Introduction à l'analyse numérique

Enseignement des mathématiques

Introduction à l'analyse numérique

Jacques Rappaz
Marco Picasso

Presses polytechniques et universitaires romandes

Les auteurs et l'éditeur remercient l'Ecole polytechnique fédérale de Lausanne
dont le soutien financier a rendu possible la publication de cet ouvrage.

LA COLLECTION «ENSEIGNEMENT DES MATHÉMATIQUES» EST ÉDITÉE SOUS LA DIRECTION DU PROFESSEUR ROBERT C. DALANG

Recherche opérationnelle pour ingénieurs I
Dominique de Werra, Thomas M. Liebling, Jean-François Hêche

Recherche opérationnelle pour ingénieurs II
Jean-François Hêche, Thomas M. Liebling, Dominique de Werra

Calcul différentiel et intégral
Jacques Douchet et Bruno Zwahlen
1 Fonctions réelles d'une variable réelle
2 Fonctions réelles de plusieurs variables réelles
3 Fonctions réelles d'une variable réelle – Exercices résolus
4 Fonctions réelles de plusieurs variables réelles – Exercices résolus

Algèbre linéaire
Aide-mémoire, exercices et applications
Robert C. Dalang et Amel Chaabouni

Analyse avancée pour ingénieurs
Bernard Dacorogna, Chiara Tanteri

Initiation aux probabilités
Sheldon M. Ross

Cours d'Analyse
Srishti D. Chatterji
1 Analyse vectorielle
2 Analyse complexe
3 Equations différentielles

DANS LA COLLECTION «MÉTHODES MATHÉMATIQUES POUR L'INGÉNIEUR»

Introduction à la statistique
Stephan Morgenthaler

Aide-mémoire d'analyse
Heinrich Matzinger

Les Presses polytechniques et universitaires romandes sont une fondation
scientifique dont le but est principalement la diffusion des travaux
de l'Ecole polytechnique fédérale de Lausanne ainsi que d'autres
universités et écoles d'ingénieurs francophones.
Le catalogue de leurs publications peut être obtenu par courrier aux
Presses polytechniques et universitaires romandes,
EPFL – Centre Midi, CH-1015 Lausanne, par E-Mail à ppur@epfl.ch,
par téléphone au (0)21 693 41 40, ou par fax au (0)21 693 40 27.

www.ppur.org

ISBN 2-88074-363-X

Avant-Propos

La plupart des phénomènes physiques, chimiques ou biologiques issus de la technologie moderne sont régis par des systèmes complexes d'équations aux dérivées partielles. La résolution numérique de ces systèmes d'équations à l'aide d'un ordinateur nécessite des connaissances approfondies en mathématiques. Chacune de ces équations aux dérivées partielles peut être répertoriée par type. Ainsi, les équations de type elliptique décrivent les phénomènes de diffusion stationnaire, les équations de type parabolique décrivent les phénomènes de diffusion évolutive et les équations de type hyperbolique décrivent les phénomènes de transport à vitesse finie. Pour chaque type d'équation, il existe de nombreuses méthodes numériques à disposition, classées par catégories. Ainsi on parle de méthodes de différences finies, d'éléments finis, de volumes finis, de méthodes spectrales, etc. L'utilisation de ces méthodes numériques nécessite la connaissance d'outils de base tels que l'interpolation polynômiale, la dérivation et l'intégration numérique d'une fonction, la résolution de grands systèmes linéaires et non linéaires, l'intégration d'équations différentielles ordinaires.

Il existe de nombreux ouvrages consacrés à l'étude mathématique et à la résolution numérique des équations aux dérivées partielles. Ce livre n'a pas la prétention d'être un exposé exhaustif sur le sujet. Par contre, il a pour but de fournir au lecteur les notions de base afin de pouvoir aborder de tels problèmes.

Le contenu de ce livre est une version étoffée du cours polycopié que le premier auteur enseigne depuis plusieurs années aux élèves ingénieurs du premier cycle de l'Ecole Polytechnique Fédérale de Lausanne. Le livre est divisé en 14 chapitres, chaque chapitre étant lui-même divisé en sections. Les premières sections de chaque chapitre contiennent généralement des arguments simples, les dernières sections développent des notions plus complexes. L'avant-dernière section de chaque chapitre contient des exercices corrigés, de sorte que le lecteur puisse juger son niveau de compréhension. Enfin, la dernière section de chaque chapitre contient des notes bibliographiques et commentaires destinés aux ingénieurs qui utilisent la simulation numérique dans leurs travaux de recherche.

Les 10 premiers chapitres exposent les outils de base de l'analyse numérique. Ces outils sont contenus dans la plupart des logiciels de calcul scientifique disponibles dans le commerce. Les 4 derniers chapitres sont consacrés à l'analyse numérique des équations aux dérivées partielles et sont de lecture plus difficile que les précédents. Toutefois, l'ouvrage a été conçu de sorte que les premières

sections des 4 derniers chapitres restent abordables par des élèves ingénieurs de premier cycle universitaire.

Les mathématiques contenues dans ce livre sont en grande partie enseignées par les deux auteurs aux étudiants du premier cycle des sections de génie civil, génie rural, mécanique, matériaux, microtechnique, physique, informatique, systèmes de communications de l'Ecole Polytechnique Fédérale de Lausanne. La compréhension de ces mathématiques requiert une bonne connaissance préalable du calcul différentiel et intégral et de l'algèbre linéaire, matières généralement enseignées lors de la première année d'études scientifiques universitaires. Dans un souci de concision, quelques résultats sont énoncés sans démonstration. Cependant, les auteurs se sont efforcés de maintenir une certaine rigueur mathématique dans la formulation de ces résultats.

Les auteurs tiennent à remercier Monsieur le Professeur S.D. Chatterji, Directeur de la collection mathématiques des Presses Polytechniques Universitaires Romandes, pour ses encouragements et son intérêt à la publication de cet ouvrage. Le travail de dactylographie a été effectué avec grand soin par Madame J. Mosetti. Qu'elle trouve ici la reconnaissance et les remerciements des deux auteurs.

Table des matières

Chapitre 1

Problèmes d'interpolation

1.1 Position du problème

Supposons que l'on veuille chercher un polynôme p de degré $n \geq 0$ qui, pour des valeurs t_0, t_1, t_2, ..., t_n distinctes données, prenne les valeurs p_0, p_1, p_2, ..., p_n respectivement, c'est-à-dire

$$p(t_j) = p_j \qquad \text{pour } 0 \leq j \leq n. \tag{1.1}$$

Une manière apparemment simple de résoudre ce problème est d'écrire

$$p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n \tag{1.2}$$

où $a_0, a_1, a_2, \ldots, a_n$ sont des coefficients qui devront être déterminés (clairement, si les coefficients a_j, $0 \leq j \leq n$ sont connus alors le polynôme p est connu). Les $(n+1)$ relations (1.1) s'écrivent alors :

$$a_0 + a_1 t_j + a_2 t_j^2 + a_3 t_j^3 + \cdots + a_n t_j^n = p_j, \qquad 0 \leq j \leq n. \tag{1.3}$$

Puisque les valeurs t_j et p_j, $0 \leq j \leq n$, sont connues, les relations (1.3) forment un système de $(n+1)$ équations à $(n+1)$ inconnues a_0, a_1, a_2, ..., a_n. Une manière différente d'écrire (1.3) est la suivante.

Soit T la $(n+1) \times (n+1)$ matrice définie par :

$$T = \begin{bmatrix} 1 & t_0 & t_0^2 & t_0^3 & \ldots & t_0^n \\ 1 & t_1 & t_1^2 & t_1^3 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & t_2^3 & \ldots & t_2^n \\ 1 & t_3 & t_3^2 & t_3^3 & \ldots & t_3^n \\ \cdot & \cdot & \cdot & \cdot & \ldots & \cdot \\ 1 & t_n & t_n^2 & t_n^3 & \ldots & t_n^n \end{bmatrix}.$$

Définition 1.1 *Nous dirons que T est la matrice de Vandermonde associée aux points t_0, t_1, t_2, ..., t_n.*

Si $\vec{a}$ et $\vec{p}$ sont les $(n+1)$-vecteurs colonnes suivants :

$$\vec{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \qquad \vec{p} = \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix},$$

nous pouvons écrire (1.3) sous forme matricielle :

$$T\vec{a} = \vec{p}. \tag{1.4}$$

Ainsi, le problème consistant à chercher le polynôme p satisfaisant (1.1) peut se réduire à résoudre le système linéaire (1.4), c'est-à-dire à déterminer $\vec{a}$ puisque T et $\vec{p}$ sont connus.

Résoudre un système linéaire de $(n+1)$ équations à $(n+1)$ inconnues n'est pas une tâche triviale (chap. 4, 5 et 6). La méthode que nous venons de décrire pour trouver le polynôme p n'est pas une bonne méthode. Dans la suite nous présentons une méthode plus astucieuse pour construire le polynôme p.

1.2 Base de Lagrange

Il est facile de résoudre le problème (1.1) lorsque toutes les valeurs p_j sont égales à zéro sauf une, qui est fixée à 1. Soit k un entier donné entre 0 et n et supposons que l'on ait $p_k = 1$ et $p_j = 0$ pour $j \neq k$. Soit φ_k la fonction de t définie par

$$\varphi_k(t) = \frac{(t-t_0)(t-t_1)\cdots(t-t_{k-1})(t-t_{k+1})\cdots(t-t_n)}{(t_k-t_0)(t_k-t_1)\cdots(t_k-t_{k-1})(t_k-t_{k+1})\cdots(t_k-t_n)}. \tag{1.5}$$

Clairement, le numérateur de φ_k est un produit de n termes $(t-t_j)$ avec $j \neq k$ et est donc un polynôme de degré n en t. Le dénominateur de φ_k est une constante et il est alors facile de vérifier que

$$\begin{array}{ll} (i) & \varphi_k \text{ est un polynôme de degré } n, \\ (ii) & \varphi_k(t_j) = 0 \qquad \text{si } j \neq k, 0 \leq j \leq n, \\ (iii) & \varphi_k(t_k) = 1. \end{array}$$

A chaque point t_k, nous avons donc associé un polynôme φ_k de degré n valant un en t_k et zéro aux autres points t_j, $j \neq k$.

Les polynômes φ_0, φ_1, φ_2, ..., φ_n sont linéairement indépendants. En effet si α_0, α_1, α_2, ..., α_n sont $(n+1)$ nombres réels tels que $\sum_{j=0}^{n} \alpha_j \varphi_j(t) = 0$, $\forall t \in \mathbb{R}$, alors pour $t = t_k$ nous obtenons :

$$0 = \sum_{j=0}^{n} \alpha_j \underbrace{\varphi_j(t_k)}_{\substack{0 \text{ si } j \neq k \\ 1 \text{ si } j = k}} = \alpha_k,$$

et par conséquent tous les α_k, $k = 0, 1, \ldots, n$ sont identiquement nuls.

Notons maintenant $\mathbb{P}_n$ l'espace vectoriel formé par tous les polynômes de degré inférieur ou égal à n. Il est bien connu que $\mathbb{P}_n$ est un espace vectoriel de dimension $(n+1)$ et que sa base canonique est donnée par $1, t, t^2, t^3, \ldots, t^n$. Le fait que $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ soient des polynômes de degré n linéairement indépendants montre que ces derniers forment aussi une base de $\mathbb{P}_n$. Ainsi nous adopterons la définition suivante :

Définition 1.2 *Nous dirons que $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ est la base de Lagrange de $\mathbb{P}_n$ associée aux points $t_0, t_1, t_2, \ldots, t_n$.*

Exemple 1.1 Prenons $n = 2$, $t_0 = -1$, $t_1 = 0$, $t_2 = 1$. La base de Lagrange de $\mathbb{P}_2$ associée aux points -1, 0 et 1 est formée par les polynômes $\varphi_0, \varphi_1, \varphi_2$ définis par

$$\varphi_0(t) \equiv \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} = \frac{1}{2}t(t-1) = \frac{1}{2}t^2 - \frac{1}{2}t; \tag{1.6}$$

$$\varphi_1(t) \equiv \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} = -(t+1)(t-1) = 1 - t^2; \tag{1.7}$$

$$\varphi_2(t) \equiv \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)} = \frac{1}{2}(t+1)t = \frac{1}{2}t^2 + \frac{1}{2}t. \tag{1.8}$$

Les graphes de φ_0, φ_1, φ_2 sur l'intervalle $[-1, +1]$ uniquement sont représentés dans la figure 1.1.

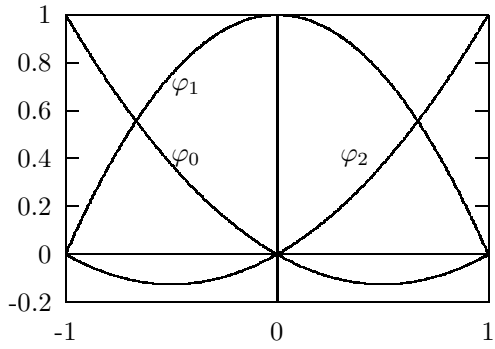

Fig. 1.1 Base de Lagrange de $\mathbb{P}_2$ associée aux points -1, 0 et 1.

1.3 Interpolation de Lagrange

Revenons au problème (1.1) consistant à chercher un polynôme p de degré n qui prenne des valeurs données p_0, p_1, p_2, ..., p_n en des points distincts donnés t_0, t_1, t_2, ..., t_n.

Soit $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ la base de Lagrange de $\mathbb{P}_n$ associée aux points t_0, t_1, t_2, ..., t_n. Alors le polynôme p cherché est défini par :

$$p(t) = p_0\varphi_0(t) + p_1\varphi_1(t) + \cdots + p_n\varphi_n(t) = \sum_{j=0}^{n} p_j\varphi_j(t). \tag{1.9}$$

En effet, puisque p est une combinaison linéaire de $(n+1)$ polynômes φ_0, φ_1, φ_2, ..., φ_n de degré n, alors p est lui-même un polynôme de degré n, c'est-à-dire $p \in \mathbb{P}_n$. D'autre part, si nous utilisons les propriétés des polynômes φ_j, nous avons pour $k = 0, 1, 2, \ldots, n$:

$$p(t_k) = \sum_{j=0}^{n} p_j \underbrace{\varphi_j(t_k)}_{\substack{0 \text{ si } j\neq k \\ 1 \text{ si } j=k}} = p_k \tag{1.10}$$

qui est bien la relation (1.1).

Il est important de remarquer que nous avons construit explicitement une solution du problème (1.1) et ceci pour n'importe quelles valeurs $p_0, p_1, \ldots, p_n$ données. Ceci montre que le système (1.4) a toujours une solution $\vec{a}$ pour n'importe quel $\vec{p}$ et ainsi la matrice de Vandermonde T est régulière. La solution (1.9) du problème (1.1) est donc unique.

Exemple 1.2 Trouver un polynôme de degré 2 qui en $t_0 = -1$ vaut $p_0 = 8$, en $t_1 = 0$ vaut $p_1 = 3$ et en $t_2 = 1$ vaut $p_2 = 6$.

D'après ce qui précède, nous avons $p(t) = 8\varphi_0(t) + 3\varphi_1(t) + 6\varphi_2(t)$ où φ_0, φ_1 et φ_2 sont donnés par (1.6)-(1.8). Nous obtenons donc :

$$\begin{aligned} p(t) &= 8\left(\frac{1}{2}t^2 - \frac{1}{2}t\right) + 3(1-t^2) + 6\left(\frac{1}{2}t^2 + \frac{1}{2}t\right) \\ &= 4t^2 - t + 3. \end{aligned}$$

1.4 Interpolation d'une fonction continue par un polynôme

Soit une fonction $f : \mathbb{R} \to \mathbb{R}$ continue donnée et soit t_0, t_1, t_2, ..., t_n, $(n+1)$ points distincts donnés.

Nous cherchons maintenant à interpoler f par un polynôme p de degré n aux points t_j, $0 \leq j \leq n$, c'est-à-dire nous cherchons un polynôme p de degré n tel que

$$p(t_j) = f(t_j), \qquad 0 \leq j \leq n. \tag{1.11}$$

Si $f(t)$ est donnée, alors en posant $p_j = f(t_j)$, $0 \leq j \leq n$ et en suivant ce qui est fait dans la section 1.3, nous obtenons $p(t) = \sum_{j=0}^{n} p_j\varphi_j(t)$ où les φ_j, $0 \leq j \leq n$, forment la base de Lagrange de $\mathbb{P}_n$ associée aux points $t_0, t_1, t_2, \ldots, t_n$. La solution du problème (1.11) est donc définie par :

$$p(t) = \sum_{j=0}^{n} f(t_j)\varphi_j(t) \qquad \forall t \in \mathbb{R}. \tag{1.12}$$

Définition 1.3 *On dira que le polynôme p défini par (1.12) est l'interpolant de f de degré n aux points $t_0, t_1, t_2, \ldots, t_n$.*

Exemple 1.3 Soit f la fonction définie par $f(t) = e^t$. Trouver l'interpolant de f de degré 2 aux points -1, 0 et 1.

Si nous reprenons la formule (1.12), nous avons $p(t) = e^{-1}\varphi_0(t) + e^0\varphi_1(t) + e\varphi_2(t)$ où $\varphi_0, \varphi_1, \varphi_2$ sont donnés par (1.6), (1.7) et (1.8). Ainsi donc nous obtenons

$$\begin{aligned} p(t) &= \frac{1}{e}\left(\frac{1}{2}t^2 - \frac{1}{2}t\right) + (1 - t^2) + e\left(\frac{1}{2}t^2 + \frac{1}{2}t\right) \\ &= \left(\frac{1}{2e} - 1 + \frac{e}{2}\right)t^2 + \left(\frac{e}{2} - \frac{1}{2e}\right)t + 1. \end{aligned}$$

La figure 1.2 montre le graphe de la fonction f et son interpolant de degré 2 aux points -1, 0 et 1.

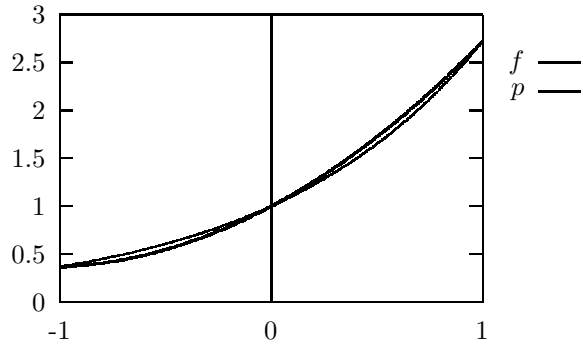

Fig. 1.2 Interpolation de la fonction f définie par $f(t) = e^t$ par un polynôme p de degré 2 aux points -1, 0 et 1.

Soit maintenant une fonction $f : [a, b] \to \mathbb{R}$ continue et donnée sur un intervalle fermé $[a, b]$. Soit n un entier positif et considérons le cas où les points t_j, $j = 0, 1, 2, \ldots, n$, sont équidistribués dans $[a, b]$, c'est-à-dire $t_j = a + jh$, $j = 0, 1, 2, \ldots, n$, avec $h = (b - a)/n$. Soit p l'interpolant de f de degré n aux points $t_0, t_1, \ldots, t_n$ que nous noterons p_n pour bien montrer qu'il dépend de l'entier n choisi au départ. D'après (1.12), p_n est défini par :

$$p_n(t) = \sum_{j=0}^{n} f(t_j)\varphi_j(t), \tag{1.13}$$

où $\varphi_0, \varphi_1, \ldots, \varphi_n$ est la base de Lagrange de $\mathbb{P}_n$ associée à $t_0, t_1, \ldots, t_n$. On peut montrer le résultat suivant :

Théorème 1.1 *Supposons que f soit $(n+1)$ fois continûment dérivable sur l'intervalle $[a,b]$. Alors si p_n est défini par (1.13) nous avons :*

$$\max_{t\in[a,b]} |f(t) - p_n(t)| \leq \frac{1}{2(n+1)} \left(\frac{b-a}{n}\right)^{(n+1)} \max_{t\in[a,b]} |f^{(n+1)}(t)| \qquad (1.14)$$

où $f^{(n+1)}(t) = d^{n+1}f(t)/dt^{n+1}$.

L'inégalité (1.14) est une estimation d'erreur entre la fonction f et son interpolant p_n de degré n aux points $t_0, t_1, t_2, \ldots, t_n$ équirépartis dans $[a,b]$. A priori nous pourrions penser que cette erreur converge vers zéro lorsque n tend vers l'infini puisque nous avons

$$\lim_{n\to\infty} \frac{1}{2(n+1)} \left(\frac{b-a}{n}\right)^{(n+1)} = 0.$$

En réalité, cette affirmation est souvent fausse car $\max_{t\in[a,b]} |f^{(n+1)}(t)|$ peut croître très rapidement avec n. Ce phénomène est illustré dans l'exemple suivant.

Exemple 1.4 (Runge) Soit $f(t) = 1/(1+25t^2)$ que l'on considère sur l'intervalle $[-1,+1]$. La fonction $f(t)$ est infiniment dérivable sur l'intervalle $[-1,+1]$ et $|f^{(n)}(1)|$ devient très rapidement grand lorsque n tend vers l'infini. La figure 1.3 montre son interpolant p_n de degré n aux points $t_j = -1 + 2j/n$, $j = 0, 1, \ldots, n$, pour $n = 5$ et $n = 10$.

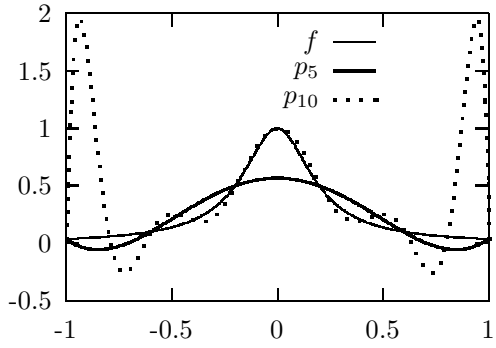

Fig. 1.3 Interpolant de $f(t) = 1/(1+25t^2)$ de degré 5 et 10 avec des points équirépartis.

Nous observons que, au voisinage des extrémités de l'intervalle $[-1,+1]$, l'interpolant présente de grandes oscillations (instabilités numériques). Nous concluons donc qu'il n'est pas indiqué d'interpoler une fonction par un polynôme de degré n élevé en des points $t_0, t_1, \ldots, t_n$ équidistribués.

Par contre, si nous choisissons les points dits de Tchebycheff

$$t_j = a + \frac{(b-a)}{2}\left(1 + \cos\frac{(2j+1)\pi}{2(n+1)}\right), \qquad j = 0, 1, 2, \ldots, n,$$

pour construire l'interpolant p_n de f, alors l'erreur $\max_{t\in[a,b]} |f(t) - p_n(t)|$ tend vers zéro lorsque n tend vers l'infini, comme le montre la figure 1.4.

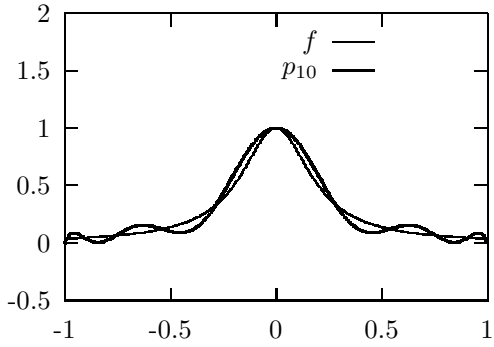

Fig. 1.4 Interpolant de $f(t) = 1/(1 + 25t^2)$ de degré 10 avec des points de Tchebycheff.

1.5 Interpolation d'Hermite

Les problèmes d'interpolation que nous venons de considérer sont des ***problèmes d'interpolation de Lagrange*** ; ils font intervenir les valeurs de polynômes et de fonctions en certains points, mais ne tiennent pas compte des dérivées. Il existe d'autres problèmes d'interpolation pour lesquels les valeurs de $p(t)$ et de la dérivée $p'(t)$ sont données en certains points. On parle dans ce cas d'interpolation d'Hermite. Pour illustrer notre propos, nous présentons un seul exemple qui est l'interpolation d'Hermite par des cubiques (polynômes de degré 3).

Soit $t_0 < t_1$ deux points donnés et soit p_0, p_1, p'_0, p'_1 quatre nombres réels donnés. Nous cherchons un polynôme p de degré 3 tel que

$$p(t_0) = p_0, \qquad p(t_1) = p_1, \tag{1.15}$$

$$p'(t_0) = p'_0, \qquad p'(t_1) = p'_1, \tag{1.16}$$

où $p'(t)$ est la dérivée de p au point t.

Les conditions (1.15) imposent la valeur de p en t_0 et t_1 ; les conditions (1.16) imposent la valeur de la dérivée p' de p en t_0 et t_1. Un polynôme de degré 3 s'écrit sous la forme

$$p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3;$$

nous avons donc fixé dans (1.15), (1.16) quatre conditions pour déterminer quatre coefficients a_0, a_1, a_2 et a_3. Nous pourrions ramener le problème à la résolution d'un système linéaire de 4 équations à 4 inconnues a_0, a_1, a_2 et a_3. Si nous montrons que ce système linéaire a toujours une solution pour n'importe quelles valeurs p_0, p_1, p'_0, p'_1, alors cette solution est unique.

Pour construire p, nous commençons par construire une base $\varphi_0, \varphi_1, \psi_0, \psi_1$ des polynômes de degré 3 associée aux points t_0 et t_1.

i) On construit φ_0 tel que φ_0 soit un polynôme de degré 3 et

$$\varphi_0(t_0) = 1, \qquad \varphi_0'(t_0) = \varphi_0(t_1) = \varphi_0'(t_1) = 0.$$

On vérifie que l'on a :

$$\varphi_0(t) = -\frac{(t-t_1)^2(2t+t_1-3t_0)}{(t_0-t_1)^3}. \tag{1.17}$$

ii) De même, on construit φ_1 tel que φ_1 soit un polynôme de degré 3 et

$$\varphi_1(t_1) = 1, \qquad \varphi_1'(t_1) = \varphi_1(t_0) = \varphi_1'(t_0) = 0.$$

On obtient :

$$\varphi_1(t) = -\frac{(t-t_0)^2(2t+t_0-3t_1)}{(t_1-t_0)^3}. \tag{1.18}$$

iii) On construit ψ_0 tel que ψ_0 soit un polynôme de degré 3 et

$$\psi_0'(t_0) = 1, \qquad \psi_0(t_0) = \psi_0(t_1) = \psi_0'(t_1) = 0.$$

On vérifie que l'on a :

$$\psi_0(t) = \frac{(t-t_1)^2(t-t_0)}{(t_0-t_1)^2}. \tag{1.19}$$

iv) De même, on construit ψ_1 tel que ψ_1 soit un polynôme de degré 3 et

$$\psi_1'(t_1) = 1, \qquad \psi_1(t_1) = \psi_1(t_0) = \psi_1'(t_0) = 0.$$

On obtient :

$$\psi_1(t) = \frac{(t-t_0)^2(t-t_1)}{(t_1-t_0)^2}. \tag{1.20}$$

Nous pouvons vérifier que ces quatre polynômes $\varphi_0, \varphi_1, \psi_0, \psi_1$ sont linéairement indépendants.

Définition 1.4 *Les quatre polynômes $\varphi_0, \varphi_1, \psi_0, \psi_1$ forment une base de $\mathbb{P}_3$ que nous appellerons base d'Hermite de type cubique associée à t_0 et t_1.*

Dans la figure 1.5, nous avons représenté φ_0, φ_1, ψ_0 et ψ_1 sur l'intervalle $[t_0, t_1]$ uniquement.

Ayant construit la base d'Hermite $\varphi_0, \varphi_1, \psi_0, \psi_1$ des polynômes de degré 3 associée aux points t_0, t_1, nous vérifions facilement que si p est le polynôme de degré 3 défini par

$$p(t) = p_0\varphi_0(t) + p_1\varphi_1(t) + p_0'\psi_0(t) + p_1'\psi_1(t), \tag{1.21}$$

alors p satisfait les relations (1.15) (1.16).

Si nous prenons maintenant une fonction f une fois continûment dérivable sur l'intervalle $[t_0, t_1]$ et si nous construisons le polynôme p défini par

$$p(t) = f(t_0)\varphi_0(t) + f(t_1)\varphi_1(t) + f'(t_0)\psi_0(t) + f'(t_1)\psi_1(t),$$

alors nous dirons que p est l'interpolant d'Hermite de f par des cubiques sur $[t_0, t_1]$. Nous avons naturellement

$$\begin{aligned} p(t_0) &= f(t_0), \qquad & p(t_1) &= f(t_1), \\ p'(t_0) &= f'(t_0), \qquad & p'(t_1) &= f'(t_1). \end{aligned}$$

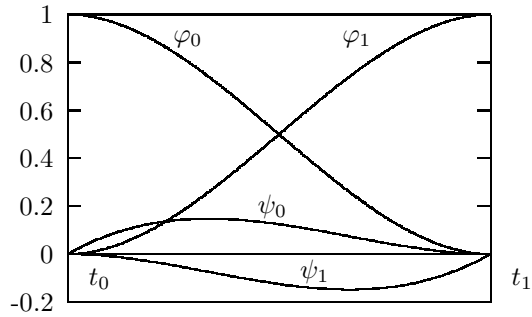

Fig. 1.5 Base d'Hermite de type cubique.

1.6 Interpolation par intervalles

L'interpolation d'une fonction par des polynômes de degré élevé en des points équidistribués peut engendrer des instabilités numériques comme nous l'avons vu dans la section 1.4. De plus, l'interpolation d'une fonction par des polynômes de degré élevé n'est pas justifiée lorsque la fonction à interpoler n'est pas régulière. C'est la raison pour laquelle l'interpolation par intervalles est souvent utilisée.

Soit f une fonction continue donnée sur un intervalle $[a,b]$ et soit $(N+1)$ points $x_0 = a < x_1 < x_2 < x_3 < \ldots < x_N = b$ dans l'intervalle $[a,b]$. Pour chaque intervalle $[x_i, x_{i+1}]$, il est possible de choisir $(n-1)$ points intérieurs équirépartis notés

$$x_{i,1} < x_{i,2} < x_{i,3} < \ldots < x_{i,n-1}.$$

En posant $t_0 = x_i$, $t_j = x_{i,j}$ avec $1 \leq j \leq n-1$, $t_n = x_{i+1}$, nous pouvons interpoler f aux points t_j, $0 \leq j \leq n$, par un polynôme de degré n comme nous l'avons fait dans la section 1.4. Dans la suite nous définissons

$$h = \max_{0 \leq i \leq N-1} |x_{i+1} - x_i|,$$

et nous construisons une fonction $f_h : x \in [a,b] \longrightarrow f_h(x) \in \mathbb{R}$ telle que f_h restreinte à cet intervalle $[x_i, x_{i+1}]$ soit justement ce polynôme d'interpolation de degré n.

Définition 1.5 *On dira que f_h est l'interpolant de degré n par intervalles de la fonction f.*

Nous démontrons le résultat suivant :

Théorème 1.2 *Soit n un entier positif donné, soit $f : [a,b] \to \mathbb{R}$ une fonction donnée que nous supposons $(n+1)$ fois continûment dérivable sur l'intervalle $[a,b]$ et soit f_h son interpolant de degré n par intervalles. Alors il existe une constante C (indépendante du choix des x_i, $1 \leq i \leq N-1$) telle que*

$$\max_{x \in [a,b]} |f(x) - f_h(x)| \leq C h^{n+1}. \tag{1.22}$$

Démonstration
En utilisant le théorème 1.1 sur l'intervalle $[x_i, x_{i+1}]$ en lieu et place de l'intervalle $[a, b]$, nous obtenons :

$$\max_{t\in[x_i,x_{i+1}]} |f(t) - f_h(t)| \leq \frac{1}{2(n+1)} \left(\frac{x_{i+1} - x_i}{n}\right)^{(n+1)} \max_{t\in[x_i,x_{i+1}]} |f^{(n+1)}(t)|.$$

Ainsi nous avons

$$\max_{t\in[x_i,x_{i+1}]} |f(t) - f_h(t)| \leq C \left(\max_{0\leq j\leq N-1} |x_{j+1} - x_j|\right)^{n+1} \tag{1.23}$$

où la constante C est donnée par

$$C = \frac{1}{2(n+1)n^{(n+1)}} \max_{t\in[a,b]} |f^{(n+1)}(t)|$$

et où $i = 0, 1, 2, \ldots, N-1$. Ainsi (1.22) est une conséquence immédiate de (1.23). ■

Une interprétation du théorème 1.2 est la suivante. Si on se donne un entier positif n et si on prend N points $x_1, x_2, \ldots, x_N$ avec N de plus en plus grand de façon à ce que h soit de plus en plus petit, alors $\max_{x\in[a,b]} |f(x)-f_h(x)|$ converge vers zéro lorsque h tend vers zéro. Par exemple, si on pose $h = (b - a)/N$ et si $x_i = a + ih$ avec $i = 0, 1, 2, \ldots, N$, on aura

$$\max_{x\in[a,b]} |f(x) - f_h(x)| \leq Ch^{n+1}.$$

En pratique, on prendra N grand et n petit ($n = 1$ ou 2 ou 3 ou 4).

Exemple 1.5 Considérons le cas où $f(x) = x^{1.7} + 0.1e^{3x}\sin(13x)$, $a = 0$, $b = 0.8$, $N = 4$ (pour les besoins de la figure), $x_0 = 0$, $x_1 = 0.2$, $x_2 = 0.4$, $x_3 = 0.6$, $x_4 = 0.8$. Si $n = 1$, il n'y a pas de point intérieur aux intervalles $[x_i, x_{i+1}]$ et l'interpolation sur chaque intervalle se fait par des polynômes de degré 1. La figure 1.6 montre le graphe de l'interpolant par intervalles.

Si $n = 2$, nous choisissons pour point intérieur à $[x_i, x_{i+1}]$ le point milieu $x_{i,1} = \frac{x_i+x_{i+1}}{2}$ et l'interpolation sur chaque intervalle se fera par des polynômes de degré 2. La figure 1.7 montre le graphe de l'interpolant par intervalles.

L'interpolation de Lagrange par intervalles met en évidence des sauts de la dérivée première en chaque point x_i (fig. 1.6 et 1.7). Une manière de construire un interpolant par intervalles plus lisse est d'utiliser l'interpolation d'Hermite avec des cubiques sur chaque intervalle $[x_i, x_{i+1}]$ (sect. 1.5). Clairement si f est une fonction C^1 sur l'intervalle $[a, b]$ et si, sur chaque intervalle $[x_i, x_{i+1}]$, nous interpolons f par un polynôme de degré 3 comme nous l'avons déjà fait dans la section 1.5 sur l'intervalle $[t_0, t_1]$, alors en chaque point x_i la fonction qui interpole f prend la valeur $f(x_i)$ et sa dérivée première prend la valeur $f'(x_i)$. L'interpolant ainsi construit est une fonction C^1 sur $[a, b]$ (fig. 1.8).

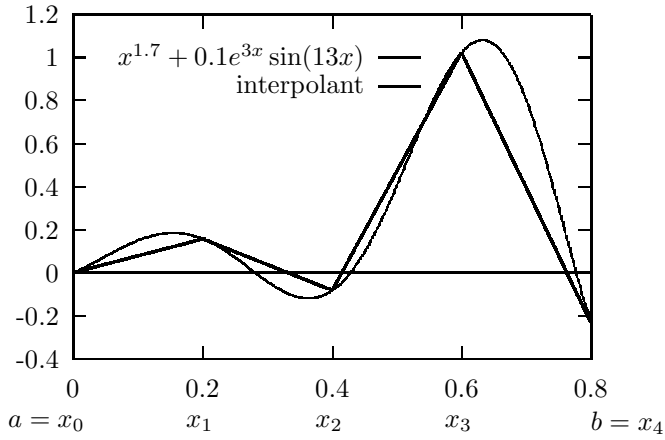

Fig. 1.6 Interpolation par intervalles de f par des polynômes de degré 1.

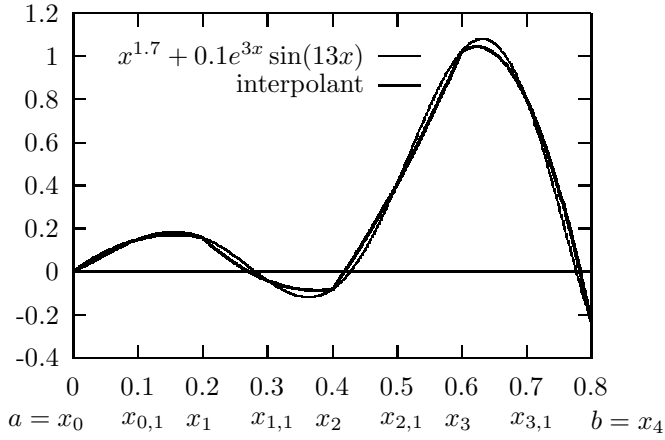

Fig. 1.7 Interpolation par intervalles de f par des polynômes de degré 2.

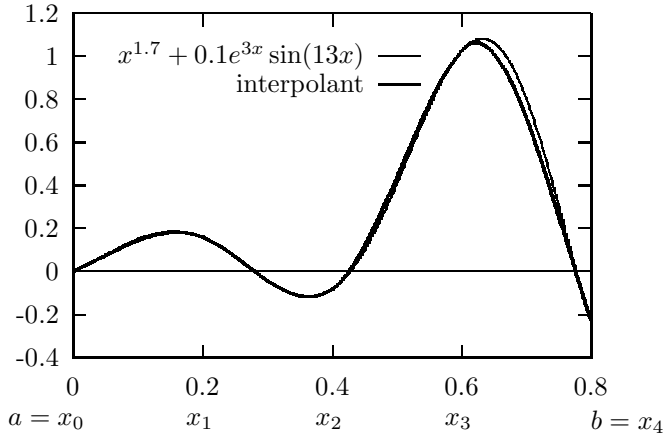

Fig. 1.8 Interpolation par intervalles de f par des polynômes d'Hermite cubiques.

1.7 Exercices

Exercice 1.1 On suppose que l'on connaisse les valeurs $f(t)$ d'une fonction continue aux valeurs de t entières seulement, c'est-à-dire on suppose connues les valeurs

$$f(k),\ k = 0,\ \pm 1,\ \pm 2,\ \ldots$$

Si $t \in \mathbb{R}$, t non entier, on définit une approximation $p(t)$ de $f(t)$ en interpolant la fonction f par un polynôme de degré 3 aux quatre points entiers les plus proches de t.

Ecrire un algorithme qui, pour t donné, fournit $p(t)$.

Solution
Soit $t \in \mathbb{R}$, t non entier et soit $k = E[t]$ la partie entière de t. Puisque $t \in]k, k+1[$ il s'agit de définir le polynôme p dont le graphe passe par les points

$$\big(k-1, f(k-1)\big),\quad \big(k, f(k)\big),\quad \big(k+1, f(k+1)\big),\quad \big(k+2, f(k+2)\big).$$

Tableau 1.1 Algorithme fournissant $p = p(t)$ à partir de t.

Algorithme	**Commentaires**
$\psi_0(x) := -\frac{1}{6}x(x-1)(x-2)$ $\psi_1(x) := \frac{1}{2}(x+1)(x-1)(x-2)$ $\psi_2(x) := -\frac{1}{2}(x+1)x(x-2)$ $\psi_3(x) := \frac{1}{6}(x+1)x(x-1)$	Définition des fonctions $\psi_0,\ \psi_1,\ \psi_2,\ \psi_3$
Lire t $k := E[t]$	Calcul de k
Si $t = k$ $p := f_k$;	Si t est entier
Sinon $p := f_{k-1} * \psi_0(t-k)$ $+f_k * \psi_1(t-k)$ $+f_{k+1} * \psi_2(t-k)$ $+f_{k+2} * \psi_3(t-k)$	Sinon on calcule $p(t)$

Conformément aux résultats de la section 1.4, p est le polynôme défini par

$$\begin{aligned} p(t) =& f(k-1)\varphi_0(t) + f(k)\varphi_1(t) \\ &+ f(k+1)\varphi_2(t) + f(k+2)\varphi_3(t), \end{aligned} \tag{1.24}$$

où φ_0, φ_1, φ_2, φ_3 est la base de Lagrange des polynômes de degré 3 associée aux points $k-1$, k, $k+1$, $k+2$. En utilisant l'égalité (1.5), un calcul simple donne :

$$\begin{aligned} \varphi_0(t) &= -\frac{1}{6}(t-k)(t-k-1)(t-k-2), \\ \varphi_1(t) &= \frac{1}{2}(t-k+1)(t-k-1)(t-k-2), \\ \varphi_2(t) &= -\frac{1}{2}(t-k+1)(t-k)(t-k-2), \\ \varphi_3(t) &= \frac{1}{6}(t-k+1)(t-k)(t-k-1). \end{aligned}$$

Effectuons le changement de variable $x = t-k$. Puisque $t \in]k, k+1[$, alors $x \in]0,1[$ et nous définissons les fonctions ψ_0, ψ_1, ψ_2, ψ_3 par

$$\begin{aligned} \psi_0(x) &= \varphi_0(x+k), \qquad & \psi_1(x) &= \varphi_1(x+k), \\ \psi_2(x) &= \varphi_2(x+k), \qquad & \psi_3(x) &= \varphi_3(x+k), \end{aligned}$$

soit encore

$$\begin{aligned} \psi_0(x) &= -\frac{1}{6}x(x-1)(x-2), \\ \psi_1(x) &= \frac{1}{2}(x+1)(x-1)(x-2), \\ \psi_2(x) &= -\frac{1}{2}(x+1)x(x-2), \\ \psi_3(x) &= \frac{1}{6}(x+1)x(x-1). \end{aligned}$$

L'égalité (1.24) s'écrit maintenant :

$$\begin{aligned} p(t) =& f(k-1)\psi_0(t-k) + f(k)\psi_1(t-k) \\ &+ f(k+1)\psi_2(t-k) + f(k+2)\psi_3(t-k). \end{aligned}$$

L'algorithme correspondant est présenté dans le tableau 1.1. Les paramètres d'entrée sont $t \in \mathbb{R}$ et les valeurs $f_k = f(t_k)$. Le paramètre de sortie p est la valeur $p(t)$.

Exercice 1.2 Soit $t_0 < t_1$ deux nombres réels distincts et soit ε tel que $0 < \varepsilon < t_1 - t_0$.

1. Expliciter un polynôme p_ε de degré 3 tel que

$$\begin{aligned} p_\varepsilon(t_0) &= p_\varepsilon(t_0+\varepsilon) = 1, \\ p_\varepsilon(t_1) &= p_\varepsilon(t_1+\varepsilon) = 0. \end{aligned}$$

2. Si $\varphi(t) = \lim_{\varepsilon \to 0} p_\varepsilon(t)$, montrer que $\varphi(t_0) = 1$, $\varphi(t_1) = \varphi'(t_0) = \varphi'(t_1) = 0$, et ainsi φ est une fonction de base des polynômes de degré 3 pour l'interpolation d'Hermite (sect. 1.5).

Solution
1. Soit φ_0, φ_1, φ_2, φ_3 la base de Lagrange de $\mathbb{P}_3$ associée aux points t_0, $t_0 + \varepsilon$, t_1, $t_1 + \varepsilon$. En utilisant le résultat (1.9) nous avons :

$$p_\varepsilon(t) = \varphi_0(t) + \varphi_1(t).$$

D'autre part, l'égalité (1.5) donne :

$$\varphi_0(t) = \frac{(t - t_0 - \varepsilon)(t - t_1)(t - t_1 - \varepsilon)}{(-\varepsilon)(t_0 - t_1)(t_0 - t_1 - \varepsilon)},$$
$$\varphi_1(t) = \frac{(t - t_0)(t - t_1)(t - t_1 - \varepsilon)}{\varepsilon(t_0 - t_1 + \varepsilon)(t_0 - t_1)}.$$

En réduisant les deux fractions ci-dessus à un dénominateur commun, nous obtenons :

$$\begin{aligned} p_\varepsilon(t) &= \frac{(t - t_1)(t - t_1 - \varepsilon)}{\varepsilon(t_0 - t_1)(t_0 - t_1 - \varepsilon)(t_0 - t_1 + \varepsilon)} \\ &\qquad \times \Big((t_0 - t_1 - \varepsilon)(t - t_0) - (t_0 - t_1 + \varepsilon)(t - t_0 - \varepsilon) \Big) \\ &= \frac{(t - t_1)(t - t_1 - \varepsilon)(3t_0 - t_1 - 2t + \varepsilon)}{(t_0 - t_1)(t_0 - t_1 - \varepsilon)(t_0 - t_1 + \varepsilon)}. \end{aligned}$$

2. Par définition de φ, nous avons :

$$\varphi(t) = \lim_{\varepsilon \to 0} p_\varepsilon(t) = \frac{(t - t_1)^2(3t_0 - t_1 - 2t)}{(t_0 - t_1)^3},$$

et donc, en dérivant, nous obtenons :

$$\begin{aligned} \varphi'(t) &= \frac{2(t - t_1)(3t_0 - t_1 - 2t) - 2(t - t_1)^2}{(t_0 - t_1)^3} \\ &= 6\frac{(t - t_1)(t_0 - t)}{(t_0 - t_1)^3}. \end{aligned}$$

Nous avons donc bien $\varphi(t_0) = 1$, $\varphi(t_1) = \varphi'(t_0) = \varphi'(t_1) = 0$, ce qui prouve que φ est une des quatre fonctions de base des polynômes d'Hermite de degré 3 associée aux points t_0 et t_1 (sect. 1.5).

Exercice 1.3 Soit f une fonction continue donnée sur l'intervalle $[-1, +1]$ et soit p le polynôme de degré 2 qui interpole f en les points $-1, 0, +1$. Exprimer $\int_{-1}^{+1} p(t)dt$ en fonction de $f(-1), f(0)$, et $f(+1)$. Vérifier que la formule ainsi obtenue coïncide avec la formule de Simpson de la section 3.4.

Solution
En utilisant le résultat (1.9), le polynôme p est défini par

$$p(t) = f(-1)\varphi_0(t) + f(0)\varphi_1(t) + f(+1)\varphi_2(t),$$

où φ_0, φ_1, φ_2 est la base de Lagrange de $\mathbb{P}_2$ associée aux points $-1, 0, +1$ et est explicitée dans l'exemple 1.1. Nous avons donc

$$\int_{-1}^{+1} p(t)dt = f(-1)\int_{-1}^{+1} \varphi_0(t)dt + f(0)\int_{-1}^{+1} \varphi_1(t)dt + f(+1)\int_{-1}^{+1} \varphi_2(t)dt.$$

Un calcul simple donne

$$\int_{-1}^{+1} \varphi_0(t)dt = \frac{1}{3}, \qquad \int_{-1}^{+1} \varphi_1(t)dt = \frac{4}{3}, \qquad \int_{-1}^{+1} \varphi_2(t)dt = \frac{1}{3},$$

et par conséquent

$$\int_{-1}^{+1} p(t)dt = \frac{1}{3}\Big(f(-1) + 4f(0) + f(+1)\Big).$$

Il semble donc naturel d'approcher $\int_{-1}^{+1} f(t)dt$ par la quantité $J(f)$ définie par

$$J(f) = \frac{1}{3}\Big(f(-1) + 4f(0) + f(+1)\Big).$$

Dans le chapitre 3 nous appellerons la quantité $J(f)$ ***formule de quadrature***. Par construction, la formule de quadrature $J(f)$ intègre exactement les polynômes de degré deux au sens où

$$\int_{-1}^{+1} q(t)dt = J(q)$$

pour tout polynôme q de degré deux. Cette formule de quadrature s'appellera formule de Simpson (sect. 3.4).

1.8 Notes bibliographiques et remarques

Nous avons présenté deux exemples d'interpolation polynômiale, à savoir l'interpolation de Lagrange et l'interpolation d'Hermite. Il existe naturellement d'autres types d'interpolation : interpolation trigonométrique, splines, etc.

A titre d'exemple, considérons l'interpolation d'une fonction $f : [a, b] \to \mathbb{R}$ par une fonction que nous appellerons spline cubique et que nous noterons S. Si $t_0 = a < t_1 < \ldots < t_n = b$, sont $(n + 1)$ points donnés dans l'intervalle $[a, b]$ alors il est possible de trouver une fonction $S : [a, b] \to \mathbb{R}$ continue, telle que $S(t_0) = f(t_0)$, ..., $S(t_n) = f(t_n)$, de dérivées première et seconde continues, polynômiale de degré 3 sur chaque intervalle $[t_{j-1}, t_j]$, $1 \leq j \leq n$ (voir par exemple [25] [28] pour plus de détails).

L'exemple de la section 1.4 mettant en évidence la divergence de l'interpolation est l'exemple de Runge. Un traitement mathématique complet de cet exemple se trouve dans [16].

L'interpolation par intervalles est un des ingrédients de la méthode des éléments finis (chap. 10 à 14). En particulier le théorème 1.2 est le genre de résultats théoriques qui permet de montrer la convergence de la méthode des éléments finis.

Nous reviendrons sur certains aspects de l'interpolation polynômiale dans les chapitres 2 et 3.

De nombreux logiciels scientifiques grand public (par exemple Maple^{TM}, Mathematica^{TM}, Matlab^{TM}) abordent les problèmes d'interpolation. Finalement la plupart des logiciels de Conception Assistée par Ordinateurs (CAO) utilisent des méthodes issues de la théorie de l'interpolation [26].

Chapitre 2

Dérivation numérique

2.1 Dérivées numériques d'ordre 1 et erreur de troncature

Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ supposée continue et de première dérivée f' continue. Si $x_0 \in \mathbb{R}$ est un réel donné, nous pouvons écrire :

$$\begin{aligned} f'(x_0) &= \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0)-f(x_0-h)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0+h/2)-f(x_0-h/2)}{h}. \end{aligned} \tag{2.1}$$

Une idée pour calculer numériquement la première dérivée f' de f au point x_0 consiste donc à se donner une valeur h positive assez petite et à calculer

$$\frac{\Delta_h f(x_0)}{h} \qquad \text{ou} \qquad \frac{\nabla_h f(x_0)}{h} \qquad \text{ou} \qquad \frac{\delta_h f(x_0)}{h} \tag{2.2}$$

après avoir défini les quantités

$$\Delta_h f(x_0) \underset{def}{=} f(x_0+h) - f(x_0), \tag{2.3}$$

$$\nabla_h f(x_0) \underset{def}{=} f(x_0) - f(x_0-h), \tag{2.4}$$

$$\delta_h f(x_0) \underset{def}{=} f(x_0+h/2) - f(x_0-h/2). \tag{2.5}$$

Lorsque $h > 0$ est donné, l'objet mathématique Δ_h est un opérateur ; à toute fonction continue $f : \mathbb{R} \to \mathbb{R}$ donnée, il fait correspondre une autre fonction continue $\Delta_h f$ définie par $\Delta_h f(x) = f(x+h) - f(x)$. Des considérations semblables sont valables pour les opérateurs ∇_h et δ_h.

Définition 2.1 *Lorsque $h > 0$ est donné, les opérateurs Δ_h, ∇_h et δ_h sont appelés opérateur de différence première respectivement progressive, rétrograde et centrée.*

Nous vérifions maintenant le résultat suivant :

Théorème 2.1 *Les opérateurs de différence première Δ_h, ∇_h et δ_h sont linéaires.*

Démonstration
Montrons le résultat pour l'opérateur Δ_h. Pour ce faire, choisissons deux nombres réels quelconques α et β et deux fonctions continues $f, g : \mathbb{R} \to \mathbb{R}$ quelconques. Nous vérifions sans difficulté que

$$\begin{aligned}\Delta_h(\alpha f + \beta g)(x) &= \alpha f(x+h) + \beta g(x+h) - \alpha f(x) - \beta g(x) \\ &= \alpha \Delta_h f(x) + \beta \Delta_h g(x) \qquad \forall x \in \mathbb{R}.\end{aligned}$$

Le même raisonnement s'applique aux opérateurs ∇_h et δ_h. ■

Si f est une fonction deux fois continûment dérivable, son développement limité au deuxième ordre au voisinage du point x_0 s'écrit :

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(\xi)h^2, \tag{2.6}$$

où ξ est un point de l'intervalle $[x_0, x_0 + h]$. Des relations (2.6) et (2.3) nous obtenons :

$$\left| f'(x_0) - \frac{\Delta_h f(x_0)}{h} \right| = \frac{1}{2}|f''(\xi)|h. \tag{2.7}$$

Nous pouvons donc énoncer le résultat suivant :

Théorème 2.2 *Si $f : \mathbb{R} \to \mathbb{R}$ est deux fois continûment dérivable, si $x_0 \in \mathbb{R}$ est fixé et si $h_0 > 0$ est un nombre positif donné, il existe une constante C telle que*

$$\left| f'(x_0) - \frac{\Delta_h f(x_0)}{h} \right| \leq Ch, \qquad \forall h \leq h_0. \tag{2.8}$$

Démonstration
Posons

$$C = \frac{1}{2} \max_{x \in [x_0, x_0 + h_0]} |f''(x)|.$$

De la relation (2.7) nous obtenons bien l'inégalité (2.8) car si $h \leq h_0$ alors $\xi \in [x_0, x_0 + h_0]$ et donc $\frac{1}{2}|f''(\xi)| \leq C$. ■

Nous obtenons de la même façon un résultat semblable si $\Delta_h f(x_0)$ est remplacé par $\nabla_h f(x_0)$.

Par contre, si nous approchons $f'(x_0)$ par la valeur $\frac{\delta_h f(x_0)}{h}$, nous obtenons une meilleure approximation. En effet, supposons f trois fois continûment dérivable et considérons les développements limités :

$$f\left(x_0 + h/2\right) = f(x_0) + f'(x_0)\frac{h}{2} + \frac{f''(x_0)}{2!}\left(\frac{h}{2}\right)^2 + \frac{f'''(\xi)}{3!}\left(\frac{h}{2}\right)^3, \tag{2.9}$$

$$f\left(x_0 - h/2\right) = f(x_0) - f'(x_0)\frac{h}{2} + \frac{f''(x_0)}{2!}\left(\frac{h}{2}\right)^2 - \frac{f'''(\eta)}{3!}\left(\frac{h}{2}\right)^3, \tag{2.10}$$

où ξ est un point de l'intervalle $[x_0, x_0+h/2]$ et η est un point de $[x_0-h/2, x_0]$. En soustrayant (2.10) à (2.9) et en utilisant la relation (2.5), nous avons :

$$\begin{aligned}\left|f'(x_0) - \frac{\delta_h f(x_0)}{h}\right| &= \left|\frac{f'''(\xi)+f'''(\eta)}{6}\right| \frac{h^2}{8} \\ &\leq \frac{|f'''(\xi)|+|f'''(\eta)|}{2}\frac{h^2}{24}.\end{aligned} \tag{2.11}$$

Si h_0 est un nombre positif fixé et si nous définissons

$$C = \frac{1}{24} \max_{x\in[x_0-h_0/2, x_0+h_0/2]} |f'''(x)|,$$

nous déduisons à partir de (2.11) le résultat suivant :

Théorème 2.3 *Si $f : \mathbb{R} \to \mathbb{R}$ est trois fois continûment dérivable, si $x_0 \in \mathbb{R}$ est fixé et si $h_0 > 0$ est un nombre positif donné, il existe une constante C telle que*

$$\left|f'(x_0) - \frac{\delta_h f(x_0)}{h}\right| \leq Ch^2, \qquad \forall h \leq h_0. \tag{2.12}$$

Les théorèmes 2.2 et 2.3 nous assurent que, si f est assez régulière, les quantités $\frac{\Delta_h f(x_0)}{h}$ et $\frac{\delta_h f(x_0)}{h}$ convergent vers $f'(x_0)$ lorsque h tend vers zéro. Dans le premier cas, la convergence est d'ordre h alors que dans le deuxième cas, la convergence est d'ordre h^2.

Définition 2.2 *On dit que $\Delta_h f(x_0)/h$ et $\nabla_h f(x_0)/h$ sont des formules de différences finies progressives et rétrogrades pour l'approximation de $f'(x_0)$. Les différences*

$$\left|f'(x_0) - \frac{\Delta_h f(x_0)}{h}\right| \qquad \text{et} \qquad \left|f'(x_0) - \frac{\nabla_h f(x_0)}{h}\right|$$

sont appelées ***erreur de troncature****. Elles sont d'ordre h et on dit que les formules de différences finies sont consistantes à l'ordre 1 en h.*

De même la formule de différences finies centrées $\delta_h f(x_0)/h$ pour l'approximation de $f'(x_0)$ est consistante à l'ordre 2 en h car l'erreur de troncature

$$\left|f'(x_0) - \frac{\delta_h f(x_0)}{h}\right|$$

est d'ordre h^2. Elle est ainsi plus précise que les formules de différences finies progressives et rétrogrades.

2.2 Dérivée numérique d'ordre 1 et erreur d'arrondis

Soit c un nombre réel donné ayant une représentation décimale infinie. Un calculateur en virgule flottante ne pourra retenir qu'un nombre fini de chiffres

pour donner une représentation décimale finie de c que nous noterons $\tilde{c}$. Ainsi par exemple, $c = 1/3$ deviendra $\tilde{c} = 0.333333$ qui est bien sûr une approximation numérique de c.

Définition 2.3 *Nous dirons qu'un nombre $\tilde{c}$ est donné avec N chiffres significatifs (en représentation décimale) s'il est donné avec N chiffres comptés à partir du premier chiffre non nul.*

Exemple 2.1 Les nombres :

$$\begin{aligned} 0.333333 &= 0.333333\ 10^{0} \\ 34.2456 &= 0.342456\ 10^{2} \\ 0.000345033 &= 0.345033\ 10^{-3} \\ 3.42550\ 10^{18} &= 0.342550\ 10^{19} \end{aligned}$$

sont donnés avec 6 chiffres significatifs.

