisent néanmoins à l'introduction d'une terminologie courante, en lien avec le conditionnement d'une matrice, qui vise à traduire le fait que la résolution numérique d'un système linéaire donné pourra être sujette, ou pas, à d'importants problèmes d'erreurs sur la solution obtenue. Ainsi, on dit qu'une matrice inversible est *« bien conditionnée »* (relativement à une norme matricielle) si son conditionnement est proche de l'unité. Au contraire, elle dite *« mal conditionnée »* si son conditionnement est très grand devant 1. Les matrices unitaires (ou orthogonales) étant très bien conditionnées, on comprend l'intérêt justifié de faire intervenir ces matrices plutôt que d'autres dans diverses méthodes numériques matricielles.

Références de l'annexe

[Tau49] O. Taussky. A recurring theorem on determinants. *Amer. Math. Monthly*, 56(10) :672–676, 1949.

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

Annexe B

Rappels d'analyse

Dans cette annexe, on rappelle en les démontrant quelques résultats d'analyse auxquels on fait appel dans les chapitres 5, 6 et 7.

Théorème B.1 *(« 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 $f(c)=y$. □

Le théorème des valeurs intermédiaires permet de démontrer le théorème suivant.

Théorème B.2 *(« théorème de Rolle[1] ») 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. Notons $m=\inf\limits_{x\in[a,b]} f(x)$ et $M=\sup\limits_{x\in[a,b]} f(x)$. 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 $\dfrac{f(x)-f(c)}{x-c}\leq 0$, et si $x<c$, on obtient $\dfrac{f(x)-f(c)}{x-c}\geq 0$. 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 $f'(c)=0$. □

Le théorème de Rolle permet à son tour de prouver le résultat suivant, appelé le *théorème des accroissements finis*.

Théorème B.3 *(« 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*

$$f'(c)=\frac{f(b)-f(a)}{b-a}.$$

1. Michel Rolle (21 avril 1652 - 8 novembre 1719) était un mathématicien français. S'il inventa la notation $\sqrt[n]{x}$ pour désigner la racine $n^{\text{ème}}$ 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.

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

$$f'(c) = \frac{f(b)-f(a)}{b-a}.$$

□

On déduit directement l'*inégalité des accroissements finis* du théorème B.3. 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.4 ***(« 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*

$$\forall x \in]a,b[,\ |f'(x)| \leq M,$$

alors on a

$$|f(b) - f(a)| \leq M|b-a|.$$

Le théorème suivant constitue une généralisation du théorème des accroissements finis.

Théorème B.5 ***(« formule de Taylor[2]–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*

$$f(b) = f(a) + \frac{f'(a)}{1!}(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(b-a)^{n+1},$$

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 d'autre part $\varphi(a) = \varphi(b) = 0$. 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 $A = f^{(n+1)}(c)$. □

Le résultat ci-dessous est une autre conséquence du théorème des valeurs intermédiaires.

2. 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.

Théorème B.6 *(« théorème de la moyenne ») Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et $f : [a,b] \to \mathbb{R}$ une fonction 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

$$m\,(b-a) \leq \int_a^b f(t)\,\mathrm{d}t \leq M\,(b-a).$$

La conclusion s'obtient grâce au théorème des valeurs intermédiaires. □

Théorème B.7 *(« théorème de la moyenne généralisé ») Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et $f : [a,b] \to \mathbb{R}$ une fonction continue sur $[a,b]$ et $g : [a,b] \to \mathbb{R}$ une fonction 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

$$m\,g(x) \leq f(x)\,g(x) \leq M\,g(x),\ \forall x \in [a,b].$$

En intégrant ces inégalités entre a et b, il vient

$$m\int_a^b g(t)\,\mathrm{d}t \leq \int_a^b f(t)\,g(t)\,\mathrm{d}t \leq M\int_a^b g(t)\,\mathrm{d}t.$$

Si l'intégrale de g entre a et b est nulle, le résultat est trivialement vérifié. Sinon, on a

$$m \leq \frac{\int_a^b f(t)\,g(t)\,\mathrm{d}t}{\int_a^b g(t)\,\mathrm{d}t} \leq M,$$

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 sur $]a,b[$, plutôt que continue sur $[a,b]$.

SECONDE FORMULE (f continue décroissante, g bornée) ?

Théorème B.8 *(« théorème de la moyenne discrète ») Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et $f : [a,b] \to \mathbb{R}$ une fonction continue sur $[a,b]$, x_j, $j = 0,\ldots,n$, $n+1$ points de $[a,b]$ et δ_j, $j = 0,\ldots,n$, $n+1$ constantes toutes de 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

$$m\sum_{j=0}^{n} \delta_j \leq \sum_{j=0}^{n} \delta_j\,f(x_j) \leq M\sum_{j=0}^{n} \delta_j.$$

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

$$F(\underline{x}) \leq \sum_{j=0}^{n} \delta_j\,f(x_j) \leq F(\overline{x}),$$

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. □