Définition 2.4 *Soit c un nombre réel et soit $\tilde{c}$ sa valeur approchée par un calculateur en virgule flottante travaillant avec N chiffres (décimaux) significatifs.*

- *La quantité $|c - \tilde{c}|$ est appelée erreur d'arrondis sur c.*
- *La quantité $\eta = 10^{-N}$ est appelée précision relative du calculateur.*

Un simple calcul nous montre que $|c - \tilde{c}| \leq 5|c|\eta$ si le calculateur arrondit correctement le nombre c. Par exemple, lorsque $c = 1/3$ et que l'on dispose de 6 chiffres significatifs, alors on aura $\tilde{c} = 0.333333$ et l'erreur d'arrondis sur c sera $|c - \tilde{c}| = \frac{1}{3}10^{-6} \leq 5|c|\eta$ puisque $\eta = 10^{-6}$.

Etudions maintenant un exemple illustrant l'importance des erreurs d'arrondis lorsqu'on évalue $\Delta_h f(x_0)/h$ pour approcher numériquement $f'(x_0)$.

Exemple 2.2 Calculons $\Delta_h f(x_0)/h$ lorsque $f(x) = x^2$, $x_0 = 7$, $h = 0.06$ ou $h = 0.01$. Si nous avons un calculateur qui ne dispose que de 3 chiffres significatifs, nous obtenons pour $h = 0.06$ et $h = 0.01$ les valeurs suivantes :

$$\begin{aligned} h = 0.06 : \frac{\Delta_h f(x_0)}{h} &= \frac{(7.06)^2 - (7.00)^2}{0.0600} \simeq \frac{49.8 - 49.0}{0.0600} \simeq 13.3, \\ h = 0.01 : \frac{\Delta_h f(x_0)}{h} &= \frac{(7.01)^2 - (7.00)^2}{0.0100} \simeq \frac{49.1 - 49.0}{0.0100} \simeq 10.0. \end{aligned}$$

Puisque $f'(x_0) = 14$ nous concluons que l'erreur obtenue par la formule aux différences $\Delta_h f(x_0)/h$ est plus grande pour $h = 0.01$ que pour $h = 0.06$. Ce phénomène est lié aux erreurs d'arrondis. Il suffit de prendre une machine disposant de 6 chiffres significatifs pour obtenir la conclusion inverse. En effet, nous obtenons dans ce cas les valeurs suivantes :

$$\begin{aligned} h = 0.06 : \frac{\Delta_h f(x_0)}{h} &= \frac{(7.06000)^2 - (7.00000)^2}{0.0600000} = \frac{49.8436 - 49.0000}{0.0600000} = 14.0600, \\ h = 0.01 : \frac{\Delta_h f(x_0)}{h} &= \frac{(7.01000)^2 - (7.00000)^2}{0.0100000} = \frac{49.1401 - 49.0000}{0.0100000} = 14.0100. \end{aligned}$$

Dans la suite, nous ne ferons pas une théorie sur les erreurs d'arrondis. Disons simplement que lorsque η est la précision relative du calculateur utilisé alors l'erreur absolue $|c - \tilde{c}|$ obtenue sur l'évaluation $\tilde{c}$ d'un nombre c est de l'ordre de grandeur (nous noterons $\sim$) de $|c|\eta$ (en fait on a vu plus haut que $|c - \tilde{c}| \leq 5|c|\eta$). Ainsi, si nous cherchons à calculer

$$\frac{\Delta_h f(x_0)}{h} \underset{def}{=} \frac{f(x_0 + h) - f(x_0)}{h}$$

avec une valeur h supposée donnée sans erreur d'arrondis (exemple : $h = 10^{-7}$, 10^{-8}, ...) nous obtenons :

$$\begin{gathered}
\text{erreur absolue commise sur l'évaluation de } f(x_0 + h) \\
\sim \eta \, |f(x_0 + h)|; \\
\text{erreur absolue commise sur l'évaluation de } f(x_0) \\
\sim \eta \, |f(x_0)|; \\
\text{erreur absolue commise sur l'évaluation de } \Delta_h f(x_0) \\
\sim \eta \, \Big(|f(x_0 + h)| + |f(x_0)| \Big) \simeq 2\eta \, |f(x_0)|; \\
\text{erreur absolue commise sur l'évaluation de } \Delta_h f(x_0)/h \\
\sim 2\eta \, \frac{|f(x_0)|}{h}.
\end{gathered}$$

En conclusion, l'erreur d'arrondis commise sur l'évaluation de $\Delta_h f(x_0)/h$ est de l'ordre de $2\eta \, |f(x_0)|/h$. Remarquons toutefois que les calculs sont très approximatifs et sont donc des indications d'ordre de grandeur probable ! Si nous calculons l'erreur relative e_r due aux erreurs d'arrondis sur l'évaluation de $\Delta_h f(x_0)/h$, nous obtenons :

$$e_r \sim \left| \frac{2\eta f(x_0)/h}{\Delta_h f(x_0)/h} \right| = \frac{2\eta |f(x_0)|}{|f(x_0 + h) - f(x_0)|} \simeq \frac{2\eta |f(x_0)|}{|f'(x_0)| h};$$

cette erreur relative augmente lorsque h diminue.

Exemple 2.3 Considérons à nouveau les données de l'exemple 2.2. Nous avons $f(x_0) = 49$, $f'(x_0) = 14$ et

$$e_r \simeq \frac{7\eta}{h}.$$

Si nous désirons, avec $h = 0.01$, obtenir une erreur relative liée aux erreurs d'arrondis de 10^{-3} au plus ($e_r = 10^{-3}$) sur l'évaluation de $\Delta_h f(x_0)/h$, nous devrons avoir $\eta \simeq 10^{-5}/7 \simeq 10^{-6}$ et donc $N = 6$. Il faut donc choisir une machine qui calcule avec au moins 6 chiffres significatifs.

Les considérations de cette section restent inchangées si on procède à l'évaluation de $\nabla_h f(x_0)/h$ ou $\delta_h f(x_0)/h$.

2.3 Dérivée numérique d'ordre 1 et erreurs

Supposons $f : \mathbb{R} \to \mathbb{R}$ deux fois continûment dérivable et soit $x_0 \in \mathbb{R}$ et h un nombre positif assez petit. Pour calculer une approximation de $f'(x_0)$, nous choisissons la formule aux différences progressives, i.e. $f'(x_0) \simeq \Delta_h f(x_0)/h$. Nous avons vu que l'erreur de troncature commise est approximativement égale à $E_t^h = \frac{1}{2}|f''(x_0)|h$ (formule (2.7)). Par contre, l'erreur d'arrondis est approximativement donnée par $E_a^h = 2\eta|f(x_0)|/h$ où η est la précision relative du calculateur. Ainsi, si au lieu de calculer $f'(x_0)$, nous calculons $\Delta_h f(x_0)/h$ avec un calculateur de précision relative η, nous pouvons nous attendre à une erreur totale $E^h = E_a^h + E_t^h$, somme des erreurs d'arrondis et de troncature. L'erreur totale E^h a donc l'expression suivante :

$$E^h = \frac{1}{2}|f''(x_0)|h + 2\eta\frac{|f(x_0)|}{h}. \tag{2.13}$$

Il est possible de calculer (théoriquement) h de façon à obtenir la plus petite erreur possible E^h. En effet, soit g la fonction définie par

$$g(x) = ax + \frac{b}{x} \qquad \forall x \in \mathbb{R},$$

où $a = \frac{1}{2}|f''(x_0)|$, $b = 2\eta|f(x_0)|$. Clairement la valeur optimale de h sera donnée par le réel $\bar{x}$ qui réalise le minimum de $g(x)$ pour $x > 0$. Pour déterminer $\bar{x}$, il suffit de calculer

$$g'(x) = a - \frac{b}{x^2}$$

et de vérifier que $g'(\bar{x}) = 0$ pour $\bar{x} = \sqrt{b/a}$ et $g''(\bar{x}) > 0$. Nous concluons que $\bar{x}$ est l'unique minimum de $g(x)$ pour $x > 0$ et la valeur de h correspondant à la plus petite valeur de E^h est donc

$$h = 2\sqrt{\frac{\eta|f(x_0)|}{|f''(x_0)|}}. \tag{2.14}$$

Il est clair que les calculs ci-dessus n'ont qu'une valeur théorique et peu de valeur pratique. Ils indiquent que si on peut faire une estimation grossière de la valeur p définie par

$$p = 2\sqrt{\frac{|f(x_0)|}{|f''(x_0)|}}$$

(ce qui n'est pas toujours le cas) et si on veut calculer une approximation numérique de $f'(x_0)$ en utilisant la formule aux différences $\Delta_h f(x_0)/h$ avec un calculateur ne disposant que de N chiffres significatifs, on a intérêt à prendre h de l'ordre de grandeur de $p \cdot 10^{-N/2}$.

Les mêmes considérations restent valables si nous approchons $f'(x_0)$ par la formule aux différences rétrogrades, i.e. $f'(x_0) \simeq \nabla_h f(x_0)/h$. Par contre, si nous approchons $f'(x_0)$ par la formule aux différences centrées, i.e. $f'(x_0) \simeq \delta_h f(x_0)/h$, l'erreur totale aura pour expression (voir (2.11) pour l'erreur de troncature) :

$$E^h \simeq \frac{1}{24}|f'''(x_0)|h^2 + 2\eta\frac{|f(x_0)|}{h}$$

qui prendra son minimum pour

$$h = 2\left(\frac{3\eta|f(x_0)|}{|f'''(x_0)|}\right)^{1/3}.$$

2.4 Dérivées numériques d'ordre supérieur

Les opérateurs aux différences introduits dans la section 2.1 peuvent être généralisés de la façon suivante.

Soit m un entier plus grand que 1, on définit récursivement :

$$\Delta_h^m f = \Delta_h(\Delta_h^{m-1} f), \tag{2.15}$$
$$\nabla_h^m f = \nabla_h(\nabla_h^{m-1} f), \tag{2.16}$$
$$\delta_h^m f = \delta_h(\delta_h^{m-1} f). \tag{2.17}$$

Ainsi, par exemple

$$\begin{aligned}\delta_h^2 f(x) =& \delta_h\Big(\delta_h f(x)\Big) = \delta_h\left(f(x+h/2) - f(x-h/2)\right)\\ =& \delta_h f(x+h/2) - \delta_h f(x-h/2)\\ =& f(x+h/2+h/2) - f(x+h/2-h/2)\\ & - [f(x-h/2+h/2) - f(x-h/2-h/2)]\\ =& f(x+h) - 2f(x) + f(x-h).\end{aligned} \tag{2.18}$$

De façon similaire à ce qui a été fait dans le cas où $m = 1$, nous vérifions que les opérateurs Δ_h^m, ∇_h^m et δ_h^m sont linéaires.

Il est possible de démontrer que, si f est une fonction assez régulière (f de classe C^{m+1} si on prend des différences progressives ou rétrogrades ou f de classe C^{m+2} si on prend des différences centrées) et si $x_0 \in \mathbb{R}$ est donné, alors les quantités

$$\frac{\Delta_h^m f(x_0)}{h^m}, \qquad \frac{\nabla_h^m f(x_0)}{h^m}, \qquad \frac{\delta_h^m f(x_0)}{h^m}$$

sont des approximations de la m-ième dérivée $f^{(m)}(x_0)$ de f au point x_0, d'ordre h, h et h^2 respectivement lorsque h tend vers zéro. Nous pouvons ainsi énoncer les résultats suivants, qui généralisent les théorèmes 2.2 et 2.3 :

Théorème 2.4 *Si m est un entier positif, si $f : \mathbb{R} \to \mathbb{R}$ est $(m+1)$ fois continûment dérivable, si $x_0 \in \mathbb{R}$ et $h_0 > 0$ sont des nombres donnés, alors il existe une constante C telle que*

$$\left|f^{(m)}(x_0) - \frac{\Delta_h^m f(x_0)}{h^m}\right| \leq Ch \qquad \forall h \leq h_0, \tag{2.19}$$
$$\left|f^{(m)}(x_0) - \frac{\nabla_h^m f(x_0)}{h^m}\right| \leq Ch \qquad \forall h \leq h_0. \tag{2.20}$$

Théorème 2.5 *Si m est un entier positif, si $f : \mathbb{R} \to \mathbb{R}$ est $(m+2)$ fois continûment dérivable, si $x_0 \in \mathbb{R}$ et $h_0 > 0$ sont des nombres donnés, alors il existe une constante C telle que*

$$\left| f^{(m)}(x_0) - \frac{\delta_h^m f(x_0)}{h^m} \right| \leq Ch^2 \qquad \forall h \leq h_0. \tag{2.21}$$

Les problèmes de diffusions d'espèces, de déformations élastiques, de propagations d'ondes, d'écoulements de fluides, etc. font intervenir des dérivées deuxième ou quatrième. Ainsi, les formules aux différences finies centrées pour l'approximation de $f''(x_0)$ (i.e. $m = 2$) et $f^{IV}(x_0)$ (i.e. $m = 4$) sont très souvent utilisées par les ingénieurs et s'écrivent :

$$f''(x_0) \simeq \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2}, \tag{2.22}$$

$$f^{IV}(x_0) \simeq \frac{f(x_0+2h) - 4f(x_0+h) + 6f(x_0) - 4f(x_0-h) + f(x_0-2h)}{h^4}. \tag{2.23}$$

Ces formules conduisent à une erreur de troncature d'ordre h^2 si f est assez régulière.

Comme nous l'avons vu dans la section précédente, les erreurs d'arrondis augmentent lorsque h diminue (contrairement aux erreurs de troncature qui diminuent lorsque h diminue). Par contre, il est important de remarquer que les erreurs d'arrondis augmentent lorsque m augmente !

2.5 Dérivées numériques et interpolation

Donnons-nous maintenant une fonction continue $f : \mathbb{R} \to \mathbb{R}$, un point $x_0 \in \mathbb{R}$ et un nombre positif petit h. Considérons les points $x_j = x_0 + jh$ avec $j = 0, 1, 2, 3, \ldots$ Si m est un entier positif, il est possible de construire le polynôme suivant (appelé ***polynôme de Newton***) :

$$\begin{aligned} p_m(x) = f(x_0) &+ \frac{\Delta_h f(x_0)}{h}(x - x_0) + \frac{\Delta_h^2 f(x_0)}{2!h^2}(x - x_0)(x - x_1) \\ &+ \frac{\Delta_h^3 f(x_0)}{3!h^3}(x - x_0)(x - x_1)(x - x_2) + \cdots \\ &+ \frac{\Delta_h^m f(x_0)}{m!h^m}(x - x_0)(x - x_1)\cdots(x - x_{m-1}). \end{aligned} \tag{2.24}$$

Nous vérifions facilement que p_m est un polynôme de degré m et, si nous calculons successivement $p_m(x_0), p_m(x_1), \ldots$, nous obtenons :

$$\begin{aligned} p_m(x_0) &= f(x_0), \\ p_m(x_1) &= f(x_0) + \frac{\Delta_h f(x_0)}{h}(x_1 - x_0) = f(x_0) + \frac{f(x_1) - f(x_0)}{h}h = f(x_1), \\ p_m(x_2) &= f(x_0) + \frac{\Delta_h f(x_0)}{h}(x_2 - x_0) + \frac{\Delta_h^2 f(x_0)}{2h^2}(x_2 - x_0)(x_2 - x_1) \\ &= f(x_0) + \Delta_h f(x_0) \cdot 2 + \Delta_h^2 f(x_0) \\ &= f(x_0) + 2(f(x_1) - f(x_0)) + (f(x_2) - 2f(x_1) + f(x_0)) = f(x_2). \end{aligned}$$

En fait, nous pouvons montrer que $p_m(x_j) = f(x_j)$, $j = 0, 1, 2, \ldots, m$ et puisque p_m est un polynôme de degré m, alors p_m est l'unique polynôme de degré m qui interpole f dans les $(m+1)$ points $x_0, x_1, x_2, \ldots, x_m$ (chap. 1). Il est facile de voir que si nous dérivons m fois la relation (2.24), nous obtenons

$$\frac{d^m}{dx^m} p_m(x) = \frac{\Delta_h^m f(x_0)}{h^m}.$$

Nous avons ainsi partiellement montré le résultat suivant :

Théorème 2.6 *Si p_m est le polynôme de degré m qui interpole f dans les points $x_j = x_0 + jh$ avec $j = 0, 1, 2, \cdots, m$, alors on a :*

$$(i) \quad p_m(x) = f(x_0) + \frac{\Delta_h f(x_0)}{h}(x - x_0) + \frac{\Delta_h^2 f(x_0)}{2! h^m}(x - x_0)(x - x_1) + \cdots + \frac{\Delta_h^m f(x_0)}{m! h^m}(x - x_0)(x - x_1)\cdots(x - x_{m-1}); \tag{2.25}$$

$$(ii) \quad \frac{d^m}{dx^m} p_m(x_0) = \frac{\Delta_h^m f(x_0)}{h^m}. \tag{2.26}$$

Nous terminons cette section par deux remarques.

Remarque 2.1 Si la fonction f est $(m+1)$ fois continûment dérivable, nous pouvons utiliser le théorème 1.1 pour établir l'estimation d'erreur suivante entre f et p_m :

$$\max_{x \in [x_0, x_0 + mh]} |f(x) - p_m(x)| \leq \frac{1}{2(m+1)} h^{m+1} \max_{x \in [x_0, x_0 + mh]} |f^{(m+1)}(x)|.$$

Remarquer l'analogie entre le polynôme de Newton p_m (2.24) et le polynôme obtenu par développement de Taylor de f autour de $x = x_0$.

Remarque 2.2 Il est possible d'établir des résultats semblables à ceux énoncés dans le théorème 2.6, mais avec les opérateurs ∇_h^m et δ_h^m. Par exemple, il est facile de montrer que si q_2 est le polynôme de degré 2 qui interpole la fonction f en les points $x_0 - h$, x_0 et $x_0 + h$, alors

$$\frac{d^2}{dx^2} q_2(x_0) = \frac{\delta_h^2 f(x_0)}{h^2} = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2}.$$

Dans la figure 2.1 nous avons représenté, pour une fonction f donnée, son graphe et celui de q_2. Nous observons que les courbures des graphes de f et de q_2 au point $(x_0, f(x_0))$ sont proches lorsque h est petit.

2.6 Extrapolation de Richardson

Il est possible de trouver des formules de dérivation numérique plus précises que celles que nous avons considérées jusqu'à présent. Supposons par exemple

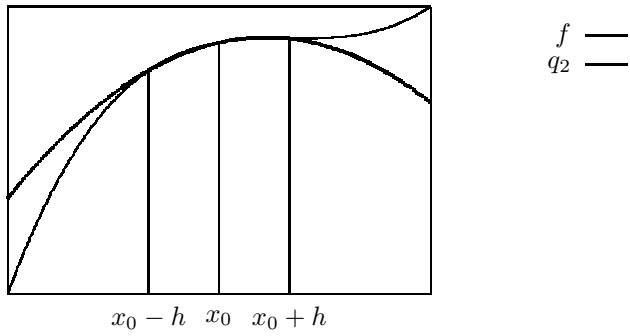

Fig. 2.1 Interpolation de f par un polynôme q_2 de degré 2 aux points $x_0 - h$, x_0, $x_0 + h$.

la fonction f cinq fois continûment dérivable et considérons le développement limité à l'ordre 5. Nous obtenons :

$$f(x_0 + \frac{h}{2}) = f(x_0) + f'(x_0)\frac{h}{2} + \frac{f''(x_0)}{2!}\left(\frac{h}{2}\right)^2 + \frac{f'''(x_0)}{3!}\left(\frac{h}{2}\right)^3 + \frac{f^{IV}(x_0)}{4!}\left(\frac{h}{2}\right)^4 + \frac{f^V(\xi)}{5!}\left(\frac{h}{2}\right)^5,$$

$$f(x_0 - \frac{h}{2}) = f(x_0) - f'(x_0)\frac{h}{2} + \frac{f''(x_0)}{2!}\left(\frac{h}{2}\right)^2 - \frac{f'''(x_0)}{3!}\left(\frac{h}{2}\right)^3 + \frac{f^{IV}(x_0)}{4!}\left(\frac{h}{2}\right)^4 - \frac{f^V(\eta)}{5!}\left(\frac{h}{2}\right)^5,$$

où ξ est un point de l'intervalle $[x_0, x_0 + \frac{h}{2}]$ et η est un point de $[x_0 - \frac{h}{2}, x_0]$. Par soustraction nous avons donc :

$$\frac{f(x_0 + h/2) - f(x_0 - h/2)}{h} = f'(x_0) + \frac{f'''(x_0)}{24}h^2 + \frac{f^V(\xi) + f^V(\eta)}{5!2^5}h^4$$

soit, par définition de $\delta_h f(x_0)$:

$$\frac{\delta_h f(x_0)}{h} = f'(x_0) + \frac{f'''(x_0)}{24}h^2 + O(h^4), \tag{2.27}$$

où ici $O(h^4)$ désigne un reste d'ordre h^4 lorsque h tend vers zéro. Si nous substituons h par $h/2$ dans (2.27), nous obtenons aussi :

$$\frac{\delta_{h/2} f(x_0)}{h/2} = f'(x_0) + \frac{f'''(x_0)}{24}\frac{h^2}{4} + O(h^4). \tag{2.28}$$

En soustrayant quatre fois (2.28) à (2.27) nous avons donc :

$$\frac{\delta_h f(x_0)}{h} - \frac{8\delta_{h/2} f(x_0)}{h} = -3f'(x_0) + O(h^4)$$

et finalement

$$f'(x_0) = \frac{8\delta_{h/2} f(x_0) - \delta_h f(x_0)}{3h} + O(h^4). \tag{2.29}$$

Par définition de l'opérateur δ_h, formule (2.5), nous obtenons :

$$\begin{aligned} &8\delta_{h/2} f(x_0) - \delta_h f(x_0) \\ &\quad = 8f(x_0 + h/4) - 8f(x_0 - h/4) - f(x_0 + h/2) + f(x_0 - h/2), \end{aligned} \tag{2.30}$$

et donc (2.30) dans (2.29) nous assure que

$$\frac{8f(x_0 + h/4) - 8f(x_0 - h/4) + f(x_0 - h/2) - f(x_0 + h/2)}{3h}$$

est une approximation de $f'(x_0)$; l'erreur de troncature est d'ordre h^4. Le procédé pour obtenir cette formule est appelé ***méthode d'extrapolation de Richardson***. Il est possible de généraliser cette méthode afin d'obtenir des formules d'ordre 6, 8, ..., en h pour approcher $f'(x_0)$. Il suffira de tenir compte de $\delta_{h/4} f(x_0)$, $\delta_{h/8} f(x_0)$, ...

2.7 Exercices

Exercice 2.1 Démontrer l'estimation (2.21) pour $m = 4$, i.e. si f est 6 fois continûment dérivable, alors il existe une constante C telle que, $\forall h \leq h_0$,

$$\left| f^{(4)}(x_0) - \frac{\delta_h^4 f(x_0)}{h^4} \right| \leq Ch^2.$$

Solution
Par l'égalité (2.17), nous avons

$$\delta_h^4 f(x_0) = \delta_h\big(\delta_h^3 f(x_0)\big) = \delta_h\Big(\delta_h\big(\delta_h^2 f(x_0)\big)\Big) = \delta_h^2\big(\delta_h^2 f(x_0)\big).$$

Puisque l'opérateur δ_h^2 est linéaire et en utilisant la formule (2.18), nous obtenons

$$\begin{aligned} \delta_h^4 f(x_0) &= \delta_h^2 f(x_0 + h) - 2\delta_h^2 f(x_0) + \delta_h^2 f(x_0 - h) \\ &= f(x_0 + 2h) - 4f(x_0 + h) + 6f(x_0) \\ &\qquad - 4f(x_0 - h) + f(x_0 - 2h). \end{aligned} \tag{2.31}$$

D'autre part, le développement limité à l'ordre 6 de la fonction f autour du point x_0 nous assure que

$$\begin{aligned}
f(x_0+2h) &= f(x_0) + f'(x_0)\frac{(2h)^1}{1!} + f''(x_0)\frac{(2h)^2}{2!} + f'''(x_0)\frac{(2h)^3}{3!} \\
&\quad + f^{(4)}(x_0)\frac{(2h)^4}{4!} + f^{(5)}(x_0)\frac{(2h)^5}{5!} + f^{(6)}(\eta_1)\frac{(2h)^6}{6!}, \\
f(x_0-2h) &= f(x_0) - f'(x_0)\frac{(2h)^1}{1!} + f''(x_0)\frac{(2h)^2}{2!} - f'''(x_0)\frac{(2h)^3}{3!} \\
&\quad + f^{(4)}(x_0)\frac{(2h)^4}{4!} - f^{(5)}(x_0)\frac{(2h)^5}{5!} + f^{(6)}(\eta_2)\frac{(2h)^6}{6!}, \\
f(x_0+h) &= f(x_0) + f'(x_0)\frac{h^1}{1!} + f''(x_0)\frac{h^2}{2!} + f'''(x_0)\frac{h^3}{3!} \\
&\quad + f^{(4)}(x_0)\frac{h^4}{4!} + f^{(5)}(x_0)\frac{h^5}{5!} + f^{(6)}(\eta_3)\frac{h^6}{6!}, \\
f(x_0-h) &= f(x_0) - f'(x_0)\frac{h^1}{1!} + f''(x_0)\frac{h^2}{2!} - f'''(x_0)\frac{h^3}{3!} \\
&\quad + f^{(4)}(x_0)\frac{h^4}{4!} - f^{(5)}(x_0)\frac{h^5}{5!} + f^{(6)}(\eta_4)\frac{h^6}{6!},
\end{aligned}$$

où $\eta_1 \in]x_0, x_0+2h[$, $\eta_2 \in]x_0-2h, x_0[$, $\eta_3 \in]x_0, x_0+h[$ et $\eta_4 \in]x_0-h, x_0[$. Après substitution dans (2.31), nous obtenons

$$\begin{aligned}
\delta_h^4 f(x_0) &= f^{(4)}(x_0)h^4 \\
&\quad + \left(\frac{64}{6!}\left(f^{(6)}(\eta_1) + f^{(6)}(\eta_2)\right) - \frac{4}{6!}\left(f^{(6)}(\eta_3) + f^{(6)}(\eta_4)\right)\right)h^6.
\end{aligned}$$

Soit $h_0 > 0$ un nombre arbitraire et soit

$$C = \frac{17}{90}\max_{x\in[x_0-2h_0, x_0+2h_0]}\left|f^{(6)}(x)\right|.$$

Pour tout $h \leq h_0$, nous avons donc

$$\left|\frac{\delta_h^4 f(x_0)}{h^4} - f^{(4)}(x_0)\right| \leq Ch^2.$$

Exercice 2.2 Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction trois fois continûment dérivable donnée, soit $x_0 \in \mathbb{R}$ et $h > 0$ donnés. Soit $x_1 = x_0 + h$, $x_2 = x_0 + 2h$ et soit g la fonction définie par :

$$g(x) = f(x_0) + \frac{\Delta_h f(x_0)}{h}(x - x_0) + \frac{\Delta_h^2 f(x_0)}{2h^2}(x - x_0)(x - x_1).$$

1. Vérifier que $g(x_j) = f(x_j)$ pour $j = 0, 1, 2$ et en déduire qu'il existe $\xi_0 \in [x_0, x_1]$ et $\xi_1 \in [x_1, x_2]$ tels que

$$f'(\xi_0) = g'(\xi_0) \quad , \quad f'(\xi_1) = g'(\xi_1).$$

2. Soit r la fonction définie par $r(x) = f(x) - g(x)$. Déduire du point 1 qu'il existe $\eta \in [\xi_0, \xi_1]$ tel que $r''(\eta) = 0$ et donc

$$r''(x) = \int_\eta^x r'''(t)dt = \int_\eta^x f'''(t)dt.$$

3. Déduire du point 2 que

$$|f(x) - g(x)| \leq 2h^3 \max_{t\in[x_0,x_2]} |f'''(t)| \qquad \text{si } x \in [x_0, x_2].$$

Comparer avec le développement de Taylor.

Solution

1. En vertu de la définition du polynôme g, il est clair que $g(x_0) = f(x_0)$. D'autre part, par définition de l'opérateur Δ_h , nous avons

$$g(x_1) = f(x_0) + \Delta_h f(x_0) = f(x_0 + h) = f(x_1),$$

et

$$\begin{aligned} g(x_2) &= f(x_0) + 2\Delta_h f(x_0) + \Delta_h^2 f(x_0) \\ &= 2f(x_0 + h) - f(x_0) + \Delta_h^2 f(x_0). \end{aligned}$$

Par définition de l'opérateur Δ_h^2, nous avons

$$\begin{aligned} \Delta_h^2 f(x_0) &= \Delta_h \Big(f(x_0 + h) - f(x_0) \Big) \\ &= \Delta_h \Big(f(x_0 + h) \Big) - \Delta_h \Big(f(x_0) \Big) \\ &= f(x_0 + 2h) - 2f(x_0 + h) + f(x_0), \end{aligned}$$

et nous obtenons bien $g(x_2) = f(x_0 + 2h) = f(x_2)$. Soit r la fonction définie par $r(x) = f(x) - g(x)$. Puisque $r(x_0) = r(x_1) = r(x_2) = 0$ et puisque r est continûment dérivable, nous pouvons utiliser le théorème de Rolle pour obtenir

$$\begin{aligned} &\exists \xi_0 \in [x_0, x_1] \text{ tel que } r'(\xi_0) = 0 \text{ i.e. } f'(\xi_0) = g'(\xi_0), \\ &\exists \xi_1 \in [x_1, x_2] \text{ tel que } r'(\xi_1) = 0 \text{ i.e. } f'(\xi_1) = g'(\xi_1). \end{aligned}$$

2. Puisque $r'(\xi_0) = r'(\xi_1) = 0$ et puisque r' est continûment dérivable, nous pouvons à nouveau utiliser le théorème de Rolle pour obtenir

$$\exists \eta \in [\xi_0, \xi_1] \text{ tel que } r''(\eta) = 0.$$

Par conséquent, puisque r''' est continue, nous avons

$$r''(x) = r''(x) - r''(\eta) = \int_\eta^x r'''(t)dt.$$

Puisque g est un polynôme de degré deux il est clair que $r'''(t) = f'''(t) - g'''(t) = f'''(t)$ et donc

$$r''(x) = \int_\eta^x f'''(t)dt.$$

3. Considérons par exemple le cas où $x \in [x_0, x_1]$. Le cas où $x \in [x_1, x_2]$ se traite de manière analogue. D'après le point 1, $r(x_0) = 0$ et donc

$$f(x) - g(x) = r(x) = r(x) - r(x_0) = \int_{x_0}^{x} r'(s)ds.$$

Par conséquent

$$|f(x) - g(x)| \leq \int_{x_0}^{x} |r'(s)|ds \leq \int_{x_0}^{x_1} |r'(s)|ds \leq h \max_{x_0 \leq s \leq x_1} |r'(s)|. \tag{2.32}$$

Soit $s \in [x_0, x_1]$. D'après le point 1, $r'(\xi_0) = 0$ et donc

$$r'(s) = r'(s) - r'(\xi_0) = \int_{\xi_0}^{s} r''(t)dt.$$

Par conséquent

$$|r'(s)| \leq |\int_{\xi_0}^{s} |r''(t)|dt| \leq \int_{x_0}^{x_1} |r''(t)|dt \leq h \max_{x_0 \leq t \leq x_1} |r''(t)|. \tag{2.33}$$

Soit $t \in [x_0, x_1]$. D'après le point 2

$$r''(t) = \int_{\eta}^{t} f'''(u)du,$$

et par conséquent

$$|r''(t)| \leq |\int_{\eta}^{t} |f'''(u)|du| \leq \int_{x_0}^{x_2} |f'''(u)|du \leq 2h \max_{x_0 \leq u \leq x_2} |f'''(u)|. \tag{2.34}$$

Les inégalités (2.32), (2.33) et (2.34) impliquent donc

$$|f(x) - g(x)| \leq 2h^3 \max_{x_0 \leq s \leq x_2} |f'''(s)|.$$

Comparons ce résultat avec la formule de Taylor. Soit $x \in [x_0, x_2]$ donné, puisque f est trois fois continûment dérivable, il existe $\xi \in [x_0, x]$ tel que

$$f(x) = G(x) + \frac{f'''(\xi)}{6}(x - x_0)^3, \tag{2.35}$$

où G est le polynôme de degré deux défini par

$$G(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2.$$

Par conséquent, lorsque $x \in [x_0, x_2]$, nous avons

$$\begin{aligned} |f(x) - G(x)| &= |\frac{f'''(\xi)}{6}(x - x_0)^3| \\ &\leq \frac{(x_2 - x_0)^3}{6} \max_{x_0 \leq t \leq x_2} |f'''(t)| \\ &= \frac{4}{3}h^3 \max_{x_0 \leq t \leq x_2} |f'''(t)|. \end{aligned}$$

Pour résumer, nous avons donc à disposition deux polynômes de degré deux, g et G, permettant d'approcher une fonction f trois fois continûment dérivable, au voisinage d'un point x_0 (plus précisément entre x_0 et x_0+2h). L'erreur maximale entre la fonction f et les polynômes g et G est d'ordre trois en h sur l'intervalle $[x_0, x_0+2h]$. Le polynôme G fait appel aux derivées première et seconde de f au point x_0. Par contre, le polynôme g fait appel à des approximations numériques de ces dérivées.

2.8 Notes bibliographiques et remarques

Pour une introduction aux erreurs d'arrondis, voir par exemple [6] ou [25] .

L'exercice 2.2 montre un résultat proche de celui énoncé dans le théorème 1.1 avec $n = 2$. Le même raisonnement s'applique pour n quelconque, voir par exemple [6] ou [28].

Nous utiliserons les formules de dérivation numérique dans les chapitres 10 à 14, lors de la mise en œuvre des méthodes de différences finies.

Chapitre 3

Intégration numérique. Formules de quadrature

3.1 Généralités

Nous avons maintenant pour but de calculer numériquement des intégrales définies. Soit $f : x \in [a, b] \to f(x) \in \mathbb{R}$ une fonction continue donnée sur un intervalle $[a, b]$. Nous désirons approcher numériquement la quantité

$$\int_a^b f(x)dx. \tag{3.1}$$

Pour ce faire, nous commençons par partitionner l'intervalle $[a, b]$ en petits intervalles $[x_i, x_{i+1}]$, $i = 0, 1, 2, \ldots, N-1$, c'est-à-dire nous choisissons des points x_i, $i = 0, 1, 2, \ldots, N$ tels que

$$a = x_0 < x_1 < x_2 < x_3 < \cdots\cdots < x_{N-1} < x_N = b. \tag{3.2}$$

Soit

$$h = \max_{0 \le i \le N-1} |x_{i+1} - x_i| \tag{3.3}$$

le réel positif caractérisant la finesse de la partition. Il est clair que, lorsque N augmente, nous pouvons placer les points x_i de sorte à ce que h soit petit. Lorsqu'aucune raison nous incite à choisir des intervalles de longueurs différentes, nous posons

$$h = \frac{b-a}{N} \qquad \text{et} \qquad x_i = a + ih, \quad i = 0, 1, \ldots, N.$$

Etant donné la partition (3.2), il est naturel d'écrire :

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx. \tag{3.4}$$

Ce sont ainsi les intégrales

$$\int_{x_i}^{x_{i+1}} f(x)dx$$

que nous allons approcher dans la suite par des formules appelées ***formules de quadrature***. Mentionnons encore que souvent, pour donner des formules de quadrature sur un intervalle standard (par exemple l'intervalle $[-1,+1]$), on exécute un changement de variable de la forme

$$t = 2\frac{x - x_i}{x_{i+1} - x_i} - 1 \tag{3.5}$$

qui, à $x \in [x_i, x_{i+1}]$, fait correspondre $t \in [-1,+1]$. Avec ce changement de variables, nous obtenons :

$$x = x_i + (x_{i+1} - x_i)\frac{t+1}{2}, \tag{3.6}$$

et par suite

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{x_{i+1} - x_i}{2} \int_{-1}^{+1} g_i(t)dt, \tag{3.7}$$

où la fonction g_i est définie par

$$g_i(t) = f\left(x_i + (x_{i+1} - x_i)\frac{t+1}{2}\right), \qquad t \in [-1,+1]. \tag{3.8}$$

Nous sommes maintenant en mesure de définir la notion de formule de quadrature pour approcher numériquement $\int_{-1}^{+1} g(t)dt$, g étant une fonction continue donnée sur $[-1,+1]$.

Définition 3.1 *Si g est une fonction continue sur $[-1,+1]$, la formule de quadrature*

$$J(g) \underset{def}{=} \sum_{j=1}^{M} \omega_j g(t_j) \tag{3.9}$$

est définie par la donnée de M points $-1 \leq t_1 < t_2 < \cdots < t_M \leq 1$ appelés points d'intégration et de M nombres réels ω_1, ω_2, ..., ω_M appelés poids de la formule de quadrature. Ces M points et ces M poids devront être cherchés de façon à ce que $J(g)$ soit une approximation numérique de $\int_{-1}^{+1} g(t)dt$.

Nous remarquons que la formule de quadrature (3.9) est linéaire. En effet, si g et ℓ sont deux fonctions continues données sur l'intervalle $[-1,+1]$ et si α et β sont deux nombres réels, nous vérifions facilement que

$$J(\alpha g + \beta \ell) = \alpha J(g) + \beta J(\ell).$$

Nous présentons maintenant un exemple de formule de quadrature.

Exemple 3.1 Un exemple classique est la formule à 2 points ($M = 2$) suivante :

$$t_1 = -1, t_2 = +1, \omega_1 = 1, \omega_2 = 1$$

et donc

$$J(g) = g(-1) + g(1). \tag{3.10}$$

Nous remarquons que $J(g)$ correspond à l'aire du trapèze hachuré de la figure 3.1. Par conséquent, approcher $\int_{-1}^{+1} g(t)dt$ par $J(g)$ correspond à approcher l'aire sous le graphe de g par l'aire du trapèze hachuré. Pour cette raison, la formule de quadrature (3.10) est appelée ***formule du trapèze***.

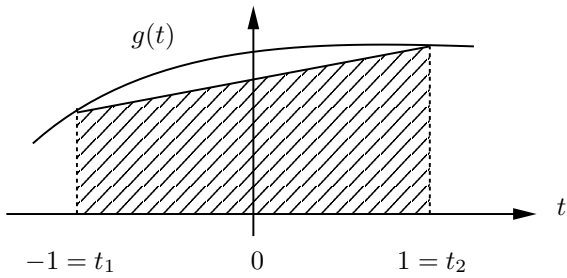

Fig. 3.1 Formule du trapèze sur $[-1, +1]$.

Dans les sections suivantes, nous construirons d'autres formules de quadrature que la formule du trapèze. Auparavant nous supposons ces formules de quadratures données et nous décrivons leur utilisation pour approcher (3.1).

Dans l'égalité (3.7) nous approchons $\int_{-1}^{+1} g_i(t)dt$ par $J(g_i)$. Ainsi la quantité $\int_{x_i}^{x_{i+1}} f(x)dx$ est approchée par la valeur suivante :

$$\frac{x_{i+1} - x_i}{2} \sum_{j=1}^{M} \omega_j f\left(x_i + (x_{i+1} - x_i)\frac{t_j + 1}{2}\right). \tag{3.11}$$

De retour à (3.4), nous allons donc approcher $\int_a^b f(x)dx$ par la formule dite ***formule composite*** :

$$L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2} \sum_{j=1}^{M} \omega_j f\left(x_i + (x_{i+1} - x_i)\frac{t_j + 1}{2}\right). \tag{3.12}$$

Exemple 3.2 Considérons à nouveau la formule du trapèze (3.10) de l'exemple 3.1, c'est-à-dire $t_1 = -1$, $t_2 = 1$, $\omega_1 = \omega_2 = 1$. La formule composite (3.12) s'écrit :

$$L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2}\Big(f(x_i) + f(x_{i+1})\Big). \tag{3.13}$$

La formule (3.13) est facile à interpréter graphiquement : la quantité $L_h(f)$ correspond à l'aire hachurée de la figure 3.2.

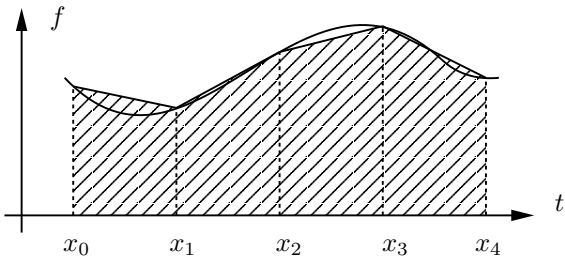

Fig. 3.2 Formule du trapèze pour approcher $\int_a^b f(x)dx$ dans le cas où $N = 4$.

En règle générale nous pouvons procéder de la manière suivante pour approcher la quantité $\int_a^b f(x)dx$ par la quantité $L_h(f)$: on définit une formule de quadrature par la donnée de M points $t_1, t_2, \ldots, t_M$ et de M poids $\omega_1, \omega_2, \ldots, \omega_M$ (ces points et ces poids sont généralement répertoriés dans des tables numériques ou des logiciels de calculs) ; on partitionne l'intervalle $[a, b]$ en intervalles $[x_i, x_{i+1}]$ (les points x_i satisfaisant (3.2)) et on calcule $L_h(f)$ par la formule composite (3.12).

Avant de montrer comment construire des formules de quadrature, définissons une propriété désirable de $J(g)$.

Définition 3.2 *On dira que la formule de quadrature*

$$J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$$

pour calculer numériquement $\int_{-1}^{+1} g(t)dt$ *est exacte pour les polynômes de degré* $r \geq 0$ *si*

$$J(p) = \int_{-1}^{+1} p(t)dt$$

pour tout polynôme p *de degré* $\leq r$.

Lorsque la formule de quadrature $J(\cdot)$ satisfait la propriété de la définition 3.2, il est possible d'estimer l'erreur entre la valeur exacte $\int_a^b f(x)dx$ et la valeur approchée $L_h(f)$, pour autant que f soit assez régulière. On peut montrer le résultat suivant :

Théorème 3.1 *Supposons que la formule de quadrature*

$$J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$$

pour calculer numériquement $\int_{-1}^{+1} g(t)dt$ *soit exacte pour des polynômes de degré* r. *Soit* f *une fonction donnée sur l'intervalle* $[a, b]$, *soit* $L_h(f)$ *la formule*

composite définie par (3.12) et soit h la quantité définie par (3.3). Alors, si la fonction f est assez régulière (i.e. $(r+1)$ fois continûment dérivable sur l'intervalle $[a,b]$), il existe une constante C indépendante du choix des points x_i telle que

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^{r+1}. \tag{3.14}$$

Exemple 3.3 Comme exemple d'application du théorème 3.1, considérons à nouveau la formule du trapèze (3.10) ainsi que la formule composite $L_h(f)$ (3.13) qui en découle (voir exemples 3.1 et 3.2).

Clairement si p est un polynôme de degré 1, c'est-à-dire si p s'écrit sous la forme

$$p(t) = \alpha t + \beta$$

où $\alpha, \beta \in \mathbb{R}$, il est facile de vérifier que lorsque la formule de quadrature est définie par (3.10), alors

$$\int_{-1}^{+1} p(t)dt = J(p).$$

Ainsi la formule du trapèze (3.10) pour calculer numériquement $\int_{-1}^{+1} g(t)dt$ est exacte pour des polynômes de degré 1 ($r = 1$ dans les hypothèses du théorème 3.1).

Si l'intervalle $[a,b]$ est divisé en N parties égales, i.e. $h = (b-a)/N$, $x_i = a+ih$ avec $i = 0, 1, 2, \ldots, N$ et si f est une fonction deux fois continûment dérivable sur l'intervalle $[a,b]$, alors le théorème 3.1 fournit l'estimation d'erreur suivante :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^2, \tag{3.15}$$

où C est une constante qui ne dépend pas de N et donc pas de h. L'estimation (3.15) indique qu'en principe, lorsqu'on utilise la formule (3.13) pour approcher numériquement $\int_a^b f(x)dx$, l'erreur est divisée par quatre chaque fois que N est multiplié par deux !

En fait, l'inégalité (3.14) montre que, lorsque la partition est fine (h petit), l'erreur obtenue en approchant $\int_a^b f(x)dx$ par $L_h(f)$ est petite. Cette erreur devient d'autant plus petite avec h que r est grand. Il est donc légitime de chercher des points d'intégration t_j et des poids ω_j, $1 \leq j \leq M$, de sorte que la formule de quadrature $J(\cdot)$ soit exacte pour des polynômes de degré r aussi élevé que possible.

3.2 Poids d'une formule de quadrature

Dans cette section, nous supposons donnés M points d'intégration distincts dans l'intervalle $[-1, +1]$

$$-1 \leq t_1 < t_2 < t_3 < \cdots < t_M \leq 1$$

et nous cherchons à déterminer les poids $\omega_1, \omega_2, \ldots, \omega_M$ de sorte que la formule de quadrature $J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$ soit exacte pour des polynômes de degré r aussi élevé que possible.

Pour réaliser cet objectif, considérons la base de Lagrange $\varphi_1, \varphi_2, \ldots, \varphi_M$ de $\mathbb{P}_{M-1}$ associée aux points t_1, t_2, ..., t_M (définition 1.2). Par définition, φ_j est le polynôme de degré $M-1$ défini par :

$$\varphi_j(t) = \frac{(t-t_1)(t-t_2)\cdots(t-t_{j-1})(t-t_{j+1})\cdots(t-t_M)}{(t_j-t_1)(t_j-t_2)\cdots(t_j-t_{j-1})(t_j-t_{j+1})\cdots(t_j-t_M)}, \quad (3.16)$$

pour $j = 1, 2, \ldots, M$. Soit $g : t \in [-1, +1] \to g(t) \in \mathbb{R}$ une fonction continue donnée. Son interpolant $\tilde{g}$ de degré $M-1$ aux points $t_1, \ldots, t_M$ (définition 1.3) est défini par :

$$\tilde{g}(t) = \sum_{j=1}^{M} g(t_j)\varphi_j(t).$$

Il semble naturel de remplacer $\int_{-1}^{+1} g(t)dt$ par $\int_{-1}^{+1} \tilde{g}(t)dt$. Puisque

$$\int_{-1}^{+1} \tilde{g}(t)dt = \sum_{j=1}^{M} g(t_j) \int_{-1}^{+1} \varphi_j(t)dt,$$

nous constatons immédiatement qu'il suffit de poser

$$\omega_j = \int_{-1}^{+1} \varphi_j(t)dt$$

pour que $J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$ soit une approximation de $\int_{-1}^{+1} g(t)dt$.

Nous obtenons ainsi le théorème suivant.

Théorème 3.2 *Soit $t_1 < t_2 < \cdots < t_M$, M points distincts de l'intervalle $[-1, +1]$ et soit $\varphi_1, \varphi_2, \cdots, \varphi_M$ la base de Lagrange de $\mathbb{P}_{M-1}$ associée à ces M points. Alors la formule de quadrature*

$$J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$$

est exacte pour les polynômes de degré $M-1$ si et seulement si

$$\omega_j = \int_{-1}^{+1} \varphi_j(t)dt, \qquad j = 1, 2, \ldots, M. \quad (3.17)$$

Démonstration
i) Montrons que si la formule de quadrature $J(\cdot)$ est exacte pour les polynômes de degré $M-1$, alors on a les relations (3.17). Puisque

$$J(p) = \sum_{j=1}^{M} \omega_j p(t_j) = \int_{-1}^{+1} p(t)dt,$$

pour tout polynôme $p \in \mathbb{P}_{M-1}$, nous pouvons choisir $p = \varphi_k$, $k = 1, 2, \ldots, M$ et nous obtenons :

$$J(\varphi_k) = \sum_{j=1}^{M} \omega_j \varphi_k(t_j) = \int_{-1}^{+1} \varphi_k(t) dt.$$

Puisque $\varphi_k(t_j) = 0$ si $j \neq k$ et $\varphi_k(t_k) = 1$, nous avons bien :

$$\omega_k = \int_{-1}^{+1} \varphi_k(t) dt.$$

ii) Montrons maintenant que si les relations (3.17) sont vraies, alors la formule de quadrature est exacte pour des polynômes de degré $M - 1$.

Soit p un polynôme quelconque de degré $M - 1$ que nous développons dans la base de Lagrange de $\mathbb{P}_{M-1}$ associée aux points $t_1, t_2, \ldots, t_M$, i.e. :

$$p(t) = \sum_{j=1}^{M} p(t_j) \varphi_j(t).$$

Ainsi donc

$$\begin{aligned} \int_{-1}^{+1} p(t) dt &= \sum_{j=1}^{M} p(t_j) \int_{-1}^{+1} \varphi_j(t) dt \\ &= \sum_{j=1}^{M} p(t_j) \omega_j = J(p). \end{aligned}$$ ■

Remarque 3.1 Les relations (3.17) nous permettent donc de calculer les poids $\omega_1, \omega_2, \ldots, \omega_M$, d'une formule de quadrature, étant donné les points d'intégration $t_1, t_2, \ldots, t_M$. De plus, $\sum_{j=1}^{M} \varphi_j(t)$ est le polynôme de degré $M-1$ qui vaut 1 aux M points t_1, t_2, ..., t_M, et est donc la fonction identique à 1. Par conséquent, nous obtenons, en utilisant (3.17)

$$\sum_{j=1}^{M} \omega_j = \int_{-1}^{+1} \left(\sum_{j=1}^{M} \varphi_j(t) \right) dt = \int_{-1}^{+1} dt = 2,$$

ce qui prouve que la somme des poids calculés par (3.17) est toujours égale à 2.

Exemple 3.4 Considérons à nouveau l'exemple 3.1, c'est-à-dire $M = 2$, $t_1 = -1$ et $t_2 = +1$ (formule du trapèze) et explicitons la base de Lagrange φ_1, φ_2 associée aux points t_1, t_2 :

$$\varphi_1(t) = \frac{t - t_2}{t_1 - t_2} = \frac{(1 - t)}{2} \qquad \text{et} \qquad \varphi_2(t) = \frac{t - t_1}{t_2 - t_1} = \frac{(t + 1)}{2}.$$

Les relations (3.17) s'écrivent :

$$\omega_1 = \int_{-1}^{+1} \varphi_1(t) dt = 1 \qquad \text{et} \qquad \omega_2 = \int_{-1}^{+1} \varphi_2(t) dt = 1,$$

qui sont bien les valeurs définies dans l'exemple 3.1.

Le théorème 3.2 nous assure que les formules de quadrature construites grâce à (3.17) sont exactes pour les polynômes de degré $M-1$. Dans la suite, nous verrons qu'il se peut que ces formules de quadrature soient exactes pour des polynômes de degré r, avec r plus grand que $M-1$.

3.3 Formule du rectangle

La formule du rectangle est une formule à un seul point ($M=1$) :

$$t_1 = 0.$$

La base de Lagrange de $\mathbb{P}_0$ associée à $t_1 = 0$ est donnée par :

$$\varphi_1(t) = 1, \qquad \forall t \in [-1, +1].$$

Ainsi donc la relation (3.17) nous donne

$$\omega_1 = \int_{-1}^{+1} \varphi_1(t)dt = 2$$

et la formule du rectangle devient :

$$J(g) = 2g(0). \tag{3.18}$$

On interprète la formule du rectangle (3.18) de la façon suivante : elle consiste à remplacer $\int_{-1}^{+1} g(t)dt$ par l'aire du rectangle de base $[-1, +1]$ et de hauteur $g(0)$ (fig. 3.3), d'où son nom. Selon le théorème 3.2, cette formule de quadrature est exacte pour des polynômes de degré zéro, mais en fait elle est meilleure ; elle est exacte pour des polynômes $p \in \mathbb{P}_1$. En effet soit $p \in \mathbb{P}_1$ défini par

$$p(t) = \alpha t + \beta$$

où $\alpha, \beta \in \mathbb{R}$. Il est alors facile de vérifier que $\int_{-1}^{+1} p(t)dt = 2\beta = 2p(0)$.

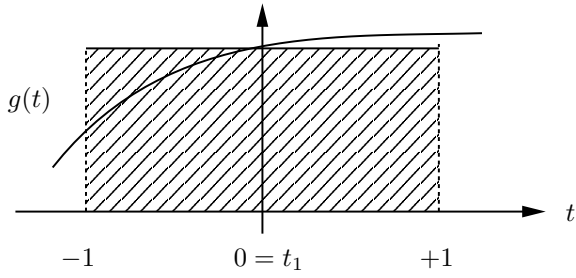

Fig. 3.3 Formule du rectangle sur $[-1, +1]$.

Si nous utilisons la formule du rectangle dans la formule composite (3.12), nous obtenons

$$L_h(f) = \sum_{i=0}^{N-1} (x_{i+1} - x_i) f\left(\frac{x_i + x_{i+1}}{2}\right) \tag{3.19}$$

et l'estimation (3.14) du théorème 3.1 devient :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^2. \tag{3.20}$$

L'interprétation géométrique de (3.19) est la suivante : on somme les aires des rectangles dont la base est le segment $[x_i, x_{i+1}]$ et dont la hauteur est $f(\xi_i)$, où ξ_i est le point milieu de $[x_i, x_{i+1}]$.

3.4 Formule de Simpson

La formule de Simpson est une formule à trois points : $M = 3$, $t_1 = -1$, $t_2 = 0$, $t_3 = +1$. La base de Lagrange φ_1, φ_2, φ_3 de $\mathbb{P}_2$ associée à ces trois points s'écrit (exemple 1.1) :

$$\varphi_1(t) = \frac{1}{2}(t^2 - t), \qquad \varphi_2(t) = 1 - t^2, \qquad \varphi_3(t) = \frac{1}{2}(t^2 + t).$$

Les relations (3.17) deviennent alors :

$$\omega_1 = \int_{-1}^{+1} \varphi_1(t)dt = \frac{1}{3}, \quad \omega_2 = \int_{-1}^{+1} \varphi_2(t)dt = \frac{4}{3}, \quad \omega_3 = \int_{-1}^{+1} \varphi_3(t)dt = \frac{1}{3}.$$

La formule de Simpson s'écrit donc :

$$J(g) = \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1). \tag{3.21}$$

Elle est une moyenne pondérée entre la formule du trapèze (poids 1/3) et la formule du rectangle (poids 2/3). Si nous utilisons cette formule de quadrature dans (3.12), nous obtenons :

$$L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{6} \left(f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2} \right) + f(x_{i+1}) \right). \tag{3.22}$$

D'après le théorème 3.2, la formule de Simpson est exacte pour des polynômes de degré 2. En fait, elle est même exacte pour des polynômes de degré 3. En effet, si $g(t) = t^3$, alors $J(g) = 0$ et $\int_{-1}^{+1} g(t)dt = \int_{-1}^{+1} t^3 dt = 0$. L'estimation (3.14) du théorème 3.1 devient donc :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^4. \tag{3.23}$$

La formule de Simpson donne une erreur d'ordre h^4. C'est une formule souvent utilisée dans la pratique car $L_h(f)$ converge très rapidement vers $\int_a^b f(x)dx$ lorsque h tend vers zéro.

3.5 Formules de Gauss-Legendre

L'idée des formules de Gauss-Legendre est de placer au mieux les points d'intégration $t_1, t_2, \ldots, t_M$ de sorte que la formule de quadrature $J(p) = \sum_{j=1}^{M} \omega_j p(t_j)$ soit égale à $\int_{-1}^{+1} p(t)dt$ pour des polynômes p de degré r aussi grand que possible.

En effet, rappelons que si f est une fonction donnée et si $L_h(f)$ est l'approximation de $\int_a^b f(x)dx$ définie par (3.12) alors, en vertu du théorème 3.1, plus r est grand et plus l'erreur entre $\int_a^b f(x)dx$ et $L_h(f)$ tend rapidement vers zéro avec h.

Commençons par la définition suivante :

Définition 3.3 *Le polynôme de Legendre de degré M est défini par*

$$L_M(t) = \frac{1}{2^M M!} \frac{d^M}{dt^M}(t^2-1)^M. \tag{3.24}$$

Ainsi donc nous avons, si $t \in \mathbb{R}$:

$$L_0(t) = 1, \qquad L_1(t) = t, \qquad L_2(t) = \frac{3t^2-1}{2}, \qquad \cdots$$

Les polynômes de Legendre L_0, L_1, L_2, ..., vérifient de nombreuses propriétés. Entre autres, nous démontrons certaines propriétés qui nous seront utiles dans la suite.

Théorème 3.3 *Les polynômes de Legendre L_0, L_1, L_2, ..., vérifient les propriétés suivantes :*

i) L_0, L_1, ..., L_M forment une base de $\mathbb{P}_M$.

ii) Si $i \neq j$ alors $\int_{-1}^{+1} L_i(t)L_j(t)dt = 0$ (propriété d'orthogonalité).

iii) L_M a exactement M zéros réels distincts tous compris dans l'intervalle ouvert $]-1,+1[$. Ces zéros sont appelés les points de Gauss.

Démonstration
i) On vérifie facilement que $L_j(t)$ est un polynôme de degré j exactement et ainsi $L_0, L_1, L_2, \ldots, L_M$ sont linéairement indépendants. Ils forment donc une base de $\mathbb{P}_M$.
ii) Supposons $i > j$. On obtient alors en intégrant par partie

$$\begin{aligned}
\int_{-1}^{+1} L_i(t)L_j(t)dt &= \frac{1}{2^{(i+j)}i!j!} \int_{-1}^{+1} \frac{d^i}{dt^i}(t^2-1)^i \frac{d^j}{dt^j}(t^2-1)^j dt \\
&= \frac{1}{2^{(i+j)}i!j!} \left\{ \left| \frac{d^{i-1}}{dt^{i-1}}(t^2-1)^i \frac{d^j}{dt^j}(t^2-1)^j \right|_{t=-1}^{t=1} \right. \\
&\qquad \left. - \int_{-1}^{+1} \frac{d^{i-1}}{dt^{i-1}}(t^2-1)^i \frac{d^{j+1}}{dt^{j+1}}(t^2-1)^j dt \right\}.
\end{aligned}$$

Puisque $(t^2-1)^i$ a un zéro d'ordre i en 1 et en -1, la $(i-1)$-ième dérivée de $(t^2-1)^i$ s'annule en $t=1$ et en $t=-1$. Ainsi nous obtenons

$$\int_{-1}^{+1} L_i(t)L_j(t)dt = \frac{(-1)}{2^{(i+j)}i!j!}\int_{-1}^{+1}\frac{d^{i-1}}{dt^{i-1}}(t^2-1)^i\frac{d^{j+1}}{dt^{j+1}}(t^2-1)^j dt.$$

En intégrant par partie j fois comme ci-dessus, nous obtenons :

$$\begin{aligned}\int_{-1}^{+1} L_i(t)L_j(t)dt &= \frac{(-1)^j}{2^{(i+j)}i!j!}\int_{-1}^{+1}\frac{d^{i-j}}{dt^{i-j}}(t^2-1)^i\underbrace{\frac{d^{2j}}{dt^{2j}}(t^2-1)^j}_{(2j)!}dt\\ &= \frac{(-1)^j(2j)!}{2^{(i+j)}i!j!}\int_{-1}^{+1}\frac{d^{i-j}}{dt^{i-j}}(t^2-1)^i dt\\ &= \frac{(-1)^j(2j)!}{2^{(i+j)}i!j!}\left|\frac{d^{i-j-1}}{dt^{i-j-1}}(t^2-1)^i\right|_{t=-1}^{t=1} = 0.\end{aligned}$$

iii) Soit $t_1, t_2, \ldots, t_s$ les points strictement compris entre -1 et $+1$ en lesquels L_M change de signe. Clairement ces points seront des zéros de L_M et on a donc $s \leq M$. Si on pose

$$p(t) = (t-t_1)(t-t_2)(t-t_3)\ldots(t-t_s)$$

on a $p \in \mathbb{P}_s$ et, puisque p change aussi de signe en les points t_j, $1 \leq j \leq s$, on obtient $p(t)L_M(t) \geq 0$, $\forall t \in [-1,+1]$ ou $p(t)L_M(t) \leq 0$, $\forall t \in [-1,+1]$. Dans tous les cas, puisque $p(t)L_M(t)$ n'est pas identiquement nul, on a

$$\int_{-1}^{+1} p(t)L_M(t)dt \neq 0.$$

En utilisant la partie i), on voit qu'il existe $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_s$ tels que

$$p(t) = \sum_{j=0}^{s}\alpha_j L_j(t)$$

et en utilisant la partie ii) on obtient

$$\int_{-1}^{+1} p(t)L_M(t)dt = \sum_{j=0}^{s}\alpha_j\int_{-1}^{+1} L_j(t)L_M(t)dt = \alpha_s\int_{-1}^{+1} L_s(t)L_M(t)dt.$$

Puisque $\int_{-1}^{+1} p(t)L_M(t)dt$ est non nul, on a nécessairement $s = M$ et donc les M zéros de L_M sont $t_1, t_2, \ldots, t_M$. ■

Définition 3.4 *Nous dirons que la formule de quadrature*

$$J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$$

est la formule de Gauss-Legendre à M points si

i) les points d'intégration $t_1 < t_2 < \cdots < t_M$ sont les M zéros du polynôme de Legendre L_M c'est-à-dire les M points de Gauss (voir propriété (iii) du théorème 3.3) ;

ii) les poids $\omega_1, \omega_2, \ldots, \omega_M$ sont définis par les relations (3.17), c'est-à-dire

$$\omega_j = \int_{-1}^{+1} \varphi_j(t)dt, \qquad j = 1, 2, \ldots, M,$$

où $\varphi_1, \varphi_2, \ldots, \varphi_M$ est la base de Lagrange de $\mathbb{P}_{M-1}$ associée aux M points de Gauss.

Nous sommes maintenant en mesure de démontrer le résultat suivant.

Théorème 3.4 *La formule de Gauss-Legendre à M points (M entier ≥ 1) est exacte pour les polynômes de degré $r = 2M - 1$.*

Démonstration
Soit $J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$ la formule de Gauss-Legendre à M points et soit p un polynôme de degré $2M - 1$. Clairement, nous pouvons définir pour $t \in \mathbb{R}$:

$$\tilde{p}(t) = \sum_{j=1}^{M} p(t_j)\varphi_j(t),$$

où $\varphi_1, \varphi_2, \ldots, \varphi_M$ est la base de Lagrange de $\mathbb{P}_{M-1}$ associée aux points de Gauss $t_1, t_2, \ldots, t_M$. Autrement dit, le polynôme $\tilde{p}$ est donc l'interpolant de p de degré $M - 1$ aux M points de Gauss $t_1, t_2, \ldots, t_M$.

Considérons maintenant le polynôme q défini par :

$$q(t) = p(t) - \tilde{p}(t) \qquad \forall t \in \mathbb{R}.$$

Le polynôme q est un polynôme de degré $2M - 1$ qui s'annule en les points $t_1, t_2, \ldots, t_M$, i.e. $q(t_j) = 0$ si $j = 1, 2, \ldots, M$. Ainsi q est divisible par le polynôme v de degré M défini par :

$$v(t) = (t - t_1)(t - t_2)(t - t_3) \cdots (t - t_M) \qquad \forall t \in \mathbb{R},$$

c'est-à-dire qu'il existe un polynôme w de degré $M - 1$ tel que

$$q(t) = v(t)w(t) \qquad \forall t \in \mathbb{R}.$$

Puisque v est un polynôme de degré M qui s'annule en les M zéros de L_M qui lui-même est aussi un polynôme de degré M, il existe un nombre réel α tel que

$$v(t) = \alpha L_M(t) \qquad \forall t \in \mathbb{R}.$$

Puisque w est un polynôme de degré $M-1$, il existe $\beta_0, \beta_1, \beta_2, \ldots, \beta_{M-1} \in \mathbb{R}$ (propriété (i) du théorème 3.3) tels que

$$w(t) = \sum_{k=0}^{M-1} \beta_k L_k(t).$$

Ainsi donc, en utilisant la propriété (ii) du théorème 3.3, nous avons :

$$\int_{-1}^{+1} q(t)dt = \int_{-1}^{+1} v(t)w(t)dt = \alpha \sum_{k=0}^{M-1} \beta_k \int_{-1}^{+1} L_M(t)L_k(t)dt = 0.$$

Par définition de q nous avons prouvé que

$$\int_{-1}^{+1} p(t)dt = \int_{-1}^{+1} \tilde{p}(t)dt,$$

et enfin, par définition de $\tilde{p}$, nous obtenons

$$\int_{-1}^{+1} p(t)dt = \sum_{j=1}^{M} p(t_j) \int_{-1}^{+1} \varphi_j(t)dt = \sum_{j=1}^{M} \omega_j p(t_j) = J(p).$$

Cette dernière relation est exactement ce que nous voulions montrer. ■

Remarque 3.2 Les poids ω_j, $j = 1, 2, ..., M$, d'une formule de Gauss-Legendre à M points sont tous positifs car φ_j^2 est un polynôme de degré $2M-2$, ce qui implique, en utilisant le théorème 3.4, que

$$0 < \int_{-1}^{+1} \varphi_j^2(t)dt = J(\varphi_j^2) = \sum_{k=1}^{M} \omega_k \varphi_j^2(t_k) = \omega_j.$$

Remarque 3.3 La formule de Gauss-Legendre à M points est optimale au sens où il existe un polynôme p de degré $2M$ tel que

$$\int_{-1}^{+1} p(t)dt \neq J(p).$$

En effet, il suffit de prendre $p(t) = \prod_{j=1}^{M}(t-t_j)^2$ pour obtenir $J(p) = 0$ alors que $\int_{-1}^{+1} p(t)dt > 0$.

Les points de Gauss et les poids correspondants sont donnés dans des tables numériques adéquates ou dans des logiciels d'intégration numérique. Connaissant une formule de Gauss à M points, nous pouvons calculer, pour une fonction f définie sur $[a, b]$, la quantité $L_h(f)$ donnée par la formule composite (3.12). Tenant compte des théorèmes 3.1 et 3.4 nous avons :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^{2M}, \tag{3.25}$$

où ici f est supposée $2M$ fois continûment dérivable et C est une constante qui ne dépend pas des points x_i, $i = 0, 1, \ldots, N$ choisis pour partitionner $[a, b]$.

Exemple 3.5 Formule de Gauss-Legendre à un seul point
On a $L_1(t) = t$ et donc le seul zéro de L_1 est donné par $t_1 = 0$. Nous retrouvons dans ce cas-là la formule du rectangle qui est d'ordre h^2 (voir (3.20)).

Exemple 3.6 Formule de Gauss-Legendre à deux points
Nous avons $L_2(t) = \frac{1}{2}(3t^2 - 1)$ et donc les deux zéros de L_2 sont donnés par

$$t_1 = -\frac{1}{\sqrt{3}} \qquad \text{et} \qquad t_2 = \frac{1}{\sqrt{3}}.$$

La base de Lagrange φ_1, φ_2 de $\mathbb{P}_2$ associée aux points t_1, t_2 est définie par

$$\varphi_1(t) = \frac{1 - \sqrt{3}t}{2} \qquad \text{et} \qquad \varphi_2(t) = \frac{\sqrt{3}t + 1}{2}$$

et ainsi

$$\omega_1 = \int_{-1}^{+1} \varphi_1(t)dt = 1 \qquad \text{et} \qquad \omega_2 = \int_{-1}^{+1} \varphi_2(t)dt = 1.$$

La formule de Gauss-Legendre à deux points s'écrit donc :

$$J(g) = g\left(-1/\sqrt{3}\right) + g\left(1/\sqrt{3}\right),$$

et la formule (3.12) devient :

$$L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1} - x_i}{2} \left\{ f\left(x_i + \frac{\sqrt{3}-1}{2\sqrt{3}}(x_{i+1} - x_i)\right) + f\left(x_i + \frac{\sqrt{3}+1}{2\sqrt{3}}(x_{i+1} - x_i)\right) \right\}.$$

Si f est quatre fois continûment dérivable, les théorèmes 3.1 et 3.4 nous assurent l'existence d'une constante C, indépendante du choix des points x_i, telle que :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^4.$$

La formule de Gauss-Legendre à deux points converge donc au même ordre que la formule de Simpson (sect. 3.4).

3.6 Exercices

Exercice 3.1 Soit α un nombre réel donné tel que $0 < \alpha < 1$, soit $t_1 = -1$, $t_2 = -\alpha$, $t_3 = \alpha$, $t_4 = +1$, et soit $\omega_1, \omega_2, \omega_3, \omega_4$ 4 nombres réels. Nous considérons la formule de quadrature définie par

$$J(g) = \sum_{j=1}^{4} \omega_j g(t_j),$$

où g est une fonction continue quelconque donnée sur $[-1, +1]$.

1. Trouver $\omega_1, \omega_2, \omega_3, \omega_4$ en fonction de α tel que $J(p) = \int_{-1}^{+1} p(t)dt$, pour tout polynôme p de degré 3.
2. Existe-t-il α tel que $J(p) = \int_{-1}^{+1} p(t)dt$, pour tout polynôme p de degré r, avec $r > 3$? Si oui, quelle est la valeur maximale de r et que valent alors α et $\omega_1, \omega_2, \omega_3, \omega_4$?

Solution
1. D'après la formule (3.17) on obtient

$$\omega_i = \int_{-1}^{1} \varphi_i(t)dt,$$

où les fonctions φ_i, $i = 1, 2, 3, 4$ sont les polynômes de la base de Lagrange de $\mathbb{P}_3$ associée aux points t_1, t_2, t_3, t_4. En utilisant l'égalité (1.5) nous avons

$$\varphi_1(t) = \frac{t+\alpha}{-1+\alpha} \cdot \frac{t-\alpha}{-1-\alpha} \cdot \frac{t-1}{-1-1},$$
$$\varphi_2(t) = \frac{t+1}{-\alpha+1} \cdot \frac{t-\alpha}{-\alpha-\alpha} \cdot \frac{t-1}{-\alpha-1},$$

et donc

$$\omega_1 = \int_{-1}^{1} \varphi_1(t)dt = \frac{1}{3} \cdot \frac{1-3\alpha^2}{1-\alpha^2},$$
$$\omega_2 = \int_{-1}^{1} \varphi_2(t)dt = \frac{1}{3} \cdot \frac{2}{1-\alpha^2}.$$

Pour des raisons de symétrie, $\omega_4 = \omega_1$ et $\omega_3 = \omega_2$. Par construction, $J(p)$ intègre exactement tout polynôme p de degré 3.

2. Tout polynôme p de degré r peut s'écrire $p(t) = at^r + q(t)$ où q est un polynôme de degré $r-1$ et $a \in \mathbb{R}$. Par conséquent

$$J(p) = a \sum_{j=1}^{4} \omega_j t_j^r + J(q)$$

et

$$\int_{-1}^{+1} p(t)dt = a \int_{-1}^{+1} t^r dt + \int_{-1}^{+1} q(t)dt.$$

Donc, pour que $J(p) = \int_{-1}^{+1} p(t)dt$ pour tout polynôme p de degré r, il suffit que $J(q) = \int_{-1}^{+1} q(t)dt$ pour tout polynôme q de degré $r-1$ et que

$$J(t^r) = \int_{-1}^{+1} t^r dt. \tag{3.26}$$

Nous procédons donc par étapes pour déterminer le degré maximal du polynôme pour lequel la formule de quadrature est exacte.

Nous avons montré au point 1 que la formule de quadrature est exacte pour les polynômes de degré 3. D'autre part

$$J(t^4) = \sum_{j=1}^{4} \omega_j t_j^4 = \frac{2}{3}\,\frac{1-3\alpha^2+2\alpha^4}{1-\alpha^2}, \qquad \text{et} \qquad \int_{-1}^{1} t^4 dt = \frac{2}{5}.$$

En égalant $J(t^4)$ à $\int_{-1}^{+1} t^4 dt$ on obtient $\alpha = 1/\sqrt{5}$ et ainsi, pour cette valeur de α, la formule de quadrature est exacte pour tout polynôme de degré 4. Nous pouvons maintenant calculer les poids de la formule de quadrature. Nous obtenons $\omega_1 = \omega_4 = \frac{1}{6}$ et $\omega_2 = \omega_3 = \frac{5}{6}$, et la formule de quadrature s'écrit

$$J(g) = \frac{1}{6}\Big(g(-1) + g(1)\Big) + \frac{5}{6}\Big(g\left(-1/\sqrt{5}\right) + g\left(1/\sqrt{5}\right)\Big).$$

De même, il est facile de vérifier que cette formule est aussi exacte pour le polynôme t^5 puisque

$$J(t^5) = 0 = \int_{-1}^{1} t^5 dt.$$

Par contre, elle ne l'est plus pour le polynôme t^6. En effet,

$$\begin{aligned} J(t^6) &= \frac{1}{6}\Big((-1)^6 + (1)^6\Big) + \frac{5}{6}\Big(\left(-1/\sqrt{5}\right)^6 + \left(1/\sqrt{5}\right)^6\Big) = \frac{26}{75} \\ &\neq \int_{-1}^{1} t^6 dt = \frac{2}{7}. \end{aligned}$$

Ainsi, la réponse à la deuxième question de l'exercice est positive et on a $r = 5$.

Exercice 3.2 Soit $0 < \alpha \leq 1$ un nombre réel donné, soit $t_1 = -\alpha$, $t_2 = 0$, $t_3 = \alpha$, et soit $\omega_1, \omega_2, \omega_3$, trois nombres réels. Nous considérons la formule de quadrature définie par

$$J(g) = \sum_{j=1}^{3} \omega_j g(t_j),$$

où g est une fonction continue donnée sur $[-1, 1]$.

1. Trouver ω_1, ω_2, ω_3 en fonction de α de sorte que $J(p) = \int_{-1}^{1} p(t)dt$, pour tout polynôme p de degré 2.
2. Montrer qu'avec de tels poids $J(p) = \int_{-1}^{1} p(t)dt$ pour tout polynôme p de degré 3.
3. Existe-t-il α tel que la formule de quadrature soit exacte pour les polynômes de degré 5 ? Si oui, calculer α et comparer avec les zéros du polynôme de Legendre de degré 3.

Solution
1. D'après la formule (3.17) nous avons

$$\omega_i = \int_{-1}^{1} \varphi_i(t)dt,$$

où les fonctions φ_i, $i = 1, 2, 3$ sont les polynômes de la base de Lagrange de $\mathbb{P}_2$ associée aux points t_1, t_2, t_3. En utilisant l'égalité (1.5) nous avons

$$\varphi_1(t) = \frac{t}{\alpha} \cdot \frac{t-\alpha}{2\alpha},$$
$$\varphi_2(t) = \frac{t+\alpha}{\alpha} \cdot \frac{t-\alpha}{-\alpha}.$$

et donc

$$\omega_1 = \int_{-1}^{1} \varphi_1(t)dt = \frac{1}{3\alpha^2},$$
$$\omega_2 = \int_{-1}^{1} \varphi_2(t)dt = \frac{-2}{3\alpha^2} + 2.$$

Pour des raisons de symétrie, $\omega_3 = \omega_1$. Par construction, $J(p)$ intègre exactement tout polynôme p de degré 2.

2. Pour que la formule de quadrature soit exacte pour un polynôme de degré 3, il suffit de vérifier que $J(t^3) = \int_{-1}^{+1} t^3 dt$. En effet

$$J(t^3) = \omega_1(-\alpha)^3 + \omega_2 \cdot 0 + \omega_3\alpha^3 = 0,$$

et, d'autre part

$$\int_{-1}^{+1} t^3 dt = \frac{1}{4}\left[t^4\right]_{t=-1}^{t=+1} = 0.$$

3. La formule de quadrature est exacte pour un polynôme de degré 4 si $J(t^4) = \int_{-1}^{+1} t^4 dt$. Puisque

$$J(t^4) = \omega_1(-\alpha)^4 + \omega_2 \cdot 0 + \omega_3\alpha^4 = 2\omega_1\alpha^4,$$

et puisque $\int_{-1}^{+1} t^4 dt = 2/5$, il suffit donc que $\omega_1\alpha^4 = 1/5$, soit encore $\alpha = \sqrt{3/5}$. Finalement nous vérifions facilement que

$$J(t^5) = \int_{-1}^{+1} t^5 dt = 0$$

et donc la formule de quadrature est exacte pour un polynôme de degré 5 lorsque $\alpha = \sqrt{3/5}$. D'après la formule (3.24), le polynôme de Legendre L_3 est donné par

$$L_3(t) = \frac{5}{2}t\left(t^2 - \frac{3}{5}\right).$$

Il s'annule donc pour $t = 0$ et $t = \pm\sqrt{3/5}$. La formule de quadrature correspondant au choix $\alpha = \sqrt{3/5}$ est donc la formule de Gauss-Legendre à 3 points. Elle est bien exacte pour les polynômes de degré $r = 2 \cdot 3 - 1 = 5$ en vertu du théorème 3.4.

3.7 Notes bibliographiques et remarques

Nous avons présenté dans la section 3.5 les formules de quadrature de Gauss-Legendre dont tous les points d'intégration sont strictement compris entre -1 et $+1$. Les formules de quadrature de Gauss-Legendre-Lobatto (voir par exemple [6]) tiennent compte des points -1 et $+1$ ainsi que des zéros de la première dérivée du polynôme de Legendre L_M. Le cas $M = 1$ donne la formule du trapèze alors que le cas $M = 2$ donne la formule de Simpson. L'exercice 3.1 traite le cas $M = 3$ car $L'_3(t) = (15t^2 - 3)/2$ et les zéros de L'_3 sont ainsi $t = -1/\sqrt{5}$ et $t = 1/\sqrt{5}$. Il est possible de montrer que les formules de Gauss-Legendre-Lobatto intègrent exactement les polynômes de degré $2M - 1$.

Nous avons uniquement présenté les formules de Gauss définies à partir des polynômes de Legendre. Il existe également des formules de Gauss basées sur les polynômes de Tchebycheff.

Dans le cas où l'intervalle d'intégration est semi-infini ou infini il existe des formules de Gauss basées sur les polynômes de Laguerre ou d'Hermite (voir par exemple [25]).

Pour un traitement mathématique complet de l'intégration gaussienne nous renvoyons à [6].

De nombreux logiciels scientifiques grand public (par exemple Maple^{TM}, Mathematica^{TM}, Matlab^{TM}) possèdent des fonctions d'intégration numérique.

Nous utiliserons les formules d'intégration numérique dans les chapitres 10 à 14, lors de la mise en œuvre des méthodes d'éléments finis.

Chapitre 4

Résolution de systèmes linéaires. Elimination de Gauss. Systèmes mal conditionnés. Systèmes surdéterminés

4.1 Position du problème

Dans ce chapitre, nous considérons un système d'équations linéaires d'ordre N de la forme

$$A\vec{x} = \vec{b}. \tag{4.1}$$

Ici A est une $N \times N$ matrice régulière d'ordre N de coefficients a_{ij}, $1 \leq i, j \leq N$, donnés, $\vec{b}$ est un vecteur colonne à N composantes b_j, $1 \leq j \leq N$, données et $\vec{x}$ est un vecteur colonne à N composantes x_j, $1 \leq j \leq N$, inconnues. Dans la suite, nous utilisons les notations matricielles standards, i.e.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}, \qquad \vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \qquad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}.$$

Le système (4.1) peut être écrit explicitement sous la forme d'un système de N équations à N inconnues $x_1, x_2, \ldots, x_N$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N & = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N & = b_2, \\ \vdots & \\ a_{N1}x_1 + a_{N2}x_2 + \cdots + a_{NN}x_N & = b_N. \end{cases} \tag{4.2}$$

Définition 4.1 *On dira que la matrice A est triangulaire supérieure (respectivement triangulaire inférieure) si $a_{ij} = 0$ pour tout couple i, j tel que $1 \leq j < i \leq N$ (resp. $1 \leq i < j \leq N$).*

Définition 4.2 *Si A est une matrice triangulaire supérieure (resp. triangulaire inférieure), on dira que (4.1) ou (4.2) est un système triangulaire supérieur (resp. triangulaire inférieur).*

Supposons un instant que la matrice A soit triangulaire supérieure. Nous constatons alors que le déterminant de A est le produit des valeurs diagonales a_{ii} et, puisque la matrice A est supposée régulière, nous avons $a_{ii} \neq 0$, $i = 1, 2, \ldots, N$. Ainsi, quitte à diviser chaque équation de (4.2) par le terme diagonal, il n'est pas restrictif de supposer que $a_{ii} = 1$, $i = 1, 2, \ldots, N$. Dans ce cas, la matrice A est triangulaire supérieure avec des valeurs 1 dans sa diagonale et, de (4.2), nous déduisons successivement les inconnues $x_N, x_{N-1}, \ldots, x_1$. En effet, nous avons :

$$x_N = b_N$$

et pour $i = N-1, N-2, \ldots, 3, 2, 1$:

$$x_i = b_i - \sum_{j=i+1}^{N} a_{ij} x_j. \tag{4.3}$$

Dans le cas où la matrice A est régulière mais non nécessairement triangulaire supérieure, la méthode d'élimination de Gauss aura pour but de transformer le système $A\vec{x} = \vec{b}$ en un système équivalent (c'est-à-dire possédant la même solution) triangulaire supérieur avec des valeurs 1 sur la diagonale.

4.2 Elimination de Gauss sur un exemple

Considérons un exemple avec $N = 3$ pour montrer comment faire la transformation d'un système linéaire en un système triangulaire supérieur équivalent. Dans cet exemple, nous choisissons :

$$A = \begin{bmatrix} 4 & 8 & 12 \\ 3 & 8 & 13 \\ 2 & 9 & 18 \end{bmatrix}, \qquad \vec{b} = \begin{bmatrix} 4 \\ 5 \\ 11 \end{bmatrix}.$$

Le système $A\vec{x} = \vec{b}$ devient dans ce cas :

$$\begin{cases} 4x_1 + 8x_2 + 12x_3 = 4, \\ 3x_1 + 8x_2 + 13x_3 = 5, \\ 2x_1 + 9x_2 + 18x_3 = 11. \end{cases} \tag{4.4}$$

Première étape La première opération consiste à diviser la première équation de (4.4) par $a_{11} = 4$ (appelé premier pivot) pour obtenir :

$$x_1 + 2x_2 + 3x_3 = 1. \tag{4.5}$$

Ensuite nous soustrayons 3 fois l'équation (4.5) à la deuxième équation de (4.4) et 2 fois l'équation (4.5) à la troisième équation de (4.4). Nous obtenons un système équivalent à (4.4) qui est (en répétant (4.5)) :

$$\left\{\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & = & 1, \\ & & 2x_2 & + & 4x_3 & = & 2, \\ & & 5x_2 & + & 12x_3 & = & 9. \end{array}\right. \tag{4.6}$$

Nous observons que les deux dernières équations de (4.6) font apparaître seulement deux inconnues x_2 et x_3 (l'inconnue x_1 a été éliminée) et nous pouvons recommencer le procédé en laissant inchangée la première équation.

Deuxième étape Nous divisons la deuxième équation de (4.6) par 2 (le deuxième pivot). Nous obtenons :

$$x_2 + 2x_3 = 1. \tag{4.7}$$

Nous soustrayons 5 fois (4.7) à la troisième équation de (4.6) et nous obtenons le système suivant :

$$\left\{\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & = & 1, \\ & & x_2 & + & 2x_3 & = & 1, \\ & & & & 2x_3 & = & 4, \end{array}\right. \tag{4.8}$$

qui est équivalent au système (4.4).

Dernière étape Finalement, il suffit de diviser la troisième équation de (4.8) par le troisième pivot, qui est encore ici 2, pour obtenir :

$$\left\{\begin{array}{rcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & = & 1, \\ & & x_2 & + & 2x_3 & = & 1, \\ & & & & x_3 & = & 2. \end{array}\right. \tag{4.9}$$

De (4.9), il est facile de déduire successivement les inconnues x_3, x_2, x_1 comme nous l'avons décrit dans (4.3). Nous obtenons :

$$x_3 = 2, \qquad x_2 = -3, \qquad x_1 = 1.$$

4.3 Algorithme d'élimination

Nous présentons maintenant un algorithme qui, effectué par un ordinateur, permet de réaliser l'élimination dont le mécanisme a été décrit dans la section précédente. Pour réaliser cet objectif, nous appelons $A^{(i)}$ la matrice et $\vec{b}^{(i)}$ le second membre obtenus avant la i-ième étape de l'élimination. Ainsi le tableau

$A^{(i)}$ a la forme suivante :

$$A^{(i)} = \begin{bmatrix} 1 & a_{12}^{(i)} & a_{13}^{(i)} & a_{14}^{(i)} & \cdots & & a_{1i}^{(i)} & \cdots & a_{1N}^{(i)} \\ 0 & 1 & a_{23}^{(i)} & a_{24}^{(i)} & \cdots & & a_{2i}^{(i)} & \cdots & a_{2N}^{(i)} \\ 0 & 0 & 1 & a_{34}^{(i)} & \cdots & & a_{3i}^{(i)} & \cdots & a_{3N}^{(i)} \\ 0 & 0 & 0 & 1 & \cdots & & a_{4i}^{(i)} & \cdots & a_{4N}^{(i)} \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & 1 & & & \\ & & 0 & & & & a_{ii}^{(i)} & \cdots & a_{iN}^{(i)} \\ & & & & & & \vdots & & \vdots \\ & & & & & & a_{Ni}^{(i)} & \cdots & a_{NN}^{(i)} \end{bmatrix}. \tag{4.10}$$

La i-ième étape de l'élimination consistera à passer du tableau $A^{(i)}$ au tableau $A^{(i+1)}$ et du tableau $\vec{b}^{(i)}$ au tableau $\vec{b}^{(i+1)}$ par les opérations suivantes :

i-ième étape Nous divisons la i-ième ligne de $A^{(i)}$ par le i-ième pivot $a_{ii}^{(i)}$ pour obtenir

$$a_{ij}^{(i+1)} = a_{ij}^{(i)} / a_{ii}^{(i)}, \qquad j = i+1, i+2, \ldots, N. \tag{4.11}$$

Nous faisons de même avec le second membre :

$$b_i^{(i+1)} = b_i^{(i)} / a_{ii}^{(i)}. \tag{4.12}$$

Nous prenons ensuite la k-ième ligne de $A^{(i)}$, $k = i+1, i+2, \ldots, N$, à laquelle nous retranchons $a_{ki}^{(i)}$ fois la ligne (4.11). Nous obtenons pour $k = i+1, i+2, \ldots, N$:

$$a_{kj}^{(i+1)} = a_{kj}^{(i)} - a_{ki}^{(i)} * a_{ij}^{(i+1)}, \qquad j = i+1, i+2, \ldots, N. \tag{4.13}$$

De même pour le second membre :

$$b_k^{(i+1)} = b_k^{(i)} - a_{ki}^{(i)} * b_i^{(i+1)}. \tag{4.14}$$

Une fois la i-ième étape réalisée, il n'est plus nécessaire d'avoir en mémoire les nombres $a_{kj}^{(i)}$, $i+1 \leq k, j \leq N$. Ainsi nous pouvons économiser la place occupée dans la mémoire de l'ordinateur en écrivant (4.11) sous la forme :

$$a_{ij} := a_{ij} / a_{ii}, \qquad j = i+1, i+2, \ldots, N, \tag{4.15}$$

ce qui signifie, par convention, que le membre de droite a_{ij}/a_{ii} est calculé en premier lieu et qu'ensuite a_{ij} est remplacé par cette valeur calculée. Il en est de même pour (4.12), (4.13) et (4.14). Avec cette convention, nous pouvons écrire l'algorithme d'élimination de Gauss qui rendra triangulaire supérieur (avec des valeurs diagonales 1) un système de N équations avec N inconnues. Cet algorithme est décrit dans le tableau 4.1.

Dans le tableau 4.1, la notation

$$\left[\begin{array}{l} \text{Faire } i = 1 \text{ à } N-1 \\ \cdots \end{array} \right.$$

Tableau 4.1 Algorithme d'élimination de Gauss.

entrées : a_{ij}, $1 \leq i, j \leq N$ et b_j, $1 \leq j \leq N$
représentant la matrice A et le second membre $\vec{b}$
du système linéaire originel
sorties : a_{ij}, $1 \leq i < j \leq N$ et b_j, $1 \leq j \leq N$
représentant la partie surdiagonale
du système triangulaire supérieur
et le nouveau second membre $\vec{b}$

Algorithme	**Commentaires**
Faire $i = 1$ à $N - 1$	Elimination de l'inconnue x_i
$p := 1/a_{ii}$	Inverse du i-ième pivot
Faire $j = i + 1$ à N	Division de la i-ième ligne par le i-ième pivot
$a_{ij} := p * a_{ij}$	(termes surdiagonaux uniquement)
$b_i := p * b_i$	Division de b_i par le i-ième pivot
Faire $k = i + 1$ à N	Elimination dans la k-ième équation
Faire $j = i + 1$ à N	Soustraction de a_{ki} fois la
$a_{kj} := a_{kj} - a_{ki} * a_{ij}$	nouvelle i-ième ligne à la k-ième ligne
$b_k := b_k - a_{ki} * b_i$	Soustraction de a_{ki} fois b_i à b_k
$p := 1/a_{NN}$	Inverse du N-ième pivot
$b_N := p * b_N$	Division de b_N par le N-ième pivot

symbolise une boucle dans laquelle on fait successivement $i = 1, 2, 3, \ldots$, jusqu'à $N - 1$.

Après avoir effectué l'algorithme décrit dans le tableau 4.1, il faut encore résoudre le système triangulaire supérieur :

$$\begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & a_{1N} \\ 0 & 1 & a_{23} & \cdots & a_{2N} \\ 0 & 0 & 1 & \cdots & a_{3N} \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_N \end{bmatrix} \tag{4.16}$$

en procédant comme dans (4.3). Dans le but d'économiser la place occupée dans la mémoire de l'ordinateur, la solution $\vec{x}$ du système linéaire (4.1) est stockée

dans le vecteur $\vec{b}$, c'est-à-dire il suffira d'ajouter la boucle suivante

$$\left[\begin{array}{l} \text{Faire } i = N-1 \text{ à } 1(\text{par pas de } -1) \\ b_i := b_i - \sum_{j=i+1}^{N} a_{ij} * b_j \end{array}\right.$$

à l'algorithme d'élimination de Gauss pour obtenir dans le vecteur $\vec{b}$ la solution $\vec{x}$ cherchée.

L'algorithme décrit dans le tableau 4.1 ne peut être exécuté jusqu'à la fin que si les pivots successifs sont non nuls. D'ailleurs, il devrait être complété par un test du type "si $a_{ii} = 0$ alors stop" lorsqu'on inverse les pivots. La question que nous devons nous poser est la suivante : Quand est-ce que les pivots a_{ii} de l'algorithme décrit dans le tableau 4.1 sont tous non nuls ? La réponse suit la définition ci-dessous.

Définition 4.3 *A_k est la sous-matrice principale d'ordre k de A si A_k est la $k \times k$ matrice de coefficients a_{ij}, $1 \leq i, j \leq k$, $1 \leq k \leq N$.*

Nous avons ainsi le résultat suivant.

Théorème 4.1 *Si toutes les sous-matrices principales A_k de la matrice de départ A sont régulières, $k = 1, 2, \ldots, N$, alors les pivots obtenus successivement dans l'élimination de Gauss (algorithme du tableau 4.1) sont tous non nuls. Inversement si tous les pivots obtenus au cours de l'élimination de Gauss sont non nuls, alors toutes les sous-matrices principales de A sont régulières.*

Démonstration
Considérons la matrice $A^{(i)}$ obtenue avant la i-ième étape de l'élimination de Gauss et décrite dans (4.10). On notera $A_k^{(i)}$ sa sous-matrice principale d'ordre k, $1 \leq k \leq N$.

Il est facile de vérifier que $A_i^{(i)}$ est une matrice triangulaire supérieure et que son déterminant est donné par le produit des valeurs diagonales, c'est-à-dire :

$$\det A_i^{(i)} = a_{ii}^{(i)}. \tag{4.17}$$

Il est également facile de vérifier que les opérations faites sur la matrice originelle A impliquent, si A_i est la sous-matrice principale d'ordre i de A :

$$\left\{\begin{array}{l} \det A_1 = \det A_1^{(1)}, \\ \det A_2 = a_{11}^{(1)} \det A_2^{(2)}, \\ \det A_3 = a_{11}^{(1)} a_{22}^{(2)} \det A_3^{(3)}, \\ \vdots \\ \det A_i = a_{11}^{(1)} a_{22}^{(2)} \ldots a_{i-1,i-1}^{(i-1)} \det A_i^{(i)}, \\ \vdots \end{array}\right. \tag{4.18}$$

Nous concluons de (4.17) et (4.18) que si dét $A_i \neq 0$ pour tout $i = 1, 2, \ldots, N$ alors les valeurs $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{NN}^{(N)}$ sont non nulles et réciproquement. ■

4.4 Nombre d'opérations pour l'élimination de Gauss

Comptons le nombre de multiplications N_m que nous avons dans l'algorithme d'élimination de Gauss. A partir du tableau 4.1, nous obtenons :

$$\begin{aligned} N_m &= \sum_{i=1}^{N-1} \Big[(N-i)+1+(N-i)(N-i+1)\Big]+1 \\ &= \sum_{i=1}^{N-1} \Big[(N-i+1)\Big]^2+1 \\ &= N^2+(N-1)^2+(N-2)^2+\cdots+2^2+1^2 \\ &= \sum_{j=1}^{N} j^2. \end{aligned}$$

Si nous montrons que

$$\sum_{j=1}^{N} j^2 = \frac{N^3}{3}+\frac{N^2}{2}+\frac{N}{6},$$

nous aurons démontré que pour N grand, $N_m = N^3/3 + O(N^2)$ où $O(N^2)$ désigne un reste d'ordre N^2 lorsque $N \to \infty$.

En effet il est facile de vérifier que

$$j^2 = \int_{j-1}^{j} x^2 dx + j - \frac{1}{3}.$$

Ainsi nous aurons

$$\begin{aligned} \sum_{j=1}^{N} j^2 &= \sum_{j=1}^{N} \int_{j-1}^{j} x^2 dx + \sum_{j=1}^{N} j - \frac{N}{3} \\ &= \int_{0}^{N} x^2 dx + \frac{(N+1)N}{2} - \frac{N}{3} \\ &= \frac{N^3}{3}+\frac{N^2}{2}+\frac{N}{6}. \end{aligned}$$

Remarquons encore que le nombre de soustractions faites dans l'algorithme d'élimination de Gauss est aussi de l'ordre de $N^3/3$. Ce qu'il faut retenir est l'affirmation suivante :

Chaque fois que l'ordre d'un système linéaire est multiplié par deux, le nombre d'opérations effectuées dans l'algorithme d'élimination de Gauss (tab. 4.1) est multiplié par huit (2^3), ceci évidemment lorsque N est grand !

4.5 Elimination de Gauss avec changement de pivot

Comme nous venons de le voir, l'algorithme d'élimination donné dans la section 4.3 ne peut être exécuté que si les pivots successifs sont non nuls, c'est-à-dire si toutes les sous-matrices principales de A sont régulières. Il est évident qu'il est impossible de traiter par cet algorithme le système suivant :

$$\left\{\begin{array}{rcrcrcl} 0x_1 & + & x_2 & + & 3x_3 & = & 1, \\ 5x_1 & + & 2x_2 & + & 3x_3 & = & 4, \\ 6x_1 & + & 8x_2 & + & x_3 & = & 1, \end{array}\right. \tag{4.19}$$

car, dans ce cas, on ne peut pas diviser la première ligne par le premier pivot qui est nul (la première sous-matrice principale est donc singulière !). On voit immédiatement que les choses se présentent mieux si on échange la première et la troisième ligne pour obtenir

$$\left\{\begin{array}{rcrcrcl} 6x_1 & + & 8x_2 & + & x_3 & = & 1, \\ 5x_1 & + & 2x_2 & + & 3x_3 & = & 4, \\ 0x_1 & + & x_2 & + & 3x_3 & = & 1. \end{array}\right. \tag{4.20}$$

En effet, maintenant nous pouvons diviser la première ligne par le nombre 6. Cette manière de faire s'appelle ***pivotage partiel*** ; elle consiste à échanger deux équations dans le but d'avoir le plus grand pivot possible en valeur absolue.

L'algorithme d'élimination de Gauss avec choix du plus grand pivot s'obtient donc en intercalant, dans l'algorithme du tableau 4.1, entre la première ligne "Faire $i = 1$ à $N-1$" et la deuxième ligne "$p := 1/a_{ii}$", la procédure décrite dans le tableau 4.2.

Il est facile de démontrer que, si la matrice initiale A est régulière, alors l'algorithme d'élimination de Gauss avec choix du plus grand pivot est exécutable jusqu'au bout, c'est-à-dire qu'à la sortie de la procédure ci-dessus, la valeur de a_{ii} est toujours non nulle.

A ce stade de l'exposé, il convient de remarquer que, lors de la résolution d'un système linéaire, il est toujours possible de multiplier une équation par un nombre non nul sans pour autant modifier ses solutions.

Ainsi, avant d'exécuter une élimination de Gauss avec choix du plus grand pivot, il convient de procéder à un ***équilibrage*** de la matrice de façon à ce que toutes les lignes (c'est-à-dire toutes les équations) aient un poids semblable. L'équilibrage consiste pour $i = 1, 2, \ldots, N$ à multiplier la i-ième équation de (4.2)

$$\sum_{j=1}^{N} a_{ij} x_j = b_i,$$

par un coefficient $r_i > 0$ choisi de telle sorte que

$$\max_{1 \leq j \leq N} r_i \mid a_{ij} \mid = 1.$$

On a vu que l'élimination de Gauss avec choix du pivot consiste à échanger des équations tout en gardant l'ordre des inconnues (on parle de pivotage

Tableau 4.2 Algorithme du choix du plus grand pivot.

Choix du plus grand pivot en valeur absolue (à intercaler entre les deux premières lignes de l'algorithme d'élimination de Gauss)	
Algorithme	**Commentaires**
$k := i$	On élimine a priori l'inconnue x_i et donc i est fixé
$m := abs(a_{ii})$	On initialise $m =\mid a_{ii} \mid$
Faire $j = i+1$ à N $s := abs(a_{ji})$ Si $m < s$ alors $k := j$ et $m := s$	Recherche du plus grand pivot en valeur absolue dans la i-ième colonne
	Après cette séquence, m est le plus grand pivot en valeur absolue et se trouve à la ligne k. (si $m = 0$ alors le système est singulier)
Si $k = i$ on saute ce qui suit	Si $k = i$ on n'a pas à modifier l'élimination de l'inconnue x_i dans l'algorithme d'élimination de Gauss
Faire $j = i$ à N $t := a_{ij}$ $a_{ij} := a_{kj}$ $a_{kj} := t$	Echange des lignes k et i
	Echange de a_{ij} et a_{kj}
$t := b_i$ $b_i := b_k$ $b_k := t$	Echange de b_i et b_k

partiel). Evidemment, on pourrait aussi penser à échanger l'ordre des inconnues en choisissant, non pas seulement le plus grand pivot en valeur absolue dans la i-ième colonne, mais aussi dans toutes les lignes à partir de la i-ième. On parle dans ce cas de ***pivotage complet*** (échange de colonnes et de lignes). Le pivotage partiel requiert un nombre d'opérations de l'ordre de N^2 alors que le pivotage complet requiert un nombre d'opérations de l'ordre de N^3 (donc coûteux !). Ce dernier n'est jamais pratiqué.

4.6 Systèmes mal conditionnés

Considérons maintenant le système de 2 équations à 2 inconnues suivant :

$$\left\{ \begin{array}{rcrcl} 4.218613x_1 & + & 6.327917x_2 & = & 10.546530 \\ 3.141592x_1 & + & 4.712390x_2 & = & 7.853982. \end{array} \right. \tag{4.21}$$

Nous vérifions aisément que ce système est régulier (déterminant de A non nul) et que la solution est donnée par

$$x_1 = x_2 = 1. \tag{4.22}$$

Considérons maintenant un système d'équations voisin (la flèche $\rightarrow$ indique un changement de décimale) :

$$\left\{ \begin{array}{rcrcl} 4.218611\overset{\downarrow}{1}x_1 & + & 6.327917x_2 & = & 10.546530 \\ 3.141594x_1 & + & 4.712390x_2 & = & 7.853980\overset{\uparrow}{.} \end{array} \right. \tag{4.23}$$

Nous vérifions encore que ce système est régulier, mais cette fois la solution est donnée par

$$x_1 = -5, \qquad x_2 = +5. \tag{4.24}$$

Nous concluons donc que, bien que les systèmes (4.21) et (4.23) soient voisins, leurs solutions sont très différentes. On parle dans ce cas de ***systèmes mal conditionnés***. L'interprétation géométrique est la suivante :

Le système (4.21) est formé de deux équations qui, dans le plan Ox_1, x_2, décrivent deux droites presque parallèles. Ainsi, résoudre le système (4.21) revient à chercher l'intersection de ces deux droites presque parallèles. Il est clair que si on perturbe un tout petit peu deux droites presque parallèles (système (4.23)), alors le point d'intersection est fortement modifié !

Résoudre un problème mal conditionné avec un ordinateur peut être une affaire délicate si l'ordinateur calcule avec trop peu de chiffres significatifs. Dans l'exemple précédent nous observons que, si l'ordinateur ne retient que 6 chiffres significatifs, il est complètement inespéré d'obtenir une solution raisonnablement proche de la solution du système (4.21). Par contre, nous prétendons qu'avec un calculateur à 10 chiffres significatifs, nous pouvons espérer obtenir une solution dont les 3 premiers chiffres significatifs, au moins, coïncident avec ceux de la solution. Dans la suite nous présentons rapidement une théorie permettant de valider cette affirmation.

Commençons par deux définitions.

Définition 4.4 *Soit A une $N \times N$ matrice. Si pour un N-vecteur $\vec{y}$ de composantes y_j, $1 \leq j \leq N$ nous notons*

$$\| \vec{y} \| = \left(\sum_{j=1}^{N} y_j^2 \right)^{1/2}$$

sa norme euclidienne, alors la norme spectrale de A est définie par

$$||| A |||= \max_{\vec{y}\neq 0} \frac{\| A\vec{y} \|}{\| \vec{y} \|}. \tag{4.25}$$

Remarque 4.1 Si $\vec{x}$ est un N-vecteur quelconque alors, par définition, on a :

$$\| A\vec{x} \| \leq ||| A ||| \cdot \| \vec{x} \| . \tag{4.26}$$

Nous sommes maintenant en mesure de montrer le résultat suivant.

Théorème 4.2 *Soit A une $N \times N$ matrice et soit ω la plus grande valeur propre de $A^T A$ où A^T est la matrice A transposée. Alors on a $||| A |||= \sqrt{\omega}$.*

Démonstration
Tout d'abord, nous remarquons que $A^T A$ est une matrice symétrique et donc que ses valeurs propres $\omega_1, \omega_2, \omega_3, \ldots, \omega_N$ sont toutes réelles. De plus si $\vec{\varphi}_j \neq 0$ est un vecteur propre de $A^T A$ associé à la valeur propre ω_j, alors nous avons $A^T A\vec{\varphi}_j = \omega_j \vec{\varphi}_j$ et par suite $\| A\vec{\varphi}_j \|^2 = \vec{\varphi}_j^T A^T A\vec{\varphi}_j = \omega_j \vec{\varphi}_j^T \vec{\varphi}_j = \omega_j \| \vec{\varphi}_j \|^2$, ce qui prouve que $\omega_j \geq 0$ pour tout $j = 1, 2, \ldots, N$. On dit que $A^T A$ est semi-définie positive. Ainsi donc $\omega = \max_{1\leq j\leq N} \omega_j$ est une valeur réelle non négative et il en est de même pour $\sqrt{\omega}$.

Si D est la $N \times N$ matrice diagonale de valeurs diagonales $\omega_1, \omega_2, \ldots, \omega_N$, alors nous savons qu'il existe une $N \times N$ matrice orthogonale Q telle que

$$A^T A = Q^T D Q. \tag{4.27}$$

Ainsi nous obtenons la suite d'égalités suivantes :

$$\begin{aligned} ||| A |||^2 &= \max_{\vec{y}\neq 0} \frac{\| A\vec{y} \|^2}{\| \vec{y} \|^2} = \max_{\vec{y}\neq 0} \frac{\vec{y}^T A^T A\vec{y}}{\vec{y}^T \vec{y}} = \max_{\vec{y}\neq 0} \frac{\vec{y}^T Q^T DQ\vec{y}}{\vec{y}^T Q^T Q\vec{y}} \\ &= \max_{\vec{z}\neq 0} \frac{\vec{z}^T D\vec{z}}{\vec{z}^T \vec{z}} = \max_{\vec{z}\neq 0} \frac{\sum_{j=1}^N \omega_j z_j^2}{\sum_{j=1}^N z_j^2}. \end{aligned} \tag{4.28}$$

Si $\vec{z}$ est un N-vecteur quelconque non nul alors

$$\sum_{j=1}^N \omega_j z_j^2 \leq \sum_{j=1}^N \omega z_j^2 = \omega \sum_{j=1}^N z_j^2. \tag{4.29}$$

De plus, il existe un N-vecteur non nul qui satisfait

$$\sum_{j=1}^N \omega_j z_j^2 = \omega \sum_{j=1}^N z_j^2 \tag{4.30}$$

car, en effet, si k est tel que $\omega = \omega_k$, il suffit de prendre $z_j = 0 \ \forall j \neq k$ et $z_k = 1$.

Les relations (4.28), (4.29) et (4.30) impliquent que $|||A|||^2 = \omega$, ce qui prouve le théorème 4.2. ■

Définition 4.5 *Soit A une $N \times N$ matrice régulière. On appellera nombre de condition spectral de A le nombre positif suivant :*

$$\chi(A) = ||| A ||| \cdot ||| A^{-1} ||| \tag{4.31}$$

où A^{-1} est la matrice inverse de A.

Lorsque A est symétrique nous avons le résultat suivant.

Théorème 4.3 *Soit A une $N \times N$ matrice symétrique régulière de valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_N$. Alors nous avons :*

$$\chi(A) = \frac{\max\limits_{1 \le j \le N} | \lambda_j |}{\min\limits_{1 \le j \le N} | \lambda_j |}.$$

Démonstration
Lorsque A est symétrique nous avons $A^T A = A^2$. Ainsi les valeurs propres de $A^T A$ sont données par $\lambda_1^2, \lambda_2^2, \ldots, \lambda_N^2$ et en utilisant le théorème 4.2 nous obtenons :

$$||| A ||| = \max_{1 \le j \le N} |\lambda_j|. \tag{4.32}$$

Le même raisonnement peut être appliqué à A^{-1} qui a les valeurs propres $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_N^{-1}$. Ainsi

$$||| A^{-1} ||| = \max_{1 \le j \le N} |\lambda_j^{-1}| = \frac{1}{\min\limits_{1 \le j \le N} |\lambda_j|}. \tag{4.33}$$

Les relations (4.32) et (4.33), ainsi que la définition 4.5 démontrent le théorème 4.3. ■

Considérons maintenant un système régulier de N équations à N inconnues

$$A\vec{x} = \vec{b}. \tag{4.34}$$

Si $\vec{\delta b}$ est une perturbation de $\vec{b}$ et si on résout, au lieu de (4.34),

$$A\vec{y} = \vec{b} + \vec{\delta b}, \tag{4.35}$$

on obtient par linéarité

$$\vec{y} = \vec{x} + \vec{\delta x} \qquad \text{avec} \qquad A\vec{\delta x} = \vec{\delta b}. \tag{4.36}$$

La question maintenant est de savoir s'il est possible de majorer l'erreur relative $\|\vec{\delta x}\|/\|\vec{x}\|$ sur la solution du système en fonction de l'erreur relative $\|\vec{\delta b}\|/\|\vec{b}\|$ commise sur le second membre. La réponse est donnée dans l'affirmation suivante.

Théorème 4.4 *Soit A une $N \times N$ matrice régulière, $\vec{b}$ un N-vecteur donné non nul et $\vec{\delta b}$ un N-vecteur qui perturbe $\vec{b}$. Alors si $\vec{x}$ et $\vec{\delta x}$ sont des N-vecteurs tels que $A\vec{x} = \vec{b}$ et $A\vec{\delta x} = \vec{\delta b}$ nous avons :*

$$\frac{\|\vec{\delta x}\|}{\|\vec{x}\|} \leq \chi(A) \frac{\|\vec{\delta b}\|}{\|\vec{b}\|} \tag{4.37}$$

où $\chi(A)$ est le nombre de condition spectral de A.

Démonstration
En utilisant la relation (4.26) avec A^{-1} en lieu et place de A, ainsi que (4.36) nous obtenons :

$$\frac{\|\vec{\delta x}\|}{\|\vec{x}\|} = \frac{\|A^{-1}\vec{\delta b}\|}{\|\vec{x}\|} \leq |||A^{-1}||| \frac{\|\vec{\delta b}\|}{\|\vec{x}\|}. \tag{4.38}$$

D'autre part, les relations (4.26) et (4.34) conduisent à :

$$\|\vec{b}\| = \|A\vec{x}\| \leq ||| A ||| \cdot \|\vec{x}\|,$$

et donc

$$\frac{1}{\|\vec{x}\|} \leq \frac{||| A |||}{\|\vec{b}\|}. \tag{4.39}$$

Les relations (4.38) et (4.39) ainsi que la définition 4.5 démontrent le théorème 4.4. ■

Nous avons donc prouvé (théorème 4.4) que l'erreur relative $\|\vec{\delta x}\|/\|\vec{x}\|$ sur la solution $\vec{x}$ est majorée par l'erreur relative $\|\vec{\delta b}\|/\|\vec{b}\|$ sur le second membre $\vec{b}$ multipliée par le facteur $\chi(A)$. Ce facteur $\chi(A)$ est, lorsque la matrice A est symétrique, le rapport entre sa plus grande valeur propre en valeur absolue et sa plus petite (théorème 4.3).

En fait l'estimation (4.37) est souvent pessimiste (au sens où $\|\vec{\delta x}\|/\|\vec{x}\|$ est souvent nettement plus petit que $\chi(A) \cdot \|\vec{\delta b}\|/\|\vec{b}\|$) mais non améliorable (au sens où il existe des exemples pour lesquels $\|\vec{\delta x}\|/\|\vec{x}\| = \chi(A) \cdot \|\vec{\delta b}\|/\|\vec{b}\|$) (exercice 4.1). Si $\|\vec{\delta b}\|/\|\vec{b}\|$ est de l'ordre de la précision relative $\eta = 10^{-p}$ du calculateur (définition 2.4 où p est le nombre de chiffres significatifs), alors $\|\vec{\delta x}\|/\|\vec{x}\|$ pourrait, au pire, être égal à

$$\chi(A) \cdot \eta = 10^{\log_{10} \chi(A)} \cdot 10^{-p} = 10^{\log_{10} \chi(A) - p}.$$

Nous conviendrons donc de la règle suivante.

Si on calcule la solution de (4.34) avec un ordinateur à p chiffres significatifs en valeur décimale, on ne pourra pas garantir a priori plus de

$$[p - \log_{10} \chi(A)]$$

chiffres significatifs sur la solution (ici $[\cdot]$ désigne la partie entière).

Appliquons cette règle au système linéaire (4.21). Il est facile de vérifier que $\chi(A)$ est de l'ordre de 10^7, par conséquent nous pouvons perdre jusqu'à 7 chiffres significatifs lors de la résolution de (4.21) ! Il faut donc bien un ordinateur calculant avec 10 chiffres significatifs pour être sûr d'obtenir les 3 premiers chiffres de la solution !

4.7 Systèmes surdéterminés. Méthode des moindres carrés

Soit A une $M \times N$ matrice dont le nombre de lignes M est plus grand que le nombre de colonnes N. Si $\vec{b}$ est un M-vecteur donné, on peut vouloir déterminer un N-vecteur $\vec{x}$ satisfaisant la relation $A\vec{x} = \vec{b}$. En fait, ce dernier système a plus d'équations (M équations) que d'inconnues (N inconnues); on dit qu'on est en présence d'un ***système surdéterminé*** et dans bien des cas il n'a pas de solution. Cependant, il se pourrait qu'il existe $\vec{x}$ tel que $A\vec{x} \simeq \vec{b}$ et dans ce cas, nous aimerions trouver un N-vecteur $\vec{x}$ qui rende minimum $\| A\vec{x} - \vec{b} \|$. Le problème peut donc se formuler ainsi :

Trouver un N-vecteur $\vec{x}$ tel que pour tout N-vecteur $\vec{y}$ on ait :

$$\| A\vec{x} - \vec{b} \| \leq \| A\vec{y} - \vec{b} \| . \tag{4.40}$$

On dira dans ce cas que l'on cherche une solution de $A\vec{x} = \vec{b}$ ***au sens des moindres carrés***.

Nous allons démontrer le résultat suivant :

Théorème 4.5 *Supposons que A soit une $M \times N$ matrice ($M \geq N$) de rang N. Alors il existe un et un seul vecteur $\vec{x}$ satisfaisant (4.40). De plus, ce vecteur $\vec{x}$ est donné comme la solution du système $A^T A\vec{x} = A^T\vec{b}$.*

Démonstration
Remarquons tout d'abord que la matrice B définie par $B = A^T A$ est une $N \times N$ matrice symétrique. Si, pour un N-vecteur $\vec{z}$ quelconque, nous calculons $\vec{z}^T B\vec{z}$, nous aurons

$$\vec{z}^T B\vec{z} = \vec{z}^T A^T A\vec{z} = \| A\vec{z} \|^2 . \tag{4.41}$$

De plus, la relation $\vec{z}^T B\vec{z} = 0$ implique $A\vec{z} = 0$ et donc, puisque A est supposée de rang N, $\vec{z} = 0$. Cette dernière affirmation avec (4.41) garantissent que $A^T A$ est régulière (elle est même symétrique définie positive). Ainsi il existe un et un seul N-vecteur $\vec{x}$ tel que

$$A^T A\vec{x} = A^T\vec{b}. \tag{4.42}$$

Montrons que, lorsque $\vec{x}$ est solution de (4.42), alors pour tout N-vecteur $\vec{y}$ différent de $\vec{x}$ on a :

$$\| A\vec{x} - \vec{b} \| < \| A\vec{y} - \vec{b} \| . \tag{4.43}$$

En posant $\vec{z} = \vec{x} - \vec{y}$, on a $\vec{z} \neq 0$ puisque $\vec{y} \neq \vec{x}$ et de plus

$$\begin{aligned}
\| A\vec{y} - \vec{b} \|^2 &= \| (A\vec{x} - \vec{b}) - A\vec{z} \|^2 \\
&= \Big((A\vec{x} - \vec{b}) - A\vec{z}\Big)^T \Big((A\vec{x} - \vec{b}) - A\vec{z}\Big) \\
&= \| A\vec{x} - \vec{b} \|^2 - \vec{z}^T A^T (A\vec{x} - \vec{b}) - (A\vec{x} - \vec{b})^T A\vec{z} + \vec{z}^T A^T A\vec{z} \\
&= \| A\vec{x} - \vec{b} \|^2 - 2\vec{z}^T A^T (A\vec{x} - \vec{b}) + \vec{z}^T A^T A\vec{z} \\
&= \| A\vec{x} - \vec{b} \|^2 - 2\vec{z}^T (A^T A\vec{x} - A^T\vec{b}) + \| A\vec{z} \|^2 .
\end{aligned}$$

Comme $\vec{x}$ est solution de (4.42), on obtient :

$$\| A\vec{y} - \vec{b} \|^2 = \| A\vec{x} - \vec{b} \|^2 + \| A\vec{z} \|^2 .$$

Puisque A est de rang N, on a nécessairement $A\vec{z} \neq 0$ lorsque $\vec{z} \neq 0$. Ainsi la relation (4.43) est bien vérifiée et, a fortiori, la relation (4.40) est satisfaite.

Si nous supposons maintenant que $\vec{p}$ est un N-vecteur tel que $\|A\vec{p} - \vec{b}\| \leq \|A\vec{y} - \vec{b}\|$ pour tout N-vecteur $\vec{y}$, alors nécessairement nous avons $\vec{p} = \vec{x}$. En effet, si $\vec{p}$ était différent de $\vec{x}$ nous aurions, en utilisant (4.43), $\|A\vec{x} - \vec{b}\| < \|A\vec{p} - \vec{b}\|$, ce qui contredit ce que nous avons supposé sur $\vec{p}$. Ce raisonnement montre l'unicité affirmée dans l'énoncé du théorème. ■

Nous avons vu que si A est de rang N, alors $A^T A$ est une matrice symétrique, définie positive. Ainsi, si nous voulons résoudre $A\vec{x} = \vec{b}$ au sens des moindres carrés, il suffira de résoudre le système linéaire de N équations et à N inconnues suivant :

$$A^T A\vec{x} = A^T \vec{b}.$$

Nous présentons dans les chapitres 5 et 6 des méthodes adaptées à la résolution de systèmes linéaires dont la matrice est symétrique définie positive.

Remarquons que si A est une $M \times N$ matrice de rang N avec $M > N$, et si $\vec{b}$ est un M-vecteur donné, nous pouvons définir, pour tout N-vecteur $\vec{y}$ quelconque, la quantité

$$\vec{r}(\vec{y}) = A\vec{y} - \vec{b}.$$

Le vecteur $\vec{r}$ est appelé ***résidu***. Si $\vec{x}$ est solution de $A\vec{x} = \vec{b}$ au sens des moindres carrés, alors on a

$$\| \vec{r}(\vec{x}) \| \leq \| \vec{r}(\vec{y}) \|, \qquad \forall \vec{y} \in \mathbb{R}^N, \tag{4.44}$$

et, d'après le théorème 4.5, on obtient $A^T A\vec{x} = A^T \vec{b}$.

On aurait pu donner des poids différents aux équations du système surdéterminé en choisissant des nombres positifs $p_1, p_2, \ldots, p_M$ et en minimisant la quantité

$$\sum_{i=1}^{M} p_i r_i^2(\vec{y}) \qquad \text{au lieu de} \qquad \sum_{i=1}^{M} r_i^2(\vec{y}), \tag{4.45}$$

comme nous l'avons fait dans (4.44). Nous dirons dans ce cas que nous cherchons une solution de $A\vec{x} = \vec{b}$ ***au sens des moindres carrés avec poids*** $p_1, p_2, \ldots, p_M$. Si D est la $M \times M$ matrice diagonale donnée par

$$D = \text{diag}\,(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_M}),$$

il est facile de voir que

$$\sum_{i=1}^{M} p_i r_i^2(\vec{y}) = \vec{r}^T(\vec{y}) D^2 \vec{r}(\vec{y}) = \| D(A\vec{y} - \vec{b}) \|^2 . \tag{4.46}$$

Ainsi, on cherche un N-vecteur $\vec{x}$ tel que $\| DA\vec{x} - D\vec{b} \|$ soit minimal, ce qui revient à remplacer dans ce qui précède A par DA et $\vec{b}$ par $D\vec{b}$. On sera donc conduit à résoudre

$$A^T D^2 A\vec{x} = A^T D^2 \vec{b} \tag{4.47}$$

pour résoudre $A\vec{x} = \vec{b}$ au sens des moindres carrés avec poids $p_1, p_2, \ldots, p_M$.

4.8 Exercices

Exercice 4.1 Le but de cet exercice est d'exhiber un exemple pour lequel la majoration (4.37) est atteinte.

Soit A une $N \times N$ matrice symétrique régulière dont les valeurs propres $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N$ (répétées selon leur multiplicité) sont numérotées de sorte à ce que l'on ait :

$$| \lambda_1 | \geq | \lambda_2 | \geq | \lambda_3 | \geq \cdots \geq | \lambda_N | .$$

Si $\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N$ est une base de vecteurs propres correspondant à ces valeurs propres, on aura

$$A\vec{\varphi}_j = \lambda_j \vec{\varphi}_j, \qquad 1 \leq j \leq N. \tag{4.48}$$

Soit ε un réel quelconque. Soit $\vec{b}$ et $\delta\vec{b}$ les N-vecteurs définis par $\vec{b} = \vec{\varphi}_1$ et $\delta\vec{b} = \varepsilon\vec{\varphi}_N$. Soit $\vec{x}$ et $\delta\vec{x}$ les N-vecteurs tels que $A\vec{x} = \vec{b}$ et $A\delta\vec{x} = \delta\vec{b}$. Montrer que

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} = \chi(A)\frac{\|\delta\vec{b}\|}{\|\vec{b}\|} \tag{4.49}$$

où $\chi(A)$ est le nombre de condition spectral de A.

Solution
Par définition, les vecteurs $\vec{x}$ et $\delta\vec{x}$ sont solution de $A\vec{x} = \vec{\varphi}_1$ et $A\delta\vec{x} = \varepsilon\vec{\varphi}_N$. Puisque A est régulière toutes ses valeurs propres sont non nulles. En utilisant (4.48), nous avons donc

$$\vec{x} = \frac{1}{\lambda_1}\vec{\varphi}_1 \qquad \text{et} \qquad \delta\vec{x} = \frac{\varepsilon}{\lambda_N}\vec{\varphi}_N,$$

et, par conséquent

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} = \frac{|\lambda_1|}{|\lambda_N|}\frac{\|\varepsilon\vec{\varphi}_N\|}{\|\vec{\varphi}_1\|} = \frac{|\lambda_1|}{|\lambda_N|}\frac{\|\delta\vec{b}\|}{\|\vec{b}\|}. \tag{4.50}$$

D'autre part, puisque A est symétrique régulière, nous pouvons utiliser le résultat du théorème 4.3, c'est-à-dire

$$\chi(A) = \frac{|\lambda_1|}{|\lambda_N|}. \tag{4.51}$$

Les égalités (4.50) et (4.51) permettent de conclure.

Exercice 4.2 L'évolution d'une grandeur physique (par exemple la masse d'un isotope radioactif) est gouvernée par la relation

$$y(t) = \alpha e^{-\beta t}, \tag{4.52}$$

où α et β sont deux constantes positives à déterminer. Des mesures fournissent, pour certains temps $t = t_1, t_2, \ldots, t_N$, des valeurs $y(t) = y_1, y_2, \ldots, y_N$. En prenant le logarithme de (4.52), nous avons $\ln y(t) = \ln \alpha - \beta t$. En posant encore $\tilde{\alpha} = \ln \alpha$, nous cherchons donc à déterminer $\tilde{\alpha}$ et β tels que

$$\ln y_i = \tilde{\alpha} - \beta t_i, \qquad 1 \leq i \leq N. \tag{4.53}$$

Les mesures n'étant pas infiniment précises nous demandons à déterminer $\tilde{\alpha}$ et β pour qu'ils soient solution de (4.53) au sens des moindres carrés.

Solution
L'égalité (4.53) s'écrit sous forme matricielle $A\vec{x} = \vec{b}$ avec

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix} \qquad \vec{x} = \begin{bmatrix} \tilde{\alpha} \\ -\beta \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_N \end{bmatrix}.$$

D'après le théorème 4.5, il suffit de résoudre $A^T A\vec{x} = A^T\vec{b}$ où

$$A^T A = \begin{bmatrix} N & \sum_{i=1}^{N} t_i \\ \sum_{i=1}^{N} t_i & \sum_{i=1}^{N} t_i^2 \end{bmatrix} \qquad A^T\vec{b} = \begin{bmatrix} \sum_{i=1}^{N} \ln y_i \\ \sum_{i=1}^{N} t_i \ln y_i \end{bmatrix}$$

pour obtenir $\tilde{\alpha}$ et $-\beta$.

4.9 Notes bibliographiques et remarques

Il est possible d'estimer le nombre de condition spectral d'une matrice sans calculer explicitement la plus grande et la plus petite valeur propre de $A^T A$ à condition de résoudre quelques système linéaires, voir par exemple [11], [19] ou [2].

Dans certains cas particuliers il est possible de calculer explicitement le nombre de condition spectral à partir du calcul des valeurs propres. Considérons par exemple la $N \times N$ matrice tridiagonale A définie par

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & 0 & & -1 & 2 \end{bmatrix}.$$

Si $h = \pi/(N+1)$, nous pouvons vérifier en utilisant la formule trigonométrique

$$\sin(\alpha - \beta) + \sin(\alpha + \beta) = 2 \sin \alpha \cos \beta,$$

que les valeurs propres de A sont données par

$$\lambda_k = 2 - 2\cos kh, \qquad 1 \leq k \leq N, \tag{4.54}$$

et les vecteurs propres correspondants par

$$\vec{\varphi}_k = (\sin kh, \sin 2kh, \ldots, \sin Nkh)^T, \qquad 1 \leq k \leq N.$$

Puisque la matrice A est symétrique et que ses valeurs propres sont positives, elle est définie positive et son nombre de condition spectral $\chi(A)$ est donc donné par

$$\chi(A) = \frac{1 - \cos Nh}{1 - \cos h} = \frac{1 + \cos h}{1 - \cos h}.$$

Lorsque N est grand, $\cos h$ se comporte comme $1 - h^2/2$ et donc $\chi(A) \simeq 4/h^2 = O(N^2)$. Il convient donc de remarquer que, plus l'ordre N de la matrice A est grand, et plus le système linéaire $A\vec{x} = \vec{b}$ est mal conditionné.

Dans la section 4.7, nous avons minimisé la quantité $\|\vec{r}(\vec{x})\|$ où $\vec{r}(\vec{x})$ est le résidu défini par $\vec{r}(\vec{x}) = A\vec{x} - \vec{b}$ et où $\|\cdot\|$ est la norme euclidienne. Nous aurions pu minimiser le résidu dans une autre norme comme, par exemple, la norme $\|\cdot\|_\infty$ définie par $\|\vec{y}\|_\infty = \max_{1 \leq i \leq M} |y_i|$. Ainsi nous aurions obtenu le problème qui consiste à chercher un N-vecteur $\vec{x}$ tel que

$$\|A\vec{x} - \vec{b}\|_\infty \leq \|A\vec{y} - \vec{b}\|_\infty,$$

pour tout N-vecteur $\vec{y}$. Ce problème est un problème de programmation linéaire qui peut être résolu au moyen de l'algorithme du simplexe, voir par exemple [7].

Chapitre 5

Décomposition LU. Décomposition de Cholesky

5.1 Décomposition LU

Soit A une $N \times N$ matrice dont toutes les sous-matrices principales sont régulières (définition 4.3). Dans le chapitre 4, nous avons vu qu'il est possible de résoudre un système linéaire du type $A\vec{x} = \vec{b}$ en utilisant l'algorithme d'élimination de Gauss sans choix du pivot (tab. 4.1). Dans ce chapitre nous présentons une autre méthode basée sur le résultat suivant :

Théorème 5.1 *Si A est une $N \times N$ matrice dont toutes les sous-matrices principales sont régulières, alors il existe une décomposition unique*

$$A = LU \tag{5.1}$$

où L est une matrice triangulaire inférieure et U est une matrice triangulaire supérieure avec des valeurs 1 dans sa diagonale [1]*. De plus, la matrice U est celle obtenue par l'algorithme d'élimination de Gauss (tab. 4.1).*

Démonstration
Soit $A^{(i)}$ la matrice obtenue avant la i-ième étape de l'élimination de Gauss (sect. 4.3). La matrice $A^{(i)}$ est définie par l'expression (4.10). Soit maintenant la matrice $S^{(i)}$ définie par :

[1] en anglais L correspond à Lower matrix et U à Upper matrix

$$S^{(i)} = \begin{bmatrix} 1 & 0 & 0 & & & 0 & & & \\ 0 & 1 & 0 & & & & & 0 & \\ 0 & 0 & 1 & & & & & & \\ & & & \ddots & & \vdots & & & \\ & & & & 1 & 0 & & & \\ & & & & & s_i^{(i)} & & & \\ & & & & & s_{i+1}^{(i)} & 1 & & \\ & & 0 & & & s_{i+2}^{(i)} & 0 & 1 & \\ & & & & & \vdots & & & \ddots \\ & & & & & s_N^{(i)} & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}$$

i-ième colonne

i-ième ligne

où $s_i^{(i)} = 1/a_{ii}^{(i)}$ et $s_k^{(i)} = -a_{ki}^{(i)}/a_{ii}^{(i)}$, $k = i+1, i+2, \ldots, N$. Remarquons que cette définition a un sens puisque les valeurs $a_{ii}^{(i)}$ sont non nulles en vertu du théorème 4.1. En faisant le produit matriciel de $S^{(i)}$ par $A^{(i)}$, nous vérifions que

$$A^{(i+1)} = S^{(i)} \cdot A^{(i)}. \tag{5.2}$$

La matrice $A^{(N+1)}$ est égale à la matrice U triangulaire supérieure avec des 1 dans sa diagonale obtenue par l'algorithme d'élimination de Gauss. Ainsi, en utilisant (5.2), nous obtenons :

$$\begin{aligned} U = A^{(N+1)} &= S^{(N)}A^{(N)} = S^{(N)}S^{(N-1)}A^{(N-1)} = \ldots \\ &= S^{(N)}S^{(N-1)} \ldots S^{(1)}A^{(1)}. \end{aligned} \tag{5.3}$$

Puisque $A^{(1)} = A$ et puisque les matrices $S^{(j)}$ sont régulières, nous obtenons en posant $L^{(j)} = [S^{(j)}]^{-1}$:

$$A = L^{(1)}L^{(2)}L^{(3)} \ldots L^{(N)}U. \tag{5.4}$$

Or l'inverse d'une matrice triangulaire inférieure régulière est une matrice triangulaire inférieure. De même le produit de deux matrices triangulaires inférieures est aussi une matrice triangulaire inférieure. Nous obtenons donc bien

$$A = LU,$$

où L est la matrice triangulaire inférieure donnée par $L^{(1)}L^{(2)}L^{(3)} \ldots L^{(N)}$.

L'unicité de la décomposition LU découle indirectement du thèorème 4.1. On peut néanmoins montrer l'unicité de façon directe en utilisant l'argument suivant. Supposons qu'il y ait deux décompositions LU de A, c'est-à-dire supposons que A puisse s'écrire sous la forme

$$A = L_1U_1 = L_2U_2,$$

où L_1, L_2 sont des matrices triangulaires inférieures et U_1, U_2 sont des matrices triangulaires supérieures avec des 1 sur la diagonale. Puisque A est régulière, alors L_1 et L_2 sont aussi régulières (U_1 et U_2 le sont toujours) et on a la relation

$$L_2^{-1}L_1 = U_2U_1^{-1}. \tag{5.5}$$

La matrice $L_2^{-1}L_1$ est triangulaire inférieure alors que la matrice $U_2U_1^{-1}$ est triangulaire supérieure avec des 1 sur la diagonale. Ainsi l'égalité (5.5) implique que les deux matrices $L_2^{-1}L_1$ et $U_2U_1^{-1}$ sont des matrices diagonales avec des 1 sur la diagonale. Les matrices $L_2^{-1}L_1$ et $U_2U_1^{-1}$ coïncident donc avec la matrice identité ce qui prouve que $L_2 = L_1$ et $U_2 = U_1$. ■

Nous allons maintenant présenter un algorithme permettant d'obtenir la décomposition LU de la matrice A.

Pour commencer, prenons l'exemple d'une matrice A de coefficients a_{ij}, $1 \leq i, j \leq 3$ composée de 3 lignes et 3 colonnes données. Nous devons déterminer dans ce cas les coefficients ℓ_{ij}, $1 \leq j \leq i \leq 3$ et u_{ij}, $1 \leq i < j \leq 3$ qui satisfont le produit matriciel suivant :

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}}_{U}. \tag{5.6}$$

En multipliant L par la première colonne de U, nous obtenons la première colonne de A :

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = \begin{bmatrix} \ell_{11} \\ \ell_{21} \\ \ell_{31} \end{bmatrix}. \tag{5.7}$$

En multipliant la première ligne de L par les deuxième et troisième colonnes de U, nous obtenons respectivement les deuxième et troisième termes de la première ligne de A :

$$a_{12} = \ell_{11}u_{12} \qquad \text{et} \qquad a_{13} = \ell_{11}u_{13},$$

c'est-à-dire, puisque ℓ_{11} est déjà connu :

$$u_{12} = a_{12}/\ell_{11}, \qquad u_{13} = a_{13}/\ell_{11}. \tag{5.8}$$

Les relations (5.7) et (5.8) donnent donc la première colonne de L et la première ligne de U. Celles-ci étant connues, nous pouvons déterminer la deuxième colonne de L en multipliant les deuxième et troisième lignes de L par la deuxième colonne de U (qui est connue) et nous obtenons :

$$\ell_{21}u_{12} + \ell_{22} = a_{22} \qquad \text{et} \qquad \ell_{31}u_{12} + \ell_{32} = a_{32}$$

soit encore :

$$\ell_{22} = a_{22} - \ell_{21}u_{12} \qquad \text{et} \qquad \ell_{32} = a_{32} - \ell_{31}u_{12}. \tag{5.9}$$

Pour déterminer la deuxième ligne de U, il suffit de multiplier la deuxième ligne de L par la troisième colonne de U et nous obtenons :

$$\ell_{21}u_{13} + \ell_{22}u_{23} = a_{23}$$

soit encore :

$$u_{23} = (a_{23} - \ell_{21}u_{13})/\ell_{22}. \tag{5.10}$$

Les relations (5.9) et (5.10) déterminent la deuxième colonne de L et la deuxième ligne de U. Enfin ℓ_{33} est donné par

$$\ell_{33} = a_{33} - \ell_{31}u_{13} - \ell_{32}u_{23}. \tag{5.11}$$

Considérons maintenant le cas où A est une $N \times N$ matrice. Supposons connues les $k-1$ premières colonnes de L et les $k-1$ premières lignes de U, $(2 \leq k < N)$. Il est alors possible de calculer la k-ième colonne de L et la k-ième ligne de U. En effet, pour déterminer la k-ième colonne de L, il suffit de multiplier la i-ième ligne de L $(k \leq i)$ avec la k-ième colonne de U pour obtenir :

$$a_{ik} = \ell_{i1}u_{1k} + \ell_{i2}u_{2k} + \cdots + \ell_{i,k-1}u_{k-1,k} + \ell_{ik}$$

soit encore :

$$\ell_{ik} = a_{ik} - \sum_{j=1}^{k-1} \ell_{ij}u_{jk}. \tag{5.12}$$

Pour déterminer la k-ième ligne de U, il suffit de multiplier la k-ième ligne de L par la i-ième colonne de U $(k+1 \leq i)$ pour obtenir :

$$a_{ki} = \ell_{k1}u_{1i} + \ell_{k2}u_{2i} + \cdots + \ell_{kk}u_{ki}$$

soit encore :

$$u_{ki} = \frac{1}{\ell_{kk}} \left(a_{ki} - \sum_{j=1}^{k-1} \ell_{kj}u_{ji} \right). \tag{5.13}$$

D'autre part, il est possible de stocker les deux matrices L et U dans le tableau de la matrice A de la façon suivante :

$$\begin{bmatrix} \ell_{11} & u_{12} & u_{13} & \cdots & u_{1N} \\ \ell_{21} & \ell_{22} & u_{23} & \cdots & u_{2N} \\ \ell_{31} & \ell_{32} & \ell_{33} & \cdots & u_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{N1} & \ell_{N2} & \ell_{N3} & \cdots & \ell_{NN} \end{bmatrix}. \tag{5.14}$$

Nous pouvons donc effectuer la décomposition LU de A en ne définissant qu'un seul $N \times N$ tableau qui, au début de l'algorithme, contient la matrice A et, à la sortie de l'algorithme, contient l'information LU sous la forme (5.14). L'algorithme de décomposition LU est présenté dans le tableau 5.1.

Comme dans le cas de l'algorithme d'élimination de Gauss (tab. 4.1 et sect. 4.4), le nombre d'opérations de l'algorithme de décomposition LU se comporte comme N^3 lorsque N est grand.

5.2 Utilité de la décomposition LU

Considérons la situation où l'on doit résoudre m systèmes linéaires

$$A\,\vec{x}^{(\ell)} = \vec{b}^{(\ell)}, \qquad \ell = 1, 2, \ldots, m, \tag{5.15}$$

Tableau 5.1 Algorithme de décomposition LU.

entrées : a_{ij}, $1 \leq i, j \leq N$ sont les coefficients de la matrice A

sorties : a_{ij}, $1 \leq j \leq i \leq N$ sont les coefficients de la matrice L

et a_{ij}, $1 \leq i < j < N$ sont les coefficients surdiagonaux de la matrice U

Algorithme	**Commentaires**
Faire $i = 2$ à N	Construction de la première ligne de U
$a_{1i} := a_{1i}/a_{11}$	(la $1^{\text{ère}}$ col. de L est la $1^{\text{ère}}$ col. de A)
Faire $k = 2$ à $N-1$	Colonnes de L et lignes de U
$a_{kk} := a_{kk} - \sum_{j=1}^{k-1} a_{kj} * a_{jk}$	Construction du pivot ℓ_{kk}
Faire $i = k+1$ à N	Construction de la k-ième colonne de L
$a_{ik} := a_{ik} - \sum_{j=1}^{k-1} a_{ij} * a_{jk}$	
$a_{ki} := \frac{1}{a_{kk}}\left(a_{ki} - \sum_{j=1}^{k-1} a_{kj} * a_{ji}\right)$	Construction de la k-ième ligne de U
$a_{NN} := a_{NN} - \sum_{j=1}^{N-1} a_{Nj} * a_{jN}$	Construction du pivot ℓ_{NN}

où, pour $\ell = 2, 3, \ldots, m$, $\vec{b}^{(\ell)}$ dépend des solutions précédentes $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots$, $\vec{x}^{(\ell-1)}$. Par exemple si $m = 2$, supposons que l'on veuille résoudre

$$A\,\vec{x}^{(1)} = \vec{b}^{(1)},$$

puis

$$A\vec{x}^{(2)} = \|\vec{x}^{(1)}\|^2 \vec{x}^{(1)},$$

où $\vec{b}^{(1)}$ est un N-vecteur donné et $\vec{b}^{(2)} = \|\vec{x}^{(1)}\|^2 \vec{x}^{(1)}$. Dans ce cas, il serait coûteux et absurde de faire m fois l'élimination de Gauss car les opérations sur la matrice A seraient m fois les mêmes. Il suffit donc d'effectuer la décomposition LU de la matrice A et (5.15) devient

$$LU\vec{x}^{(\ell)} = \vec{b}^{(\ell)}, \qquad \ell = 1, 2, \ldots, m. \tag{5.16}$$

En posant

$$\vec{y}^{(\ell)} = U\vec{x}^{(\ell)},$$

il suffit de résoudre successivement pour $\ell = 1, 2, \ldots, m$:

$$L\vec{y}^{(\ell)} = \vec{b}^{(\ell)}, \tag{5.17}$$

puis

$$U\vec{x}^{(\ell)} = \vec{y}^{(\ell)}. \tag{5.18}$$

Le système (5.17) est un système triangulaire inférieur alors que (5.18) est un système triangulaire supérieur. Résoudre les deux systèmes linéaires (5.17) et (5.18) sont donc deux opérations peu coûteuses. Nous verrons dans les chapitres suivants des méthodes numériques nécessitant la résolution de systèmes linéaires du type (5.15).

Un autre exemple pour lequel la décomposition LU est utile est le suivant. Supposons que l'on ait effectué la décomposition LU d'une $N \times N$ matrice A. Alors nous avons $\det(A) = \det(L) \cdot \det(U)$. De plus, puisque U n'a que des 1 dans sa diagonale, nous avons $\det(U) = 1$. Le déterminant de L est donné par $\det(L) = \ell_{11}.\ell_{22}.\ell_{33} \ldots \ell_{NN}$ puisque L est triangulaire. Ainsi

$$\det(A) = \prod_{j=1}^{N} \ell_{jj}, \tag{5.19}$$

et on conclut que si on veut calculer $\det(A)$, il suffit d'utiliser (5.19) après avoir effectué une décomposition LU de A. Il est important de mentionner qu'il ne faut *jamais utiliser la méthode des mineurs de façon récursive* pour calculer un déterminant. En effet, cette méthode donne lieu à $N!$ (N factoriel) opérations. Il est amusant de constater que si $N = 100$, alors $N! = 100! \simeq 10^{158}$. Sur un ordinateur personnel ou une station de travail ayant un cycle d'horloge de 200 MHz, on peut exécuter approximativement 10^9 opérations par seconde en virgule flottante, ce qui conduit à un temps de 10^{149} secondes ou de 3 10^{141} années pour exécuter ces 100! opérations ; ce temps dépasse l'âge de l'univers. En effectuant la décomposition LU de A, puis en appliquant la formule (5.19), le calcul du déterminant d'une matrice A d'ordre 100 exigera sur un tel ordinateur un temps de calcul de l'ordre de la fraction de seconde.

5.3 Décomposition LU avec changement de pivot

Supposons que la matrice A soit régulière, mais qu'au cours de la décomposition LU, pour un k donné ($1 \leq k < N$), le pivot a_{kk} soit nul. Par conséquent, une division par zéro a lieu dans l'algorithme du tableau 5.1. Comme nous l'avons fait pour l'élimination de Gauss, il convient dans ce cas d'échanger la k-ième ligne de A avec une j-ième, $j > k$, afin d'obtenir un pivot non nul. Clairement si nous échangeons des lignes de A, nous n'obtiendrons pas une décomposition LU de A, mais une décomposition LU de la matrice A dans laquelle nous avons permuté des lignes. A ce stade, remarquons que les permutations des lignes de A correspondent aux permutations des lignes de L. Ainsi, avant de résoudre un système $A\vec{x} = \vec{b}$ avec cette décomposition LU avec changement de pivot, il conviendra d'exécuter en premier lieu les permutations correspondantes sur le second membre $\vec{b}$. Il est donc important qu'à chaque étape de la décomposition LU avec changement de pivot, on mémorise les permutations de lignes

que l'on a exécutées. Pour ce faire, nous définissons un N-vecteur $\vec{p}$ (dit de permutation) et nous posons

$$p_k = j$$

pour indiquer que la k-ième ligne a été échangée avec la j-ième.

5.4 Matrices symétriques définies positives. Décomposition de Cholesky

Rappelons que le symbole ^{T} qui accompagne une matrice ou un vecteur signifie que l'on a affaire à son transposé. Rappelons encore la définition suivante.

Définition 5.1 *Une $N \times N$ matrice A est dite symétrique définie positive si :*

i) $A = A^T$ *(A est symétrique),*

ii) $\vec{y}^T A\vec{y} \geq 0$ *pour tout N-vecteur $\vec{y}$,*

iii) $\vec{y}^T A\vec{y} = 0$ *si et seulement si* $\vec{y} = 0$.

Nous avons le résultat suivant.

Théorème 5.2 *Si A est une $N \times N$ matrice symétrique définie positive, alors toutes ses sous-matrices principales sont symétriques définies positives et sont donc régulières.*

Démonstration
Soit A une $N \times N$ matrice symétrique définie positive et soit A_k sa sous-matrice principale d'ordre k (définition 4.3). Il est évident que A_k est symétrique. Soit $\vec{z}$ un k-vecteur quelconque, calculons $\vec{z}^T A_k \vec{z}$. Si $\vec{y}$ est le N-vecteur construit de la façon suivante :

$$\vec{y} = \begin{bmatrix} \vec{z} \\ 0 \end{bmatrix} \begin{matrix} \} k \text{ premières composantes} \\ \} (N-k) \text{ composantes nulles,} \end{matrix}$$

nous vérifions aisément que

$$\vec{z}^T A_k \vec{z} = \vec{y}^T A\vec{y}. \tag{5.20}$$

Les propriétés (ii) et (iii) de la définition 5.1 et la relation (5.20), nous assurent que $\vec{z}^T A_k \vec{z} \geq 0$ et $\vec{z}^T A_k \vec{z} = 0$ si et seulement si $\vec{z} = 0$. Ainsi donc A_k est symétrique définie positive. ■

Du théorème 5.2, nous pouvons démontrer le théorème suivant.

Théorème 5.3 (Théorème de Cholesky) *Si A est une matrice symétrique définie positive, il existe une et une seule matrice triangulaire inférieure à valeurs diagonales positives notée L telle que $A = LL^T$.*

Démonstration
Supposons que A soit symétrique définie positive. Alors en utilisant les théorèmes 5.2 et 5.1, il existe une et une seule décomposition LU de A que nous noterons

$$A = \tilde{L}U \tag{5.21}$$

où U est une $N \times N$ matrice triangulaire supérieure avec des valeurs 1 dans sa diagonale et $\tilde{L}$ est une matrice triangulaire inférieure.

Si A_k, $\tilde{L}_k$ et U_k sont les sous-matrices principales d'ordre k de A, $\tilde{L}$ et U respectivement, il est facile de vérifier que $A_k = \tilde{L}_k U_k$. Soit $\tilde{\ell}_{kk}$, $1 \leq k \leq N$, les éléments diagonaux de $\tilde{L}$. Puisque, en vertu du théorème 5.2, A_k est symétrique définie positive, nous avons, compte tenu de (5.19) :

$$\det A_k = \tilde{\ell}_{11}.\tilde{\ell}_{22}.\tilde{\ell}_{33} \ldots \tilde{\ell}_{kk} > 0. \tag{5.22}$$

La relation (5.22) prise successivement pour $k = 1, 2, 3, \ldots, N$, donne :

$$\tilde{\ell}_{jj} > 0, \qquad j = 1, 2, \ldots, N. \tag{5.23}$$

Soit maintenant D la $N \times N$ matrice diagonale définie par

$$D \underset{def}{=} \text{diag } \left(\sqrt{\tilde{\ell}_{11}}, \sqrt{\tilde{\ell}_{22}}, \sqrt{\tilde{\ell}_{33}}, \ldots, \sqrt{\tilde{\ell}_{NN}} \right)$$

(qui a bien un sens puisque nous avons les relations (5.23)) et soit E la $N \times N$ matrice définie par

$$E = D^2 = \text{ diag } \left(\tilde{\ell}_{11}, \tilde{\ell}_{22}, \tilde{\ell}_{33}, \ldots, \tilde{\ell}_{NN} \right). \tag{5.24}$$

Il est facile de vérifier que la $N \times N$ matrice $\hat{L}$ définie par

$$\hat{L} = \tilde{L}E^{-1} \tag{5.25}$$

est une matrice triangulaire inférieure avec des valeurs 1 dans sa diagonale.

En remplaçant (5.25) dans (5.21) nous obtenons :

$$A = \hat{L}EU$$

et, puisque A est symétrique, nous avons aussi

$$A = A^T = U^T E \hat{L}^T. \tag{5.26}$$

Comme $\hat{L}^T$ est une matrice triangulaire supérieure avec des 1 dans sa diagonale, comme $U^T E$ est une matrice triangulaire inférieure et puisque la décomposition LU de A est unique (théorème 5.1), nous concluons des relations (5.21) et (5.26) que

$$\hat{L}^T = U.$$

Ainsi donc nous avons :

$$A = \hat{L}E\hat{L}^T = \hat{L}DD\hat{L}^T. \tag{5.27}$$

Il suffit d'introduire la $N \times N$ matrice triangulaire inférieure L définie par

$$L = \hat{L}D \tag{5.28}$$

pour obtenir la décomposition de Cholesky

$$A = LL^T. \tag{5.29}$$

Cette décomposition est unique. En effet, supposons qu'il existe une autre décomposition

$$A = MM^T$$

où M est triangulaire inférieure avec des valeurs positives dans sa diagonale. Nous avons alors

$$LL^T = MM^T$$

et par suite

$$L^T M^{-T} = L^{-1} M \tag{5.30}$$

où M^{-T} dénote l'inverse de la matrice transposée de M.

Puisque L^T et M^{-T} sont des matrices triangulaires supérieures, alors $L^T M^{-T}$ est une matrice triangulaire supérieure. Nous montrons de la même façon que $L^{-1}M$ est une matrice triangulaire inférieure. Ainsi, en utilisant (5.30), nous constatons que $L^{-1}M$ est une matrice diagonale.

De la relation (5.30) nous vérifions que

$$\frac{\ell_{jj}}{m_{jj}} = \frac{m_{jj}}{\ell_{jj}}, \qquad 1 \leq j \leq N,$$

et donc que cette matrice diagonale est la $N \times N$ matrice identité I (puisque ℓ_{jj} et m_{jj} sont positifs). Nous obtenons ainsi $L^{-1}M = I$ et donc $M = L$. ■

Remarque 5.1 Dans le théorème 5.3, nous avons noté L la matrice triangulaire inférieure qui satisfait $A = LL^T$. Evidemment, cette matrice L n'est en principe pas la même que celle obtenue lors de la décomposition LU (nous avons noté cette matrice $\tilde{L}$ dans la démonstration du théorème 5.3).

Nous sommes maintenant en mesure de décrire l'algorithme de Cholesky. Il suffit de reprendre l'algorithme du tableau 5.1, de constater que la k-ième colonne de L est divisée par $\sqrt{a_{kk}}$ et de tenir compte de la relation $a_{kj} = a_{jk}$ puisque A est symétrique. L'algorithme ainsi obtenu est présenté dans le tableau 5.2.

Remarque 5.2 Si, dans l'algorithme du tableau 5.2, les racines carrées sont effectuées sur des nombres négatifs, alors la matrice A n'est pas symétrique définie positive. Il convient donc d'introduire des tests dans cet algorithme.

En conclusion, lorsqu'on veut résoudre un système linéaire

$$A\vec{x} = \vec{b},$$

où la matrice A est symétrique définie positive, on fait une décomposition de Cholesky $A = LL^T$ et on résout successivement

$$L\vec{y} = \vec{b}, \qquad \text{puis} \qquad L^T\vec{x} = \vec{y}.$$

Tableau 5.2 Algorithme de Cholesky.

entrées : $(a_{ij})_{1\leq j\leq i\leq N}$ représentant la partie triangulaire inférieure de A ;
(A est symétrique définie positive)
sorties : $(a_{ij})_{1\leq j\leq i\leq N}$ représentant L qui satisfait $A = LL^T$

Algorithme	**Commentaires**
$a_{11} := \sqrt{a_{11}}$	Construction de ℓ_{11}
Faire $i = 2$ à N	Construction de la première colonne de L
$a_{i1} := a_{i1}/a_{11}$	
Faire $k = 2$ à $N-1$	Parcours des colonnes de L
$a_{kk} := \Big(a_{kk} - \sum_{j=1}^{k-1} a_{kj}^2\Big)^{1/2}$	Construction de ℓ_{kk}
Faire $i = k+1$ à N	Construction de la k-ième col. de L
$a_{ik} := \frac{1}{a_{kk}}\Big(a_{ik} - \sum_{j=1}^{k-1} a_{ij} * a_{kj}\Big)$	
$a_{NN} := \Big(a_{NN} - \sum_{j=1}^{N-1} a_{Nj}^2\Big)^{1/2}$	Construction de ℓ_{NN}

5.5 Matrices de bande

Commençons par une définition.

Définition 5.2 *Soit A une $N \times N$ matrice de coefficients a_{ij}, $1 \leq i, j \leq N$ et soit ℓ un entier positif inférieur à N. On dira que A est une matrice de bande de demi-largeur ℓ si on a $a_{ij} = 0$ pour tout i, j satisfaisant $1 \leq i, j \leq N$ et $\mid i - j \mid \geq \ell$.*

Une matrice de bande de demi-largeur ℓ est illustrée dans la figure 5.1. Une matrice de bande de demi-largeur $\ell = 1$ est une matrice ***diagonale***. Une matrice de bande de demi-largeur $\ell = 2$ est une matrice ***tridiagonale***.

Nous vérifions facilement, en reprenant les algorithmes LU (relations (5.12) (5.13)) et LL^T, le résultat suivant :

Théorème 5.4 *Soit A une $N \times N$ matrice de bande de demi-largeur ℓ dont toutes les sous-matrices principales sont régulières (ou symétrique définie positive). Alors la décomposition LU de A (ou la décomposition LL^T si A est symétrique définie positive) donne lieu à des matrices triangulaires qui sont aussi de bande de demi-largeur ℓ. Le nombre d'opérations pour faire la décomposition LU ou LL^T est de l'ordre de $N\ell^2$ lorsque N est grand.*

Clairement, les décompositions LU et LL^T d'une matrice de bande ne requièrent pas la mémorisation des éléments nuls qui sont en dehors de la bande. Le stockage de la matrice A peut se faire, par exemple, diagonale par diagonale comme le montre l'exemple 5.1.

$$A = \begin{bmatrix} * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & 0 & 0 \\ 0 & * & * & * & * & * & * & * & * & * & 0 \\ 0 & 0 & * & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * \end{bmatrix}$$

ℓ lignes

ℓ colonnes

Fig. 5.1 Matrice de bande de demi-largeur ℓ. Ici $*$ mentionne la place d'un élément non nécessairement nul et 0 mentionne la place d'un élément nécessairement nul.

Exemple 5.1 Soit la $N \times N$ matrice tridiagonale A définie par

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Nous avons vu dans la section 4.9 que la matrice A est symétrique définie positive et donc il est possible de faire une décomposition de Cholesky $A = LL^T$. D'après le théorème 5.4, nous pouvons utiliser deux vecteurs $\vec{d}$ et $\vec{e}$ de composantes d_j,

$1 \leq j \leq N$ et e_j, $1 \leq j \leq N-1$ pour mémoriser L de la façon suivante :

$$L = \begin{bmatrix} d_1 & & & & \\ e_1 & d_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & d_{N-1} & \\ & & & e_{N-1} & d_N \end{bmatrix}.$$

En faisant le produit LL^T et en l'égalant à A, nous obtenons successivement :

$$\begin{array}{c} d_1^2 = 2; \quad e_1 d_1 = -1; \quad e_1^2 + d_2^2 = 2; \quad e_2 d_2 = -1; \\ e_2^2 + d_3^2 = 2; \quad e_3 d_3 = -1; \quad \cdots; \quad e_{N-1}^2 + d_N^2 = 2, \end{array}$$

soit encore $d_1 = \sqrt{2}$, et pour $j = 1, 2, \ldots, N-1$:

$$e_j = -1/d_j \qquad \text{et} \qquad d_{j+1} = \sqrt{2 - e_j^2}.$$

Dans ce cas l'algorithme de décomposition LL^T de A se met sous la forme suivante :

$$\begin{array}{l} d_1 = \sqrt{2} \\ \left[\begin{array}{l} \text{Faire } j = 1 \text{ à } N-1 \\ e_j := -1/d_j \\ d_{j+1} := \sqrt{2 - e_j^2}. \end{array} \right. \end{array}$$

Remarquons que, dans le cadre de cet exemple simple, il est possible de donner explicitement les valeurs de e_j et d_j, soit :

$$d_j = \sqrt{\frac{j+1}{j}} \qquad \text{et} \qquad e_j = -\sqrt{\frac{j}{j+1}}.$$

5.6 Exercices

Exercice 5.1 Soit A la $N \times N$ matrice tridiagonale de l'exemple 5.1 et soit B une $N \times N$ matrice de coefficients $(b_{ij})_{1 \leq i,j \leq N}$ donnés. Ecrire un algorithme pour calculer $A^{-1}B$ en n'utilisant qu'un seul tableau à N^2 éléments, à savoir le tableau $(b_{ij})_{1 \leq i,j \leq N}$.

Les entrées de l'algorithme seront les coefficients $(b_{ij})_{1 \leq i,j \leq N}$ de la matrice B. Les sorties seront les coefficients $(b_{ij})_{1 \leq i,j \leq N}$ de la matrice $A^{-1}B$.

Solution
Soit C la $N \times N$ matrice définie par $C = A^{-1}B$. Déterminer la matrice C revient à résoudre $AC = B$. Pour $1 \leq j \leq N$, notons $\vec{B}_j$ et $\vec{C}_j$ les j-ièmes colonnes des

matrices B et C, respectivement. Par définition du produit matriciel, résoudre $AC = B$ revient à résoudre $A\vec{C}_j = \vec{B}_j$, $1 \leq j \leq N$. Dans notre algorithme, ces N systèmes linéaires sont résolus simultanément. Nous utilisons la décomposition de Cholesky $A = LL^T$ de la matrice A, voir exemple 5.1. Ensuite nous résolvons N systèmes linéaires triangulaires inférieurs $L\vec{X}_j = \vec{B}_j$, $1 \leq j \leq N$, puis N systèmes linéaires triangulaires supérieurs $L^T\vec{C}_j = \vec{X}_j$, $1 \leq j \leq N$. Nous stockons les matrices B, X et C dans le même tableau $(b_{ij})_{1 \leq i,j \leq N}$. L'algorithme est décrit dans le tableau 5.3.

Tableau 5.3 Calcul de $A^{-1}B$ dans le cadre de l'exercice 5.1.

entrées : b_{ij}, $1 \leq i, j \leq N$ sont les coefficients de la matrice B ;	
sorties : b_{ij}, $1 \leq i, j \leq N$ sont les coefficients de la matrice C ;	
d_i, $1 \leq i < N$ sont les coefficients diagonaux de la matrice L ;	
e_i, $1 \leq i < N-1$ sont les coefficients sous-diagonaux de la matrice L.	
Algorithme	**Commentaires**
$d_1 = \sqrt{2}$ Faire $i = 1$ à $N-1$ $e_i := -1/d_i$ $d_{i+1} := \sqrt{2 - e_i^2}$	Calcul de la matrice L
Faire $j = 1$ à N $b_{1j} := b_{1j}/d_1$ Faire $i = 2$ à N Faire $j = 1$ à N $b_{ij} := (b_{ij} - e_{i-1} * b_{i-1,j})/d_i$	Résolution de $LX = B$
Faire $j = 1$ à N $b_{Nj} := b_{Nj}/d_N$ Faire $i = N-1$ à 1 pas de -1 Faire $j = 1$ à N $b_{ij} := (b_{ij} - e_i * b_{i+1,j})/d_i$	Résolution de $L^TC = X$

Exercice 5.2 Soit $\vec{a}$ un N-vecteur de composantes a_i, $i = 1, \ldots, N$, supposées toutes non nulles, soit $\vec{b}$ et $\vec{c}$ deux $(N-1)$-vecteurs de composantes b_i et c_i,

$i = 1, \ldots, N-1$. Considérons la $N \times N$ matrice A définie par

$$A = \begin{bmatrix} a_1 & & & & b_1 \\ & a_2 & & 0 & b_2 \\ & & \ddots & & \vdots \\ & 0 & & a_{N-1} & b_{N-1} \\ c_1 & c_2 & \ldots & c_{N-1} & a_N \end{bmatrix}.$$

On dit que la matrice A a une structure ***flèche***. Considérons deux matrices L et U de la forme

$$L = \begin{bmatrix} d_1 & & & & \\ & d_2 & & 0 & \\ & & \ddots & & \\ & 0 & & d_{N-1} & \\ f_1 & f_2 & \ldots & f_{N-1} & d_N \end{bmatrix}, \quad U = \begin{bmatrix} 1 & & & & e_1 \\ & 1 & & 0 & e_2 \\ & & \ddots & & \vdots \\ & 0 & & 1 & e_{N-1} \\ & & & & 1 \end{bmatrix}. \tag{5.31}$$

1. Si on suppose que A est une matrice régulière, démontrer qu'il existe une unique décomposition de A de la forme $A = LU$, où L et U ont l'allure donnée en (5.31).
2. Expliciter l'algorithme de décomposition $A = LU$.

Solution

1. Puisque les nombres a_i sont supposés non nuls, alors les sous-matrices principales A_k d'ordre k de A sont régulières pour $k = 1, 2, \ldots, N-1$. Puisque A est supposée régulière, alors toutes les sous-matrices principales de A sont régulières. Ainsi, en utilisant le théorème 5.1, la décomposition LU de A existe et est unique. En considérant les relations (5.12) et (5.13), il est facile de se convaincre que les matrices L et U prennent la forme (5.31).

2. Identifions la première colonne de A avec la première colonne de LU. Nous obtenons $d_1 = a_1$ et $f_1 = c_1$. Identifions maintenant la première ligne de A avec la première ligne de LU. Nous obtenons $b_1 = d_1 e_1$. En identifiant les $N-1$ premières lignes et colonnes de A et LU nous obtenons donc

$$d_i = a_i, \qquad f_i = c_i, \qquad e_i = \frac{b_i}{d_i}, \qquad i = 1, 2, \ldots, N-1.$$

Identifions finalement le dernier coefficient diagonal de A et LU. Nous obtenons

$$a_N = f_1 e_1 + \cdots + f_{N-1} e_{N-1} + d_N,$$

soit encore

$$d_N = a_N - \sum_{j=1}^{N-1} f_j e_j.$$

L'algorithme correspondant est présenté dans le tableau 5.4.

Tableau 5.4 Algorithme de décomposition LU dans le cadre de l'exercice 5.2.

<table>
<tr><td colspan="2">
entrées : a_i, $1 \leq i \leq N$, b_i et c_i, $1 \leq i \leq N-1$
sont les coefficients de la matrice A ;
sorties : a_i, $1 \leq i \leq N$, c_i, $1 \leq i \leq N-1$
sont les coefficients de la matrice L
($\vec{a}$ représente $\vec{d}$ et $\vec{c}$ représente $\vec{f}$) ;
b_i, $1 \leq i \leq N-1$ sont les coefficients de la matrice U
($\vec{b}$ représente $\vec{e}$).
</td></tr>
<tr><th>Algorithme</th><th>Commentaires</th></tr>
<tr><td>
$\left[\begin{array}{l}\text{Faire } i = 1 \text{ à } N-1 \\ b_i := b_i / a_i\end{array}\right.$
$a_N := a_N - \sum_{j=1}^{N-1} b_j * c_j$
</td><td>
Les $(N-1)$ premières colonnes de L sont celles de A. On ne calcule que la dernière colonne de U et le dernier élément diagonal de L.
</td></tr>
</table>

5.7 Notes bibliographiques et remarques

Dans la décomposition LU, nous avons prescrit les valeurs diagonales de U égales à 1. Ce choix est évidemment arbitraire. Dans la littérature, ce sont souvent les valeurs diagonales de L qui sont prescrites comme étant égales à 1, voir par exemple [11]. Cet autre choix ne modifie pas fondamentalement les principes et algorithmes que nous avons vus dans ce chapitre.

Un grand nombre de bibliothèques numériques sont à disposition pour exécuter les décompositions que nous venons de voir. Citons par exemple LAPACK [2]. Cette bibliothèque est accessible depuis certains logiciels grand public, par exemple Matlab^{TM} ou NAG^{TM}.

Pour une introduction à la résolution de systèmes linéaires dans un environnement de calcul parallèle, nous renvoyons le lecteur à la littérature référencée dans [11].

Chapitre 6

Résolution de systèmes linéaires par des méthodes itératives

6.1 Généralités. Méthodes de Jacobi et de Gauss-Seidel

Dans ce chapitre nous considérons, comme dans les chapitres 4 et 5, un système d'équations linéaires d'ordre N de la forme

$$A\vec{x} = \vec{b}, \tag{6.1}$$

où A est une $N \times N$ matrice régulière de coefficients $(a_{ij})_{1 \leq i,j \leq N}$ donnés et $\vec{b}$ est un N-vecteur de composantes $(b_i)_{1 \leq i \leq N}$ données également. Nous avons vu que pour résoudre numériquement le système (6.1), c'est-à-dire pour trouver le N-vecteur $\vec{x}$ dont les composantes $(x_i)_{1 \leq i \leq N}$ sont inconnues, il suffit de procéder à une élimination de Gauss avec ou sans changement de pivot (chap. 4), ou éventuellement, de procéder à une décomposition LU de A (ou LL^T si A est symétrique définie positive), puis résoudre deux systèmes triangulaires (chap. 5). Dans tous les cas, si la matrice A est pleine, le nombre d'opérations nécessaires à la mise en œuvre de ces algorithmes est de l'ordre de N^3, ce qui peut être énorme lorsque N est grand. Dans ce chapitre nous allons étudier d'autres méthodes dites ***itératives*** pour résoudre (6.1). Elles consisteront à construire une suite de N-vecteurs $\vec{x}^0, \vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n, \ldots$, telle que

$$\lim_{n \to \infty} \|\vec{x} - \vec{x}^n\| = 0. \tag{6.2}$$

Remarquons au passage que les indices j de $\vec{x}^j$ sont placés en haut du symbole $\vec{x}$ pour ne pas confondre $\vec{x}^j$ avec sa i-ième composante x_i^j.

Les ***méthodes de décomposition*** sont un premier exemple de méthodes itératives. Supposons que l'on ait décomposé A en une différence de deux $N \times N$

matrices K et M, c'est-à-dire que l'on ait

$$A = K - M. \tag{6.3}$$

Si K est une matrice régulière, nous pouvons transformer (6.1) en utilisant (6.3) de la façon suivante :

$$K\vec{x} = M\vec{x} + \vec{b} \tag{6.4}$$

soit encore

$$\vec{x} = K^{-1}M\vec{x} + K^{-1}\vec{b}. \tag{6.5}$$

L'égalité (6.5) nous suggère la méthode itérative suivante :

- on se donne un N-vecteur $\vec{x}^0$ quelconque ;
- pour $n = 0, 1, 2, 3, \ldots$, on calcule

$$\vec{x}^{n+1} = K^{-1}M\vec{x}^n + K^{-1}\vec{b}. \tag{6.6}$$

Dans la pratique, lorsque nous voulons calculer $\vec{x}^{n+1}$ à partir de $\vec{x}^n$, nous commençons par calculer le vecteur $\vec{c} = M\vec{x}^n + \vec{b}$, puis nous résolvons le système $K\vec{x}^{n+1} = \vec{c}$. De façon évidente, ce dernier système doit être plus facile à résoudre que le système (6.1) pour que nous gagnions quelque chose. Nous ferons donc en sorte que la décomposition (6.3) donne lieu à une matrice K telle que le système $K\vec{x}^{n+1} = \vec{c}$ soit facile à résoudre, ce qui est par exemple le cas si K est diagonale ou triangulaire. Pour ce faire écrivons la matrice A sous la forme

$$A = D - E - F \tag{6.7}$$

où D est la matrice diagonale formée des éléments $(a_{ii})_{1\leq i\leq N}$, $-E$ est la matrice formée des éléments sous-diagonaux de A, c'est-à-dire $e_{ij} = -a_{ij}$ si $1 \leq j < i \leq N$, $e_{ij} = 0$ si $1 \leq i \leq j \leq N$ et enfin $-F$ est la matrice formée des éléments surdiagonaux de A, c'est-à-dire $f_{ij} = -a_{ij}$ si $1 \leq i < j \leq N$, $f_{ij} = 0$ si $1 \leq j \leq i \leq N$. Schématiquement nous avons donc

$$A = \begin{bmatrix} \ddots & & & & \\ & \ddots & & -F & \\ & & D & & \\ & -E & & \ddots & \\ & & & & \ddots \end{bmatrix}.$$

Si nous supposons que $a_{ii} \neq 0$ pour tout $i = 1, 2, \ldots, N$, alors D et $(D - E)$ sont régulières et nous pouvons choisir par exemple $K = D$ ou $K = D - E$.

La *méthode de Jacobi* consiste à poser $K = D$ et $M = E + F$. On a bien $A = K - M$ et puisque $K^{-1} = D^{-1} = \text{ diag } (1/a_{11}, 1/a_{22}, 1/a_{33}, \ldots, 1/a_{NN})$, on peut écrire l'égalité (6.6) composante par composante de la manière suivante :

$$x_i^{n+1} = \frac{1}{a_{ii}} \left(-\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}x_j^n + b_i \right), \qquad 1 \leq i \leq N. \tag{6.8}$$

Les relations (6.8) permettent de calculer explicitement les composantes x_i^{n+1}, $1 \leq i \leq N$, du vecteur $\vec{x}^{n+1}$ à partir des composantes x_i^n, $1 \leq i \leq N$, du vecteur $\vec{x}^n$. La matrice J définie par

$$J = K^{-1}M = D^{-1}(E+F) \tag{6.9}$$

est appelée ***matrice de Jacobi***.

La *méthode de Gauss-Seidel* consiste à poser $K = D - E$ et $M = F$. On a bien $A = K - M$ et l'égalité (6.6) peut aussi s'écrire :

$$(D-E)\vec{x}^{n+1} = F\vec{x}^n + \vec{b} \tag{6.10}$$

qui est un système triangulaire inférieur pour l'inconnue $\vec{x}^{n+1}$. On vérifie sans difficultés que la relation (6.10) écrite composante par composante devient

$$x_i^{n+1} = \frac{1}{a_{ii}} \left(-\sum_{j<i} a_{ij} x_j^{n+1} - \sum_{j>i} a_{ij} x_j^n + b_i \right), \quad 1 \leq i \leq N. \tag{6.11}$$

Si les composantes $(x_i^n)_{1 \leq i \leq N}$ du vecteur $\vec{x}^n$ sont connues, les relations (6.11) permettent de calculer successivement $x_1^{n+1}, x_2^{n+1}, x_3^{n+1}, \ldots$, de la manière suivante :

$$\begin{aligned}
x_1^{n+1} &= \frac{1}{a_{11}} \left(-\sum_{j=2}^{N} a_{1j} x_j^n + b_1 \right) \\
x_2^{n+1} &= \frac{1}{a_{22}} \left(-a_{21} x_1^{n+1} - \sum_{j=3}^{N} a_{2j} x_j^n + b_2 \right) \\
x_3^{n+1} &= \frac{1}{a_{33}} \left(-a_{31} x_1^{n+1} - a_{32} x_2^{n+1} - \sum_{j=4}^{N} a_{3j} x_j^n + b_3 \right) \\
&\vdots \\
&\text{etc.}
\end{aligned}$$

La matrice G définie par

$$G = K^{-1}M = (D-E)^{-1}F \tag{6.12}$$

est appelée ***matrice de Gauss-Seidel***. A priori, la méthode de Gauss-Seidel devrait être plus performante que la méthode de Jacobi puisqu'on tient compte, au fur et à mesure, des valeurs x_i^{n+1} déjà calculées.

Qu'en est-il de la convergence de ces méthodes ? Avant d'énoncer un résultat général de convergence, nous donnons les définitions suivantes :

Définition 6.1 *Nous dirons que la méthode itérative (6.6) est convergente si, pour tout second membre $\vec{b}$, la relation (6.2) est vraie pour n'importe quel vecteur de départ $\vec{x}^0$.*

Définition 6.2 *Si B est une $N \times N$ matrice de valeurs propres complexes $\lambda_1, \lambda_2, \ldots, \lambda_N$, nous appelons rayon spectral de B la quantité*

$$\rho(B) = \max_{1 \leq j \leq N} |\lambda_j|,$$

où $|\lambda_j|$ est le module de λ_j, $1 \leq j \leq N$.

Nous donnons sans démonstration le résultat suivant :

Théorème 6.1 *La méthode itérative (6.6) est convergente si et seulement si le rayon spectral $\rho(K^{-1}M)$ de la matrice $K^{-1}M$ est strictement inférieur à 1.*

Nous allons maintenant montrer un résultat de convergence pour la méthode de Gauss-Seidel.

Théorème 6.2 *Si A est une matrice symétrique définie positive, alors la méthode de Gauss-Seidel est convergente.*

Démonstration
Tout d'abord remarquons que, si A est symétrique définie positive, alors ses termes diagonaux a_{ii}, $1 \leq i \leq N$, sont positifs et la matrice $D - E$ est régulière. Ainsi la méthode de Gauss-Seidel est bien définie.

Soit λ une valeur propre complexe quelconque de la matrice de Gauss-Seidel $(D-E)^{-1}F$. Il suffit de montrer que $|\lambda| < 1$ et d'utiliser le théorème 6.1 pour conclure.

Si $\vec{\varphi}$ est un N-vecteur non nul de composantes complexes $\varphi_1, \varphi_2, \ldots, \varphi_N$ qui satisfait

$$(D-E)^{-1}F\vec{\varphi} = \lambda\vec{\varphi}, \tag{6.13}$$

(un tel vecteur $\vec{\varphi}$ existe puisque c'est un vecteur propre de $(D-E)^{-1}F$ correspondant à la valeur propre λ), alors on a

$$F\vec{\varphi} = \lambda(D-E)\vec{\varphi}$$

et par suite

$$A\vec{\varphi} = (D-E-F)\vec{\varphi} = (1-\lambda)(D-E)\vec{\varphi}. \tag{6.14}$$

De la relation (6.14) on conclut que λ est différent de 1 car sinon on aurait $A\vec{\varphi} = 0$ et donc $\vec{\varphi} = 0$, ce qui est une contradiction avec le fait que la matrice A soit régulière. Multiplions maintenant la relation (6.14) à gauche par $\vec{\varphi}^*$ où $\vec{\varphi}^*$ est le conjugué complexe de $\vec{\varphi}^T$. Nous avons

$$\vec{\varphi}^* A\vec{\varphi} = (1-\lambda)\vec{\varphi}^*(D-E)\vec{\varphi}. \tag{6.15}$$

Clairement parlant $\vec{\varphi}^* A\vec{\varphi}$ est un nombre réel positif car A est symétrique définie positive et $\vec{\varphi}$ est non nul. On conclut que le nombre

$$\alpha = (1-\lambda)\vec{\varphi}^*(D-E)\vec{\varphi} \tag{6.16}$$

est un nombre réel positif. De même en prenant le conjugué complexe de $(1-\lambda)\vec{\varphi}^*(D-E)\vec{\varphi}$ nous obtenons, puisque $E^T = F$, la relation :

$$\alpha = (1-\bar{\lambda})\vec{\varphi}^*(D-F)\vec{\varphi}, \tag{6.17}$$

où $\bar{\lambda}$ est le conjugué complexe de λ. En multipliant (6.16) par $(1-\bar{\lambda})$ et (6.17) par $(1-\lambda)$, puis en additionnant on trouve

$$(2-\lambda-\bar{\lambda})\alpha = |1-\lambda|^2 \vec{\varphi}^*(2D-E-F)\vec{\varphi} \tag{6.18}$$

où $|1-\lambda|$ est le module de $(1-\lambda)$. Puisque $A = D-E-F$ et puisque $\alpha = \vec{\varphi}^* A\vec{\varphi}$ nous obtenons de (6.18) :

$$(2-\lambda-\bar{\lambda}-|1-\lambda|^2)\alpha = |1-\lambda|^2 \vec{\varphi}^* D\vec{\varphi}. \tag{6.19}$$

La matrice D est symétrique définie positive (car c'est la diagonale d'une matrice symétrique définie positive) et donc $\vec{\varphi}^* D\vec{\varphi}$ est positif. Puisque $\lambda \neq 1$ et $\alpha > 0$, nous obtenons de (6.19) que

$$2-\lambda-\bar{\lambda}-|1-\lambda|^2 > 0. \tag{6.20}$$

En développant $|1-\lambda|^2 = (1-\lambda)(1-\bar{\lambda}) = 1-\lambda-\bar{\lambda}+|\lambda|^2$, nous obtenons finalement

$$1-|\lambda|^2 > 0,$$

ce qui prouve que $|\lambda| < 1$. ■

Nous pouvons démontrer avec le même type d'arguments que ceux développés dans la démonstration précédente le résultat suivant.

Théorème 6.3 *Si A et $2D-A$ sont des matrices symétriques définies positives, alors la méthode de Jacobi est convergente.*

6.2 Un exemple

Définissons la $N \times N$ matrice tridiagonale A par

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}. \tag{6.21}$$

Nous verrons que cette matrice joue un rôle important lorsqu'on veut résoudre numériquement une équation différentielle du deuxième ordre avec des conditions aux limites prescrites (chap. 10). Dans la section 4.9, nous avons déjà introduit cette matrice et nous avons vérifié qu'elle est symétrique définie positive. Ainsi, d'après le théorème 6.2, la méthode de Gauss-Seidel pour résoudre $A\vec{x} = \vec{b}$ est convergente.

Soit λ_k, $k = 1, \ldots, N$, les valeurs propres de la matrice A, données par l'expression (4.54). Puisque $D = 2I$, les valeurs propres de la matrice $2D - A$ sont $4 - \lambda_k = 2 + 2\cos(k\pi/(N+1))$, $k = 1, \ldots, N$. Puisque toutes ces valeurs

propres sont positives, la matrice $2D - A$ est donc aussi symétrique définie positive. Ainsi, d'après le théorème 6.2, la méthode de Jacobi pour résoudre $A\vec{x} = \vec{b}$ est convergente.

Nous allons maintenant étudier en détail le rayon spectral de la matrice de Jacobi.

Lorsque la matrice A est donnée par (6.21), nous avons $D = 2I$ et ainsi nous vérifions immédiatement que la matrice de Jacobi $J = D^{-1}(E+F)$ peut s'écrire

$$J = D^{-1}(E + F - D + D) = D^{-1}(D - A) = I - D^{-1}A = I - \frac{1}{2}A.$$

Les valeurs propres de la matrice J sont donc $1 - \lambda_k/2 = \cos(k\pi/(N+1))$, $k = 1, \ldots, N$. Par définition, le rayon spectral de J étant la plus grande valeur propre de J en valeur absolue, nous avons donc

$$\rho(J) = \cos\frac{\pi}{N+1}. \tag{6.22}$$

La relation (6.22) nous montre que $\rho(J) < 1$ et, en utilisant le théorème 6.1, nous retrouvons le résultat qui affirme que la méthode itérative de Jacobi pour résoudre $A\vec{x} = \vec{b}$ est convergente.

Cependant, il est facile de conclure de la relation (6.22) et du développement $\cos x = 1 - x^2/2 + O(x^4)$ lorsque x tend vers zéro, qu'il existe une constante C indépendante de N telle que

$$|1 - \rho(J)| \leq C\frac{1}{N^2}, \qquad \forall N > 1. \tag{6.23}$$

Nous constatons donc que, plus l'ordre N de la matrice A est grand, et plus le rayon spectral $\rho(J)$ de la matrice J est proche de 1, en restant toujours strictement inférieur à 1. Si nous admettons que $\rho(J)$ est une mesure de la vitesse de convergence de la méthode itérative de Jacobi ("$\rho(J)$ proche de 1" signifie que la vitesse est très lente alors que "$\rho(J)$ proche de zéro" signifie que la vitesse est très rapide), alors nous constatons que la méthode de Jacobi converge très lentement lorsque N est grand. Pire encore, la méthode peut ne pas converger à cause des erreurs d'arrondis !

Nous pouvons faire une analyse semblable pour la méthode de Gauss-Seidel. En effet, la matrice de Gauss-Seidel étant donnée par $G = (D - E)^{-1}F$, il est possible de montrer que, dans le cas où la matrice A est définie par (6.21), nous avons la relation $\rho(G) = \rho(J)^2$. Ce résultat est d'ailleurs démontré dans le cas particulier où $N = 2$ dans l'exercice 6.1. Nous en déduisons les mêmes conclusions que celles ci-dessus, i.e. $|1 - \rho(G)| \leq C/N^2$, $\forall N > 1$.

6.3 Méthodes de relaxation, méthode SSOR

Soit A une $N \times N$ matrice régulière dont tous les termes diagonaux a_{ii}, $1 \leq i \leq N$, sont non nuls. Nous écrivons la matrice A sous la forme (6.7), i.e.

$$A = D - E - F.$$

Si ω est un nombre réel non nul (appelé paramètre de relaxation), nous pouvons aussi écrire

$$A = \frac{1}{\omega}D - E - \left(\frac{1-\omega}{\omega}D + F\right),$$

et par suite, nous posons $K = \omega^{-1}D - E$, $M = \omega^{-1}(1-\omega)D + F$ et nous avons bien $A = K - M$. La méthode itérative (6.6) devient ainsi

$$\left(\frac{1}{\omega}D - E\right)\vec{x}^{n+1} = \left(\frac{1-\omega}{\omega}D + F\right)\vec{x}^{n} + \vec{b}. \tag{6.24}$$

Cette méthode, appelée ***méthode de relaxation***, nécessite à chaque pas la résolution d'un système triangulaire, tout comme la méthode de Gauss-Seidel. D'ailleurs, on retrouve la méthode de Gauss-Seidel lorsqu'on choisit $\omega = 1$. Si $\omega < 1$ on parle de ***sous-relaxation*** alors que si $\omega > 1$ on parle de ***surrelaxation***.

Définissons la matrice de la méthode itérative

$$G_\omega = \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right). \tag{6.25}$$

Il est possible, dans certains cas, de faire une analyse du rayon spectral $\rho(G_\omega)$ de cette matrice en fonction de ω et de montrer qu'il existe un nombre ω_{opt} tel que $\rho(G_{\omega_{opt}}) < \rho(G_\omega)$ pour tout ω différent de ω_{opt}. Plus précisément, nous avons le résultat suivant :

Théorème 6.4 *Si A est une matrice tridiagonale définie positive, alors la méthode de Jacobi et la méthode de relaxation sont convergentes lorsque $0 < \omega < 2$. De plus, il existe un et un seul paramètre de relaxation optimal ω_{opt} égal à*

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(J)^2}}, \tag{6.26}$$

où $\rho(J)$ est le rayon spectral de la matrice J de la méthode de Jacobi. Le graphe de $\rho(G_\omega)$ en fonction de ω est représenté dans la figure 6.1.

Dans le cadre du théorème 6.4, nous constatons que ω_{opt} est plus grand que 1 et ainsi il vaut mieux utiliser une méthode de surrelaxation ($\omega > 1$). De plus, si nous ne connaissons qu'approximativement le paramètre optimal ω_{opt}, nous avons intérêt à le surévaluer plutôt qu'à le sous-évaluer. En effet, si nous posons $f(\omega) = \rho(G_\omega)$, nous constatons dans la figure 6.1 que

$$\lim_{\substack{\omega \to \omega_{opt} \\ \omega < \omega_{opt}}} f'(\omega) = -\infty \qquad \text{et} \qquad \lim_{\substack{\omega \to \omega_{opt} \\ \omega > \omega_{opt}}} f'(\omega) = 1,$$

ce qui montre notre affirmation. Dans (6.24) nous avons privilégié les rôles des matrices E et F en laissant E à gauche et en déplaçant F à droite de l'égalité. En fait, nous aurions pu faire jouer des rôles semblables à E et F en définissant une étape intermédiaire comme suit :

$$\left(\frac{1}{\omega}D - E\right)\vec{y}^{n} = \left(\frac{1-\omega}{\omega}D + F\right)\vec{x}^{n} + \vec{b}, \tag{6.27}$$

$$\left(\frac{1}{\omega}D - F\right)\vec{x}^{n+1} = \left(\frac{1-\omega}{\omega}D + E\right)\vec{y}^{n} + \vec{b}, \tag{6.28}$$

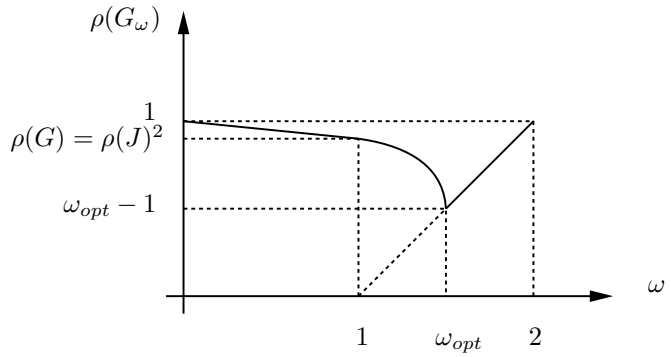

Fig. 6.1 Rayon spectral $\rho(G_\omega)$ en fonction de ω lorsque A est tridiagonale, symétrique définie positive.

où $\vec{y}^n$ est le vecteur intermédiaire qui permet de calcul $\vec{x}^{n+1}$ à partir de $\vec{x}^n$. Cette méthode itérative est bien connue sous l'abréviation anglaise SSOR (*symmetric successive overrelaxation*). Il existe une abondante littérature sur le choix du paramètre ω suivant les applications que l'on rencontre lorsqu'on discrétise des équations aux dérivées partielles (chap. 10 à 14) ; nous ne dirons rien de plus sur cette méthode itérative.

6.4 Méthodes du gradient et du gradient conjugué

Dans cette section, nous supposons la matrice A symétrique définie positive et nous étudions un autre type de méthodes itératives pour résoudre numériquement $A\vec{x} = \vec{b}$, le N-vecteur $\vec{b}$ étant donné.

Commençons par définir, pour tout N-vecteur $\vec{y}$, la quantité

$$\mathcal{L}(\vec{y}) = \frac{1}{2}\vec{y}^T A\vec{y} - \vec{b}^T\vec{y}. \tag{6.29}$$

Il est clair que $\mathcal{L}(\vec{y})$ est un nombre réel. En fait, $\mathcal{L}$ est une application de $\mathbb{R}^N$ dans $\mathbb{R}$. Nous avons le résultat suivant.

Théorème 6.5 *Si A est une $N \times N$ matrice symétrique définie positive et si $\vec{x}$ est solution de $A\vec{x} = \vec{b}$ alors, pour tout N-vecteur $\vec{y}$ différent de $\vec{x}$, on a :*

$$\mathcal{L}(\vec{x}) < \mathcal{L}(\vec{y}).$$

Démonstration
Soit $\vec{x}$ tel que $A\vec{x} = \vec{b}$ et soit $\vec{y}$ un N-vecteur différent de $\vec{x}$. Par hypothèse, le vecteur $\vec{z} = \vec{x} - \vec{y}$ est donc différent de zéro. Par définition de $\mathcal{L}$ nous avons :

$$\begin{aligned}\mathcal{L}(\vec{y}) = \mathcal{L}(\vec{x} - \vec{z}) &= \frac{1}{2}(\vec{x} - \vec{z})^T A(\vec{x} - \vec{z}) - \vec{b}^T(\vec{x} - \vec{z}) \\ &= \frac{1}{2}\vec{x}^T A\vec{x} - \vec{b}^T\vec{x} - \frac{1}{2}\vec{z}^T A\vec{x} - \frac{1}{2}\vec{x}^T A\vec{z} + \frac{1}{2}\vec{z}^T A\vec{z} + \vec{b}^T\vec{z}.\end{aligned}$$

En tenant compte de la symétrie de A, nous avons $\vec{x}^T A\vec{z} = \vec{z}^T A\vec{x}$ ainsi que $\vec{b}^T\vec{z} = \vec{z}^T\vec{b}$, si bien que

$$\begin{aligned}\mathcal{L}(\vec{y}) &= \mathcal{L}(\vec{x}) - \vec{z}^T A\vec{x} + \vec{z}^T\vec{b} + \frac{1}{2}\vec{z}^T A\vec{z} \\ &= \mathcal{L}(\vec{x}) - \vec{z}^T\left(A\vec{x} - \vec{b}\right) + \frac{1}{2}\vec{z}^T A\vec{z},\end{aligned}$$

et, puisque $A\vec{x} = \vec{b}$, nous obtenons

$$\mathcal{L}(\vec{y}) = \mathcal{L}(\vec{x}) + \frac{1}{2}\vec{z}^T A\vec{z}.$$

La matrice A étant symétrique définie positive et le vecteur $\vec{z}$ étant non nul, nous avons $\vec{z}^T A\vec{z} > 0$ et ainsi $\mathcal{L}(\vec{y}) > \mathcal{L}(\vec{x})$. ■

Puisque la solution $\vec{x}$ du système $A\vec{x} = \vec{b}$ réalise le minimum de $\mathcal{L}$, nous allons construire une méthode itérative qui cherche, à chaque itération, à diminuer $\mathcal{L}$. Supposons donc que l'on ait obtenu $\vec{x}^n$ et calculons $\vec{x}^{n+1}$ de sorte à ce que $\mathcal{L}(\vec{x}^{n+1}) < \mathcal{L}(\vec{x}^n)$.

Une idée légitime consiste à choisir un vecteur $\vec{w}^{n+1}$ non nul et à poser

$$\vec{x}^{n+1} = \vec{x}^n + \alpha^{n+1}\vec{w}^{n+1} \tag{6.30}$$

où α^{n+1} est un nombre réel qui minimise la quantité $f(\alpha)$ définie par

$$f(\alpha) = \mathcal{L}\left(\vec{x}^n + \alpha\vec{w}^{n+1}\right). \tag{6.31}$$

Cette manière de calculer $\vec{x}^{n+1}$ à partir de $\vec{x}^n$ est appelée ***méthode de descente*** ; le vecteur $\vec{w}^{n+1}$ est appelé ***direction de descente***. Dans la figure 6.2, nous avons représenté les lignes de niveau de la fonction $\mathcal{L}$ lorsque $N = 2$, le minimum de $\mathcal{L}$ correspondant au vecteur $\vec{x}$.

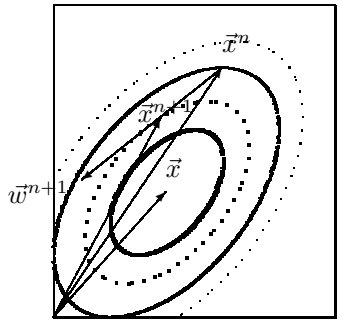

Fig. 6.2 Lignes de niveau de $\mathcal{L}$ lorsque $N = 2$.

Explicitons maintenant le calcul de α^{n+1}. Pour trouver le nombre α^{n+1} tel que $f(\alpha^{n+1}) \leq f(\alpha)$, $\forall\alpha \in \mathbb{R}$, nous annulons la dérivée $f'(\alpha)$. En utilisant (6.29)

nous avons

$$\mathcal{L}(\vec{y}) = \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}y_i y_j - \sum_{i=1}^{N} b_i y_i \tag{6.32}$$

et par suite, en tenant compte de la symétrie de A :

$$\frac{\partial}{\partial y_i}\mathcal{L}(\vec{y}) = \sum_{j=1}^{N} a_{ij}y_j - b_i. \tag{6.33}$$

Il suffit d'utiliser la définition (6.31) pour montrer que

$$\begin{aligned} f'(\alpha) &= \sum_{i=1}^{N} w_i^{n+1}\frac{\partial}{\partial y_i}\mathcal{L}(\vec{x}^n + \alpha\vec{w}^{n+1}) \\ &= \sum_{i=1}^{N} w_i^{n+1}\left(\sum_{j=1}^{N} a_{ij}\left(x_j^n + \alpha w_j^{n+1}\right) - b_i\right) \\ &= (\vec{w}^{n+1})^T\left(A\left(\vec{x}^n + \alpha\vec{w}^{n+1}\right) - \vec{b}\right). \end{aligned} \tag{6.34}$$

Clairement, si nous voulons que α^{n+1} soit tel que $f'(\alpha^{n+1}) = 0$, nous déduisons de (6.34) que α^{n+1} est donné par :

$$\alpha^{n+1} = \frac{(\vec{w}^{n+1})^T\left(\vec{b} - A\vec{x}^n\right)}{(\vec{w}^{n+1})^T A\vec{w}^{n+1}}. \tag{6.35}$$

En définissant encore

$$\vec{r}^n = \vec{b} - A\vec{x}^n \tag{6.36}$$

qui est appelé ***résidu*** à l'étape n, la méthode de la descente peut se résumer ainsi lorsqu'on veut calculer $\vec{x}^{n+1}$ à partir de $\vec{x}^n$:

- on choisit une direction de descente $\vec{w}^{n+1}$;
- on calcule

$$\alpha^{n+1} = \frac{(\vec{w}^{n+1})^T\vec{r}^n}{(\vec{w}^{n+1})^T A\vec{w}^{n+1}}; \tag{6.37}$$

– on calcule $\vec{x}^{n+1} = \vec{x}^n + \alpha^{n+1}\vec{w}^{n+1}$.

Toute la question est maintenant de bien choisir la direction de descente $\vec{w}^{n+1}$ de sorte à ce que la méthode converge.

Remarque 6.1 On montre que le résidu à l'étape $n+1$ est orthogonal à la direction de descente $\vec{w}^{n+1}$. En effet on a

$$\vec{r}^{n+1} = \vec{b} - A\vec{x}^{n+1} = \vec{b} - A(\vec{x}^n + \alpha^{n+1}\vec{w}^{n+1}) = \vec{r}^n - \alpha^{n+1}A\vec{w}^{n+1}.$$

En utilisant (6.37) on vérifie immédiatement que $(\vec{w}^{n+1})^T\vec{r}^{n+1} = 0$.

Méthode du gradient ou méthode de la plus grande pente

Le choix le plus naturel pour minimiser $\mathcal{L}$ consiste à choisir la direction de la plus grande pente comme direction de descente. Puisque la direction de la plus grande pente au point $\vec{x}^n$ est orthogonale à la ligne de niveau de $\mathcal{L}$, nous avons $\vec{w}^{n+1} = \overrightarrow{\text{grad}}\,\mathcal{L}(\vec{x}^n)$. La i-ième composante de $\overrightarrow{\text{grad}}\,\mathcal{L}(\vec{y})$ étant définie par $\partial\mathcal{L}(\vec{y})/\partial y_i$, nous obtenons directement de (6.33) et de (6.36) :

$$\vec{w}^{n+1} = \overrightarrow{\text{grad}}\,\mathcal{L}(\vec{x}^n) = A\vec{x}^n - \vec{b} = -\vec{r}^n. \tag{6.38}$$

Remarquons que nous venons de démontrer que le résidu $\vec{r}^n$ à l'étape n est orthogonal à la ligne de niveau de $\mathcal{L}$ passant par $\vec{x}^n$. Il est clair que si le résidu $\vec{r}^n$ est nul, alors $\vec{x}^n$ est la solution cherchée. Il n'est donc pas restrictif de supposer dans la suite $\vec{r}^n \neq 0$. Avec ce choix de direction de descente l'égalité (6.37) devient :

$$\alpha^{n+1} = -\frac{\|\vec{r}^n\|^2}{(\vec{r}^n)^T A\vec{r}^n}. \tag{6.39}$$

Remarquons encore que $\vec{r}^{n+1}$ peut s'exprimer en fonction de $\vec{r}^n$. En effet, en utilisant (6.38), nous avons :

$$\begin{aligned}\vec{r}^{n+1} &= \vec{b} - A\vec{x}^{n+1} = \vec{b} - A\left(\vec{x}^n + \alpha^{n+1}\vec{w}^{n+1}\right)\\ &= \vec{r}^n + \alpha^{n+1}A\vec{r}^n.\end{aligned} \tag{6.40}$$

Introduisons un vecteur intermédiaire $\vec{z}^{n+1}$ défini par $\vec{z}^{n+1} = -A\vec{r}^n$. Nous obtenons $\vec{r}^{n+1} = \vec{r}^n - \alpha^{n+1}\vec{z}^{n+1}$ et l'algorithme de la plus grande pente s'exprime de la façon suivante :

– on choisit un vecteur de départ $\vec{x}^0$ et on calcule $\vec{r}^0 = \vec{b} - A\vec{x}^0$;

– pour $n = 0, 1, 2, \ldots$, on calcule $\vec{x}^{n+1}$ et $\vec{r}^{n+1}$ par les relations suivantes :

$$\vec{z}^{n+1} = -A\vec{r}^n, \tag{6.41}$$

$$\alpha^{n+1} = \frac{\|\vec{r}^n\|^2}{(\vec{r}^n)^T\vec{z}^{n+1}}, \tag{6.42}$$

$$\vec{x}^{n+1} = \vec{x}^n - \alpha^{n+1}\vec{r}^n, \tag{6.43}$$

$$\vec{r}^{n+1} = \vec{r}^n - \alpha^{n+1}\vec{z}^{n+1}, \text{ (si } \vec{r}^{n+1} = 0 : \text{stop)}. \tag{6.44}$$

Remarquons que l'opération la plus coûteuse de l'algorithme ci-dessus est la multiplication de la matrice A par le vecteur $\vec{r}^n$ dans (6.41).

Il est possible de montrer le résultat suivant.

Théorème 6.6 *Rappelons que A est une matrice symétrique définie positive. Alors la méthode de la plus grande pente converge.*

Bien que cette méthode soit convergente, elle converge souvent lentement. Il se peut même, dans les applications pratiques, qu'elle ne converge pas à cause des erreurs d'arrondis commises lors des calculs.

Nous allons donc améliorer cette méthode en proposant la méthode du gradient conjugué.

Méthode du gradient conjugué

Supposons que l'on ait obtenu $\vec{x}^n$ et utilisons la méthode de descente (6.37) pour calculer $\vec{x}^{n+1}$ à partir de $\vec{x}^n$. L'idée de la méthode du gradient conjugué est de corriger la direction $-\vec{r}^n$ prise précédemment, par l'ancienne direction $\vec{w}^n$, de sorte à ce que l'erreur entre $\vec{x}$ et $\vec{x}^{n+1}$ soit la plus petite possible dans la norme $\|\cdot\|_A$ définie par $\|\vec{y}\|_A = (\vec{y}^T A\vec{y})^{1/2}$ pour tout N-vecteur $\vec{y}$. Plus précisément nous choisissons dans (6.37) :

$$\vec{w}^{n+1} = -\vec{r}^n + \beta^n \vec{w}^n, \tag{6.45}$$

où β^n est un nombre calculé de sorte à ce que la quantité $\|\vec{x} - \vec{x}^{n+1}\|_A$ soit la plus petite possible. Nous n'explicitons pas les calculs, mais nous pourrions montrer qu'un tel objectif conduit à l'expression suivante pour β^n :

$$\beta^n = \frac{(\vec{r}^n)^T A\vec{w}^n}{(\vec{w}^n)^T A\vec{w}^n}. \tag{6.46}$$

Introduisons le vecteur intermédiaire $\vec{z}^n$ défini par $\vec{z}^n = A\vec{w}^n$. Nous vérifions que

$$\vec{r}^n = \vec{b} - A\vec{x}^n = \vec{b} - A\left(\vec{x}^{n-1} + \alpha^n \vec{w}^n\right) = \vec{r}^{n-1} - \alpha^n \vec{z}^n. \tag{6.47}$$

Finalement, l'algorithme de descente (6.37) avec les choix (6.45) et (6.46) devient :

- on choisit un vecteur de départ $\vec{x}^0$ et on calcule $\vec{r}^0 = \vec{b} - A\vec{x}^0$; on exécute ensuite le premier pas de la méthode de la plus grande pente, c'est-à-dire on calcule :

$$\vec{w}^1 = -\vec{r}^0 \qquad \text{et} \qquad \vec{z}^1 = A\vec{w}^1, \tag{6.48}$$

$$\alpha^1 = \frac{(\vec{r}^0)^T \vec{w}^1}{(\vec{w}^1)^T \vec{z}^1}, \tag{6.49}$$

$$\vec{x}^1 = \vec{x}^0 + \alpha^1 \vec{w}^1. \tag{6.50}$$

– pour $n = 1, 2, 3, \ldots$, on calcule $\vec{r}^n$, β^n, $\vec{w}^{n+1}$, $\vec{z}^{n+1}$, α^{n+1}, $\vec{x}^{n+1}$ par les relations suivantes :

$$\vec{r}^n = \vec{r}^{n-1} - \alpha^n \vec{z}^n, \text{ (si } \vec{r}^n = 0 : \text{stop)}, \tag{6.51}$$

$$\beta^n = \frac{(\vec{r}^n)^T \vec{z}^n}{(\vec{w}^n)^T \vec{z}^n}, \tag{6.52}$$

$$\vec{w}^{n+1} = -\vec{r}^n + \beta^n \vec{w}^n, \tag{6.53}$$

$$\vec{z}^{n+1} = A\vec{w}^{n+1}, \tag{6.54}$$

$$\alpha^{n+1} = \frac{(\vec{r}^n)^T \vec{w}^{n+1}}{(\vec{w}^{n+1})^T \vec{z}^{n+1}}, \tag{6.55}$$

$$\vec{x}^{n+1} = \vec{x}^n + \alpha^{n+1} \vec{w}^{n+1}. \tag{6.56}$$

L'algorithme (6.48)-(6.56) est appelé ***méthode du gradient conjugué***. Notons qu'on pourrait mémoriser le produit scalaire $(\vec{w}^n)^T \vec{z}^n$ afin de ne le calculer qu'une seule fois et non deux. Ici encore, lors de chaque itération de l'algorithme, l'opération la plus coûteuse est le produit matrice-vecteur (6.54).

Il est possible de démontrer que le résidu $\vec{r}^n$ obtenu à la n^e étape de l'algorithme (6.48)-(6.56) est orthogonal à tous les précédents $\vec{r}^0, \vec{r}^1, \ldots, \vec{r}^{n-1}$. Ainsi nous obtenons le résultat suivant.

Théorème 6.7 *Rappelons que A est une $N \times N$ matrice symétrique définie positive. Alors la méthode du gradient conjugué fournit la solution $\vec{x}$ en au plus N itérations. Ainsi il existe $n \leq N$ tel que $\vec{r}^n = 0$.*

L'algorithme du gradient conjugué peut paraître parfait puisqu'il converge en un nombre fini d'itérations. Malheureusement, dans les applications pratiques cet algorithme est souvent mal conditionné, ce qui a pour effet de dégrader la convergence.

En fait, on constate (et on peut démontrer) que l'erreur obtenue au cours des itérations de la méthode du gradient conjugué diminue d'autant plus rapidement que le nombre de condition spectral $\chi(A)$ de la matrice A est petit. Ainsi, une idée de base pour améliorer cette méthode consiste à considérer le système linéaire $C^{-1}A\vec{x} = C^{-1}\vec{b}$ au lieu de $A\vec{x} = \vec{b}$. Ici la matrice C est une matrice symétrique définie positive choisie de sorte à ce que $\chi(C^{-1}A)$ soit nettement plus petit que $\chi(A)$. La matrice $C^{-1}A$ n'étant pas symétrique nous allons modifier le système $C^{-1}A\vec{x} = C^{-1}\vec{b}$, pour pouvoir utiliser la méthode du gradient conjugué. Pour ce faire nous procédons à une décomposition de Cholesky de C en posant

$$C = LL^T,$$

où L est une matrice triangulaire inférieure. Ainsi le système linéaire $C^{-1}A\vec{x} = C^{-1}\vec{b}$ devient $L^{-T}L^{-1}A\vec{x} = L^{-T}L^{-1}\vec{b}$ puis $L^{-1}AL^{-T}L^T\vec{x} = L^{-1}\vec{b}$. Posons

$$A^\star = L^{-1}AL^{-T}, \qquad \vec{x}^\star = L^T\vec{x} \qquad \text{et} \qquad \vec{b}^\star = L^{-1}\vec{b}.$$

Ainsi résoudre le système linéaire $C^{-1}A\vec{x} = C^{-1}\vec{b}$ est équivalent à résoudre le système linéaire

$$A^\star \vec{x}^\star = \vec{b}^\star.$$

Puisque la matrice $A^\star$ est symétrique définie positive nous pouvons, en théorie, résoudre ce dernier système par la méthode du gradient conjugué. Cette manière de procéder est appelée ***méthode du gradient conjugué préconditionné***. Nous ne discutons pas dans ce chapitre, ni du choix de la matrice C, appelée ***matrice de préconditionnement***, ni de la mise en œuvre de la méthode. Par contre, nous donnons des références bibliographiques dans la section 6.6.

6.5 Exercices

Exercice 6.1 On considère la matrice $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

1. Montrer que la matrice G_ω définie par (6.25) est donnée par

$$G_\omega = \begin{bmatrix} 1-\omega & \dfrac{\omega}{2} \\ \dfrac{\omega}{2}(1-\omega) & 1-\omega+\dfrac{\omega^2}{4} \end{bmatrix}. \tag{6.57}$$

2. Soit J la matrice de Jacobi définie par (6.9). Montrer que $\rho(J) = 1/2$. En déduire que le coefficient ω_{opt} défini par (6.26) vaut $\omega_{opt} = 8 - 4\sqrt{3}$.
3. Montrer que l'on a

$$\begin{aligned} &\rho(G_\omega) = (1-\omega+\frac{\omega^2}{8}) + \frac{\omega}{8}\sqrt{\omega^2-16\omega+16} && \text{si } \omega \in [0, \omega_{opt}], \\ &\rho(G_\omega) = \omega - 1 && \text{si } \omega \in [\omega_{opt}, 2]. \end{aligned}$$

   Tracer le graphe de la fonction $\omega \to \rho(G_\omega)$.
4. Montrer que, si f est la fonction définie par $f(\omega) = \rho(G_\omega)$, alors on a

$$\lim_{\substack{\omega \to \omega_{opt} \\ \omega < \omega_{opt}}} f'(\omega) = -\infty \qquad \text{et} \qquad \lim_{\substack{\omega \to \omega_{opt} \\ \omega > \omega_{opt}}} f'(\omega) = 1.$$

Solution

1. Notons $G_\omega = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$. La définition (6.25) donne l'égalité suivante :

$$G_\omega = \begin{bmatrix} \dfrac{2}{\omega} & 0 \\ -1 & \dfrac{2}{\omega} \end{bmatrix}^{-1} \begin{bmatrix} 2\dfrac{1-\omega}{\omega} & 1 \\ 0 & 2\dfrac{1-\omega}{\omega} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}.$$

Il s'agit donc de résoudre les deux systèmes linéaires

$$\begin{bmatrix} \dfrac{2}{\omega} & 0 \\ -1 & \dfrac{2}{\omega} \end{bmatrix} \begin{bmatrix} g_{11} \\ g_{21} \end{bmatrix} = \begin{bmatrix} 2\dfrac{1-\omega}{\omega} \\ 0 \end{bmatrix} \qquad \text{et} \qquad \begin{bmatrix} \dfrac{2}{\omega} & 0 \\ -1 & \dfrac{2}{\omega} \end{bmatrix} \begin{bmatrix} g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 2\dfrac{1-\omega}{\omega} \end{bmatrix},$$

ce qui est une tâche aisée puisque la matrice est triangulaire. Nous obtenons bien $g_{11} = 1 - \omega$, $g_{21} = \omega(1-\omega)/2$, puis $g_{12} = \omega/2$, $g_{22} = 1 - \omega + \omega^2/4$. Nous avons donc bien montré que la matrice G_ω est donnée par (6.57).

2. Par définition, la matrice de Jacobi est donnée par :

$$J = D^{-1}(E+F) = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

et donc, pour tout nombre λ, nous avons

$$\det(J - \lambda I) = \det \begin{bmatrix} -\lambda & \frac{1}{2} \\ \frac{1}{2} & -\lambda \end{bmatrix} = \lambda^2 - \frac{1}{4}.$$

Donc, $\det(J - \lambda I) = 0$ pour $\lambda = \pm 1/2$ et par conséquent $\rho(J) = 1/2$. On déduit de (6.26) que $\omega_{opt} = 8 - 4\sqrt{3}$.

3. En utilisant le point 1, nous avons

$$\begin{aligned} \det(G_\omega - \lambda I) &= \det \begin{bmatrix} 1-\omega-\lambda & \frac{\omega}{2} \\ \frac{\omega}{2}(1-\omega) & 1-\omega+\frac{\omega^2}{4}-\lambda \end{bmatrix} \\ &= \lambda^2 - \lambda\left(2(1-\omega) + \frac{\omega^2}{4}\right) + (1-\omega)^2. \end{aligned} \tag{6.58}$$

Les valeurs propres de G_ω sont donc les racines λ_1 et λ_2 du trinôme du second degré en ω, défini ci-dessus. Le discriminant de ce trinôme vaut

$$\begin{aligned} \Delta &= \left(2(1-\omega) + \frac{\omega^2}{4}\right)^2 - 4(1-\omega)^2 \\ &= \omega^2\left(1 - \omega + \frac{\omega^2}{16}\right) \\ &= \frac{\omega^2}{16}\left(\omega - (8+4\sqrt{3})\right)\left(\omega - (8-4\sqrt{3})\right). \end{aligned}$$

Le signe du discriminant dépend clairement de ω. Puisque $0 \leq \omega \leq 2$, nous devons donc considérer deux cas. Si $0 \leq \omega \leq 8 - 4\sqrt{3} = \omega_{opt}$, alors $\Delta \geq 0$ et les deux racines (réelles) du trinôme sont

$$\begin{aligned} \lambda_1 &= (1-\omega) + \frac{\omega^2}{8} + \frac{\omega}{2}\sqrt{1 - \omega + \frac{\omega^2}{16}}, \\ \lambda_2 &= (1-\omega) + \frac{\omega^2}{8} - \frac{\omega}{2}\sqrt{1 - \omega + \frac{\omega^2}{16}}, \end{aligned}$$

et donc

$$\begin{aligned}\rho(G_\omega) &= \max\{|\lambda_1|, |\lambda_2|\} \\ &= (1-\omega) + \frac{\omega^2}{8} + \frac{\omega}{8}\sqrt{\omega^2 - 16\omega + 16}.\end{aligned}$$

Si $\omega_{opt} < \omega \leq 2$, alors $\Delta < 0$ et les deux racines (complexes) du trinôme sont

$$\begin{aligned}\lambda_1 &= (1-\omega) + \frac{\omega^2}{8} + i\frac{\omega}{2}\sqrt{-(1-\omega+\frac{\omega^2}{16})}, \\ \lambda_2 &= (1-\omega) + \frac{\omega^2}{8} - i\frac{\omega}{2}\sqrt{-(1-\omega+\frac{\omega^2}{16})},\end{aligned}$$

et donc

$$\begin{aligned}\rho(G_\omega) &= \max\{|\lambda_1|, |\lambda_2|\} \\ &= \omega - 1.\end{aligned}$$

Le graphe de la fonction $\omega \to \rho(G_\omega)$ est représenté dans la figure 6.3.

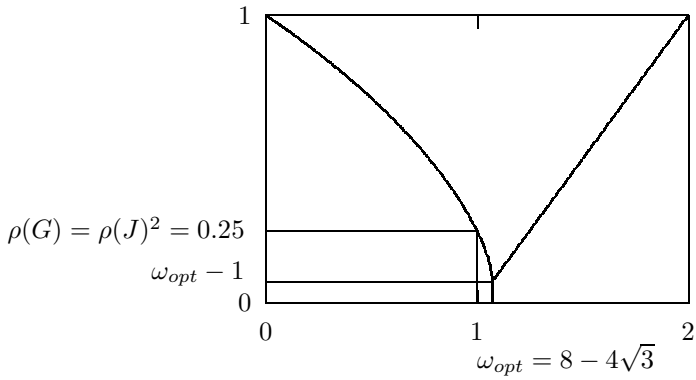

Fig. 6.3 Le graphe de la fonction $\omega \to \rho(G_\omega)$ dans le cadre de l'exercice 6.1.

4. Posons $f(\omega) = \rho(G_\omega)$. Si $\omega > \omega_{opt}$, on a bien évidemment $f'(\omega) = 1$ et donc

$$\lim_{\substack{\omega \to \omega_{opt} \\ \omega > \omega_{opt}}} f'(\omega) = 1.$$

Si $\omega < \omega_{opt}$, on a

$$f'(\omega) = -1 + \frac{\omega}{4} + \frac{1}{8}\sqrt{\omega^2 - 16\omega + 16} + \frac{\omega(2\omega - 16)}{16\sqrt{\omega^2 - 16\omega + 16}}$$

et donc

$$\lim_{\substack{\omega \to \omega_{opt} \\ \omega < \omega_{opt}}} f'(\omega) = -\infty.$$

Exercice 6.2 Il s'agit de résoudre le système linéaire $A\vec{x} = \vec{b}$ où A et $\vec{b}$ sont définis par

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \qquad \vec{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

1. Vérifier que A est symétrique définie positive.
2. Effectuer deux itérations de la méthode du gradient conjugué en partant de $\vec{x}^0 = \begin{bmatrix} 3/2 \\ 2 \end{bmatrix}$. Vérifier que $\vec{x}^2 = \vec{x}$, où $\vec{x}$ est la solution du système linéaire $A\vec{x} = \vec{b}$.
3. Soit $\mathcal{L}$ la grandeur définie en (6.29). Représenter graphiquement les lignes de niveau de $\mathcal{L}$ ainsi que les vecteurs $\vec{x}^0$, $\vec{x}^1$, $\vec{x}^2$.

Solution

1. La matrice A est clairement symétrique. De plus, pour tout nombre λ nous avons

$$\det(A - \lambda I) = \det \begin{bmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{bmatrix} = (2-\lambda)^2 - 1,$$

et donc $\det(A - \lambda I) = 0$ pour $\lambda = 1$ et $\lambda = 3$. Les valeurs propres de A sont donc réelles, strictement positives. La matrice A est donc bien symétrique définie positive.

2. Après calcul nous avons

$$\vec{r}^0 = b - A\vec{x}^0 = \begin{bmatrix} 0 \\ -3/2 \end{bmatrix}, \qquad \vec{w}^1 = -\vec{r}^0,$$

$$\vec{z}^1 = A\vec{w}^1 = \begin{bmatrix} -3/2 \\ 3 \end{bmatrix}, \qquad \alpha^1 = \frac{(\vec{r}^0)^T \vec{w}^1}{(\vec{w}^1)^T \vec{z}^1} = -\frac{1}{2},$$

$$\vec{x}^1 = \vec{x}^0 + \alpha^1 \vec{w}^1 = \begin{bmatrix} 3/2 \\ 5/4 \end{bmatrix}, \qquad \vec{r}^1 = \vec{r}^0 - \alpha^1 \vec{z}^1 = \begin{bmatrix} -3/4 \\ 0 \end{bmatrix},$$

$$\beta^1 = \frac{(\vec{r}^1)^T \vec{z}^1}{(\vec{w}^1)^T \vec{z}^1} = \frac{1}{4}, \qquad \vec{w}^2 = -\vec{r}^1 + \beta^1 \vec{w}^1 = \begin{bmatrix} 3/4 \\ 3/8 \end{bmatrix},$$

$$\vec{z}^2 = A\vec{w}^2 = \begin{bmatrix} 9/8 \\ 0 \end{bmatrix}, \qquad \alpha^2 = \frac{(\vec{r}^1)^T \vec{w}^2}{(\vec{w}^2)^T \vec{z}^2} = -\frac{2}{3},$$

$$\vec{x}^2 = \vec{x}^1 + \alpha^2 \vec{w}^2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Nous avons donc bien $\vec{x}^2 = \vec{x}$ qui est la solution du système linéaire $A\vec{x} = \vec{b}$, ce qui n'est pas surprenant puisque, en vertu du théorème 6.7, la méthode du gradient conjugué sur une matrice 2×2 symétrique définie positive converge en au plus 2 itérations.

3. Soit $\vec{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ un vecteur quelconque. La quantité $\mathcal{L}$ est donnée par

$$\begin{aligned} \mathcal{L}(\vec{y}) &= \frac{1}{2}\vec{y}^T A\vec{y} - \vec{b}^T\vec{y} \\ &= y_1^2 + y_2^2 - y_1y_2 - y_1 - y_2 \\ &= \frac{3}{4}(y_1 - y_2)^2 + \frac{1}{4}(y_1 + y_2)^2 - y_1 - y_2. \end{aligned}$$

Les lignes de niveau de $\mathcal{L}$ sont donc des ellipes centrées au point $(1, 1)$ et dont les axes sont donnés par la première et la deuxième bissectrice ; ces ellipses sont représentées dans la figure 6.4.

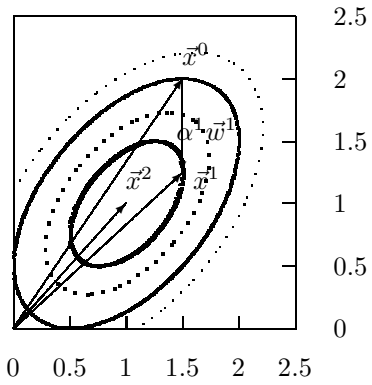

Fig. 6.4 Lignes de niveau de $\mathcal{L}$ dans le cadre de l'exercice 6.2.

6.6 Notes bibliographiques et remarques

Le nombre d'itérations nécessaires à la résolution d'un système linéaire par la méthode du gradient conjugué dépend, en général, de la taille de la matrice. Comme nous l'avons déjà indiqué à la fin de la section 6.4, il convient de ***préconditionner*** le système linéaire. Les méthodes de Jacobi, Gauss-Seidel et IC (Incomplete Cholesky, voir par exemple [11, 24]) sont les préconditionneurs les plus usuels, mais sont d'une efficacité limitée. Il existe des ***préconditionneurs optimaux*** au sens où le nombre d'itérations de la méthode du gradient conjugué ne dépend plus de la taille de la matrice. Ces méthodes sont néanmoins

difficiles à mettre en œuvre et relèvent, aujourd'hui encore, du domaine de la recherche.

Il existe des méthodes proches de la méthode du gradient conjugué adaptées aux matrices non symétriques. Les méthodes GMRES et BICG-Stab sont les exemples les plus connus, voir par exemple [11, 24, 20].

Les méthodes multiniveaux (par exemple la méthode multigrille, voir [24]) sont les méthodes les plus efficaces pour résoudre des systèmes linéaires dont la matrice est symétrique définie positive. Il est possible de montrer que la complexité de ces méthodes est optimale au sens où le nombre d'opérations est proportionnel au nombre d'inconnues du système linéaire. La mise en œuvre de ces méthodes peut toutefois s'avérer difficile dans le cadre de problèmes industriels.

Chapitre 7

Méthodes numériques pour le calcul des valeurs propres d'une matrice symétrique

7.1 Généralités

Dans ce chapitre nous cherchons à calculer numériquement les valeurs propres et les vecteurs propres d'une $N \times N$ matrice A. Il s'agit donc de trouver des nombres complexes λ_1, λ_2, ..., λ_N, ainsi que des vecteurs complexes non nuls $\vec{\varphi}_1$, $\vec{\varphi}_2$, ..., $\vec{\varphi}_N$ tels que

$$A\vec{\varphi}_j = \lambda_j \vec{\varphi}_j, \qquad 1 \leq j \leq N.$$

Le calcul des valeurs et vecteurs propres d'une $N \times N$ matrice est, en toute généralité, une affaire délicate puisqu'il équivaut à chercher les zéros d'un polynôme de degré N. En effet, calculer les valeurs propres de A, revient à trouver les zéros de son polynôme caractéristique défini par

$$p(\lambda) = \det\,(\lambda I - A), \tag{7.1}$$

où I est ici la $N \times N$ matrice identité. Réciproquement, si nous avions à chercher les zéros d'un polynôme de degré N du type

$$p(\lambda) = \lambda^N + a_{N-1}\lambda^{N-1} + a_{N-2}\lambda^{N-2} + \cdots + a_1\lambda + a_0, \tag{7.2}$$

où ici les a_j, $1 \leq j \leq N-1$, sont des nombres réels, nous serions amenés à chercher les valeurs propres λ de sa matrice compagnon (matrice de Frobenius)

définie par

$$A = \begin{bmatrix} -a_{N-1} & -a_{N-2} & -a_{N-3} & \cdots & \cdots & -a_0 \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ & & 1 & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 1 & 0 \end{bmatrix} \tag{7.3}$$

dont le vecteur propre associé $\vec{\varphi}$ est donné par ses composantes $\varphi_j = \lambda^{N-j}$, $1 \leq j \leq N$. Or la recherche des zéros d'un polynôme de degré N (avec N grand) n'est pas une opération triviale puisque nous savons que, pour $N \geq 5$, il n'existe pas de formule explicite pour les exprimer ! Nous aurons donc recours à des méthodes ***itératives*** utilisant des opérations algébriques élémentaires. Ces méthodes ne donnent pas, en principe, le résultat exact après un nombre fini de pas et il est donc impératif de savoir sous quelles conditions elles convergent. Un grand nombre de méthodes sont à disposition lorsque la matrice A est symétrique, aussi nous n'aborderons que ce cas dans ce chapitre. Le cas des matrices non symétriques est plus délicat. Le lecteur intéressé pourra néanmoins consulter les références proposées à la fin du chapitre.

Supposons donc que A soit une $N \times N$ matrice symétrique dont toutes les valeurs propres $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N$ (répétées selon leur multiplicité) sont numérotées de sorte que l'on ait

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_N|. \tag{7.4}$$

Soit $\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N$ une base de vecteurs propres orthonormalisée de $\mathbb{R}^N$ correspondant à ces valeurs propres, c'est-à-dire

$$A\vec{\varphi}_j = \lambda_j \vec{\varphi}_j, \qquad 1 \leq j \leq N, \tag{7.5}$$

et

$$\vec{\varphi}_j^T \vec{\varphi}_k = \delta_{jk}, \qquad 1 \leq j, k \leq N, \tag{7.6}$$

où δ_{kj} est le symbole de Kronecker qui vaut 1 si $j = k$ et 0 sinon. Si Q est la $N \times N$ matrice dont les colonnes sont successivement $\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N$, alors Q est orthogonale ($Q^T = Q^{-1}$) et on a, puisque $\vec{\varphi}_k^T A \vec{\varphi}_j = \lambda_j \delta_{jk}$, $1 \leq j, k \leq N$, la relation suivante :

$$Q^T A Q = D; \tag{7.7}$$

la matrice D étant la matrice diagonale donnée par $D = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_N)$.

La première méthode que nous exposons est une méthode bien connue, il s'agit de la ***méthode de la puissance***. Cette méthode permet d'obtenir une approximation de la plus grande valeur propre en valeur absolue λ_1, ainsi qu'une approximation du vecteur propre correspondant $\vec{\varphi}_1$.

7.2 Méthode de la puissance

Soit $\vec{x}^{(0)}$ un vecteur non nul donné. La méthode de la puissance consiste à construire la suite de vecteurs $(\vec{x}^{(n)})_{n=1}^{\infty}$ et la suite de nombres réels $(\mu^{(n)})_{n=1}^{\infty}$ à l'aide des relations suivantes :

$$\vec{x}^{(n)} = A\vec{x}^{(n-1)}, \qquad n = 1, 2, \ldots, \tag{7.8}$$

$$\mu^{(n)} = \frac{\vec{x}^{(n)T} A \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|^2} = \frac{\vec{x}^{(n)T} \vec{x}^{(n+1)}}{\vec{x}^{(n)T} \vec{x}^{(n)}}, \qquad n = 1, 2, \ldots. \tag{7.9}$$

Il est clair que, si nous itérons (7.8), nous obtenons $\vec{x}^{(n)} = A^n \vec{x}^{(0)}$, ce qui justifie le nom donné à la méthode. Nous allons maintenant démontrer le résultat suivant.

Théorème 7.1 *Soit λ_1 la plus grande valeur propre de A en valeur absolue et soit $\vec{\varphi}_1$ le vecteur propre correspondant. Si $|\lambda_1| > |\lambda_k|$, $k = 2, 3, \ldots, N$, et si $\vec{x}^{(0)}$ n'est pas choisi orthogonalement à $\vec{\varphi}_1$, i.e. $\vec{\varphi}_1^T \vec{x}^{(0)} \neq 0$, alors on a*

$$\lim_{n\to\infty} \frac{\vec{\varphi}_k^T \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} = 0, \qquad k = 2, 3, \ldots, N, \tag{7.10}$$

et

$$\lim_{n\to\infty} \mu^{(n)} = \lambda_1. \tag{7.11}$$

Démonstration
Puisque $\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N$ est une base orthonormalisée de $\mathbb{R}^N$, nous pouvons décomposer $\vec{x}^{(0)}$ sous la forme

$$\vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j \vec{\varphi}_j \tag{7.12}$$

où nous avons noté

$$\alpha_j = \vec{\varphi}_j^T \vec{x}^{(0)}, \qquad 1 \leq j \leq N. \tag{7.13}$$

En utilisant (7.8) pour calculer $\vec{x}^{(1)}$, nous avons

$$\vec{x}^{(1)} = A\vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j A\vec{\varphi}_j = \sum_{j=1}^{N} \alpha_j \lambda_j \vec{\varphi}_j. \tag{7.14}$$

En itérant le processus, nous vérifions sans difficulté que

$$\vec{x}^{(n)} = \sum_{j=1}^{N} \alpha_j \lambda_j^n \vec{\varphi}_j. \tag{7.15}$$

Le théorème de Pythagore nous assure que

$$\|\vec{x}^{(n)}\| = \left(\sum_{j=1}^{N} \alpha_j^2 \lambda_j^{2n} \right)^{1/2}, \tag{7.16}$$

et par conséquent, pour $k = 2, 3, \ldots, N$, nous obtenons :

$$\frac{\vec{\varphi}_k^T \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} = \frac{\alpha_k \lambda_k^n}{\left(\sum_{j=1}^{N} \alpha_j^2 \lambda_j^{2n}\right)^{1/2}} = \frac{\alpha_k \left(\frac{\lambda_k}{\lambda_1}\right)^n}{\left(\sum_{j=1}^{N} \alpha_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2n}\right)^{1/2}}. \tag{7.17}$$

En tenant compte du fait que, par hypothèse, $\alpha_1 = \vec{\varphi}_1^T \vec{x}^{(0)}$ est différent de zéro et que $|\lambda_1| > |\lambda_k|$ pour $k = 2, \ldots, N$, nous déduisons la relation (7.10) de la relation (7.17).

En utilisant (7.9) et (7.15), nous effectuons facilement les calculs suivants :

$$\begin{aligned} \mu^{(n)} &= \frac{\vec{x}^{(n)T} \vec{x}^{(n+1)}}{\vec{x}^{(n)T} \vec{x}^{(n)}} = \frac{\sum_{j=1}^{N} \alpha_j^2 \lambda_j^{2n+1}}{\sum_{j=1}^{N} \alpha_j^2 \lambda_j^{2n}} \\ &= \lambda_1 \frac{\sum_{j=1}^{N} \alpha_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2n+1}}{\sum_{j=1}^{N} \alpha_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2n}}. \end{aligned} \tag{7.18}$$

Puisque, par hypothèse α_1 est non nul et $|\lambda_1| > |\lambda_k|$ pour tout $k = 2, 3, \ldots, N$, nous obtenons

$$\lim_{n \to \infty} \frac{\sum_{j=1}^{N} \alpha_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2n+1}}{\sum_{j=1}^{N} \alpha_j^2 \left(\frac{\lambda_j}{\lambda_1}\right)^{2n}} = \frac{\alpha_1^2}{\alpha_1^2} = 1,$$

et par conséquent (7.18) implique (7.11). ■

Remarque 7.1 La quantité

$$\mu^{(n)} = \frac{\vec{x}^{(n)T} A \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|^2}$$

est appelée ***quotient de Rayleigh***. Sous les hypothèses du théorème 7.1, la relation (7.11) indique que les quotients de Rayleigh convergent vers la plus grande valeur propre de A en valeur absolue, alors que la relation (7.10) indique que les vecteurs unités

$$\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|}$$

deviennent, lorsque n tend vers l'infini, orthogonaux à $\vec{\varphi}_2, \vec{\varphi}_3, \ldots, \vec{\varphi}_N$, c'est-à-dire colinéaires à $\vec{\varphi}_1$.

Remarque 7.2 L'hypothèse $\vec{\varphi}_1^T \vec{x}^{(0)} \neq 0$ n'est pas déterminante dans la pratique. En effet, si, par malchance, $\vec{x}^{(0)}$ était choisi othogonalement à $\vec{\varphi}_1$, les erreurs d'arrondis au cours des itérations produiraient, à un moment ou un autre, un vecteur non orthogonal à $\vec{\varphi}_1$!

Par contre, l'hypothèse $|\lambda_1| > |\lambda_k|$ pour tout $k = 2, 3, \ldots, N$ est nettement plus importante. Si maintenant nous supposons par exemple que

$$|\lambda_1| = |\lambda_2| > |\lambda_k| \qquad \text{pour } k = 3, 4, \ldots, N,$$

nous devons alors considérer les deux cas suivants :

- ou bien λ_1 est valeur propre de multiplicité deux (i.e. $\lambda_1 = \lambda_2$) et il lui correspond un sous-espace propre (plan) engendré par deux vecteurs propres linéairement indépendants correspondant à λ_1 ;
- ou bien $\lambda_1 = -\lambda_2$ et ces deux valeurs propres ont une multiplicité égale à un.

Dans le premier cas, la méthode de la puissance reste utilisable et il n'est pas difficile de montrer, comme nous l'avons fait dans le théorème 7.1, que

$$\lim_{n\to\infty} \frac{\vec{\varphi}_k^T \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} = 0, \qquad k = 3, 4, \ldots, N,$$

et $\lim_{n\to\infty} \mu^{(n)} = \lambda_1 = \lambda_2$.

Dans le deuxième cas, la méthode de la puissance ne s'applique plus dans sa version originelle. Le remède consiste à substituer la matrice A par la matrice $A + \varepsilon I$ où I est la $N \times N$ matrice identité et ε est un nombre réel positif. Cette substitution a pour conséquence de décaler les valeurs propres de ε, les valeurs propres de $A + \varepsilon I$ étant données par $\lambda_j + \varepsilon$, $1 \leq j \leq N$. Il existe donc une valeur de ε pour laquelle $|\lambda_1 + \varepsilon|$ est différent de $|\lambda_2 + \varepsilon|$. Ainsi en utilisant les relations (7.8), (7.9) avec $A + \varepsilon I$ en lieu et place de A, nous obtenons la convergence de $\mu^{(n)}$ vers $\lambda_1 + \varepsilon$ ou $\lambda_2 + \varepsilon$ lorsque n tend vers l'infini.

Remarque 7.3 A partir de la relation (7.18) il n'est pas difficile de montrer qu'il existe une constante C (indépendante de n) telle que

$$|\lambda_1 - \mu^{(n)}| \leq C \left(\frac{\lambda_2}{\lambda_1}\right)^{2n},$$

ce qui montre que la convergence de $\mu^{(n)}$ vers λ_1 est d'autant plus rapide que le rapport λ_2/λ_1 est petit !

7.3 Méthode de la puissance inverse

Soit A une $N \times N$ matrice symétrique dont les valeurs propres sont $\lambda_1, \lambda_2, \ldots, \lambda_N$, les vecteurs orthonormalisés correspondants sont $\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N$ et soit μ un nombre réel tel que

$$\mu \neq \lambda_j, \qquad 1 \leq j \leq N. \tag{7.19}$$

L'hypothèse (7.19) implique que la matrice $A-\mu I$ est régulière. Par conséquent $(A-\mu I)^{-1}$ existe et ses valeurs propres ω_j, $1 \leq j \leq N$, sont données par

$$\omega_j = (\lambda_j - \mu)^{-1}, \qquad 1 \leq j \leq N. \tag{7.20}$$

Supposons maintenant qu'il existe k tel que

$$|\lambda_k - \mu| < |\lambda_j - \mu|, \qquad j = 1, 2, \ldots, N; j \neq k. \tag{7.21}$$

Emettre l'hypothèse (7.21) est équivalent à supposer que la valeur propre λ_k la plus proche de μ est de multiplicité égale à un et que $2\mu - \lambda_k$ n'est pas une valeur propre de A. Compte tenu de (7.20) et (7.21), nous avons donc

$$|\omega_k| > |\omega_j|, \qquad j = 1, 2, \ldots, N; j \neq k. \tag{7.22}$$

Clairement, la méthode de la puissance sur la matrice $(A-\mu I)^{-1}$ permet de déterminer ω_k et, en utilisant (7.20), nous obtenons λ_k. La méthode permet également d'obtenir le vecteur propre correspondant, puisque les vecteurs propres de A sont ceux de $(A-\mu I)^{-1}$. Remplaçons donc, dans (7.8), la matrice A par la matrice $(A-\mu I)^{-1}$. A partir de $\vec{x}^{(0)}$ nous pouvons donc calculer $\vec{x}^{(n)}$, $n = 1, 2, \ldots$, de la manière suivante :

$$\vec{x}^{(n)} = (A-\mu I)^{-1}\vec{x}^{(n-1)}. \tag{7.23}$$

En procédant comme dans la démonstration du théorème 7.1, c'est-à-dire en décomposant $\vec{x}^{(0)}$ dans la base des vecteurs propres de A, nous obtenons

$$\vec{x}^{(n)} = (A-\mu I)^{-n}\vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j \omega_j^n \vec{\varphi}_j, \tag{7.24}$$

où α_j est donné par (7.13). Définissons le quotient de Rayleigh

$$\mu^{(n)} = \frac{\vec{x}^{(n)T} A \vec{x}^{(n)}}{\|\vec{x}^{(n)}\|^2}, \tag{7.25}$$

et procédons de manière similaire à ce qui a été fait pour montrer (7.18). A partir de (7.24) et (7.25), nous obtenons :

$$\begin{aligned} \mu^{(n)} - \mu &= \frac{\vec{x}^{(n)T}(A-\mu I)\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|^2} \\ &= (\lambda_k - \mu)\frac{\sum\limits_{j=1}^{N} \alpha_j^2 \left(\dfrac{\omega_j}{\omega_k}\right)^{2n-1}}{\sum\limits_{j=1}^{N} \alpha_j^2 \left(\dfrac{\omega_j}{\omega_k}\right)^{2n}}. \end{aligned} \tag{7.26}$$

Si α_k est différent de zéro, nous avons donc en utilisant (7.22) :

$$\lim_{n\to\infty} \mu^{(n)} = \lambda_k. \tag{7.27}$$

Remarquons encore que

$$\frac{\omega_j}{\omega_k} = \frac{\lambda_k - \mu}{\lambda_j - \mu} \tag{7.28}$$

et en utilisant la relation (7.26) nous déduisons que, plus μ est proche de λ_k, plus $|\omega_k|$ est grand et donc plus la convergence est rapide dans (7.27). La ***méthode de la puissance inverse*** consiste à modifier μ au cours des itérations, de sorte à ce qu'il coïncide avec le quotient de Rayleigh.

Ainsi la méthode de la puissance inverse consiste, après avoir choisi un vecteur $\vec{x}^{(0)}$, approximation du vecteur propre $\vec{\varphi}_k$, à calculer,

$$\mu^{(n-1)} = \frac{\vec{x}^{(n-1)T} A \vec{x}^{(n-1)}}{\|\vec{x}^{(n-1)}\|^2}, \tag{7.29}$$

$$\vec{x}^{(n)} = (A - \mu^{(n-1)} I)^{-1} \vec{x}^{(n-1)}, \tag{7.30}$$

pour $n = 1, 2, 3, \ldots$ En pratique, le calcul de $\vec{x}^{(n)}$ nécessite la résolution du système linéaire

$$(A - \mu^{(n-1)} I)\vec{x}^{(n)} = \vec{x}^{(n-1)}. \tag{7.31}$$

Puisque $\mu^{(n)}$ converge très rapidement vers λ_k lorsque n tend vers l'infini (convergence cubique au sens de la définition 8.1), la matrice $(A - \mu^{(n-1)} I)$ converge très rapidement vers la matrice $(A - \lambda_k I)$ qui est singulière. Ce phénomène ne porte pas à conséquence si nous utilisons la méthode d'élimination de Gauss avec changement de pivot (sect. 4.5) pour résoudre le système (7.31).

Enfin mentionnons qu'en pratique on multiplie $\vec{x}^{(n)}$ par un nombre r_n choisi tel que $\|r_n \vec{x}^{(n)}\| = 1$, de sorte à ce que les vecteurs ne grandissent pas au cours des itérations.

7.4 Méthode de Jacobi

Rappelons tout d'abord que, si A est symétrique, alors A est orthogonalement diagonalisable et prend donc la forme (7.7), c'est-à-dire

$$Q^T A Q = D, \tag{7.32}$$

où D est la matrice diagonale formée des valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_N$ de A. L'idée de la méthode de Jacobi est de construire une suite de matrices orthogonales $\{Q^{(k)}\}_{k=1}^{\infty}$ telle que

$$\lim_{k \to \infty} Q^{(1)} Q^{(2)} Q^{(3)} \ldots Q^{(k)} = Q. \tag{7.33}$$

Soit $T^{(k)}$ la matrice définie par

$$T^{(k)} = Q^{(k)T} Q^{(k-1)T} \ldots Q^{(1)T} A Q^{(1)} Q^{(2)} \ldots Q^{(k)}. \tag{7.34}$$

Les relations (7.32) et (7.33) impliquent alors que

$$\lim_{k \to \infty} T^{(k)} = D. \tag{7.35}$$

Notons $T^{(0)} = A$, nous avons

$$T^{(k)} = Q^{(k)T} T^{(k-1)} Q^{(k)}, \qquad k = 1, 2, \ldots, \tag{7.36}$$

et toute la difficulté de la méthode réside dans la construction de $T^{(k)}$ à partir de $T^{(k-1)}$. Soit $t_{ij}^{(k)}$ et $q_{ij}^{(k)}$, $1 \leq i, j \leq N$, les éléments de $T^{(k)}$ et $Q^{(k)}$ respectivement. Soit $t_{mn}^{(k-1)}$, $m \neq n$, un élément hors diagonal non nul de $T^{(k-1)}$. Nous allons construire $Q^{(k)}$ de sorte à ce que $t_{mn}^{(k)}$ soit nul. Pour ce faire, nous définissons la matrice $Q^{(k)}$ par

$$\begin{aligned} q_{mm}^{(k)} &= q_{nn}^{(k)} = \cos\theta_k, \\ q_{mn}^{(k)} &= -q_{nm}^{(k)} = \sin\theta_k, \\ q_{ii}^{(k)} &= 1 \qquad \text{si } i \neq m \text{ et } i \neq n, \\ q_{ij}^{(k)} &= 0 \qquad \text{dans tous les autres cas,} \end{aligned} \tag{7.37}$$

où ici θ_k sera choisi de sorte à ce que $t_{mn}^{(k)}$ soit nul. Il est facile de vérifier qu'une telle matrice est orthogonale (on dit qu'il s'agit d'une matrice de rotation dans le plan des coordonnées mn). A partir des relations (7.36) et (7.37), nous obtenons :

$$\begin{aligned} t_{mn}^{(k)} &= \sum_{i,j=1}^{N} q_{im}^{(k)} t_{ij}^{(k-1)} q_{jn}^{(k)} \\ &= q_{mm}^{(k)} t_{mn}^{(k-1)} q_{nn}^{(k)} + q_{mm}^{(k)} t_{mm}^{(k-1)} q_{mn}^{(k)} \\ &\quad + q_{nm}^{(k)} t_{nn}^{(k-1)} q_{nn}^{(k)} + q_{nm}^{(k)} t_{nm}^{(k-1)} q_{mn}^{(k)}. \end{aligned}$$

Puisque les matrices $T^{(k)}$ sont symétriques et en utilisant à nouveau (7.37), nous vérifions que

$$t_{mn}^{(k)} = t_{mn}^{(k-1)} \left(\cos^2\theta_k - \sin^2\theta_k\right) + \left(t_{mm}^{(k-1)} - t_{nn}^{(k-1)}\right) \left(\cos\theta_k \sin\theta_k\right)$$

soit, en utilisant des formules trigonométriques élémentaires

$$t_{mn}^{(k)} = t_{mn}^{(k-1)} \cos 2\theta_k + \frac{1}{2} \left(t_{mm}^{(k-1)} - t_{nn}^{(k-1)}\right) \sin 2\theta_k. \tag{7.38}$$

Puisque nous voulons choisir θ_k de sorte à ce que $t_{mn}^{(k)} = 0$, nous avons

$$\operatorname{cotg} 2\theta_k = \frac{t_{nn}^{(k-1)} - t_{mm}^{(k-1)}}{2t_{mn}^{(k-1)}}. \tag{7.39}$$

Comme nous l'avons fait pour obtenir (7.38), il est facile d'obtenir les autres éléments de la matrice $T^{(k)}$ en fonction de ceux de $T^{(k-1)}$. Ainsi, lorsque θ_k

satisfait (7.39), nous pouvons établir les relations suivantes :

$$\begin{aligned}
&t_{mj}^{(k)} = t_{jm}^{(k)} = t_{mj}^{(k-1)} \cos\theta_k - t_{nj}^{(k-1)} \sin\theta_k, \quad \text{si } j \neq m \text{ et } j \neq n,\\
&t_{nj}^{(k)} = t_{jn}^{(k)} = t_{mj}^{(k-1)} \sin\theta_k + t_{nj}^{(k-1)} \cos\theta_k, \quad \text{si } j \neq m \text{ et } j \neq n,\\
&t_{mm}^{(k)} = t_{mm}^{(k-1)} \cos^2\theta_k + t_{nn}^{(k-1)} \sin^2\theta_k - 2t_{mn}^{(k-1)} \sin\theta_k \cos\theta_k,\\
&t_{nn}^{(k)} = t_{mm}^{(k-1)} \sin^2\theta_k + t_{nn}^{(k-1)} \cos^2\theta_k + 2t_{mn}^{(k-1)} \sin\theta_k \cos\theta_k,\\
&t_{mn}^{(k)} = t_{nm}^{(k)} = 0,\\
&t_{ij}^{(k)} = t_{ij}^{(k-1)}, \qquad \text{pour tous les autres couples } (i,j).
\end{aligned} \tag{7.40}$$

En pratique, nous ne calculerons jamais l'angle θ_k mais seulement les valeurs $\sin\theta_k$ et $\cos\theta_k$ puisque ce sont les seules valeurs présentes dans (7.40). Pour calculer $\sin\theta_k$ et $\cos\theta_k$, nous posons

$$T = \operatorname{tg}\theta_k \qquad \text{et} \qquad \alpha = \operatorname{cotg} 2\theta_k. \tag{7.41}$$

En vertu de (7.39) et (7.41), nous avons

$$\alpha = \frac{t_{nn}^{(k-1)} - t_{mm}^{(k-1)}}{2t_{mn}^{(k-1)}} = \frac{1}{2}\left(\frac{1}{T} - T\right) = \frac{1-T^2}{2T},$$

et par conséquent nous choisissons T comme la plus petite racine en valeur absolue de l'équation

$$T^2 + 2\alpha T - 1 = 0. \tag{7.42}$$

Ainsi nous avons

$$\begin{aligned}
&T = \sqrt{1+\alpha^2} - \alpha \qquad \text{si } \alpha \geq 0,\\
&T = -\sqrt{1+\alpha^2} - \alpha \quad \text{si } \alpha < 0.
\end{aligned} \tag{7.43}$$

Après avoir calculé T, nous pouvons facilement obtenir

$$\cos\theta_k = \frac{1}{\sqrt{1+T^2}} \qquad \text{et} \qquad \sin\theta_k = \frac{T}{\sqrt{1+T^2}}. \tag{7.44}$$

Si nous posons

$$h = Tt_{mn}^{(k-1)}, \quad C = \frac{1}{\sqrt{1+T^2}}, \quad S = TC, \quad \tau = \frac{S}{1+C}, \tag{7.45}$$

alors, compte tenu du fait que $C^2 + S^2 = 1$, nous obtenons à partir de (7.40) et (7.44) :

$$\begin{aligned}
&t_{mj}^{(k)} = t_{jm}^{(k)} = t_{mj}^{(k-1)} - S\left(t_{nj}^{(k-1)} + \tau t_{mj}^{(k-1)}\right) \quad \text{si } j \neq m \text{ et } j \neq n,\\
&t_{nj}^{(k)} = t_{jn}^{(k)} = t_{nj}^{(k-1)} + S\left(t_{mj}^{(k-1)} - \tau t_{nj}^{(k-1)}\right) \quad \text{si } j \neq m \text{ et } j \neq n,\\
&t_{mm}^{(k)} = t_{mm}^{(k-1)} - h,\\
&t_{nn}^{(k)} = t_{nn}^{(k-1)} + h,\\
&t_{mn}^{(k)} = t_{nm}^{(k)} = 0;
\end{aligned} \tag{7.46}$$

les autres éléments de $T^{(k-1)}$ restant inchangés. Un pas de l'algorithme de Jacobi pour le calcul des valeurs propres de A se décompose en étapes suivantes :

- On choisit un élément hors diagonal non nul de la matrice A ; soit $a_{mn} = a_{nm}$ cet élément.
- On calcule $\alpha = (a_{nn} - a_{mm})/2a_{mn}$, puis T par les relations (7.43), puis $h = Ta_{mn}$, $C = (1 + T^2)^{-1/2}$, $S = TC$, $\tau = S/(1 + C)$.
- On modifie les m-ième et n-ième lignes et colonnes de A en utilisant les formules (7.46), c'est-à-dire en calculant :

$$\begin{aligned}
&a_{mj} = a_{jm} := a_{mj} - S\,(a_{nj} + \tau a_{mj})\,, \qquad j \neq m, j \neq n,\\
&a_{nj} = a_{jn} := a_{nj} + S\,(a_{mj} - \tau a_{nj})\,, \qquad j \neq m, j \neq n,\\
&a_{mm} := a_{mm} - h,\\
&a_{nn} := a_{nn} + h,\\
&a_{mn} = a_{nm} := 0.
\end{aligned}$$

Il suffit de reproduire ces trois étapes jusqu'à ce que A devienne diagonale (ou presque !). Notons au passage qu'un élément a_{mn} rendu nul dans un pas de l'algorithme deviendra en général non nul dans une étape ultérieure. En effet, si cela n'était pas le cas nous pourrions diagonaliser A en un nombre fini d'opérations algébriques élémentaires, ce qui contredit ce que nous avons dit dans la section 7.1. Il reste maintenant à choisir judicieusement les éléments hors diagonaux non nuls. Plusieurs techniques sont envisageables et il est hors de propos de les détailler. Signalons simplement qu'une de ces techniques consiste à choisir à chaque pas de l'algorithme l'élément hors diagonal le plus grand ; dans ce cas il n'est pas très difficile de montrer la convergence de la méthode. Une autre technique consiste à balayer, colonne par colonne, tous les coefficients de la matrice, c'est-à-dire à examiner à tour de rôle les coefficients $a_{21}, a_{31}, a_{41}, \ldots, a_{N1}, a_{32}, a_{42}, \ldots, a_{N2}, \ldots$ Dans ce cas, il faut préciser le seuil de tolérance ε en dessous duquel le coefficient a_{mn} est considéré comme étant nul (dans l'algorithme on dira que a_{mn} est nul si $|a_{mn}| < \varepsilon$). On parle dans ce cas de ***méthode de Jacobi avec seuil***. Le choix du seuil est une question délicate.

Terminons cette section en notant qu'il est aussi possible de construire les vecteurs propres de A en calculant à chaque étape la matrice $R^{(k)}$ définie par $R^{(k)} = Q^{(1)}Q^{(2)}Q^{(3)}\ldots Q^{(k)} (= R^{(k-1)}Q^{(k)})$. La méthode de Jacobi, contrairement à la méthode de la puissance ou puissance inverse, nous oblige à calculer toutes les valeurs propres de la matrice.

7.5 Exercices

Exercice 7.1 Soit A la matrice définie par

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

On considère l'algorithme de la puissance défini par les relations (7.8) et (7.9). Le vecteur initial $\vec{x}^{(0)}$ est défini par

$$\vec{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \neq 0.$$

1. Calculer les matrices A^2, A^3 et A^4. En déduire les valeurs de $\vec{x}^{(3)}$ et $\vec{x}^{(4)}$.
2. Déduire du point 1 la valeur de $\mu^{(3)}$ dans le cas où le vecteur initial $\vec{x}^{(0)}$ a une de ses deux composantes nulle. Comparer $\mu^{(3)}$ avec la plus grande valeur propre de la matrice A.
3. On prend $x_1^{(0)} = x_2^{(0)}$. Calculer $\mu^{(3)}$ et conclure.

Solution

1. La relation (7.8) implique, pour $n = 1, 2, \ldots$, que :

$$\vec{x}^{(n)} = A\vec{x}^{(n-1)} = A^2\vec{x}^{(n-2)} = \cdots = A^n\vec{x}^{(0)}.$$

Par conséquent, $\vec{x}^{(3)} = A^3\vec{x}^{(0)}$ et $\vec{x}^{(4)} = A^4\vec{x}^{(0)}$. Calculons maintenant A^3 et A^4. Nous avons :

$$A^2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}.$$

De même nous obtenons

$$A^3 = A^2A = \begin{bmatrix} 14 & -13 \\ -13 & 14 \end{bmatrix} \quad \text{et} \quad A^4 = A^3A = \begin{bmatrix} 41 & -40 \\ -40 & 41 \end{bmatrix}.$$

Nous avons donc

$$\vec{x}^{(3)} = A^3\vec{x}^{(0)} = \begin{bmatrix} 14 & -13 \\ -13 & 14 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix},$$

$$\vec{x}^{(4)} = A^4\vec{x}^{(0)} = \begin{bmatrix} 41 & -40 \\ -40 & 41 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix}.$$

2. En utilisant l'expression (7.9) ainsi que les relations ci-dessus, nous obtenons :

$$\mu^{(3)} = \frac{\vec{x}^{(3)T}\vec{x}^{(4)}}{\|\vec{x}^{(3)}\|^2} = \frac{\vec{x}^{(0)T}A^3A^4\vec{x}^{(0)}}{\vec{x}^{(0)T}A^3A^3\vec{x}^{(0)}}.$$

D'autre part nous avons :

$$A^3A^3 = \begin{bmatrix} 14 & -13 \\ -13 & 14 \end{bmatrix} \begin{bmatrix} 14 & -13 \\ -13 & 14 \end{bmatrix} = \begin{bmatrix} 365 & -364 \\ -364 & 365 \end{bmatrix},$$

$$A^3A^4 = \begin{bmatrix} 14 & -13 \\ -13 & 14 \end{bmatrix} \begin{bmatrix} 41 & -40 \\ -40 & 41 \end{bmatrix} = \begin{bmatrix} 1094 & -1093 \\ -1093 & 1094 \end{bmatrix}.$$

Par conséquent, si $x_1^{(0)} = 0$ ou $x_2^{(0)} = 0$, nous obtenons finalement :

$$\mu^{(3)} = \frac{1094}{365} \simeq 2.997.$$

Un calcul simple nous permet d'affirmer que les valeurs propres de A sont 1 et 3. La quantité $\mu^{(3)}$ est donc déjà très proche de la plus grande valeur propre de la matrice A.

3. Si $x_1^{(0)} = x_2^{(0)}$ alors on obtient $\mu^{(3)} = 1$. Ce résultat provient du fait que $\vec{x}^{(0)}$ est un vecteur propre de A correspondant à la valeur propre 1. Ainsi $\vec{x}^{(0)}$ est orthogonal au vecteur propre correspondant à la plus grande valeur propre qui est 3 dans le cas présent.

Exercice 7.2 Soit A la matrice diagonale définie par

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

On considère l'algorithme de la puissance inverse pour déterminer les valeurs propres et vecteurs propres de la matrice A. Le vecteur initial $\vec{x}^{(0)}$ est défini par

$$\vec{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix}.$$

Montrer que si $0 < x_1^{(0)} < x_2^{(0)}$, alors l'algorithme de la puissance inverse est tel que $\lim_{n\to\infty} \mu^{(n)} = 1$. De même, montrer que si $0 < x_2^{(0)} < x_1^{(0)}$, alors l'algorithme de la puissance inverse est tel que $\lim_{n\to\infty} \mu^{(n)} = 2$.

Solution
L'algorithme de la puissance inverse consiste à calculer $\mu^{(n-1)}$ par (7.29), puis à résoudre le système linéaire (7.31). Notons $x_1^{(n-1)}$, $x_2^{(n-1)}$ les composantes du vecteur $\vec{x}^{(n-1)}$, et supposons que

$$x_1^{(n-1)} \neq 0 \qquad \text{et} \qquad x_2^{(n-1)} \neq 0, \tag{7.47}$$

(l'hypothèse est vraie pour $n = 1$). Un calcul simple donne

$$\mu^{(n-1)} = \frac{2(x_1^{(n-1)})^2 + (x_2^{(n-1)})^2}{(x_1^{(n-1)})^2 + (x_2^{(n-1)})^2} = \frac{2\left(\dfrac{x_1^{(n-1)}}{x_2^{(n-1)}}\right)^2 + 1}{\left(\dfrac{x_1^{(n-1)}}{x_2^{(n-1)}}\right)^2 + 1}. \tag{7.48}$$

D'autre part le système linéaire (7.31) s'écrit

$$\begin{bmatrix} 2 - \mu^{(n-1)} & 0 \\ 0 & 1 - \mu^{(n-1)} \end{bmatrix} \begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(n-1)} \\ x_2^{(n-1)} \end{bmatrix}.$$

Notons que la matrice ci-dessus est régulière car, en vertu de (7.47), $\mu^{(n-1)}$ est différent des valeurs 1 et 2. Nous pouvons donc résoudre le système linéaire pour obtenir

$$\frac{x_1^{(n)}}{x_2^{(n)}} = \frac{x_1^{(n-1)}}{x_2^{(n-1)}} \cdot \frac{1-\mu^{(n-1)}}{2-\mu^{(n-1)}}.$$

En utilisant (7.48), nous obtenons finalement

$$\frac{x_1^{(n)}}{x_2^{(n)}} = -\left(\frac{x_1^{(n-1)}}{x_2^{(n-1)}}\right)^3.$$

Nous en déduisons que $x_1^{(n)} \neq 0$ et $x_2^{(n)} \neq 0$, et donc l'algorithme est bien défini. De plus

$$\frac{x_1^{(n)}}{x_2^{(n)}} = -\left(\frac{x_1^{(n-1)}}{x_2^{(n-1)}}\right)^3 = \left(\frac{x_1^{(n-2)}}{x_2^{(n-2)}}\right)^{3^2} = \cdots = (-1)^n \left(\frac{x_1^{(0)}}{x_2^{(0)}}\right)^{3^n}.$$

Nous concluons en utilisant (7.48) :

$$\text{si } x_1^{(0)} < x_2^{(0)} \text{ alors } \lim_{n\to\infty} \frac{x_1^{(n)}}{x_2^{(n)}} = 0 \text{ et } \lim_{n\to\infty} \mu^{(n)} = 1,$$

$$\text{si } x_2^{(0)} < x_1^{(0)} \text{ alors } \lim_{n\to\infty} \frac{x_1^{(n)}}{x_2^{(n)}} = +\infty \text{ et } \lim_{n\to\infty} \mu^{(n)} = 2.$$

7.6 Notes bibliographiques et remarques

Sous certaines hypothèses sur la matrice A, la méthode de la puissance s'applique à des matrices non symétriques. Dans ce cas la méthode converge vers la plus grande valeur propre en module.

Il existe d'autres méthodes permettant d'approcher les valeurs propres d'une matrice symétrique, voir par exemple [28, 25, 27]. Les méthodes de Lanczos et de Givens-Householder permettent de transformer une matrice symétrique en une matrice semblable, mais tridiagonale. On calcule ensuite les valeurs propres de la matrice tridiagonale en étudiant les racines du polynôme caractéristique. Notons que la méthode de Lanczos permet d'obtenir rapidement une bonne estimation de la plus grande et de la plus petite valeur propre.

Les méthodes LR et QR, voir par exemple [28, 25, 27], permettent de calculer les valeurs propres d'une matrice A non symétrique. Pour toute matrice A régulière, il existe une décomposition $A = QR$ où Q est orthogonale et R est triangulaire supérieure. De même nous avons vu dans le chapitre 5 que si la matrice A a toutes ses sous-matrices principales régulières, alors il existe une décomposition unique $A = LR$ où L est triangulaire inférieure et R triangulaire supérieure

avec des valeurs 1 sur la diagonale (dans le contexte du calcul des valeurs propres la décomposition LU se note LR et signifie Left and Right matrices en anglais). Le principe des algorithmes LR et QR est le suivant. On pose $A_1 = A$ et on effectue sa décomposition QR (ou LR), c'est-à-dire $A_1 = Q_1 R_1$. On pose ensuite $A_2 = R_1 Q_1$. Etant donné A_k, on effectue sa décomposition QR (ou LR), c'est-à-dire $A_k = Q_k R_k$, on pose ensuite $A_{k+1} = R_k Q_k$. Il est facile de vérifier que toutes les matrices A_k sont semblables. En effet, $A_2 = R_1 Q_1 = Q_1^T A Q_1$. De même

$$A_{k+1} = R_k Q_k = Q_k^T A_k Q_k = (Q_1 \dots Q_k)^T A (Q_1 \dots Q_k).$$

Il est possible de montrer que, sous certaines conditions, la matrice A_k converge vers une matrice triangulaire supérieure lorsque k tend vers l'infini, ce qui permet de déterminer les valeurs propres de la matrice A. Les méthodes LR et QR sont très utilisées dans la pratique. Néanmoins, elles nécessitent le calcul de toutes les valeurs propres de la matrice et s'avèrent donc être des méthodes très coûteuses.

La bibliothèque numérique LAPACK [2] contient un certain nombre de méthodes permettant de résoudre des problèmes de valeurs propres. Cette bibliothèque est accessible depuis certains logiciels grand public, par exemple Matlab^{TM} ou NAG^{TM}.

Chapitre 8

Equations et systèmes d'équations non linéaires

8.1 Equations non linéaires : généralités

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continue donnée dont on veut chercher numériquement un ou plusieurs zéros $\bar{x}$, i.e. $f(\bar{x}) = 0$. Les méthodes numériques pour approcher $\bar{x}$ consistent à :

- localiser grossièrement le ou les zéros de f en procédant à des évaluations qui souvent sont de type graphique ; on note x_0 cette solution grossière ;
- construire, à partir de x_0, une suite $x_1, x_2, x_3, \ldots, x_n, \ldots$, telle que

$$\lim_{n\to\infty} x_n = \bar{x}, \qquad \text{où } \bar{x} \text{ satisfait } f(\bar{x}) = 0.$$

  On dit alors que la méthode est convergente.

Exemple 8.1 (Méthode de la bissection) Supposons que l'on connaisse deux valeurs α et β telles que $f(\alpha)f(\beta) < 0$. Ainsi f change de signe entre α et β et il existe au moins un zéro de f entre α et β. On pose $x_0 = (\alpha + \beta)/2$ le point milieu de l'intervalle d'extrémités α et β. Si $f(x_0) = 0$, alors x_0 est un zéro de f et on s'arrête. Ainsi, nous pouvons supposer que $f(x_0)$ est différent de zéro et nous construisons x_1 à partir de x_0 de la manière suivante :

- si $f(x_0)f(\alpha) > 0$ alors f change de signe entre x_0 et β et on change α qui devient $\alpha := x_0$. On pose ensuite $x_1 = (\alpha + \beta)/2$;
- si $f(x_0)f(\alpha) < 0$ alors f change de signe entre x_0 et α et on change β qui devient $\beta := x_0$. On pose ensuite $x_1 = (\alpha + \beta)/2$.

Dans la procédure ci-dessus, il suffit de remplacer x_0 par x_1 et x_1 par x_2 pour construire x_2 à partir de x_1. En répétant indéfiniment cette procédure, on construit une suite $(x_n)_{n=1}^{\infty}$ qui converge vers une valeur $\bar{x}$ telle que $f(\bar{x}) = 0$.

La figure 8.1 illustre la méthode de la bissection. Clairement si $\varepsilon = |\beta - \alpha|$ où β et α sont les valeurs de départ et si M est un entier positif donné, nous

avons

$$|\bar{x} - x_M| \leq \frac{\varepsilon}{2^{M+1}},$$

puisqu'à chaque pas l'intervalle est divisé par deux. La méthode est donc convergente.

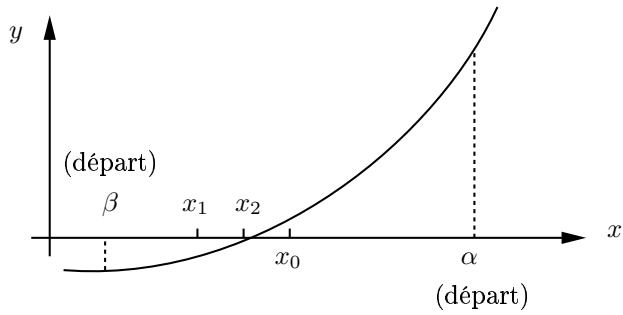

Fig. 8.1 Méthode de la bissection.

Revenons au cadre général décrit dans les premières lignes de cette section. Nous adopterons la définition suivante :

Définition 8.1 *Soit p un entier positif. On dit qu'une méthode convergente est d'ordre p s'il existe une constante C telle que*

$$|\bar{x} - x_{n+1}| \leq C|\bar{x} - x_n|^p.$$

- *Si $p = 1$ (et $C < 1$) on parle de convergence linéaire.*
- *Si $p = 2$ on parle de convergence quadratique.*
- *Si $p = 3$ on parle de convergence cubique.*
- *Si $p = 1$ et $C = C_n$, où C_n dépend de n et est tel que $\lim_{n\to\infty} C_n = 0$, on parle de convergence surlinéaire.*

Exemple 8.2 Considérons le cas où $f(x) = x - \cos x$. Trouver $\bar{x}$ tel que $f(\bar{x}) = 0$ est équivalent à trouver $\bar{x}$ tel que $\bar{x} = \cos \bar{x}$. Sur la figure 8.2, nous vérifions immédiatement qu'il n'y a qu'un seul zéro de f et que celui-ci est compris entre 0 et 1. Nous pouvons par exemple choisir $x_0 = 0.75$ comme solution grossière.

En nous inspirant de l'écriture $\bar{x} = \cos \bar{x}$, nous sommes tentés de poser

$$x_{n+1} = \cos x_n, \qquad n = 0, 1, 2, \ldots, \tag{8.1}$$

ce qui permet de construire successivement $x_1, x_2, \ldots$, à partir de x_0. Cette construction numérique s'appelle commmunément ***méthode de Picard***. Nous avons le résultat suivant.

Théorème 8.1 *Pour tout $x_0 \in \mathbb{R}$, la suite donnée par (8.1) converge vers $\bar{x}$. De plus la convergence est linéaire.*

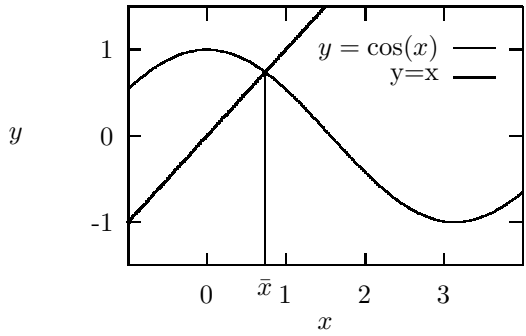

Fig. 8.2 Le point $\bar{x}$ tel que $\bar{x} = \cos \bar{x}$.

Démonstration
Il n'est pas restrictif de supposer $x_0 \in [0, 1]$. En effet, si cela n'était pas le cas, nous aurions $x_1 = \cos x_0 \in [-1, +1]$, $x_2 = \cos x_1 \in [0, 1]$ et il suffirait de partir de x_2 au lieu de x_0. Supposons donc $x_0 \in [0, 1]$. Puisque $\bar{x} = \cos \bar{x}$ nous obtenons, en utilisant (8.1) :

$$|\bar{x} - x_{n+1}| = |\cos \bar{x} - \cos x_n| = \left| \int_{\bar{x}}^{x_n} \sin t dt \right|.$$

Puisque $x_0 \in [0, 1]$, nous avons $x_1 = \cos x_0 \in [0, 1]$. Nous pouvons donc montrer que $x_{n+1} = \cos x_n \in [0, 1]$, $n = 1, 2, \ldots$, et par conséquent

$$|\bar{x} - x_{n+1}| \leq \max_{t \in [0,1]} |\sin t| \cdot |\bar{x} - x_n|.$$

Posons $\chi = \max_{t \in [0,1]} |\sin t|$, nous avons donc

$$|\bar{x} - x_{n+1}| \leq \chi |\bar{x} - x_n|, \tag{8.2}$$

et, par induction

$$|\bar{x} - x_n| \leq \chi^n |\bar{x} - x_0|. \tag{8.3}$$

La relation (8.3) et le fait que $\chi < 1$ impliquent la convergence de la suite $(x_n)_{n=1}^{\infty}$ vers $\bar{x}$. De plus, (8.2) montre que cette convergence est linéaire. ■

Considérons encore l'exemple 8.2. Nous dirons que $\bar{x}$ est un point fixe de $x \to \cos x$ car nous avons $\bar{x} = \cos \bar{x}$. La méthode (8.1) est aussi appelée ***méthode de point fixe***. Dans la section suivante, nous présentons quelques généralités sur les méthodes de point fixe.

8.2 Méthodes de point fixe : généralités

Une méthode de point fixe pour résoudre numériquement $f(x) = 0$ consiste, dans une première phase, à transformer le problème $f(x) = 0$ en un problème équivalent (admettant les mêmes solutions) du type

$$x = g(x). \tag{8.4}$$

Clairement, il existe une infinité de manières pour opérer cette transformation. Par exemple, on peut poser, en s'inspirant de l'exemple 8.2

$$g(x) = x - f(x),$$

ou plus généralement

$$g(x) = x + \alpha f(x) \tag{8.5}$$

avec $\alpha \in \mathbb{R}$, $\alpha \neq 0$ quelconque. Dans (8.5), on peut même prendre pour α une fonction de x pour autant qu'elle ne s'annule pas.

Définition 8.2 *Si $\bar{x} \in \mathbb{R}$ est tel que $\bar{x} = g(\bar{x})$, on dira que $\bar{x}$ est un point fixe de g ; l'image de $\bar{x}$ par g est $\bar{x}$ lui-même.*

Supposons donc que $\bar{x} \in \mathbb{R}$ soit un zéro de f ou, de façon équivalente, un point fixe de g. Alors une méthode de point fixe consiste à :

- évaluer une approximation x_0 du point fixe $\bar{x}$ de g ;
- calculer successivement $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$

Naturellement toute méthode de point fixe n'est pas forcément convergente. Par contre, si elle converge, c'est-à-dire si la suite $(x_n)_{n=0}^{\infty}$ a une limite que nous notons x, et si g est continue, alors cette limite x est nécessairement un point fixe de g puisque

$$x = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g(x).$$

Avant d'établir des critères de convergence de la méthode ci-dessus, nous donnons une définition concernant la fonction g.

Définition 8.3 *Soit I un intervalle de $\mathbb{R}$ et soit $g : x \in I \to g(x) \in \mathbb{R}$ une fonction. Nous dirons que g est une contraction stricte sur I s'il existe une constante $\chi < 1$ telle que*

$$|g(x) - g(y)| \leq \chi |x - y|, \qquad \forall x, y \in I. \tag{8.6}$$

Remarquons que si g est une contraction stricte, alors elle est nécessairement continue puisque, si nous prenons une suite $(y_n)_{n=1}^{\infty}$ dans I qui converge vers un élément x de I, alors nous avons $|g(x) - g(y_n)| \leq \chi |x - y_n|$ et par suite $\lim_{n\to\infty} g(y_n) = g(x)$. Nous sommes maintenant en mesure d'énoncer un premier résultat concernant la convergence des méthodes de point fixe.

Théorème 8.2 *Soit I un intervalle* fermé *de $\mathbb{R}$ et $g : I \to \mathbb{R}$ une fonction donnée. On suppose que g satisfait les deux propriétés suivantes :*

i) g est une contraction stricte sur I,

ii) $g(I) \subset I$, c'est-à-dire pour tout $x \in I$ on a $g(x) \in I$.

Alors g a un et un seul point fixe $\bar{x}$ dans I et, pour tout $x_0 \in I$, la suite $(x_n)_{n=0}^{\infty}$ donnée par

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots, \tag{8.7}$$

converge vers $\bar{x}$ lorsque n tend vers l'infini. De plus la convergence est linéaire.

Démonstration
Supposons donc que g satisfait les propriétés $i)$ et $ii)$. Si $x_0 \in I$ est quelconque et si $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, la propriété $ii)$ garantit que tous les x_n, $n = 1, 2, \ldots$, sont dans l'intervalle I. La propriété $i)$ permet d'écrire

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq \chi |x_n - x_{n-1}|$$

où $\chi < 1$ est la constante de contraction qui intervient dans la définition 8.3. En itérant n fois cette relation nous obtiendrons

$$|x_{n+1} - x_n| \leq \chi^n |x_1 - x_0|, \qquad n = 0, 1, 2, \ldots. \tag{8.8}$$

Soit maintenant un entier m positif et évaluons

$$\begin{aligned} |x_{n+m} - x_n| &= |x_{n+m} - x_{n+m-1} + x_{n+m-1} - x_{n+m-2} + x_{n+m-2} \\ &\qquad - \cdots + x_{n+1} - x_n| \\ &\leq |x_{n+m} - x_{n+m-1}| + |x_{n+m-1} - x_{n+m-2}| \\ &\qquad + \cdots + |x_{n+1} - x_n|. \end{aligned}$$

En utilisant (8.8) nous obtenons

$$\begin{aligned} |x_{n+m} - x_n| &\leq \left(\chi^{n+m-1} + \chi^{n+m-2} + \ldots + \chi^n\right) |x_1 - x_0| \\ &\leq \chi^n \left(1 + \chi + \chi^2 + \ldots + \chi^{m-1}\right) |x_1 - x_0|, \end{aligned}$$

et ainsi

$$|x_{n+m} - x_n| \leq \chi^n \frac{1 - \chi^m}{1 - \chi} |x_1 - x_0|. \tag{8.9}$$

Etant donné que χ est plus petit que 1, on déduit de (8.9) que la suite $(x_n)_{n=0}^{\infty}$ est une suite de Cauchy et donc converge. Notons $\bar{x}$ sa limite, c'est-à-dire $\lim_{n\to\infty} x_n = \bar{x}$. Alors $\bar{x}$ est nécessairement dans I puisque nous avons supposé que I est un intervalle fermé. Puisque g est continue, alors $\bar{x}$ est un point fixe de g car, pour le montrer, il suffit de prendre la limite lorsque n tend vers l'infini dans (8.7) (nous l'avons déjà dit !). La convergence est linéaire (définition 8.1) car pour tout n :

$$|\bar{x} - x_{n+1}| = |g(\bar{x}) - g(x_n)| \leq \chi |\bar{x} - x_n|. \tag{8.10}$$

De plus $\bar{x}$ est le seul point fixe de g dans I car si nous supposons qu'il y en a un autre dans I noté $\bar{y}$, nous avons $|\bar{x} - \bar{y}| = |g(\bar{x}) - g(\bar{y})| \leq \chi |\bar{x} - \bar{y}|$ qui implique $\bar{y} = \bar{x}$ (puisque χ est plus petit que 1). Nous avons donc montré le théorème 8.2. ■

Revenons à l'exemple 8.2 traité au début du chapitre et cherchons à appliquer le théorème 8.2. La fonction $g(x) = \cos x$ envoie l'intervalle fermé $I = [0, 1]$ dans lui-même. De plus nous avons, pour tout $x, y \in I$:

$$\begin{aligned} |g(x) - g(y)| = |\cos x - \cos y| &= \left| \int_x^y \sin t dt \right| \\ &\leq \max_{t \in [0,1]} |\sin t| \cdot |x - y|. \end{aligned}$$

En posant $\chi = \max_{t\in[0,1]} |\sin t|$, nous obtenons $\chi < 1$ et ainsi g est une contraction stricte sur I. Le théorème 8.2 s'applique donc et assure l'existence d'un seul point fixe $\bar{x}$ dans I, ainsi que la convergence vers $\bar{x}$ de la suite donnée par $x_{n+1} = \cos x_n$, ceci pour tout $x_0 \in [0,1]$. En tenant compte de la remarque faite au début de la démonstration du théorème 8.1, nous obtenons facilement la convergence pour tout $x_0 \in \mathbb{R}$.

Nous sommes maintenant en mesure d'énoncer un autre théorème de convergence sur les méthodes de point fixe.

Théorème 8.3 *Supposons $g : \mathbb{R} \to \mathbb{R}$ une fois continûment dérivable et soit $\bar{x}$ un point fixe de g, i.e. $\bar{x} = g(\bar{x})$. Si $|g'(\bar{x})| < 1$, alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, alors la suite donnée par*

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots, \tag{8.11}$$

converge vers $\bar{x}$ lorsque n tend vers l'infini. De plus la convergence est linéaire.

Démonstration
Si $|g'(\bar{x})| < 1$, alors par continuité de g', il existe $\varepsilon > 0$ et $\chi < 1$ tels que

$$|g'(x)| \leq \chi, \qquad \text{pour tout } x \in [\bar{x} - \varepsilon, \bar{x} + \varepsilon]. \tag{8.12}$$

Posons $I = [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$ et soit x, y deux éléments quelconques de I. En utilisant (8.12) nous avons

$$|g(x) - g(y)| = \left| \int_y^x g'(t)dt \right| \leq \max_{t\in I} |g'(t)| \cdot |x - y| \leq \chi |x - y|. \tag{8.13}$$

Si $y = \bar{x}$, nous tirons de (8.13) que

$$|g(x) - \bar{x}| = |g(x) - g(\bar{x})| \leq \chi |x - \bar{x}| \leq |x - \bar{x}| \leq \varepsilon, \tag{8.14}$$

ce qui prouve que $g(x)$ appartient à I dès que x appartient à I.

Les relations (8.13) et (8.14) impliquent que g satisfait les hypothèses du théorème 8.2. Ainsi la suite donnée par (8.11) est convergente et la convergence est linéaire. Le théorème 8.3 est donc une conséquence directe du théorème 8.2. ■

8.3 Méthode de Newton et méthode de la corde

Méthode de Newton (ou Newton-Raphson)

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction continûment dérivable et soit $\bar{x}$ un zéro simple de f, c'est-à-dire $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$. Supposons que l'on connaisse une valeur x_n proche de $\bar{x}$. Pour calculer x_{n+1} nous prenons l'intersection de l'axe Ox avec la droite tangente au graphe de f passant par le point $(x_n, f(x_n))$, comme cela est indiqué sur la construction graphique de la figure 8.3.

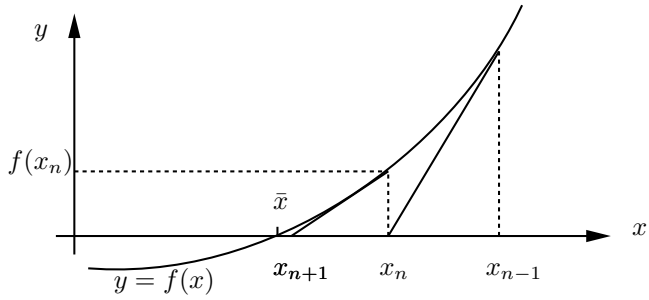

Fig. 8.3 Méthode de Newton.

Clairement, nous avons la relation $f(x_n)/(x_n - x_{n+1}) = f'(x_n)$ qui donne, lorsque x_0 est choisi proche de $\bar{x}$, la méthode de Newton :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots. \tag{8.15}$$

Nous voyons ainsi que la méthode de Newton est une méthode de point fixe pour calculer $\bar{x}$. En effet, il suffit de constater que si on pose

$$g(x) = x - \frac{f(x)}{f'(x)},$$

alors $f(x) = 0$ est équivalent à $x = g(x)$ (du moins dans un voisinage de $\bar{x}$ pour lequel $f'(x) \neq 0$) et (8.15) est équivalent à $x_{n+1} = g(x_n)$.

En vue d'utiliser le théorème 8.3, calculons $g'(x)$, puis $g'(\bar{x})$. Nous vérifions que si f est deux fois continûment dérivable :

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} \tag{8.16}$$

et par suite, puisque $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$

$$g'(\bar{x}) = 0. \tag{8.17}$$

Nous obtenons ainsi le résultat suivant.

Théorème 8.4 *Supposons f deux fois continûment dérivable et supposons que $\bar{x}$ soit tel que $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$. Alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, la suite $(x_n)_{n=0}^{\infty}$ donnée par la méthode de Newton (8.15) converge vers $\bar{x}$. De plus la convergence est quadratique.*

Démonstration
Nous avons observé que la méthode de Newton est une méthode de point fixe avec $g(x) = x - f(x)/f'(x)$ et que $|g'(\bar{x})| < 1$ en vertu de la relation (8.17). Ainsi le résultat de convergence annoncé dans ce théorème est une conséquence du théorème 8.3. A priori la convergence est linéaire. Nous allons maintenant

démontrer que la convergence est quadratique, ceci étant une conséquence du fait que $g'(\bar{x}) = 0$.

Si nous développons f autour du point x_n nous obtenons :

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(\xi_x)}{2}(x - x_n)^2$$

où ξ_x appartient à l'intervalle d'extrémités x et x_n. En choisissant $x = \bar{x}$ dans l'égalité ci-dessus, en divisant par $f'(x_n)$ et en tenant compte du fait que $f(\bar{x}) = 0$, nous avons

$$\frac{f(x_n)}{f'(x_n)} + \bar{x} - x_n + \frac{f''(\xi_{\bar{x}})}{2f'(x_n)}(\bar{x} - x_n)^2 = 0.$$

En utilisant (8.15) nous obtenons

$$|\bar{x} - x_{n+1}| = \frac{|f''(\xi_{\bar{x}})|}{2|f'(x_n)|}|\bar{x} - x_n|^2.$$

Il suffit maintenant de poser

$$C = \frac{\max\limits_{x\in[\bar{x}-\varepsilon,\bar{x}+\varepsilon]} |f''(x)|}{2 \min\limits_{x\in[\bar{x}-\varepsilon,\bar{x}+\varepsilon]} |f'(x)|}$$

pour obtenir :

$$|\bar{x} - x_{n+1}| \leq C|\bar{x} - x_n|^2.$$

Cette dernière inégalité montre que la convergence est bien quadratique. ■

Remarque 8.1 Le théorème 8.4 nous assure que si x_0 est choisi suffisamment proche de $\bar{x}$ et si f satisfait certaines hypothèses, alors la méthode de Newton converge quadratiquement. Il est possible d'affaiblir les hypothèses sur f et d'obtenir encore la convergence de la méthode. Par exemple si $f'(\bar{x}) = 0$, la méthode de Newton converge encore (pour autant que x_0 soit différent de $\bar{x}$), mais la convergence n'est plus que linéaire.

Méthode de la corde (ou Newton-corde)

Cette méthode permet d'éviter qu'à chaque itération de (8.15) on ait à évaluer $f'(x_n)$. La méthode de la corde consiste à remplacer $f'(x_n)$ par $f'(x_0)$ dans (8.15), ce qui donne :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \qquad n = 0, 1, 2, \ldots. \tag{8.18}$$

L'interprétation géométrique de cette méthode est suggérée dans la figure 8.4. Le calcul de la suite $(x_n)_{n=0}^{\infty}$ s'effectue en prenant toujours la même pente $f'(x_0)$, d'où l'appellation méthode de la corde.

Ici encore, nous posons $g(x) = x - f(x)/f'(x_0)$ et constatons que $f(\bar{x}) = 0$ si $\bar{x} = g(\bar{x})$. Ainsi (8.18) s'écrit $x_{n+1} = g(x_n)$ et la méthode de la corde est une méthode de point fixe. Remarquons toutefois que g dépend du point de départ x_0.

Nous avons le résultat suivant.

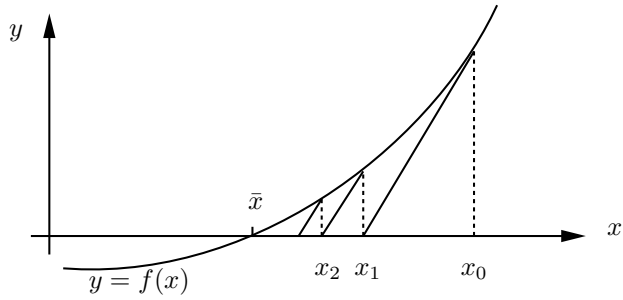

Fig. 8.4 Méthode de la corde.

Théorème 8.5 *Supposons f une fois continûment dérivable et supposons que $\bar{x}$ soit tel que $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$. Alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, la suite $(x_n)_{n=0}^{\infty}$ donnée par la méthode de la corde (8.18) converge vers $\bar{x}$. La convergence est linéaire.*

Démonstration
Puisque f est une fois continûment dérivable et puisque $f'(\bar{x}) \neq 0$, il est facile de montrer qu'il existe $\varepsilon > 0$ et $\chi < 1$ tels que si x_0 appartient à l'intervalle $I = [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$, alors on a :

$$|g'(x)| = \left|1 - \frac{f'(x)}{f'(x_0)}\right| \leq \chi, \qquad \text{pour tout } x \in I.$$

En utilisant les mêmes arguments qui nous ont permis de déduire (8.13) et (8.14) de (8.12), nous pouvons vérifier que les hypothèses du théorème 8.2 sont satisfaites. Ainsi le résultat annoncé dans le théorème 8.5 est une conséquence immédiate du théorème 8.2. ■

8.4 Systèmes non linéaires

Soit une application continue $f : \mathbb{R}^N \to \mathbb{R}^N$ (N entier positif) dont on cherche un ou des zéros $\bar{x}$, i.e. $f(\bar{x}) = 0$. En fait, si $x \in \mathbb{R}^N$, alors x a N composantes $x_1, x_2, \ldots, x_N$ et x peut être vu comme un vecteur (on ne mettra pas de flêche sur x pour ne pas alourdir les notations), c'est-à-dire

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}.$$

Si $x \in \mathbb{R}^N$, alors $f(x) \in \mathbb{R}^N$ et $f(x)$ est un N-vecteur. Chaque composante f_j, $1 \leq j \leq N$, de f est une fonction définie sur $\mathbb{R}^N$ et à valeur dans $\mathbb{R}$. Nous

noterons ainsi :

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_N(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_N) \\ f_2(x_1, x_2, \ldots, x_N) \\ \vdots \\ f_N(x_1, x_2, \ldots, x_N) \end{bmatrix}.$$

L'équation

$$f(x) = 0 \tag{8.19}$$

est un système de N équations (non linéaires en principe) à N inconnues x_1, x_2, ..., x_N et peut s'écrire

$$\begin{aligned} &f_1(x_1, x_2, \ldots, x_N) = 0, \\ &f_2(x_1, x_2, \ldots, x_N) = 0, \\ &\vdots \\ &f_N(x_1, x_2, \ldots, x_N) = 0. \end{aligned} \tag{8.20}$$

Si $f_1, f_2, \ldots, f_N$ sont continûment dérivables, alors on définit la $N \times N$ matrice jacobienne $Df(x)$ de f au point $x \in \mathbb{R}^N$ de la façon suivante :

$$Df(x) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(x) & \dfrac{\partial f_1}{\partial x_2}(x) & \ldots & \dfrac{\partial f_1}{\partial x_N}(x) \\ \dfrac{\partial f_2}{\partial x_1}(x) & \dfrac{\partial f_2}{\partial x_2}(x) & \ldots & \dfrac{\partial f_2}{\partial x_N}(x) \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_N}{\partial x_1}(x) & \dfrac{\partial f_N}{\partial x_2}(x) & \ldots & \dfrac{\partial f_N}{\partial x_N}(x) \end{bmatrix};$$

les coefficients de la matrice $Df(x)$ sont donc

$$Df(x)_{ij} = \frac{\partial f_i}{\partial x_j}(x), \qquad 1 \leq i, j \leq N. \tag{8.21}$$

La méthode de Newton se généralise aux systèmes non linéaires de la manière suivante :

$$x^{n+1} = x^n - Df(x^n)^{-1} f(x^n), \qquad n = 0, 1, 2, \ldots, \tag{8.22}$$

où le vecteur de départ x^0 est choisi suffisamment proche du vecteur $\bar{x}$, solution de $f(\bar{x}) = 0$. Remarquons que dans (8.22) nous avons mis les indices n des différents itérés x^n en haut du symbole x pour ne pas les confondre avec les indices décrivant les composantes de x.

En définissant la fonction $g : \mathbb{R}^N \to \mathbb{R}^N$ par

$$g(x) = x - Df(x)^{-1} f(x), \tag{8.23}$$

nous pouvons vérifier que si $\bar{x}$ satisfait $f(\bar{x}) = 0$ et si $Df(\bar{x})$ est une matrice régulière, alors $\bar{x}$ est aussi tel que $\bar{x} = g(\bar{x})$. Dans ce cas nous dirons encore que

$\bar{x}$ est un point fixe de g, comme dans la définition 8.2. Les théorèmes 8.2 et 8.3 restent encore vrais pour autant que nous remplacions I par un ensemble fermé, les valeurs absolues $|\cdot|$ par la norme euclidienne $\|\cdot\|$ et la valeur absolue de $g'(\bar{x})$ par la norme spectrale de $Dg(\bar{x})$. Ainsi on peut encore montrer que si f est deux fois continûment dérivable par rapport aux variables x_j, $1 \leq j \leq N$, si $\bar{x}$ est tel que $f(\bar{x}) = 0$ et si $Df(\bar{x})$ est une $N \times N$ matrice régulière, alors la suite $(x^n)_{n=0}^{\infty}$ définie par la méthode de Newton (8.22) converge vers $\bar{x}$ lorsque n tend vers l'infini, ceci pour autant que x^0 soit choisi suffisamment proche de $\bar{x}$. En d'autres termes, si la quantité

$$\|x^0 - \bar{x}\| = \left(\sum_{j=1}^{N} (x_j^0 - \bar{x}_j)^2 \right)^{1/2}$$

est suffisamment petite, alors

$$\lim_{n \to \infty} \|x^n - \bar{x}\| = 0.$$

De plus la convergence est quadratique, c'est-à-dire qu'il existe une constante C telle que

$$\| x^{n+1} - \bar{x} \| \leq C \| x^n - \bar{x} \|^2, \qquad n = 0, 1, 2, \ldots. \tag{8.24}$$

Pour calculer x^{n+1} à partir de x^n, on écrit (8.22) sous la forme suivante :

$$Df(x^n)(x^n - x^{n+1}) = f(x^n), \qquad n = 0, 1, 2, \ldots, \tag{8.25}$$

et, en pratique, on procède comme indiqué ci-dessous :

- on construit le vecteur $\vec{b} = f(x^n)$;
- on construit la matrice $A = Df(x^n)$;
- on résout le système $A\vec{y} = \vec{b}$ par élimination de Gauss (chap.4) ou décomposition LU (ou LL^T) de A (chap. 5), ou encore grâce à une méthode itérative (chap. 6) ;
- on pose $x^{n+1} = x^n - \vec{y}$.

L'inconvénient majeur de cette méthode réside dans le fait qu'à chaque pas de la méthode, on ait à construire la matrice $Df(x^n)$ et à procéder à la résolution d'un système linéaire. Pour contourner cet inconvénient, on peut utiliser ***la méthode de la corde*** qui, pour un système non linéaire, s'écrit :

$$x^{n+1} = x^n - Df(x^0)^{-1} f(x^n), \qquad n = 0, 1, 2, \ldots. \tag{8.26}$$

Ainsi on peut construire une fois pour toutes la matrice $Df(x^0)$ et en faire sa décomposition LU (ou LL^T si la matrice est symétrique définie positive). Ensuite, lors de chaque itération, on calcule x^{n+1} à partir de x^n de la façon suivante :

- on construit le vecteur $\vec{b} = f(x^n)$;
- on résout le système triangulaire $L\vec{z} = \vec{b}$;

- on résout le système triangulaire $U\vec{y} = \vec{z}$;
- on pose $x^{n+1} = x^n - \vec{y}$.

A chaque pas de la méthode de la corde, il suffit donc de résoudre deux systèmes linéaires triangulaires. Cette méthode semble donc, de prime abord, considérablement plus économique que la méthode de Newton, ce qui n'est pas forcément le cas. En effet, la convergence de la méthode de la corde est linéaire alors que la méthode de Newton converge quadratiquement. On devra donc, à précision égale, faire plus d'itérations avec la méthode de la corde qu'avec la méthode de Newton !

8.5 Exercices

Exercice 8.1 Soit λ un nombre positif donné. Il s'agit de trouver un nombre réel $\bar{x}$ tel que

$$\bar{x} = \lambda e^{\bar{x}}. \tag{8.27}$$

1. Montrer graphiquement que, si $\lambda < 1/e$, alors (8.27) a deux solutions positives $\bar{x}_1 < \bar{x}_2$, si $\lambda > 1/e$, alors (8.27) n'a pas de solution, si $\lambda = 1/e$, alors (8.27) a une seule solution.
2. Dans le cas où $\lambda < 1/e$, on propose d'approcher numériquement $\bar{x}_1$ et $\bar{x}_2$ à l'aide des méthodes suivantes. Etant donné une valeur x_0, on calcule

$$x_{n+1} = \lambda e^{x_n}, \qquad n = 0, 1, 2, \ldots, \tag{8.28}$$

ou bien

$$x_{n+1} = \ln x_n - \ln \lambda, \qquad n = 0, 1, 2, \ldots. \tag{8.29}$$

Montrer que la suite définie par (8.28) converge vers $\bar{x}_1$ lorsque x_0 est choisi proche de $\bar{x}_1$ et que la suite définie par (8.29) converge vers $\bar{x}_2$ lorsque x_0 est choisi proche de $\bar{x}_2$.

Solution
1. Le graphe de la fonction $x \to \lambda e^x$ est représenté dans la figure 8.5, pour $\lambda = 1 > 1/e$, $\lambda = 1/e$, $\lambda = 1/e^2 < 1/e$. Pour $\lambda = 1/e$, les graphes des fonctions $x \to \lambda e^x$ et $x \to x$ sont tangents en $x = 1$. On vérifie que, si $\lambda > 1/e$, alors l'équation $\bar{x} = \lambda e^{\bar{x}}$ n'a pas de solution ; si $\lambda = 1/e$, alors l'équation $\bar{x} = \lambda e^{\bar{x}}$ a une solution $\bar{x} = 1$; si $\lambda < 1/e$, alors l'équation $\bar{x} = \lambda e^{\bar{x}}$ a deux solutions, l'une strictement plus petite que 1 (il s'agit de $\bar{x}_1$), l'autre strictement plus grande (il s'agit de $\bar{x}_2$).

2. Considérons le cas où $\lambda < 1/e$. Nous cherchons à approcher $\bar{x}_1$ à l'aide du schéma (8.28) et $\bar{x}_2$ à l'aide du schéma (8.29). Considérons d'abord le schéma (8.28). Il s'agit d'une méthode de point fixe $x_{n+1} = g(x_n)$, où la fonction g est définie par $g(x) = \lambda e^x$. La dérivée de g est donc donnée par $g'(x) = \lambda e^x$ et nous avons donc, puisque $\bar{x}_1$ est un point fixe de g :

$$|g'(\bar{x}_1)| = \lambda e^{\bar{x}_1} = \bar{x}_1 < 1.$$

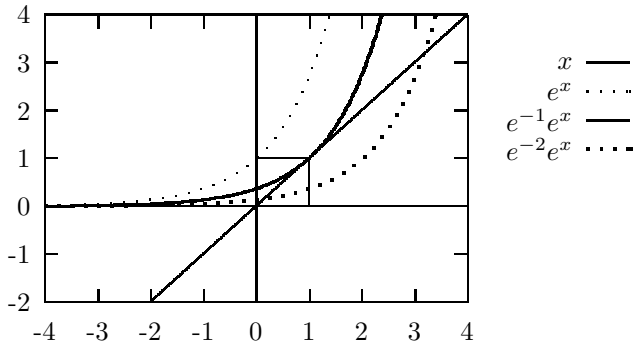

Fig. 8.5 Le graphe de la fonction $x \to \lambda e^x$ pour $\lambda = 1$, $\lambda = 1/e$, $\lambda = 1/e^2$.

Nous pouvons alors appliquer le théorème 8.3 et nous obtenons la convergence de la suite donnée par $x_{n+1} = g(x_n)$ vers $\bar{x}_1$, pour autant que x_0 soit choisi suffisamment proche de $\bar{x}_1$. En fait, nous pouvons montrer que la méthode (8.28) converge si $x_0 < \bar{x}_2$, où $\bar{x}_2$ est l'autre point fixe de g. Par contre, la suite $(x_n)_{n=0}^{\infty}$ définie par (8.28) diverge dès que x_0 est choisi plus grand que $\bar{x}_2$.

Considérons maintenant le schéma (8.29) pour approcher $\bar{x}_2$. Il s'agit encore d'une méthode de point fixe $x_{n+1} = g(x_n)$, où la fonction g est maintenant définie par $g(x) = \ln x - \ln \lambda$. Notons que $\bar{x}$ est un point fixe de g si et seulement si $\bar{x} = \ln \bar{x} - \ln \lambda$, c'est-à-dire si et seulement si

$$e^{\bar{x}} = e^{\ln \bar{x} - \ln \lambda} = \frac{\bar{x}}{\lambda},$$

soit encore si et seulement si $\bar{x}$ satisfait (8.27). La dérivée de g est maintenant donnée par $g'(x) = 1/x$ et nous avons donc, puisque $\bar{x}_2 > 1$:

$$|g'(\bar{x}_2)| = \frac{1}{\bar{x}_2} < 1.$$

Nous pouvons, une fois encore, appliquer le théorème 8.3 et nous obtenons la convergence de la suite donnée par $x_{n+1} = g(x_n)$ vers $\bar{x}_2$, pour autant que x_0 soit choisi suffisamment proche de $\bar{x}_2$. En fait, nous pouvons montrer que la méthode (8.29) converge si $x_0 > \bar{x}_1$ mais, par contre, diverge lorsque x_0 est choisi plus petit que $\bar{x}_1$.

Exercice 8.2 Il s'agit de trouver le zéro réel $\bar{x}$ du polynôme $f(x) = x^3 - x - 3$ en utilisant la méthode de Newton.

1. Expliciter la méthode de Newton dans ce cas particulier.
2. Effectuer quelques itérations de la méthode à partir du point de départ $x_0 = 1$, puis $x_0 = 0$. Interprétez les résultats obtenus.

Solution
1. La méthode de Newton s'écrit, si on pose $f(x) = x^3 - x - 3$:

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\ &= x_n - \frac{x_n^3 - x_n - 3}{3x_n^2 - 1} \\ &= \frac{2x_n^3 + 3}{3x_n^2 - 1}, \end{aligned}$$

pour $n = 1, 2, \ldots$

2. Les quantités x_0, x_1, x_2, x_3 ainsi que le graphe de f sont représentés dans les figures 8.6 et 8.7, lorsque $x_0 = 1$ et $x_0 = 0$. Nous savons, d'après le théorème 8.4, que si x_0 est choisi suffisamment proche du zéro réel $\bar{x}$ du polynôme $f(x)$, alors la méthode de Newton converge. De plus, puisque la dérivée du polynôme $f(x)$ ne s'annulle pas en $\bar{x}$, la méthode converge quadratiquement, c'est-à-dire très rapidement. C'est le cas lorsqu'on choisit $x_0 = 1$, comme sur la figure 8.6. Par contre lorsqu'on choisit $x_0 = 0$, les quantités x_n, $n = 1, 2, \ldots$, ne convergent pas vers le zéro du polynôme. Nous dirons dans ce cas que la méthode de Newton diverge.

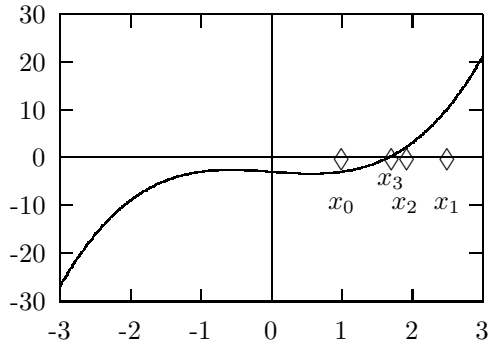

Fig. 8.6 Graphe du polynôme $f(x) = x^3 - x - 3$ ainsi que les valeurs obtenues par la méthode de Newton avec $x_0 = 1$.

Exercice 8.3 Si A est un nombre positif donné, on considère l'algorithme suivant : étant donné une valeur x_0, on calcule

$$x_{n+1} = x_n + \frac{1}{2}(A - x_n^2), \qquad n = 0, 1, 2, \ldots. \tag{8.30}$$

1. Montrer que si la suite $(x_n)_{n=0}^{\infty}$ converge, alors sa limite est soit $\sqrt{A}$ soit $-\sqrt{A}$.
2. On considère le cas où $A \in]0, 4[$. Montrer qu'il existe $\varepsilon > 0$ tel que, si $|x_0 - \sqrt{A}| \leq \varepsilon$, alors la suite $(x_n)_{n=0}^{\infty}$ converge vers $\sqrt{A}$.

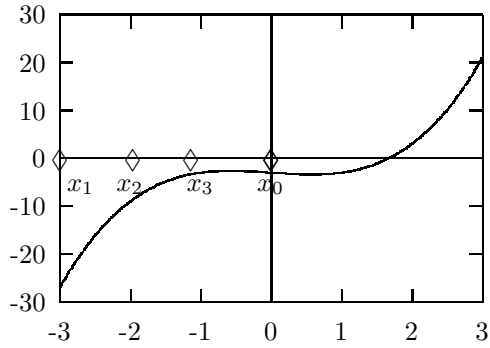

Fig. 8.7 Graphe du polynôme $f(x) = x^3 - x - 3$ ainsi que les valeurs obtenues par la méthode de Newton avec $x_0 = 0$.

3. Vérifier graphiquement que si x_0 est proche de $-\sqrt{A}$ mais différent de $-\sqrt{A}$, alors la suite $(x_n)_{n=0}^{\infty}$ ne converge pas vers $-\sqrt{A}$.
4. Vérifier que si $x_0 = 1$, alors l'algorithme (8.30) coïncide avec la méthode de Newton-corde pour résoudre $x^2 - A = 0$.
5. Proposer un algorithme plus efficace pour calculer la racine carrée d'un nombre positif A.

Solution

1. Supposons que la suite $(x_n)_{n=0}^{\infty}$ converge vers $\bar{x}$. En passant à la limite dans (8.30) nous obtenons donc

$$\bar{x} = \bar{x} + \frac{1}{2}(A - \bar{x}^2),$$

c'est-à-dire $\bar{x}^2 = A$ et donc $\bar{x} = \pm\sqrt{A}$.

2. La méthode (8.30) peut s'écrire sous la forme $x_{n+1} = g(x_n)$, la fonction g étant définie par

$$g(x) = x + \frac{1}{2}(A - x^2).$$

Considérons le cas où $\bar{x} = \sqrt{A}$ et cherchons à appliquer le théorème 8.3. Puisque $g'(x) = 1 - x$, nous avons donc

$$g'(\bar{x}) = 1 - \bar{x} = 1 - \sqrt{A}.$$

Si nous supposons que $0 < A < 4$, alors nous obtenons

$$|g'(\bar{x})| = |1 - \sqrt{A}| < 1.$$

En vertu du théorème 8.3 il existe donc $\varepsilon > 0$ tel que si $|\bar{x} - x_0| < \varepsilon$, alors la suite $(x_n)_{n=0}^{\infty}$ donnée par (8.30) converge vers $\bar{x} = \sqrt{A}$.

3. Nous avons représenté dans la figure 8.8 le graphe de la fonction g lorsque $A = 2$. Si nous choisissons $x_0 < -\sqrt{A}$, x_0 proche de $-\sqrt{A}$, alors nous vérifions

graphiquement que la suite $(x_n)_{n=0}^{\infty}$ diverge. Si nous choisissons $x_0 > -\sqrt{A}$, x_0 proche de $-\sqrt{A}$, alors nous vérifions graphiquement que la suite $(x_n)_{n=0}^{\infty}$ converge vers $\sqrt{A}$.

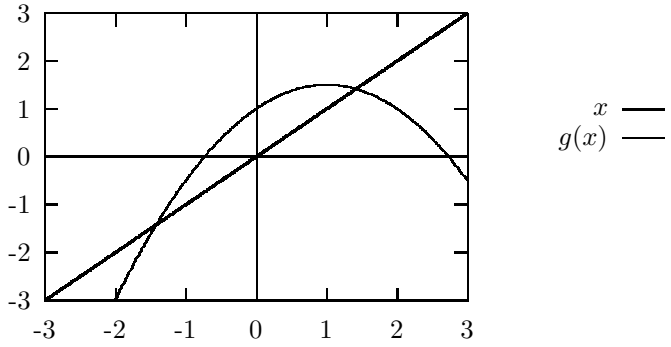

Fig. 8.8 Graphe de la fonction $g(x) = x + (A - x^2)/2$ lorsque $A = 2$.

4. Soit f la fonction définie par $f(x) = x^2 - A$. La méthode de Newton-corde s'écrit dans ce cas

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} = x_n - \frac{x_n^2 - A}{2x_0}, \qquad n = 0, 1, 2, \ldots.$$

Si nous choisissons $x_0 = 1$, cette méthode s'écrit donc

$$x_{n+1} = x_n - \frac{x_n^2 - A}{2} = x_n + \frac{1}{2}(A - x_n^2), \qquad n = 0, 1, 2, \ldots.$$

Ainsi nous concluons en affirmant que la méthode (8.30) coïncide avec la méthode de Newton-corde pour résoudre $f(x) = x^2 - A = 0$, avec $x_0 = 1$ comme point de départ.

5. Si nous choisissons la méthode de Newton pour résoudre $f(x) = x^2 - A = 0$, alors nous avons

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - A}{2x_n} = \frac{x_n^2 + A}{2x_n}, \qquad n = 0, 1, 2, \ldots.$$

Nous savons, en vertu du théorème 8.4, que cette méthode converge quadratiquement vers $\sqrt{A}$ lorsque x_0 est suffisamment proche de $\sqrt{A}$. En fait la méthode de Newton converge dans ce cas précis pour tout $x_0 > 0$.

8.6 Notes bibliographiques et remarques

Dans ce chapitre, nous avons laissé de côté les méthodes utilisées pour obtenir un bon vecteur de départ x^0, ainsi que les critères d'arrêt des différentes

méthodes. Notons simplement que le choix du vecteur de départ est souvent inspiré par des considérations d'ordre physique lorsque le problème à résoudre provient du domaine de la physique.

La méthode de Newton pour résoudre un système d'équations non linéaires peut être couplée avec les méthodes itératives que nous avons étudiées dans le chapitre 6 pour résoudre les systèmes linéaires, voir par exemple [23, 17]. Par exemple, la ***méthode de Newton-Jacobi*** consiste à effectuer un pas de la méthode Jacobi pour résoudre le système linéaire obtenu lors de la méthode de Newton, c'est-à-dire pour résoudre

$$Df(x^n)(x^n - x^{n+1}) = f(x^n), \qquad n = 0, 1, 2, \ldots.$$

Un algorithme important dans la pratique est l'algorithme ***Newton-GMRES***, voir par exemple [20, 17], qui permet de résoudre efficacement les équations de Navier-Stokes.

Les ***méthodes de quasi-Newton*** ont pour but de modifier la matrice jacobienne $Df(x^n)$ au cours des itérations pour diminuer le coût des résolutions des systèmes linéaires. Pour une description de ces méthodes, voir par exemple [23, 17].

Dans ce chapitre, nous avons montré que la méthode de Newton converge pour autant que le point de départ x^0 soit choisi suffisamment proche de la solution. Il existe des situations pour lesquelles on ne sait pas choisir a priori le point de départ x^0. Il convient alors d'utiliser des ***méthodes de continuation***, voir par exemple [1].

Chapitre 9

Equations différentielles

9.1 Equations différentielles du premier ordre : généralités

Dans ce chapitre, nous notons $\mathbb{R}^+$ l'ensemble des nombres réels non négatifs et $f : (x,t) \in \mathbb{R} \times \mathbb{R}^+ \mapsto f(x,t) \in \mathbb{R}$ une fonction donnée à deux variables x et t. Dans les applications, x est souvent une variable physique alors que t est une variable temporelle. Dans ce chapitre t sera souvent appelé le temps. Nous supposons f assez régulière, par exemple continue, et nous voulons résoudre le problème suivant.

Etant donnée une valeur $u_0 \in \mathbb{R}$ (u_0 est dite valeur initiale), trouver une fonction une fois continûment dérivable $u : t \in \mathbb{R}^+ \to u(t) \in \mathbb{R}$ qui satisfait

$$\begin{aligned} &\dot{u}(t) = f(u(t),t) \qquad \text{si } t > 0, \\ &u(0) = u_0, \end{aligned} \tag{9.1}$$

où nous avons noté $\dot{u}(t) = du(t)/dt$.

Le problème (9.1) est appelé ***problème de Cauchy***. La première équation de (9.1) est une ***équation différentielle***. La deuxième équation de (9.1) est une ***condition de Cauchy***. Une fonction u qui satisfait (9.1) est appelée ***intégrale de l'équation différentielle***.

Exemple 9.1 On se donne $f(x,t) = 3x - 3t$ et $u_0 = \alpha$ (un nombre quelconque). Le problème de Cauchy devient :

$$\begin{aligned} &\dot{u}(t) = 3u(t) - 3t \qquad \text{si } t > 0, \\ &u(0) = \alpha; \end{aligned}$$

sa solution est donnée par $u(t) = (\alpha - 1/3)e^{3t} + t + 1/3$.

Exemple 9.2 On se donne $f(x,t) = \sqrt[3]{x}$, $u_0 = 0$. Le problème de Cauchy devient

$$\begin{aligned} &\dot{u}(t) = \sqrt[3]{u(t)} \qquad \text{si } t > 0, \\ &u(0) = 0; \end{aligned}$$

on vérifie que les fonctions u définies par $u(t) = 0$, et $u(t) = \pm\sqrt{8t^3/27}$, pour tout $t \geq 0$, sont toutes trois des solutions du problème (9.1). Cet exemple nous montre que le problème (9.1) n'a pas nécessairement une solution unique.

Exemple 9.3 On se donne $f(x,t) = x^3$, $u_0 = 1$. Le problème de Cauchy devient

$$\begin{aligned} &\dot{u}(t) = u^3(t), \qquad t > 0, \\ &u(0) = 1; \end{aligned}$$

on vérifie que la solution u est donnée pour $t \in [0, 1/2[$ par $u(t) = 1/\sqrt{1-2t}$. Cet exemple nous montre que le problème (9.1) n'a pas toujours une solution pour tout $t \in [0, \infty[$ puisque ici la solution explose lorsque t tend vers la valeur $1/2$ (en effet, nous avons $\lim_{\substack{t \to 1/2 \\ t<1/2}} u(t) = +\infty$).

Les trois exemples ci-dessus montrent que l'étude mathématique de l'existence et de l'unicité de solutions du problème (9.1) peut être une affaire délicate. Dans ce chapitre, nous nous contentons de donner, sans démonstration, un résultat d'existence et d'unicité global, au sens où on peut intégrer (9.1) jusqu'à $t = \infty$. Ce résultat s'énonce :

Théorème 9.1 *Soit une fonction continue $f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ qui vérifie la propriété suivante :*

il existe une fonction continue $\ell : t \in \mathbb{R}^+ \to \ell(t) \in \mathbb{R}$ telle que pour tout $x, y \in \mathbb{R}$ et pour tout $t \in \mathbb{R}^+$ on ait :

$$(f(x,t) - f(y,t))\,(x-y) \leq \ell(t)|x-y|^2. \tag{9.2}$$

Alors le problème de Cauchy (9.1) admet une solution globale unique.

Exemple 9.4 On se donne une fonction $f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ une fois continûment dérivable et telle qu'il existe une constante K satisfaisant

$$\frac{\partial f}{\partial x}(x,t) \leq K, \qquad \forall x \in \mathbb{R}, \quad \forall t \geq 0.$$

Regardons si l'hypothèse (9.2) est vérifiée avec une telle fonction. Si $x, y \in \mathbb{R}$ et $t \in \mathbb{R}^+$ sont fixés, il existe ξ dans l'intervalle d'extrémités x, y tel que

$$f(x,t) - f(y,t) = \frac{\partial f}{\partial x}(\xi, t)(x-y).$$

Ainsi nous obtenons

$$(f(x,t) - f(y,t))\,(x-y) = \frac{\partial f}{\partial x}(\xi,t)(x-y)^2 \leq K(x-y)^2.$$

L'hypothèse (9.2) est donc bien satisfaite avec $\ell(t) = K$, $\forall t \in \mathbb{R}^+$. Nous pouvons donc appliquer le théorème 9.1 et ainsi conclure que le problème (9.1) avec une telle fonction f admet une solution globale unique. De cette remarque nous pouvons conclure que le problème de Cauchy

$$\begin{aligned} &\dot{u}(t) = -u^3(t) + e^{-t^2/2}, \qquad t > 0, \\ &u(0) = 1, \end{aligned} \tag{9.3}$$

admet une solution globale unique. En effet, la fonction f dans ce cas est donnée par $f(x,t) = -x^3 + e^{-t^2/2}$ et donc $\frac{\partial f}{\partial x}(x,t) = -3x^2 \leq 0$.

Une conséquence immédiate du théorème 9.1 est le résultat suivant :

Théorème 9.2 (Cauchy-Lipschitz) *On suppose que la fonction f est continue sur $\mathbb{R} \times \mathbb{R}^+$ et qu'il existe un réel L tel que pour tout $x, y \in \mathbb{R}$ et pour tout $t \in \mathbb{R}^+$, on ait*

$$|f(x,t) - f(y,t)| \leq L|x-y|. \tag{9.4}$$

Alors le problème (9.1) admet une solution globale unique.

Démonstration
Si la fonction f satisfait l'inégalité (9.4), alors bien évidemment nous avons

$$(f(x,t) - f(y,t))\,(x-y) \leq L\,|x-y|^2, \qquad \forall x, y \in \mathbb{R}, \forall t \in \mathbb{R}^+.$$

Il suffit ainsi de définir la fonction ℓ par $\ell(t) = L$, $\forall t \geq 0$, et d'utiliser le théorème 9.1 pour conclure. ■

Exemple 9.5 On se donne $f(x,t) = |x| + \sin x + e^{-t^2/2}$, $u_0 = 1$. On vérifie facilement que

$$\begin{aligned} |f(x,t) - f(y,t)| &= \big||x| - |y| + \sin x - \sin y\big| \\ &= \left||x| - |y| + \int_y^x \cos\theta d\theta\right| \\ &\leq \big||x| - |y|\big| + |x-y| \leq 2|x-y|. \end{aligned}$$

Par le théorème de Cauchy-Lipschitz, le problème (9.1) a une solution globale unique $u(t)$.

Dans les exemples 9.4 et 9.5 il n'est pas possible de donner une expression explicite de la solution $u(t)$; il est donc nécessaire d'utiliser une méthode numérique pour obtenir des approximations des valeurs $u(t)$ pour différents $t \in \mathbb{R}^+$. Ci-dessous, nous décrivons brièvement une de ces méthodes.

Soit $0 = t_0 < t_1 < t_2 < t_3 < \ldots < t_n < t_{n+1} < \ldots$, des points de $\mathbb{R}^+$ et supposons connue une approximation u^n de u en $t = t_n$. Nous notons dans la suite $u^n \simeq u(t_n)$. Un schéma numérique à un pas consiste à calculer u^{n+1}, approximation de u en $t = t_{n+1}$, à partir de u^n. Par exemple, le ***schéma d'Euler progressif*** est obtenu en s'inspirant de la première équation de (9.1) et s'écrit :

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = f(u^n, t_n),$$

ou, de façon équivalente :

$$u^{n+1} = u^n + (t_{n+1} - t_n) f(u^n, t_n). \tag{9.5}$$

Si nous posons $u^0 = u(0) = u_0$ (qui est une valeur connue), nous constatons immédiatement que (9.5) permet de calculer successivement u^1, puis u^2, puis u^3, etc.

Il est maintenant légitime de chercher à savoir dans quelle mesure l'erreur

$$|u(t_n) - u^n| \qquad n = 1, 2, \ldots,$$

est petite ou non. Nous essayons de répondre à cette question dans la suite de ce chapitre.

9.2 Problèmes numériquement mal posés

Pour illustrer notre propos, revenons à l'exemple 9.1 correspondant au cas où :

$$f(x,t) = 3x - 3t \qquad \text{et} \qquad u_0 = \alpha.$$

Nous avons vu que la solution du problème de Cauchy (9.1) est donnée dans ce cas par

$$u(t) = \left(\alpha - \frac{1}{3}\right) e^{3t} + t + \frac{1}{3}. \tag{9.6}$$

Si nous cherchons à résoudre (9.1) jusqu'à $t = 10$ avec une valeur de $\alpha = \frac{1}{3}$, nous obtenons $u(10) = 10 + \frac{1}{3} = \frac{31}{3}$. Par contre, si nous faisons le calcul avec $\alpha = 0.333333$ au lieu de $\alpha = \frac{1}{3}$, nous avons

$$\begin{aligned} u(10) &= (0.333333 - \tfrac{1}{3})e^{30} + 10 + \tfrac{1}{3} \\ &= -\tfrac{1}{3}10^{-6} \cdot e^{30} + \tfrac{31}{3}, \end{aligned}$$

ce qui représente une différence avec la précédente valeur de $\frac{1}{3}10^{-6} \cdot e^{30}$, évaluée à environ $\frac{1}{3}10^7$. Cet exemple nous apprend qu'une petite erreur sur la condition initiale (erreur relative d'ordre 10^{-6}) peut provoquer une très grande erreur sur $u(10)$ (erreur relative d'ordre 10^6). Ainsi, si le calculateur mis à notre disposition ne calcule qu'avec 6 chiffres significatifs (en virgule flottante), alors $\alpha = \frac{1}{3}$ devient $\alpha = 0.333333$ et il est inutile d'essayer d'inventer une méthode numérique pour calculer $u(10)$. En effet, la seule erreur sur la condition initiale provoque

déjà une erreur inadmissible sur la solution ! Nous sommes ici en présence d'un problème ***numériquement mal posé***, appelé aussi problème mal conditionné.

Sans définir de façon plus précise ce qu'est un problème numériquement mal posé, signalons toutefois que cette notion est relative au problème à résoudre, mais aussi au calculateur mis à disposition. Dans l'exemple précédent, si nous avions eu à disposition un calculateur avec 16 chiffres significatifs, l'erreur relative sur la donnée initiale aurait été de 10^{-16}, ce qui n'aurait induit qu'une erreur relative de l'ordre de 10^{-4} sur la valeur exacte $u(10)$.

Si nous choisissons $f(x,t) = -3x - 3t$ au lieu de $f(x,t) = 3x - 3t$, alors la solution du problème (9.1) avec $u_0 = \alpha$ est donnée par :

$$u(t) = (\alpha - \frac{1}{3})e^{-3t} - t + \frac{1}{3}.$$

Contrairement à l'exemple précédent, la partie exponentielle de la solution décroît très rapidement. Ce deuxième problème est ***numériquement bien posé*** car

- si $\alpha = \frac{1}{3}$ on a $u(10) = -10 + \frac{1}{3} = -\frac{29}{3}$;
- si $\alpha = 0.333333$ on a $u(10) = -\frac{1}{3}10^{-6}e^{-30} - 10 + \frac{1}{3} \simeq -\frac{29}{3} - \frac{1}{3}10^{-19}$.

L'erreur relative sur la donnée initiale ne porte donc pas à conséquence sur le résultat.

Dans la suite, nous nous intéressons uniquement à des problèmes numériquement bien posés.

9.3 Schémas d'Euler

Pour établir un schéma d'approximation du problème (9.1), nous commençons par partitionner l'axe Ot, c'est-à-dire nous choisissons des points t_0, t_1, t_2, ..., tels que

$$0 = t_0 < t_1 < t_2 < \ldots < t_n < t_{n+1} < \ldots.$$

En posant $h_n = t_{n+1} - t_n$, nous pouvons approcher

$$\dot{u}(t_n) \text{ ou } \dot{u}(t_{n+1}) \qquad \text{par} \qquad \frac{u(t_{n+1}) - u(t_n)}{h_n}.$$

Si u^n est une approximation de $u(t_n)$, ces deux approches nous suggèrent les schémas suivants :

Schéma d'Euler progressif

$$\begin{aligned} &\frac{u^{n+1} - u^n}{h_n} = f(u^n, t_n), \qquad n = 0, 1, 2, \ldots, \\ &u^0 = u_0. \end{aligned} \tag{9.7}$$

Schéma d'Euler rétrograde

$$\begin{aligned} &\frac{u^{n+1}-u^n}{h_n} = f(u^{n+1}, t_{n+1}), \qquad n = 0, 1, 2, \ldots, \\ &u^0 = u_0. \end{aligned} \tag{9.8}$$

Ces deux schémas nous permettent de calculer u^{n+1} à partir u^n et il est donc possible de déterminer successivement $u^1, u^2, u^3, \ldots$, en partant de u^0. Le schéma d'Euler progressif est un ***schéma explicite*** car il permet d'expliciter u^{n+1} en fonction de u^n :

$$u^{n+1} = u^n + h_n f(u^n, t_n).$$

Le schéma d'Euler rétrograde est un ***schéma implicite*** car il ne permet pas d'expliciter directement u^{n+1} en fonction de u^n lorsque la fonction f n'est pas triviale. En effet, on a dans ce cas :

$$u^{n+1} - h_n f(u^{n+1}, t_{n+1}) = u^n.$$

Si nous voulons calculer u^{n+1}, nous définissons la fonction

$$g(x) = x - h_n f(x, t_{n+1}) - u^n$$

et nous cherchons un zéro de $g(x)$ en prenant par exemple une méthode de Newton (sect. 8.3). Ainsi nous pouvons poser $x_0 = u^n$ et $x_{m+1} = x_m - g(x_m)/g'(x_m)$, $m = 0, 1, \ldots$ Puisque $g'(x) = 1 - h_n \partial f(x, t_{n+1})/\partial x$, nous obtenons donc dans ce cas le schéma :

$$\begin{aligned} x_0 &= u^n, \\ x_{m+1} &= x_m - \frac{x_m - h_n f(x_m, t_{n+1}) - u^n}{1 - h_n \dfrac{\partial f}{\partial x}(x_m, t_{n+1})}, \qquad m = 0, 1, \ldots. \end{aligned}$$

En vertu du théorème 8.4 nous savons que

$$\lim_{m \to \infty} x_m = u^{n+1}$$

pour autant que f soit suffisament régulière et que x_0 soit suffisamment proche de u^{n+1}, ce qui est le cas si le pas h_n est suffisamment petit.

A première vue, il semble que le schéma d'Euler progressif soit préférable au schéma d'Euler rétrograde puisque ce dernier n'est pas explicite. Cependant, nous verrons dans la suite que le schéma progressif peut engendrer des difficultés que le schéma rétrograde n'engendre pas.

Cas particulier : $f(x,t) = -\beta x$

Considérons le cas où $f(x,t) = -\beta x$, où β est un nombre réel positif donné. Le problème (9.1) devient :

$$\begin{aligned} &\dot{u}(t) = -\beta u(t), \qquad \text{si } t > 0, \\ &u(0) = u_0, \end{aligned} \tag{9.9}$$

dont la solution est trivialement $u(t) = e^{-\beta t}u_0$. Puisque β est positif, ce problème est numériquement bien posé : $u(t)$ décroît exponentiellement lorsque t tend vers l'infini.

Pour discrétiser l'axe Ot, nous choisissons un nombre réel petit $h > 0$ et nous posons $t_n = nh$ avec $n = 0, 1, 2, \ldots$ Nous allons étudier dans ce cadre les deux schémas d'Euler, à savoir le schéma d'Euler progressif et le schéma d'Euler rétrograde.

Le schéma d'Euler progressif (9.7) devient

$$u^{n+1} = (1 - \beta h)u^n, \qquad n = 0, 1, 2, \ldots \tag{9.10}$$

et par suite

$$u^n = (1 - \beta h)^n u_0, \qquad n = 0, 1, 2, \ldots. \tag{9.11}$$

Bien que la solution $u(t)$ de (9.9) tende vers zéro lorsque t tend vers l'infini, nous voyons dans (9.11) que si $u_0 \neq 0$ et $1 - \beta h < -1$, alors u^n tend vers l'infini en alternant de signe lorsque n tend vers l'infini. Nous dirons dans ce cas que le schéma d'Euler progressif est *instable*. Pour éviter ce phénomène, il convient donc d'imposer $-1 \leq 1 - \beta h$ ce qui aura pour effet de limiter h à :

$$h \leq \frac{2}{\beta}. \tag{9.12}$$

La condition (9.12) est appelée ***condition de stabilité*** ; elle limite le pas h d'avance en t lorsqu'on utilise le schéma d'Euler progressif.

Le schéma d'Euler rétrograde (9.8) devient dans le cadre de notre exemple :

$$(1 + \beta h)u^{n+1} = u^n, \qquad n = 0, 1, 2, \ldots \tag{9.13}$$

et par suite :

$$u^n = \left(\frac{1}{1 + \beta h}\right)^n u_0, \qquad n = 0, 1, 2, \ldots. \tag{9.14}$$

Dans ce cas, nous voyons que pour tout $h > 0$, nous avons

$$\lim_{n \to \infty} u^n = 0;$$

le schéma d'Euler rétrograde est donc toujours stable ; h n'a pas à être limité !

Pour clore la discussion dans le cas particulier où $f(x, t) = -\beta x$ avec $\beta > 0$, nous faisons une estimation de l'erreur lorsque nous utilisons le schéma d'Euler progressif (9.10) pour calculer une approximation de la solution u de (9.9) au temps T positif fixé. En d'autres termes, nous posons $h = T/N$ où N est un entier assez grand, nous définissons $t_n = nh$ avec $n = 0, 1, 2, \ldots, N$ et nous évaluons $|u(T) - u^N|$ lorsque u^N résulte du schéma (9.10).

Lorsque $u(t)$ est solution de (9.9), il est facile de vérifier que

$$u(t_{n+1}) = e^{-\beta h}u(t_n), \qquad n = 0, 1, 2, \ldots, N - 1, \tag{9.15}$$

et par conséquent

$$u(t_n) = \left(e^{-\beta h}\right)^n u_0, \qquad n = 0, 1, 2, \ldots, N. \tag{9.16}$$

A ce stade de l'exposé, il est intéressant de remarquer l'analogie entre les relations (9.15), (9.16) et les relations (9.10), (9.11) lorsqu'on écrit le développement de l'exponentielle autour de zéro :

$$e^{-\beta h} = 1 - \beta h + O(h^2), \tag{9.17}$$

où $O(h^2)$ désigne un reste qui tend vers zéro comme h^2 lorsque h tend vers zéro. Calculons maintenant $|u(T) - u^N|$. En utilisant (9.16) et (9.11) avec $n = N$ nous obtenons :

$$|u(T) - u^N| = |(e^{-\beta h})^N - (1 - \beta h)^N| \cdot |u_0|. \tag{9.18}$$

Choisissons $N \geq \beta T$ de sorte à ce que $(1 - \beta h) \geq 0$. Alors les relations

$$a^N - b^N = (a - b)\left(a^{N-1} + a^{N-2}b + a^{N-3}b^2 + \cdots + ab^{N-2} + b^{N-1}\right),$$

et

$$1 - \beta x \leq e^{-\beta x} \qquad \forall x \in \mathbb{R},$$

ainsi que (9.18) impliquent :

$$|u(T) - u^N| \leq |e^{-\beta h} - (1 - \beta h)| \cdot N e^{-\beta(N-1)h} |u_0|. \tag{9.19}$$

La relation $1 - \beta h \geq 0$ implique $e^{\beta h} \leq e$ et par suite

$$e^{-\beta(N-1)h} = e^{\beta h} e^{-\beta T} \leq e \cdot e^{-\beta T}.$$

De plus, le développement limité de l'exponentielle au voisinage de zéro donne

$$|e^{-\beta h} - (1 - \beta h)| \leq \frac{\beta^2 h^2}{2} = \frac{\beta^2 T^2}{2N^2}.$$

Ainsi nous déduisons de l'inégalité (9.19) que

$$|u(T) - u^N| \leq \frac{e\beta^2 T^2 e^{-\beta T}}{2} |u_0| \cdot \frac{1}{N}. \tag{9.20}$$

L'inégalité (9.20) est une estimation d'erreur entre $u(T)$ et u^N. Elle montre en particulier que $\lim_{N \to \infty} u^N = u(T)$.

Nous aurions pu obtenir des estimations du même type pour le schéma d'Euler rétrograde. Dans ce cas nous pourrions aussi remarquer l'analogie entre les relations (9.15), (9.16) et les relations (9.13), (9.14) lorsqu'on écrit le développement suivant de l'exponentielle autour de zéro :

$$e^{-\beta h} = \frac{1}{e^{\beta h}} = \frac{1}{1 + \beta h + O(h^2)} = \frac{1}{1 + \beta h} + O(h^2). \tag{9.21}$$

De façon générale nous pourrions obtenir des conclusions semblables lorsque la fonction f n'est plus nécessairement égale à $-\beta x$ mais satisfait l'hypothèse (9.4). Ainsi il est possible de démontrer le résultat suivant :

Théorème 9.3 *Supposons que $f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ soit une fonction continue et de premières dérivées partielles par rapport à x et t continues. De plus, supposons que l'hypothèse de Cauchy-Lipschitz (9.4) soit satisfaite et notons $u(t)$ l'unique solution du problème (9.1).*

Soit $T > 0$ le temps jusqu'où nous voulons intégrer (9.1), soit N un entier positif, soit $h = T/N$ le pas de temps et soit $t_n = nh$ avec $n = 0, 1, 2, \ldots, N$. Alors il existe une constante C (indépendante de N mais qui peut dépendre de T) telle que pour tout N nous avons :

$$|u(T) - u^N| \leq \frac{C}{N} = \frac{C}{T}h, \tag{9.22}$$

où les approximations $u^0, u^1, \ldots, u^N$ sont données par le schéma d'Euler progressif (9.7) ou le schéma d'Euler rétrograde (9.8). En particulier nous avons $\lim_{N\to\infty} |u(T) - u^N| = 0$.

Considérons l'estimation d'erreur entre $u(T)$ et u^N, donnée de façon générale par la relation (9.22). Puisque $1/N = h/T = O(h)$ lorsque h tend vers zéro (il faut lire $1/N$ est d'ordre h lorsque h tend vers zéro), $|u(T) - u^N| = O(h)$. Nous dirons que les schémas d'Euler sont d'ordre 1 en h. En pratique, l'erreur commise au temps $t = T$ sera, en principe, divisée par deux chaque fois que le pas de temps h est divisé par deux.

9.4 Méthodes de Runge-Kutta d'ordre 2

Considérons le problème (9.1) et supposons que l'on ait montré l'existence d'une solution u. Si nous intégrons la première équation de (9.1) entre t_n et t_{n+1}, nous obtenons

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_{n+1}} f(u(t), t)dt. \tag{9.23}$$

Soit u^n une approximation de $u(t_n)$ et soit u^{n+1} une approximation de $u(t_{n+1})$. Nous proposons le schéma obtenu en intégrant le membre de droite de (9.23) par la formule des trapèzes (sect. 3.1), c'est-à-dire :

$$u^{n+1} - u^n = \frac{1}{2}h_n\Big(f(u^n, t_n) + f(u^{n+1}, t_{n+1})\Big), \quad n = 0, 1, 2 \ldots, \tag{9.24}$$

où nous avons noté $h_n = t_{n+1} - t_n$. Le schéma (9.24) est bien évidemment un schéma implicite. Il s'agit en fait du schéma obtenu en faisant la moyenne des schémas d'Euler progressif et rétrograde. Il est possible de montrer qu'il est d'ordre 2 en h, c'est-à-dire d'ordre h^2.

Pour éviter le calcul implicite de u^{n+1} dans (9.24), nous pouvons utiliser une ***prédiction d'Euler progressive*** et remplacer u^{n+1} dans le membre de droite de (9.24) par

$$\tilde{u}^{n+1} = u^n + h_n f(u^n, t_n). \tag{9.25}$$

Nous avons construit ainsi un nouveau schéma, appelé ***méthode de Heun***. Ce schéma consiste, à partir de u^n, à calculer $\tilde{u}^{n+1}$ par (9.25), puis à calculer u^{n+1} par (9.24), après avoir remplacé u^{n+1} au membre de droite par $\tilde{u}^{n+1}$. Plus précisément, la méthode de Heun s'écrit :

$$\begin{aligned} p_1 &= f(u^n, t_n), \\ p_2 &= f(u^n + h_n p_1, t_{n+1}), \\ u^{n+1} &= u^n + \frac{h_n}{2}(p_1 + p_2). \end{aligned} \tag{9.26}$$

La méthode de Heun fait partie des ***méthodes de Runge-Kutta d'ordre 2 explicites***.

Revenons à l'égalité (9.23) et utilisons la formule du rectangle (sect. 3.3) pour intégrer le membre de droite. Notons $t_{n+1/2} = (t_n + t_{n+1})/2$ le point milieu de $[t_n, t_{n+1}]$. Nous obtenons un nouveau schéma :

$$u^{n+1} - u^n = h_n f(u^{n+1/2}, t_{n+1/2}), \tag{9.27}$$

où $u^{n+1/2}$ est une approximation de $u(t_{n+1/2})$. Ici encore, nous utilisons une prédiction d'Euler progressive pour approcher $u^{n+1/2}$, c'est-à-dire :

$$u^{n+1/2} = u^n + \frac{h_n}{2} f(u^n, t_n). \tag{9.28}$$

La méthode ainsi obtenue est aussi une méthode de Runge-Kutta d'ordre 2 explicite que nous appellerons ***méthode d'Euler modifiée***. Connaissant u^n, cette méthode s'écrit :

$$\begin{aligned} p_1 &= f(u^n, t_n), \\ p_2 &= f(u^n + \frac{h_n}{2} p_1, t_n + \frac{h_n}{2}), \\ u^{n+1} &= u^n + h_n p_2. \end{aligned} \tag{9.29}$$

Nous pourrions montrer que les méthodes de Runge-Kutta d'ordre 2 sont, comme leur nom l'indique, d'ordre 2 en h. L'étude de la stabilité de la méthode de Heun fait l'objet de l'exercice 9.2.

9.5 Méthode de Runge-Kutta classique

La méthode de Runge-Kutta classique permet, à partir de u^n (approximation de u en $t = t_n$), de calculer u^{n+1} (approximation de u en $t = t_{n+1}$) de la manière

suivante :

$$\begin{aligned} p_1 &= f(u^n, t_n), \\ p_2 &= f(u^n + \frac{h_n}{2} p_1, t_n + \frac{h_n}{2}), \\ p_3 &= f(u^n + \frac{h_n}{2} p_2, t_n + \frac{h_n}{2}), \\ p_4 &= f(u^n + h_n p_3, t_{n+1}), \\ u^{n+1} &= u^n + \frac{h_n}{6}(p_1 + 2p_2 + 2p_3 + p_4). \end{aligned} \tag{9.30}$$

La méthode de Runge-Kutta classique est clairement une méthode explicite. De plus, c'est une méthode d'ordre 4 en h dans le sens suivant. Supposons que l'on veuille intégrer numériquement (9.1) jusqu'au temps $t = T$ où T est une nombre positif donné. Pour ce faire, nous choisissons un entier positif donné N, nous posons $h = \frac{T}{N}$, $t_j = jh$, $j = 0, 1, 2, \ldots, N$, et nous utilisons le schéma (9.30). Sous certaines hypothèses de régularité sur f, il est possible de montrer que l'estimation d'erreur :

$$|u(T) - u^N| \leq Ch^4 = C\frac{T^4}{N^4}, \tag{9.31}$$

est satisfaite lorsque N tend vers l'infini. Ici C est une constante indépendante de N mais qui dépend de T. La relation (9.31) montre que la méthode de Runge-Kutta classique est d'ordre 4 en h. De plus, si f ne dépend que de t, on peut facilement vérifier que la méthode de Runge-Kutta classique coïncide avec la méthode de Simpson (sect. 3.4) pour intégrer numériquement le membre de droite de (9.23).

9.6 Systèmes différentiels du premier ordre

Soit M un entier positif et soit $\vec{f} : (\vec{x}, t) \in \mathbb{R}^M \times \mathbb{R}^+ \to \vec{f}(\vec{x}, t) \in \mathbb{R}^M$ une fonction vectorielle donnée, supposée continue. Si $\vec{u}_0$ est un vecteur à M composantes données, on pose le problème de trouver une fonction à valeurs vectorielles

$$\vec{u} : t \in \mathbb{R}^+ \to \vec{u}(t) \in \mathbb{R}^M$$

telle que

$$\begin{aligned} \dot{\vec{u}}(t) &= \vec{f}(\vec{u}(t), t), \qquad t > 0, \\ \vec{u}(0) &= \vec{u}_0. \end{aligned} \tag{9.32}$$

Clairement, (9.32) est un système différentiel de M équations à M inconnues qui sont les composantes $u_1(t), u_2(t), \ldots, u_M(t)$ de $\vec{u}(t)$. La notation $\dot{\vec{u}}(t)$ représente le vecteur dont les composantes sont $\dot{u}_1(t), \dot{u}_2(t), \ldots, \dot{u}_M(t)$. Les résultats d'existence et d'unicité de la solution de (9.32) sont identiques à ceux obtenus dans les théorèmes 9.1 et 9.2 lorsqu'on remplace les valeurs absolues $|\cdot|$ par des normes euclidiennes $\|\cdot\|$ et le produit $(f(x,t) - f(y,t))(x - y)$ par le produit scalaire $(\vec{f}(\vec{x}, t) - \vec{f}(\vec{y}, t))^T(\vec{x} - \vec{y})$.

Les schémas d'Euler progressif et rétrograde et les méthodes de Runge-Kutta présentés dans les sections 9.4 et 9.5 sont généralisables au système différentiel (9.32). Par exemple, le schéma d'Euler progressif devient :

$$\vec{u}^{n+1} = \vec{u}^n + h_n \vec{f}(\vec{u}^n, t_n). \tag{9.33}$$

Ce schéma permet donc, à partir de $\vec{u}^n$ (approximation de $\vec{u}(t_n)$), de calculer $\vec{u}^{n+1}$ (approximation de $\vec{u}(t_{n+1})$).

Les estimations d'erreur sont les mêmes que dans le cas scalaire des sections 9.4 et 9.5.

9.7 Equations différentielles d'ordre supérieur

Considérons dans un premier temps le cas d'une fonction à 3 variables

$$f : (x, y, t) \in \mathbb{R}^2 \times \mathbb{R}^+ \to f(x, y, t) \in \mathbb{R}$$

que nous supposons continue, et soit deux nombres donnés u_0 et v_0. On pose le problème de trouver une fonction

$$u : t \in \mathbb{R}^+ \to u(t) \in \mathbb{R},$$

deux fois continûment dérivable, telle que

$$\begin{aligned} &\ddot{u}(t) = f(u(t), \dot{u}(t), t), \qquad t > 0, \\ &u(0) = u_0, \qquad \dot{u}(0) = v_0, \end{aligned} \tag{9.34}$$

où nous avons noté $\ddot{u}(t) = d^2u(t)/dt^2$. Le problème (9.34) est appelé ***problème différentiel du deuxième ordre*** ; deux conditions initiales sont fixées qui, en dynamique par exemple, correspondent le plus souvent à l'état initial et à la vitesse initiale lorsque t est le temps.

Pour résoudre numériquement le problème (9.34), nous pouvons par exemple introduire une nouvelle inconnue $v(t) = \dot{u}(t)$ et ainsi remplacer (9.34) par un système du premier ordre pour les inconnues $u(t)$ et $v(t)$:

$$\begin{aligned} &\dot{u}(t) = v(t), \\ &\dot{v}(t) = f(u(t), v(t), t), \qquad t > 0, \\ &u(0) = u_0 \qquad \text{et} \qquad v(0) = v_0. \end{aligned} \tag{9.35}$$

Le problème (9.35) est un système différentiel du premier ordre de 2 équations à 2 inconnues. Ainsi nous pouvons appliquer les méthodes de la section 9.6 pour résoudre ce système.

Il existe des méthodes spécifiquement adaptées aux problèmes différentiels du deuxième ordre. Il s'agit des ***méthodes de Newmark***. Considérons par exemple le cas où la fonction f dans (9.34) est indépendante de la seconde variable. Nous notons $f(x, y, t) = g(x, t)$. Le problème (9.34) s'écrit donc :

$$\begin{aligned} &\ddot{u}(t) = g(u(t), t), \qquad t > 0, \\ &u(0) = u_0, \qquad \dot{u}(0) = v_0. \end{aligned} \tag{9.36}$$

Choisissons $h > 0$, notons $t_n = nh$, $n = 0, 1, 2, \ldots$, et soit u^n une approximation de $u(t_n)$. En nous inspirant des formules de différences finies pour l'approximation des dérivées secondes (chap. 2), nous écrivons le schéma :

$$\frac{u^{n+1} - 2u^n + u^{n-1}}{h^2} = g(u^n, t_n), \qquad n = 1, 2, \ldots, \tag{9.37}$$

$$u^0 = u_0, \tag{9.38}$$

$$u^1 = u_0 + hv_0 + \frac{1}{2}h^2 g(u_0, 0). \tag{9.39}$$

Remarquons que (9.37) permet de calculer u^{n+1} à partir de u^n et u^{n-1}. Pour interpréter (9.39), il suffit d'utiliser (9.36) et d'écrire :

$$u^1 = u(0) + h\dot{u}(0) + \frac{h^2}{2}\ddot{u}(0); \tag{9.40}$$

le membre de droite de (9.40) est la somme des trois premiers termes du développement limité à l'ordre 2 de $u(h)$ en $t = 0$.

La méthode (9.37) (9.38) (9.39) est d'ordre 2 ; elle est souvent utilisée pour résoudre numériquement l'équation des ondes (chap. 13). Cependant, cette méthode n'est pas inconditionnellement stable comme nous allons le voir sur l'exemple, simple mais instructif, de l'oscillateur harmonique.

Cas particulier : $g(x, t) = -\lambda x$

Soit $\lambda > 0$ et considérons l'équation différentielle du deuxième ordre suivante :

$$\begin{aligned} &\ddot{u}(t) = -\lambda u(t), \qquad t > 0, \\ &u(0) = u_0, \quad \dot{u}(0) = v_0, \end{aligned} \tag{9.41}$$

où u_0, v_0 sont deux nombres réels donnés. Clairement (9.41) correspond au problème (9.36) avec $g(x, t) = -\lambda x$. Il est facile de vérifier que la solution du problème (9.41) est donnée par

$$u(t) = \frac{v_0}{\sqrt{\lambda}} \sin \sqrt{\lambda} t + u_0 \cos \sqrt{\lambda} t, \tag{9.42}$$

qui est périodique de période $P = 2\pi/\sqrt{\lambda}$. D'autre part, nous pouvons vérifier sans difficultés que si nous posons

$$\alpha = \left(1 - \frac{\lambda h^2}{2}\right), \tag{9.43}$$

alors le schéma numérique (9.37) (9.38) (9.39) devient :

$$u^{n+1} = 2\alpha u^n - u^{n-1}, \qquad n = 1, 2, \ldots, \tag{9.44}$$

$$u^0 = u_0, \tag{9.45}$$

$$u^1 = \alpha u_0 + hv_0. \tag{9.46}$$

L'égalité (9.44) est appelée ***équation aux différences***. Considérons le trinôme du second degré

$$r^2 = 2\alpha r - 1, \tag{9.47}$$

appelé ***équation caractéristique*** de l'équation aux différences (9.44) (nous avons obtenu cette équation en remplaçant u^{n-1} par 1, u^n par r et u^{n+1} par r^2). Si $|\alpha|$ est différent de 1, l'équation caractéristique (9.47) admet les deux racines distinctes (complexes ou non) :

$$r_1 = \alpha + \sqrt{\alpha^2 - 1} \qquad \text{et} \qquad r_2 = \alpha - \sqrt{\alpha^2 - 1}. \tag{9.48}$$

Nous pouvons alors vérifier que si nous posons

$$u^n = a(r_1)^n + b(r_2)^n, \qquad n = 0, 1, 2, \ldots \tag{9.49}$$

avec

$$a = \frac{1}{2}u_0 + \frac{h}{2\sqrt{\alpha^2 - 1}}v_0, \qquad \text{et} \qquad b = \frac{1}{2}u_0 - \frac{h}{2\sqrt{\alpha^2 - 1}}v_0,$$

alors u^n satisfait (9.44) (9.45) (9.46). Nous sommes maintenant en mesure de comparer u^n et $u(t_n)$ qui, en vertu de (9.42), est donnée par :

$$u(t_n) = \frac{v_0}{\sqrt{\lambda}} \sin \sqrt{\lambda} t_n + u_0 \cos \sqrt{\lambda} t_n.$$

Puisque $|u(t_n)|$ reste une quantité majorée par une constante indépendante de n, il paraît naturel d'imposer $|r_1| \leq 1$ et $|r_2| \leq 1$, sans quoi $|u^n|$ dans (9.49) devrait croître indéfiniment lorsque n tend vers l'infini. Nous serions alors en présence d'instabilités d'origine numérique. La condition de stabilité pour le schéma (9.44) (9.45) (9.46) devient ainsi

$$\left| \alpha \pm \sqrt{\alpha^2 - 1} \right| \leq 1. \tag{9.50}$$

Si $|\alpha| > 1$, l'inégalité (9.50) ne peut pas être satisfaite. Si par contre $|\alpha| \leq 1$, alors $\alpha \pm \sqrt{\alpha^2 - 1} = \alpha \pm i\sqrt{1 - \alpha^2}$ (où ici i est l'unité imaginaire) et nous aurons $|\alpha \pm \sqrt{\alpha^2 - 1}| = (\alpha^2 + (1 - \alpha^2))^{1/2} = 1$. La condition de stabilité est donc $|\alpha| \leq 1$ qui, au vu de (9.43), devient :

$$h \leq \frac{2}{\sqrt{\lambda}}. \tag{9.51}$$

Par conséquent, plus la période $P = 2\pi/\sqrt{\lambda}$ de u est petite, plus le pas de temps h doit être choisi petit. Une conséquence de la relation (9.51) est qu'il faut effectuer plus de trois pas de temps par période pour que le schéma (9.44) (9.45) (9.46) ait des chances de donner une approximation correcte.

En fait, on peut montrer que sous la condition (9.51), le schéma numérique (9.44) (9.45) (9.46) est d'ordre 2 en h, c'est-à-dire si on pose $h = T/N$ avec $T > 0$ fixé, $t_n = nh$ avec $n = 0, 1, 2, \ldots, N$, alors l'erreur $|u(T) - u^N|$ est bornée par une constante (indépendante de N) multipliée par h^2 lorsque N tend vers l'infini. Il en est de même pour le schéma (9.37) (9.38) (9.39) lorsque la fonction g est suffisamment régulière.

9.8 Exercices

Exercice 9.1 On considère le problème de Cauchy :

$$\begin{aligned} &\dot{u}(t) = -(u(t))^m + \cos(t) \qquad \text{si } t > 0, \\ &u(0) = 0, \end{aligned}$$

où m est un entier impair.

1. Montrer que le problème ci-dessus possède une solution globale unique.
2. Soit h un paramètre positif donné, soit $t_n = nh$, $n = 0, 1, 2, \ldots$, et soit u^n une approximation de $u(t_n)$, $n = 0, 1, 2, \ldots$ Ecrire le schéma d'Euler rétrograde permettant de calculer u^{n+1} à partir de u^n.
3. A partir du schéma obtenu au point 2, écrire un seul pas de la méthode de Newton pour calculer une nouvelle approximation de u^1.

Solution
1. Le problème ci-dessus peut se mettre sous la forme du problème (9.1) en posant $f(x,t) = -x^m + \cos t$ et $u_0 = 0$. Puisque m est impair et

$$\frac{\partial f}{\partial x}(x,t) = -mx^{m-1} \leq 0,$$

la fonction f est décroissante et satisfait donc

$$(f(x,t) - f(y,t))(x-y) \leq 0 \qquad \forall x, y \in \mathbb{R}.$$

Par conséquent, l'hypothèse (9.2) du théorème 9.1 est satisfaite pour $\ell(t) = 0$. D'où le résultat d'existence et d'unicité.

2. Le schéma d'Euler rétrograde s'écrit

$$\begin{aligned} &u^{n+1} = u^n + h\Big(-(u^{n+1})^m + \cos(t_{n+1})\Big), \qquad n = 0, 1, 2, \ldots, \\ &u^0 = 0. \end{aligned}$$

Ce schéma est implicite car il ne permet pas de calculer u^{n+1} directement à partir de u^n.

3. Effectuons le premier pas du schéma d'Euler rétrograde. Il s'agit de trouver un nombre réel u^1 tel que

$$\begin{aligned} u^1 &= u^0 + h\Big(-(u^1)^m + \cos(t_1)\Big) \\ &= h\Big(-(u^1)^m + \cos(h)\Big). \end{aligned} \tag{9.52}$$

Pour déterminer u^1 nous devons donc chercher le zéro de la fonction g définie par

$$g(x) = x + hx^m - h\cos(h).$$

La méthode de Newton pour approcher le zéro de g s'écrit

$$\begin{aligned} x_{k+1} &= x_k - \frac{g(x_k)}{g'(x_k)} \\ &= x_k - \frac{x_k + h(x_k)^m - h\cos(h)}{1 + mh(x_k)^{m-1}}. \end{aligned}$$

Choisissons $x_0 = u^0 = 0$ comme valeur de départ. Le premier pas de la méthode de Newton s'écrit donc :

$$x_1 = h\cos(h).$$

Nous pouvons utiliser x_1 comme approximation de u^1.

Exercice 9.2 Soit $\beta > 0$ un nombre réel positif donné et considérons le problème

$$\begin{aligned} &\dot{u}(t) = -\beta u(t), \qquad \text{si } t > 0, \\ &u(0) = u_0, \end{aligned} \tag{9.53}$$

où u_0 est une valeur donnée. Soit h un paramètre positif donné, soit $t_n = nh$ et soit u^n une approximation de $u(t_n)$, $n = 0, 1, 2, \ldots$ Pour les méthodes de Heun et Runge-Kutta classique, sous quelle condition la relation

$$\lim_{n\to\infty} u^n = 0$$

a-t-elle lieu ?

Solution
Méthode de Heun : Considérons le schéma (9.26) pour résoudre numériquement le problème (9.53). Nous avons

$$\begin{aligned} p_1 &= -\beta u^n, \\ p_2 &= -\beta(u^n + hp_1) = \beta u^n(-1 + \beta h), \\ u^{n+1} &= u^n + \frac{h}{2}(p_1 + p_2) = \left(1 - \beta h + \frac{\beta^2 h^2}{2}\right) u^n. \end{aligned}$$

Par induction nous obtenons donc

$$u^n = \left(1 - \beta h + \frac{\beta^2 h^2}{2}\right)^n u_0,$$

et par conséquent $\lim_{n\to\infty} u^n = 0$ si

$$\left|1 - \beta h + \frac{\beta^2 h^2}{2}\right| < 1. \tag{9.54}$$

Notons q le polynôme défini par $q(x) = 1 - x + x^2/2$, dont le graphe est représenté dans la figure 9.1. Clairement nous avons $|q(x)| < 1$ si $0 < x < 2$. L'inégalité

(9.54) est donc satisfaite si $\beta h < 2$. Nous concluons donc en affirmant que $\lim_{n\to\infty} u^n = 0$ lorsque

$$h < \frac{2}{\beta}. \tag{9.55}$$

La condition de stabilité du schéma de Heun dans ce cas est $h \leq 2/\beta$. Cette condition est la même que celle obtenue pour le schéma d'Euler progressif, voir l'inégalité (9.12). Notons en passant que $1 - x + x^2/2$ est le développement de limité à l'ordre 2 de e^{-x}.

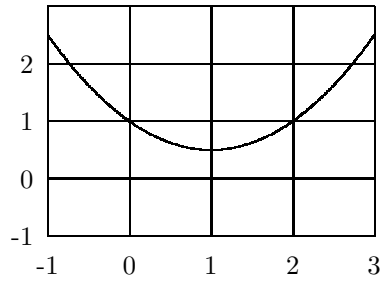

Fig. 9.1 Le graphe du polynôme $x \to 1 - x + x^2/2$.

Méthode de Runge-Kutta classique : Considérons maintenant le schéma (9.30) pour résoudre le problème (9.53). Nous avons

$$\begin{aligned}
p_1 &= -\beta u^n, \\
p_2 &= -\beta\left(u^n + \frac{h}{2}p_1\right) = \beta u^n\left(-1 + \frac{\beta h}{2}\right), \\
p_3 &= -\beta\left(u^n + \frac{h}{2}p_2\right) = \beta u^n\left(-1 + \frac{\beta h}{2} - \frac{\beta^2 h^2}{4}\right), \\
p_4 &= -\beta\left(u^n + hp_3\right) = \beta u^n\left(-1 + \beta h - \frac{\beta^2 h^2}{2} + \frac{\beta^3 h^3}{4}\right), \\
u^{n+1} &= u^n + \frac{h}{6}(p_1 + 2p_2 + 2p_3 + p_4) \\
&= \left(1 - \beta h + \frac{\beta^2 h^2}{2} - \frac{\beta^3 h^3}{6} + \frac{\beta^4 h^4}{24}\right) u^n.
\end{aligned}$$

Par induction nous obtenons donc

$$u^n = \left(1 - \beta h + \frac{\beta^2 h^2}{2} - \frac{\beta^3 h^3}{6} + \frac{\beta^4 h^4}{24}\right)^n u_0,$$

et par conséquent $\lim_{n\to\infty} u^n = 0$ si

$$\left|1 - \beta h + \frac{\beta^2 h^2}{2} - \frac{\beta^3 h^3}{6} + \frac{\beta^4 h^4}{24}\right| < 1. \tag{9.56}$$

Notons r le polynôme défini par $r(x) = 1 - x + x^2/2 - x^3/6 + x^4/24$, dont le graphe est représenté dans la figure 9.2. Clairement $|r(x)| < 1$ si $0 < x < \bar{x}$ où $\bar{x} \simeq 2.785$. L'inégalité (9.56) est donc satisfaite si $\beta h < 2.78$. Nous concluons donc que $\lim_{n\to\infty} u^n = 0$ lorsque

$$h < \frac{2.78}{\beta}.$$

La condition de stabilité de la méthode de Runge-Kutta classique dans ce cas est $h \leq \bar{x}/\beta$. Cette condition est moins restrictive que celle obtenue pour le schéma d'Euler progressif, voir l'inégalité (9.12). Notons en passant que $r(x)$ est le développement limité à l'ordre 4 de e^{-x}.

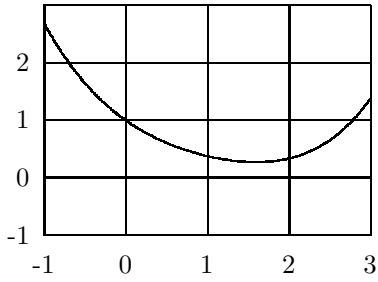

Fig. 9.2 Le graphe du polynôme $x \to 1 - x + x^2/2 - x^3/6 + x^4/24$.

9.9 Notes bibliographiques et remarques

Les méthodes que nous avons présentées dans ce chapitre pour résoudre numériquement des systèmes différentiels d'ordre 1 sont des ***méthodes à un pas*** (Euler, Runge-Kutta). Ces méthodes permettent de calculer u^{n+1} à partir de u^n. Les ***méthodes multipas*** utilisent les valeurs u^n, u^{n-1}, u^{n-2}, ..., pour calculer u^{n+1}. Citons par exemple les méthodes d'Adams, les méthodes prédicteur-correcteur, voir par exemple [12, 4].

Comme nous l'avons vu dans la section 9.2, certaines équations différentielles sont très sensibles aux perturbations numériques. On dit alors que le problème est ***raide*** (*stiff* en anglais). Il existe un certain nombre de méthodes pour résoudre ce genre de problèmes, voir par exemple [13].

Chapitre 10

Différences finies et éléments finis pour des problèmes aux limites unidimensionnels

10.1 Un problème aux limites unidimensionnel

Considérons le problème suivant : étant donné deux fonctions c et f continues sur l'intervalle $[0,1]$, trouver une fonction u deux fois continûment dérivable sur $[0,1]$ telle que

$$\begin{aligned} &-u''(x) + c(x)u(x) = f(x) \qquad \text{si } 0 < x < 1, \\ &u(0) = u(1) = 0. \end{aligned} \tag{10.1}$$

Un exemple de situation physique où ce problème est rencontré est celui du fléchissement d'une poutre de longueur unité, d'extrémités $x = 0$ et $x = 1$, étirée selon son axe par une force P, soumise à une densité linéaire de charge $f(x)$ et simplement appuyée à ses extrémités (fig. 10.1). Alors le moment fléchissant $u(x)$ au point d'abscisse x est solution du problème (10.1) avec $c(x) = P/EI(x)$, où E est le module de Young du matériau et $I(x)$ est le moment principal d'inertie de la section de la poutre au point x.

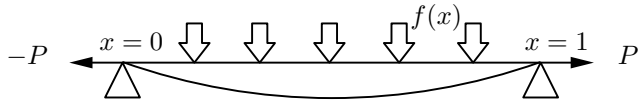

Fig. 10.1 Fléchissement d'une poutre.

Un autre exemple est celui du déplacement vertical $u(x)$ au point x d'une corde tendue entre les extrémités $x = 0$ et $x = 1$, soumise à une tension unité et à une densité de charge verticale $f(x)$; dans ce cas, on a $c(x) = 0$, $\forall x \in [0,1]$ (fig. 10.2).

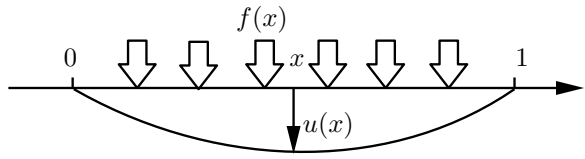

Fig. 10.2 Déplacement vertical d'une corde tendue.

Remarquons que la première équation de (10.1) est une équation différentielle du deuxième ordre dont l'intégration doit faire apparaître deux constantes. Nous avons aussi deux conditions $u(0) = 0$, $u(1) = 0$, appelées ***conditions aux limites***, qui permettent de déterminer les deux constantes d'intégration.

Si l'on suppose $c \geq 0$ sur l'intervalle $[0, 1]$, on peut montrer que le problème (10.1) a une et une seule solution. Sauf pour quelques cas très rares, il n'existe pas de formule permettant d'obtenir explicitement $u(x)$, pour tout $x \in (0, 1)$. Il convient donc de trouver un moyen d'approcher les valeurs de la solution du problème (10.1), d'aussi près que l'on veut. Une méthode pour atteindre ce but consiste à discrétiser le problème (10.1), c'est-à-dire à le transformer en un problème qui lui est proche et qui comprend un nombre fini de valeurs à calculer. Dans ce chapitre nous présentons quelques-unes de ces méthodes.

10.2 Méthode des différences finies

Soit N un entier positif; on pose $h = 1/(N+1)$ et on note $x_j = jh$, $j = 0, 1, 2, \ldots, N+1$, les points de discrétisation (fig. 10.3).

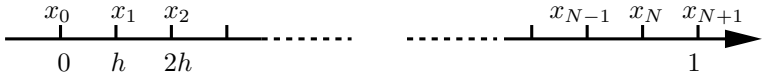

Fig. 10.3 Points de discrétisation.

Nous avons vu dans le théorème 2.5 que, si u est quatre fois continûment dérivable, alors

$$\begin{aligned} u''(x) &= \frac{\delta_h^2 u(x)}{h^2} + O(h^2) \\ &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2), \end{aligned} \tag{10.2}$$

où $O(h^2)$ désigne un reste qui, lorsque h tend vers zéro, reste borné par une constante multipliée par h^2.

Pour résoudre numériquement le problème (10.1), nous nous inspirons de (10.2) et nous calculons des valeurs u_j, sensées être proches de $u(x_j)$ et satisfaisant :

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + c(x_j)u_j = f(x_j) \qquad 1 \leq j \leq N, \tag{10.3}$$

$$u_0 = u_{N+1} = 0. \tag{10.4}$$

Le schéma (10.3) (10.4) est appelé ***problème approché*** ou encore ***problème discret***, par opposition au problème (10.1) qui est appelé ***problème continu***. Il s'agit d'un schéma d'approximation par une ***méthode de différences finies*** du problème (10.1). Résoudre (10.3) (10.4) revient à chercher un nombre fini de valeurs u_j qui devraient être des approximations de $u(x_j)$ (on note $u_j \simeq u(x_j)$), $1 \leq j \leq N$.

Si $\vec{u}$ est le N-vecteur colonne de composantes $u_1, u_2, \ldots, u_N$, si $\vec{f}$ est le N-vecteur de composantes $f(x_1), f(x_2), \ldots, f(x_N)$ et si A est la $N \times N$ matrice tridiagonale définie par

$$A = \frac{1}{h^2} \begin{bmatrix} 2+c_1h^2 & -1 & & & \\ -1 & 2+c_2h^2 & -1 & & 0 \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & 0 & & -1 & 2+c_Nh^2 \end{bmatrix}, \tag{10.5}$$

où $c_i = c(x_i)$, alors le problème approché (10.3) (10.4) est clairement équivalent à chercher $\vec{u}$ tel que

$$A\vec{u} = \vec{f}. \tag{10.6}$$

Si $c(x) \geq 0$ pour tout $x \geq 0$, on peut montrer que A est une matrice symétrique définie positive. Elle est donc régulière et soit $\vec{u}$ la solution de (10.6) que l'on peut obtenir après avoir fait une décomposition de Cholesky de la matrice A (chap. 5). Si la solution u de (10.1) est quatre fois continûment dérivable, il est possible de montrer (en utilisant des notions de stabilité et consistance que nous n'introduirons pas ici) le résultat de convergence suivant :

Théorème 10.1 *On suppose que $c(x) \geq 0$, $\forall x \in [0,1]$ et que la solution u de (10.1) est quatre fois continûment dérivable. Alors il existe une constante C indépendante de N (et donc de h) telle que*

$$\max_{1 \leq j \leq N} \mid u(x_j) - u_j \mid \leq Ch^2. \tag{10.7}$$

On constate donc que si $(u_j)_{1 \leq j \leq N}$ est solution de (10.3) et si u est solution de (10.1), on a

$$\lim_{N \to \infty} \max_{1 \leq j \leq N} \mid u(x_j) - u_j \mid = 0. \tag{10.8}$$

De plus, l'erreur est, en principe, quatre fois plus petite chaque fois qu'on double le nombre de points de discrétisation.

10.3 Méthode de Galerkin

Considérons le problème (10.1) et multiplions sa première équation par une fonction v une fois continûment dérivable sur $[0,1]$. Si nous intégrons sur l'intervalle $[0,1]$, nous obtenons :

$$-\int_0^1 u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

En intégrant par parties le premier terme, nous avons :

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

Si nous imposons à la fonction v d'être nulle en $x = 0$ et $x = 1$, alors nous en déduisons l'égalité :

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx. \tag{10.9}$$

Soit maintenant V l'ensemble de toutes les fonctions g continues, de première dérivée g' continue par morceaux et telles que $g(0) = g(1) = 0$. Ici le terme g' continue par morceaux signifie que g' existe et est continue sauf éventuellement en un nombre fini de points de l'intervalle $[0, 1]$ où g' pourrait ne pas exister mais posséderait des limites à gauche et à droite de ces points, voir par exemple les figures 10.4 et 10.5 plus loin. La somme de deux fonctions de V reste un élément de V, ainsi que le produit d'une fonction de V par un nombre réel ; ainsi V a une structure d'espace vectoriel. Nous allons maintenant chercher

$$u \in V \text{ qui satisfait (10.9) pour toute fonction } v \in V.$$

Dans la suite, ce problème est appelé problème (10.9). Le problème (10.9) est appelé ***problème faible*** ou ***problème variationnel***.

A priori, les fonctions u solutions du problème (10.9) sont moins régulières que les fonctions u solutions du problème différentiel (10.1). En effet, le problème (10.9) contient une dérivée première de la solution u alors que le problème (10.1) contient une dérivée seconde.

Puisque nous avons déduit le problème (10.9) du problème (10.1), il est évident que toute solution u de (10.1) est solution de (10.9). En fait, on peut montrer que si $c(x) \geq 0$, $\forall x \in [0, 1]$, alors le problème (10.9) a une et une seule solution u qui est celle du problème (10.1).

Nous présentons maintenant la ***méthode de Galerkin***, qui est basée sur la formulation faible (10.9), contrairement à la méthode des différences finies, qui est basée sur la formulation différentielle (10.1). La méthode de Galerkin est le point de départ des méthodes d'éléments finis et des méthodes spectrales.

Si $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont N fonctions linéairement indépendantes de V, on peut construire un sous-espace vectoriel de V, noté V_h, engendré par les combinaisons linéaires des fonctions φ_i. Ainsi V_h sera l'ensemble de toutes les fonctions g qui peuvent s'exprimer sous la forme

$$g(x) = \sum_{i=1}^{N} g_i \varphi_i(x),$$

où les g_i sont N nombres réels. Il est donc naturel de formuler une approximation du problème (10.9) de la manière suivante : trouver une fonction $u_h \in V_h$ telle

que

$$\int_0^1 u_h'(x)v_h'(x)dx + \int_0^1 c(x)u_h(x)v_h(x)dx = \int_0^1 f(x)v_h(x)dx \qquad (10.10)$$

pour toute fonction $v_h \in V_h$. On dira que (10.10) est une ***approximation de Galerkin*** de (10.9). Puisque u_h est cherché dans V_h, on peut écrire :

$$u_h(x) = \sum_{i=1}^{N} u_i\varphi_i(x),$$

où $u_1, u_2, \ldots, u_N$ sont N nombres réels à déterminer. En prenant $v_h = \varphi_j$, $1 \leq j \leq N$, dans (10.10), le problème (10.10) est alors équivalent à chercher $u_1, u_2, \ldots, u_N$ tels que

$$\sum_{i=1}^{N} u_i \left(\int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx \right) = \int_0^1 f(x)\varphi_j(x)dx \qquad (10.11)$$

pour tout $j = 1, 2, \ldots, N$. Si A est la $N \times N$ matrice de coefficients

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx, \qquad (10.12)$$

(dans le cas où $c = 0$, la matrice A est appelée ***matrice de rigidité***), si $\vec{u}$ est le N-vecteur de composantes $u_1, u_2, \ldots, u_N$ et si $\vec{f}$ est le N-vecteur dont la j^e composante est

$$f_j = \int_0^1 f(x)\varphi_j(x)dx, \qquad (10.13)$$

alors les problèmes (10.10) ou (10.11) sont équivalents à chercher $\vec{u}$ tel que

$$A\vec{u} = \vec{f}. \qquad (10.14)$$

Nous dirons que (10.10) ou (10.11) ou (10.14) sont une discrétisation du problème continu (10.1). Notons encore que, après avoir construit la matrice A et le vecteur $\vec{f}$, la méthode de Galerkin nécessite, tout comme la méthode des différences finies, la résolution d'un système linéaire.

Dans la suite de ce chapitre nous étudions plus en détail une méthode de Galerkin spécifique, en l'occurrence la ***méthode des éléments finis***. Cette méthode revient à faire un choix judicieux des fonctions φ_1, φ_2, ..., φ_N définissant V_h, de sorte que les propriétés suivantes soient vérifiées :

- La matrice A doit être une ***matrice creuse*** au sens où elle contient un grand nombre de coefficients nuls. C'est, par exemple, le cas des matrices de bande (sect. 5.5). Dans le cas où la matrice A est creuse, les méthodes numériques présentées dans les chapitre 4, 5 et 6 sont des méthodes bien adaptées pour résoudre le système linéaire (10.14).

– La solution u_h du problème (10.10) doit converger, dans un certain sens, vers la solution u du problème (10.9) lorsque le nombre N de fonctions linéairement indépendantes de V devient grand.

Avant de présenter plus en détails la méthode des éléments finis, nous énonçons un résultat général de convergence pour les méthodes de Galerkin. Munissons l'espace vectoriel V de la norme $|\cdot|_1$ définie par

$$|g|_1 = \left(\int_0^1 (g'(x))^2 dx \right)^{1/2} \text{ si } g \in V.$$

Nous obtenons alors le résultat suivant :

Théorème 10.2 *On suppose que $c(x) \geq 0$, $\forall x \in [0,1]$. Alors si u est solution de (10.9) et si u_h est solution de (10.10), nous avons l'estimation d'erreur*

$$|u - u_h|_1 \leq C \min_{v_h \in V_h} |u - v_h|_1, \tag{10.15}$$

où C est donnée par $C = 1 + \max_{x\in[0,1]} |c(x)|$ (et est donc indépendante du choix de V_h).

Démonstration
Considérons pour simplifier le cas où $c(x) = 0$, $\forall x \in [0,1]$. Si u est solution de (10.9) et si u_h est solution de (10.10) nous obtenons, par soustraction de ces deux relations :

$$\int_0^1 (u'(x) - u_h'(x))\, v_h'(x) dx = 0 \qquad \forall v_h \in V_h. \tag{10.16}$$

En posant $e(x) = u(x) - u_h(x)$, qui représente l'erreur entre u et u_h au point x, les égalités (10.16) s'écrivent :

$$\int_0^1 e'(x) v_h'(x) dx = 0, \qquad \forall v_h \in V_h. \tag{10.17}$$

Par définition de la norme $|\cdot|_1$ et de l'erreur e, nous avons

$$|e|_1^2 = \int_0^1 (e'(x))^2\, dx = \int_0^1 e'(x)(u'(x) - u_h'(x)) dx.$$

En tenant compte de (10.17) dans l'égalité ci-dessus, nous obtenons donc :

$$|e|_1^2 = \int_0^1 e'(x) u'(x) dx = \int_0^1 e'(x)(u'(x) - v_h'(x)) dx,$$

où v_h est une élément quelconque de V_h. En utilisant l'inégalité de Cauchy-Schwarz dans cette dernière expression, nous obtenons :

$$|e|_1^2 \leq \left(\int_0^1 (e'(x))^2 dx \right)^{1/2} \left(\int_0^1 (u'(x) - v_h'(x))^2 dx \right)^{1/2},$$

c'est-à-dire

$$|e|_1^2 \leq |e|_1\ |u - v_h|_1.$$

Il suffit de simplifier cette inégalité par $|e|_1$ et de prendre le minimum sur $v_h \in V_h$ pour obtenir (10.15) avec $C = 1$. ■

10.4 Méthode d'éléments finis de degré 1

Divisons l'intervalle $[0,1]$ en $N+1$ parties (N étant un entier positif) et posons $h = 1/(N+1)$, $x_i = ih$ avec $i = 0, 1, 2, \ldots, N+1$, comme dans la figure 10.3. On définit, pour $i = 1, 2, \ldots, N$, les fonctions suivantes :

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{si } x_{i-1} \leq x \leq x_i, \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}} & \text{si } x_i \leq x \leq x_{i+1}, \\ 0 & \text{si } x \leq x_{i-1} \text{ ou } x \geq x_{i+1}. \end{cases} \tag{10.18}$$

Le graphe de la fonction φ_i est représenté dans la figure 10.4. Clairement la fonction φ_i est telle que

$$\begin{aligned} &\varphi_i(x_j) = \delta_{ij}, \qquad 0 \leq j \leq N+1, \\ &\varphi_{i_{|[x_{j-1}, x_j]}} \text{ est un polynôme de degré un, } 1 \leq j \leq N+1. \end{aligned} \tag{10.19}$$

Ainsi la fonction φ_i appartient à V. Les fonctions $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont linéairement indépendantes et nous les choisissons pour engendrer l'espace V_h. Nous dirons ainsi que :

- $x_0, x_1, x_2, \ldots, x_{N+1}$ sont les ***nœuds de la discrétisation***,
- $[x_0, x_1], [x_1, x_2], \ldots, [x_N, x_{N+1}]$ sont les ***éléments géométriques***,
- $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont les fonctions de base du sous-espace V_h de type ***éléments finis de degré 1*** associées aux nœuds intérieurs $x_1, x_2, \ldots, x_N$.

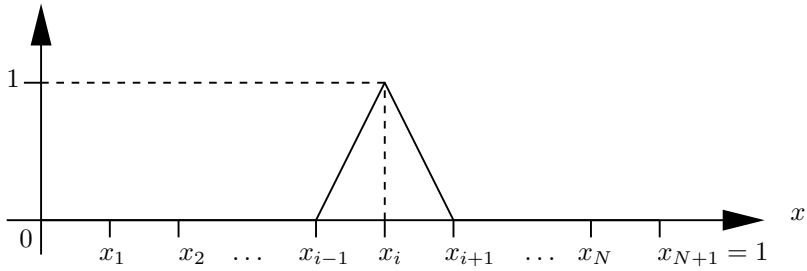

Fig. 10.4 Le graphe de la fonction φ_i.

Si $g \in V_h$, alors g est une combinaison linéaire des φ_i, i.e.

$$g(x) = \sum_{i=1}^{N} g_i \varphi_i(x),$$

et le graphe de g est représenté dans la figure 10.5. En particulier, nous remarquons, en vertu de (10.19), que $g(x_j) = g_j$, $1 \leq j \leq N$, que $g(0) = g(1) = 0$ et que g est une fonction affine sur chaque élément géométrique.

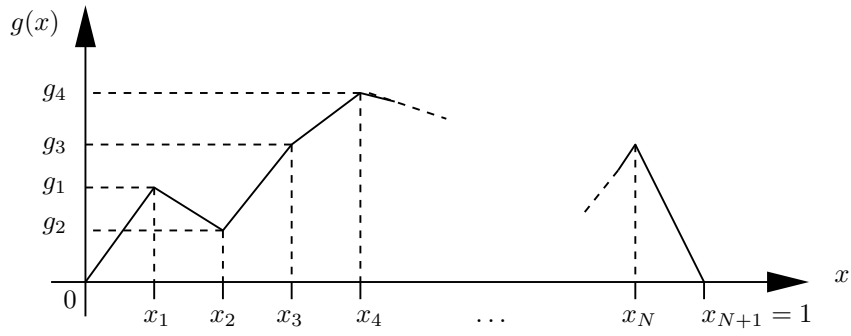

Fig. 10.5 Graphe d'une fonction g élément de V_h.

Si $u \in V$, alors la fonction définie par

$$r_h u = \sum_{i=1}^{N} u(x_i)\varphi_i$$

est l'interpolant de degré 1 par intervalle de la fonction u, voir la définition 1.5. Par construction, $r_h u \in V_h$ et nous avons bien évidemment :

$$\min_{v_h \in V_h} |u - v_h|_1 \leq |u - r_h u|_1. \tag{10.20}$$

Ainsi l'estimation d'erreur (10.15) du théorème 10.2 devient, en utilisant (10.20) :

$$|u - u_h|_1 \leq C|u - r_h u|_1. \tag{10.21}$$

Cette dernière inégalité montre que l'erreur entre u et u_h dans la norme $|\cdot|_1$ peut être contrôlée si nous savons estimer l'erreur d'interpolation entre u et $r_h u$. Cette estimation résulte d'arguments très semblables à ceux qui conduisent au théorème 1.2. Nous démontrons le résultat suivant.

Théorème 10.3 *On suppose que $c(x) \geq 0$, $\forall x \in [0,1]$. Soit u la solution de (10.9) et soit u_h la solution de (10.10) lorsque V_h est engendré par les fonctions de base (10.18). Alors nous avons l'estimation d'erreur :*

$$|u - u_h|_1 \leq Ch, \tag{10.22}$$

où C est une constante indépendante de N (et donc de h).

Démonstration
L'estimation d'erreur (10.22) est une conséquence de (10.21) à condition de montrer que

$$|u - r_h u|_1 \leq \tilde{C}h, \tag{10.23}$$

où $\tilde{C}$ est une constante indépendante de N. Posons donc maintenant

$$w = u - r_h u.$$

Puisque nous avons $r_h u(x_i) = u(x_i)$, $0 \leq i \leq N+1$, nous avons donc $w(x_i) = 0$. En utilisant le théorème de Rolle, nous en déduisons qu'il existe $\xi_i \in]x_i, x_{i+1}[$ tel que $w'(\xi_i) = 0$, $0 \leq i \leq N$. Ainsi, puisque $r_h u$ est un polynôme de degré 1 sur chaque élément géométrique $[x_i, x_{i+1}]$ nous obtenons, pour $x \in [x_i, x_{i+1}]$:

$$w'(x) = \int_{\xi_i}^{x} w''(s)ds = \int_{\xi_i}^{x} u''(s)ds.$$

Nous déduisons de cette égalité que, pour $x \in [x_i, x_{i+1}]$:

$$|w'(x)| \leq \int_{x_i}^{x_{i+1}} |u''(s)|ds.$$

En appliquant l'inégalité de Cauchy-Schwarz nous avons donc, pour $x \in [x_i, x_{i+1}]$:

$$\begin{aligned} |w'(x)| &\leq \left(\int_{x_i}^{x_{i+1}} 1^2 ds \right)^{1/2} \left(\int_{x_i}^{x_{i+1}} |u''(s)|^2 ds \right)^{1/2} \\ &\leq h^{1/2} \left(\int_{x_i}^{x_{i+1}} |u''(s)|^2 ds \right)^{1/2}. \end{aligned}$$

En élevant au carré cette dernière inégalité et en l'intégrant sur l'élément géométrique $[x_i, x_{i+1}]$, nous obtenons :

$$\int_{x_i}^{x_{i+1}} |w'(x)|^2 dx \leq h^2 \int_{x_i}^{x_{i+1}} |u''(s)|^2 ds.$$

Il suffit maintenant de sommer sur l'indice i pour avoir :

$$\begin{aligned} |u - r_h u|_1^2 = |w|_1^2 &= \int_0^1 |w'(x)|^2 dx \\ &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |w'(x)|^2 dx \leq h^2 \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |u''(s)|^2 ds \\ &= h^2 \int_0^1 |u''(s)|^2 ds. \end{aligned}$$

Nous avons donc montré l'inégalité (10.23), la constante $\tilde{C}$ étant donnée par

$$\tilde{C} = \left(\int_0^1 |u''(s)|^2 ds \right)^{1/2}. \qquad \blacksquare$$

Nous avons vu dans la section précédente que la résolution du problème (10.10) nécessitait la construction de la matrice A, du vecteur $\vec{f}$ et la résolution du système linéaire $A\vec{u} = \vec{f}$. Nous sommes donc amenés à calculer les coefficients :

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx,$$

pour $1 \leq i, j \leq N$, ainsi que les composantes

$$f_j = \int_0^1 f(x)\varphi_j(x)dx \qquad \text{pour } 1 \leq j \leq N.$$

Nous vérifions que

$$\int_0^1 \varphi_i'(x)\varphi_j'(x)dx = \begin{cases} 2/h & \text{si } i = j, \\ -1/h & \text{si } |i-j| = 1, \\ 0 & \text{autrement.} \end{cases}$$

D'autre part, pour obtenir les valeurs de $\int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx$ et de $\int_0^1 f(x)\varphi_j(x)dx$, nous pouvons utiliser une formule d'intégration numérique, en l'occurrence la formule composite des trapèzes, voir l'exemple 3.2. Nous approchons donc $\int_0^1 \ell(x)dx$ par

$$L_h(\ell) = h\left(\frac{1}{2}\ell(x_0) + \ell(x_1) + \ell(x_2) + \cdots + \ell(x_N) + \frac{1}{2}\ell(x_{N+1})\right).$$

Nous vérifions aisément que

$$L_h(c\varphi_i\varphi_j) = \begin{cases} hc(x_j) & \text{si } i = j, \\ 0 & \text{si } i \neq j, \end{cases} \tag{10.24}$$

et

$$L_h(f\varphi_j) = hf(x_j). \tag{10.25}$$

Si, dans (10.12), nous remplaçons $\int_0^1 c(x)\varphi_i(x)\varphi_j(x)dx$ par (10.24) et si nous remplaçons (10.13) par (10.25), alors nous vérifions que le système (10.14) est exactement égal à h fois le système obtenu dans (10.6). Nous concluons donc que notre méthode d'éléments finis avec intégration numérique par la formule des trapèzes est strictement équivalente à une méthode de différences finies. Cependant, contrairement à la méthode des différences finies, la méthode des éléments finis est très souple et se laisse facilement généraliser aux situations décrites ci-dessous.

Lorsque la distribution des points de discrétisation $(x_j)_{1\leq j\leq N}$ n'est pas uniforme, les fonctions φ_i peuvent toujours être définies par (10.18). Ainsi, en concentrant les nœuds aux endroits de forte variation de la solution, les fonctions φ_i peuvent engendrer un sous-espace V_h de fonctions mieux adaptées au problème considéré. Lorsque la distribution des points de discrétisation n'est plus uniforme, nous obtenons toujours l'estimation d'erreur (10.22) entre u et u_h si nous définissons h par

$$h = \max_{0\leq i\leq N} |x_{i+1} - x_i|.$$

Lorsque les fonctions de base de V_h sont définies par des polynômes de degré plus élevé que 1 sur chaque élément géométrique $[x_j, x_{j+1}]$ (sect. 1.6), nous

pouvons augmenter la précision de la méthode. Ici les notions d'interpolation par intervalle introduites dans le chapitre 1 interviennent de façon capitale et nous en donnons un exemple dans la section suivante.

Finalement, notons que la théorie des éléments finis est complètement mathématisée. Des résultats de convergence semblables à ceux du théorème 10.3 existent dans des cadres tout à fait généraux contenant bon nombre d'applications physiques.

10.5 Méthode d'éléments finis de degré 2

Divisons l'intervalle $[0,1]$ en $M+1$ parties égales (M étant un entier positif), posons $h = 1/(M+1)$, $x_i = ih$, avec $i = 0, 1, \ldots, M+1$ et $x_{i+1/2} = x_i + h/2$, avec $i = 0, 1, \ldots, M$. On définit pour $i = 1, 2, \ldots, M$, les fonctions suivantes :

$$\psi_i(x) = \begin{cases} \dfrac{(x - x_{i-1})(x - x_{i-\frac{1}{2}})}{(x_i - x_{i-1})(x_i - x_{i-\frac{1}{2}})} & \text{si } x_{i-1} \leq x \leq x_i, \\ \dfrac{(x - x_{i+1})(x - x_{i+\frac{1}{2}})}{(x_i - x_{i+1})(x_i - x_{i+\frac{1}{2}})} & \text{si } x_i \leq x \leq x_{i+1}, \\ 0 & \text{si } x \leq x_{i-1} \text{ ou } x \geq x_{i+1}; \end{cases} \tag{10.26}$$

et pour $i = 0, 1, \ldots, M$, les fonctions suivantes :

$$\psi_{i+\frac{1}{2}}(x) = \begin{cases} \dfrac{(x - x_i)(x - x_{i+1})}{(x_{i+\frac{1}{2}} - x_i)(x_{i+\frac{1}{2}} - x_{i+1})} & \text{si } x_i \leq x \leq x_{i+1}, \\ 0 & \text{si } x \leq x_i \text{ ou } x \geq x_{i+1}. \end{cases} \tag{10.27}$$

Le graphe des fonctions ψ_i et $\psi_{i+1/2}$ est représenté dans la figure 10.6. Clairement les fonctions ψ_i et $\psi_{i+1/2}$ sont telles que :

$$\begin{aligned} &\psi_i(x_j) = \delta_{ij}, \qquad 0 \leq j \leq M+1, \\ &\psi_i(x_{j+\frac{1}{2}}) = 0, \qquad 0 \leq j \leq M, \\ &\psi_{i|_{[x_{j-1}, x_j]}} \text{ est un polynôme de degré 2 }, 1 \leq j \leq M+1; \end{aligned} \tag{10.28}$$

$$\begin{aligned} &\psi_{i+\frac{1}{2}}(x_{j+\frac{1}{2}}) = \delta_{ij}, \qquad 0 \leq j \leq M, \\ &\psi_{i+\frac{1}{2}}(x_j) = 0, \qquad 0 \leq j \leq M+1, \\ &\psi_{i+\frac{1}{2}|_{[x_{j-1}, x_j]}} \text{ est un polynôme de degré 2 }, 1 \leq j \leq M+1. \end{aligned} \tag{10.29}$$

Si nous posons maintenant $N = 2M+1$, $\varphi_1 = \psi_{1/2}$, $\varphi_2 = \psi_1$, $\varphi_3 = \psi_{3/2}$, $\varphi_4 = \psi_2$, $\varphi_5 = \psi_{5/2}$, $\varphi_6 = \psi_3$, ..., $\varphi_{2M} = \psi_M$, $\varphi_{2M+1} = \psi_{M+1/2}$, alors les fonctions $\varphi_1, \varphi_2, \ldots, \varphi_N$ appartiennent à V et sont linéairement indépendantes. Nous les choisissons pour engendrer l'espace V_h que nous appelons ici encore espace de type éléments finis. Nous dirons ainsi que :

- $x_0, x_1, x_2, \ldots, x_{M+1}$ sont les ***nœuds principaux de la discrétisation***,

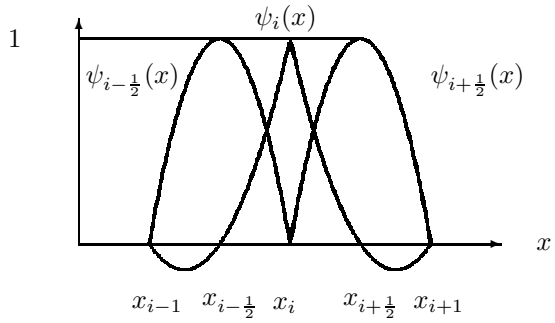

Fig. 10.6 Graphe des fonctions $\psi_{i-1/2}$, ψ_i et $\psi_{i+1/2}$.

- $[x_0, x_1], [x_1, x_2], \ldots, [x_M, x_{M+1}]$ sont les ***éléments géométriques***,
- $x_{1/2}, x_{3/2}, x_{5/2}, \ldots, x_{M+1/2}$ sont les ***nœuds intérieurs aux éléments géométriques***,
- $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont les fonctions de base du sous-espace V_h de type ***éléments finis de degré 2*** associées aux nœuds de la discrétisation.

Si $g \in V_h$, alors g est une combinaison linéaire des φ_i, i.e.

$$g(x) = \sum_{i=1}^{N} g_i \varphi_i(x),$$

et le graphe de g est représenté dans la figure 10.7. En particulier nous remarquons, en vertu de (10.28) (10.29), que $g(x_j) = g_{2j}$, $1 \leq j \leq M$, que $g(x_{j+1/2}) = g_{2j+1}$, $0 \leq j \leq M$, que $g(0) = g(1) = 0$ et que g est un polynôme de degré 2 sur chaque élément géométrique.

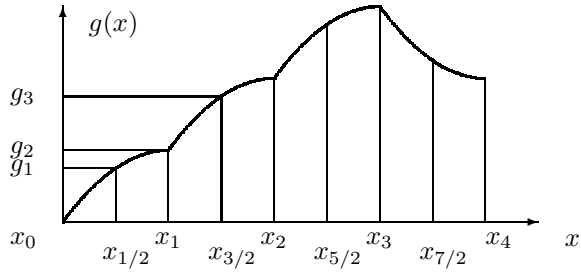

Fig. 10.7 Graphe d'une fonction g élément de V_h dans le cas où $M = 3$.

Si $u \in V$, alors la fonction définie par

$$r_h u = \sum_{j=1}^{M} u(x_j)\varphi_{2j} + \sum_{j=0}^{M} u(x_{j+\frac{1}{2}})\varphi_{2j+1}$$

est l'interpolant de degré 2 par intervalle de la fonction u, voir la définition 1.5. Par construction, $r_h u \in V_h$ et nous avons bien évidemment :

$$\min_{v_h \in V_h} |u - v_h|_1 \leq |u - r_h u|_1. \tag{10.30}$$

Si u est trois fois continûment dérivable, nous pouvons montrer de façon analogue au théorème 10.3 qu'il existe une constante C indépendante de h et N telle que

$$|u - r_h u|_1 \leq Ch^2. \tag{10.31}$$

Les relations (10.15), (10.30) et (10.31) conduisent donc au résultat suivant :

Théorème 10.4 *On suppose que $c(x) \geq 0$, $\forall x \in [0,1]$. Soit u la solution de (10.9) que nous supposons trois fois continûment dérivable. Si u_h est la solution de (10.10) lorsque V_h est engendré par les fonctions de base (10.26) (10.27), alors nous avons l'estimation d'erreur :*

$$|u - u_h|_1 \leq Ch^2, \tag{10.32}$$

où C est une constante indépendante de N (et donc de h).

Nous constatons que la méthode de type éléments finis de degré 2 est plus précise que la méthode de type éléments finis de degré 1 décrite dans la section précédente. Il suffit de se rapporter à la section 1.6 concernant l'interpolation par intervalles pour généraliser ces résultats à des méthodes de type éléments finis de degré k, où k est un entier plus grand que 2.

10.6 Approximation par différences finies d'un problème aux limites non linéaire

Revenons au problème (10.1) et supposons maintenant que c ne dépende pas seulement de x mais aussi de l'inconnue du problème u. Nous considérons donc une fonction f donnée ainsi qu'une fonction à deux variables $\tilde{c} : (x, v) \in [0,1] \times \mathbb{R} \longrightarrow \tilde{c}(x, v) \in \mathbb{R}$ et nous posons le problème aux limites suivant :

$$\begin{aligned} &-u''(x) + \tilde{c}(x, u(x)) = f(x), \qquad \text{si } 0 < x < 1, \\ &u(0) = u(1) = 0, \end{aligned} \tag{10.33}$$

où u est naturellement la fonction inconnue. Si $\tilde{c}(x, v) = c(x)v$, où $c(x)$ est une fonction donnée, nous retrouvons bien le problème (10.1).

Pour résoudre numériquement (10.33), nous considérons à nouveau la discrétisation correspondant à la figure 10.3 et, de façon semblable à (10.3), nous avons, si u_j est une approximation de $u(x_j)$:

$$\begin{aligned} &\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + \tilde{c}(x_j, u_j) = f(x_j) \qquad 1 \leq j \leq N, \\ &u_0 = u_{N+1} = 0. \end{aligned} \tag{10.34}$$

Clairement (10.34) est une approximation par différences finies de (10.33) ; nous sommes en présence d'un système non linéaire de N équations pour les N inconnues $u_1, u_2, \ldots, u_N$.

Nous choisissons maintenant de résoudre le système (10.34) par la méthode de Newton (8.22). Soit $\vec{u}$ le N-vecteur de composantes $u_1, u_2, \ldots, u_N$ et soit $F(\vec{u})$ la fonction de $\mathbb{R}^N$ dans $\mathbb{R}^N$ définie par

$$F(\vec{u}) = \begin{bmatrix} \dfrac{2u_1 - u_2}{h^2} & + & \tilde{c}(x_1, u_1) - f(x_1) \\ \dfrac{-u_1 + 2u_2 - u_3}{h^2} & + & \tilde{c}(x_2, u_2) - f(x_2) \\ & \vdots & \\ \dfrac{-u_{N-2} + 2u_{N-1} - u_N}{h^2} & + & \tilde{c}(x_{N-1}, u_{N-1}) - f(x_{N-1}) \\ \dfrac{-u_{N-1} + 2u_N}{h^2} & + & \tilde{c}(x_N, u_N) - f(x_N) \end{bmatrix}. \quad (10.35)$$

Clairement le problème (10.34) est équivalent à chercher $\vec{u}$ tel que :

$$F(\vec{u}) = 0. \quad (10.36)$$

Soit $\vec{u}^0$ une solution approchée de $\vec{u}$. La méthode de Newton pour résoudre numériquement (10.36) s'écrit :

$$\vec{u}^{n+1} = \vec{u}^n - DF(\vec{u}^n)^{-1} F(\vec{u}^n), \qquad n = 0, 1, 2, \ldots \quad (10.37)$$

Supposons que $\tilde{c}$ soit assez régulière pour définir une fonction continue d par

$$d(x, v) = \frac{\partial}{\partial v} \tilde{c}(x, v).$$

Pour simplifier l'écriture nous notons encore

$$d_j^n \underset{def}{=} d(x_j, u_j^n), \qquad 1 \leq j \leq N.$$

Il est alors facile de vérifier que la matrice $DF(\vec{u}^n)$ est donnée par :

$$DF(\vec{u}^n) = \frac{1}{h^2} \begin{bmatrix} 2 + d_1^n h^2 & -1 & & & \\ -1 & 2 + d_2^n h^2 & -1 & & \bigcirc \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ \bigcirc & & & -1 & 2 + d_N^n h^2 \end{bmatrix}. \quad (10.38)$$

Ainsi, pour calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$ dans (10.37), on calculera :

- le vecteur $F(\vec{u}^n)$ en remplaçant $u_1, u_2, \ldots, u_N$ dans (10.35) par u_1^n, u_2^n, ..., u_N^n ;
- la matrice $DF(\vec{u}^n)$ par l'expression (10.38) ;
- le vecteur $\vec{y}$ solution de $DF(\vec{u}^n)\vec{y} = F(\vec{u}^n)$ (chap. 4, 5 ou 6) ;
- le vecteur $\vec{u}^{n+1} = \vec{u}^n - \vec{y}$.

Si $\vec{u}^0$ est choisi suffisammment proche de $\vec{u}$ et si F satisfait des hypothèses raisonables, nous savons que la méthode de Newton converge quadratiquement. L'algorithme ci-dessus permet donc d'obtenir, en quelques itérations, une solution numérique du problème (10.33).

Remarquons enfin que si $\tilde{c}(x,u) = c(x)u$ comme dans (10.1), alors $d(x,u) = c(x)$ et par suite $d_j^n = c(x_j) = c_j$. Dans ce cas, la matrice $DF(\vec{u}^n)$ est égale à la matrice A donnée dans (10.5) et devient indépendante de n. Ainsi, (10.37) nous fournit pour $n = 1$ (un seul pas de la méthode de Newton) la solution de (10.36).

10.7 Exercices

Exercice 10.1 Soit $f : [0,1] \to \mathbb{R}$ une fonction continue donnée. On cherche une fonction $u : [0,1] \to \mathbb{R}$ telle que

$$\begin{aligned} -\frac{d}{dx}\left((1+x)\frac{du}{dx}(x)\right) &= f(x), \qquad \text{si } 0 < x < 1, \\ u(0) = u(1) &= 0. \end{aligned} \tag{10.39}$$

1. Donner une formulation faible du problème ci-dessus.
2. Soit N un entier positif, $h = 1/(N+1)$, $x_j = jh$, $j = 0, 1, \ldots, N+1$. On considère la méthode de Galerkin dans le cas où l'espace V_h est engendré par les fonctions de base φ_1, φ_2, ..., φ_N définies par (10.18). Calculer explicitement la matrice du système à résoudre.

Solution

1. Considérons le problème (10.39) et multiplions sa première équation par une fonction v une fois continûment dérivable sur $[0,1]$. Si nous intégrons sur l'intervalle $[0,1]$, nous obtenons :

$$-\int_0^1 \frac{d}{dx}\left((1+x)\frac{du}{dx}(x)\right) v(x)dx = \int_0^1 f(x)v(x)dx.$$

En intégrant par parties le premier terme de l'équation ci-dessus, nous avons :

$$\int_0^1 (1+x)u'(x)v'(x)dx - \Big[(1+x)u'(x)v(x)\Big]_{x=0}^{x=1} = \int_0^1 f(x)v(x)dx.$$

Si nous imposons à la fonction v d'être nulle en $x = 0$ et en $x = 1$ alors nous en déduisons l'égalité :

$$\int_0^1 (1+x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \tag{10.40}$$

Soit V l'ensemble de toutes les fonctions g continues, de première dérivée g' continue par morceaux et telles que $g(0) = g(1) = 0$. La formulation faible du problème (10.39) consiste à chercher $u \in V$ qui satisfait (10.40) pour toute fonction $v \in V$.

2. Soit V_h l'espace engendré par les fonctions de base $\varphi_1, \varphi_2, \ldots, \varphi_N$ définies par (10.18). L'approximation de Galerkin de (10.40) consiste à trouver une fonction $u_h \in V_h$ telle que

$$\int_0^1 (1+x) u_h'(x) v_h'(x) dx = \int_0^1 f(x) v_h(x) dx$$

pour toute fonction $v_h \in V_h$. En procédant comme dans la section 10.3, nous pouvons montrer que les coefficients de la matrice de rigidité A sont définis par

$$A_{ji} = \int_0^1 (1+x) \varphi_i'(x) \varphi_j'(x) dx, \qquad 1 \le i, j \le N.$$

Puisque les fonctions φ_i', $1 \le i \le N$, sont constantes sur chaque intervalle $[x_{j-1}, x_j]$, $1 \le j \le N+1$, l'intégrant ci-dessus est un polynôme de degré un par intervalle. Nous pouvons donc utiliser la formule de quadrature des trapèzes sur chaque élément géométrique pour calculer exactement l'intégrale ci-dessus.

Comme dans la section 10.4 la matrice A est symétrique et tridiagonale. Il suffit donc de déterminer les coefficients A_{ii}, $1 \le i \le N$, et $A_{i,i+1}$, $1 \le i \le N-1$. Nous avons

$$\begin{aligned} A_{ii} &= \int_{x_{i-1}}^{x_i} (1+x) \varphi_i'(x) \varphi_i'(x) dx + \int_{x_i}^{x_{i+1}} (1+x) \varphi_i'(x) \varphi_i'(x) dx \\ &= \frac{h}{2}(1 + x_{i-1} + 1 + x_i) \frac{1}{h^2} + \frac{h}{2}(1 + x_i + 1 + x_{i+1}) \frac{1}{h^2} \\ &= \frac{2}{h}(1 + ih). \end{aligned}$$

De même nous obtenons

$$\begin{aligned} A_{i,i+1} &= \int_{x_i}^{x_{i+1}} (1+x) \varphi_{i+1}'(x) \varphi_i'(x) dx \\ &= \frac{h}{2}(1 + x_i + 1 + x_{i+1}) \left(-\frac{1}{h^2} \right) \\ &= -\frac{1}{h}(1 + ih + h/2). \end{aligned}$$

Exercice 10.2 Soit $f : [0,1] \to \mathbb{R}$ une fonction continue donnée et soit α un nombre réel positif. On cherche une fonction $u : [0,1] \to \mathbb{R}$ telle que

$$\begin{aligned} -u''(x) &= f(x) \qquad \text{si } 0 < x < 1, \\ u(0) &= 0, \\ u'(1) + \alpha u(1) &= 0. \end{aligned} \tag{10.41}$$

1. Donner une formulation faible du problème ci-dessus (choisir une fonction test nulle en $x = 0$ seulement).
2. Soit N un entier positif, $h = 1/(N+1)$, $x_j = jh$, $j = 0, 1, \ldots, N+1$. Définir l'espace V_h à partir de fonctions de base affines sur chaque élément géométrique $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, N+1$.
3. Calculer explicitement la matrice et le second membre du système à résoudre lorsque celui-ci est intégré numériquement par la formule des trapèzes.

Solution
Considérons le problème (10.41) et multiplions sa première équation par une fonction v une fois continûment dérivable sur $[0, 1]$. Si nous intégrons sur l'intervalle $[0, 1]$, nous obtenons :

$$-\int_0^1 u''(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

En intégrant par parties le premier terme, nous avons :

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) = \int_0^1 f(x)v(x)dx.$$

Si nous imposons à la fonction v d'être nulle en $x = 0$, alors nous en déduisons l'égalité :

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) = \int_0^1 f(x)v(x)dx.$$

Finalement si nous utilisons la troisième équation de (10.41), nous obtenons

$$\int_0^1 u'(x)v'(x)dx + \alpha u(1)v(1) = \int_0^1 f(x)v(x)dx. \tag{10.42}$$

Soit V l'ensemble de toutes les fonctions g continues, de première dérivée g' continue par morceaux et telles que $g(0) = 0$. La formulation faible du problème (10.41) consiste à chercher $u \in V$ qui satisfait (10.42) pour toute fonction $v \in V$. En fait, il est possible de montrer que le problème (10.42) a une et une seule solution qui est celle du problème (10.41).

2. Puisque l'espace V n'impose pas que les fonctions s'annulent en $x = 1$, nous pouvons choisir pour fonctions linéairement indépendantes les fonctions φ_1, φ_2, ..., φ_N définies par (10.18), auquelles nous y ajoutons la fonction φ_{N+1} définie par

$$\varphi_{N+1}(x) = \begin{cases} \dfrac{x - x_N}{x_{N+1} - x_N} & \text{si } x_N \leq x \leq x_{N+1}, \\ 0 & \text{sinon.} \end{cases}$$

Soit maintenant V_h l'espace vectoriel engendré par ces fonctions de base φ_1, φ_2, ..., φ_N, φ_{N+1}. L'approximation de Galerkin de (10.42) consiste à trouver une fonction $u_h \in V_h$ telle que

$$\int_0^1 u_h'(x)v_h'(x)dx + \alpha u_h(1)v_h(1) = \int_0^1 f(x)v_h(x)dx \tag{10.43}$$

pour toute fonction $v_h \in V_h$. Procédons comme dans la section 10.3. Ecrivons u_h dans la base de V_h, c'est-à-dire

$$u_h = \sum_{i=1}^{N+1} u_j \varphi_j,$$

et choisissons $v_h = \varphi_1$, $v_h = \varphi_2$, ..., $v_h = \varphi_N$, $v_h = \varphi_{N+1}$. Le problème (10.43) est alors équivalent à chercher $u_1, u_2, \ldots, u_N, u_{N+1}$ tels que

$$\sum_{i=1}^{N+1} u_i \left(\int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \alpha\varphi_i(1)\varphi_j(1) \right) = \int_0^1 f(x)\varphi_j(x)dx$$

pour tout $j = 1, 2, \ldots, N, N+1$. Il s'agit donc de résoudre le système linéaire $A\vec{u} = \vec{f}$, la matrice A étant la $(N+1) \times (N+1)$ matrice dont les coefficients sont définis par

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx + \alpha\varphi_i(1)\varphi_j(1), \qquad 1 \leq i, j \leq N+1,$$

le vecteur $\vec{f}$ étant le $(N+1)$-vecteur de composantes

$$f_j = \int_0^1 f(x)\varphi_j(x)dx, \qquad 1 \leq j \leq N+1.$$

Comme dans les exemples précédents, la matrice A est tridiagonale et symétrique. Puisque les fonctions de base φ_1, φ_2, ..., φ_N sont nulles au point $x = 1$, le calcul des N premières lignes de la matrice A se fait comme dans la section 10.4. Nous obtenons donc :

$$A_{ii} = \frac{2}{h}, \qquad A_{i,i+1} = -\frac{1}{h}, \qquad 1 \leq i \leq N.$$

Le calcul du dernier coefficient diagonal donne

$$\begin{aligned} A_{N+1,N+1} &= \int_0^1 \varphi_{N+1}'(x)\varphi_{N+1}'(x)dx + \alpha\varphi_{N+1}(1)\varphi_{N+1}(1) \\ &= \frac{1}{h} + \alpha. \end{aligned}$$

Comme dans la section 10.4 nous approchons les composantes f_j du vecteur $\vec{f}$ en utilisant la formule de quadrature des trapèzes. Nous obtenons :

$$\begin{aligned} f_j &= \int_0^1 f(x)\varphi_j(x)dx \simeq hf(x_j) \qquad 1 \leq j \leq N, \\ f_{N+1} &= \int_0^1 f(x)\varphi_{N+1}(x)dx \simeq \frac{h}{2} f(x_{N+1}). \end{aligned}$$

Finalement la matrice A et le second membre $\vec{f}$ du système linéaire $A\vec{u} = \vec{f}$ sont donnés par

$$A = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1+\alpha h \end{bmatrix} \qquad \vec{f} = h\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \\ 1/2\ f(x_{N+1}) \end{bmatrix}.$$

10.8 Notes bibliographiques et remarques

Dans ce chapitre, nous n'avons présenté que deux méthodes pour résoudre le problème (10.1), en l'occurence une méthode de différences finies et une méthode d'éléments finis. La méthodes des différences finies est simple à mettre en œuvre, par contre, elle se transpose difficilement à des applications industrielles (problèmes tridimensionnels sur des géométries complexes, conditions aux limites). Le principe de la méthode des éléments finis est plus difficile à comprendre mais la méthode est généralisable à un grand nombre de problèmes industriels. Dans le chapitre suivant nous étudions ces deux méthodes dans un cadre bidimensionnel.

Dans ce chapitre, nous avons présenté la méthode de Galerkin en liaison avec la méthode des éléments finis. Il existe d'autres méthodes utilisées dans la pratique en liaison avec la méthode de Galerkin. Citons par exemple les ***méthodes spectrales*** [24] qui sont efficaces lorsque la solution du problème à résoudre est très régulière.

Il existe de nombreuses autres méthodes pour résoudre le problème (10.1). Citons par exemple les ***méthodes de collocation*** [24], les ***méthodes de volumes finis*** [24, 22]. Notons finalement que la ***tranformée de Fourier rapide*** [28, 11] ou la ***tranformée en ondelettes*** [29, 21] sont des méthodes très utilisées en traitement du signal.

Chapitre 11

Une méthode d'éléments finis pour l'approximation de problèmes elliptiques

11.1 Problèmes elliptiques et formulation variationnelle

Soit Ω un domaine polygonal dans le plan $O_{x_1x_2}$ de frontière $\partial\Omega$ et soit $\overline{\Omega} = \Omega \cup \partial\Omega$. Soit x un point de coordonnées (x_1, x_2) dans le domaine Ω. Dans la suite nous notons indifféremment x ou (x_1, x_2). Soit a_{11}, a_{12}, a_{21}, a_{22} quatre fonctions de x données (ici encore nous notons indifféremment $a_{ij}(x)$ ou $a_{ij}(x_1, x_2)$, $1 \leq i, j \leq 2$) que nous supposons continues et une fois continûment dérivables par rapport à x_1 et x_2. Soit encore $f : \overline{\Omega} \to \mathbb{R}$ une fonction continue donnée. Nous posons le problème de trouver une fonction $u : \overline{\Omega} \to \mathbb{R}$ satisfaisant les relations :

$$-\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u(x)\right) = f(x) \qquad \forall x \in \Omega, \tag{11.1}$$

$$u(x) = 0 \qquad \forall x \in \partial\Omega, \tag{11.2}$$

où la notation $\partial/\partial x_i$ désigne l'opération de dérivation partielle par rapport à la variable x_i, $i = 1$ ou 2. Nous dirons que le problème (11.1) (11.2) est un problème différentiel aux limites d'ordre 2, la condition limite étant l'équation (11.2).

Définition 11.1 *Nous dirons que le problème (11.1) (11.2) est fortement elliptique si les fonctions a_{ij}, $1 \leq i, j \leq 2$, sont telles qu'il existe un nombre positif α qui satisfasse, pour tout $x \in \Omega$ et pour tout couple de nombres réels (ξ_1, ξ_2), la relation*

$$\sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \geq \alpha(\xi_1^2 + \xi_2^2).$$

Remarquons que si le problème (11.1) (11.2) est fortement elliptique au sens de la définition ci-dessus alors, pour tout $x \in \Omega$, l'équation en ξ_1, ξ_2 donnée par

$$a_{11}(x)\xi_1^2 + (a_{12}(x) + a_{21}(x))\xi_1\xi_2 + a_{22}(x)\xi_2^2 = 1$$

est l'équation d'une ellipse (ou d'un cercle), d'où la terminologie ***problème elliptique***.

Supposons maintenant que les fonctions a_{ij} soient constantes en x et que $a_{12} = a_{21}$. Alors une condition nécessaire et suffisante pour que le problème (11.1) (11.2) soit fortement elliptique est que la matrice

$$S = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

soit symétrique définie positive.

Les problèmes elliptiques interviennent lors de la modélisation de problèmes physiques tels que les problèmes de potentiel, de déformation de membranes, d'écoulements de fluides. Dans la suite de ce chapitre nous décrivons la méthode des éléments finis, qui est souvent utilisée pour résoudre numériquement ce type de problèmes.

Mentionnons qu'un exemple typique de problème fortement elliptique est donné par $a_{11} = a_{22} = 1$ et $a_{12} = a_{21} = 0$. Dans la suite, nous nous restreignons à ce cas et nous cherchons donc une fonction $u : \overline{\Omega} \to \mathbb{R}$ satisfaisant

$$-\Delta u(x) = f(x) \qquad \forall x \in \Omega, \tag{11.3}$$

$$u(x) = 0 \qquad \forall x \in \partial\Omega, \tag{11.4}$$

où Δu est le laplacien de u, i.e. $\Delta u = \partial^2 u/\partial x_1^2 + \partial^2 u/\partial x_2^2$. Le problème (11.3) (11.4) est appelé ***problème de Poisson***.

La solution u du problème de Poisson modélise le déplacement vertical $u(x)$ au point x d'une membrane Ω tendue, attachée à $\partial\Omega$, et soumise à une densité de force verticale et proportionnelle à f (fig. 11.1).

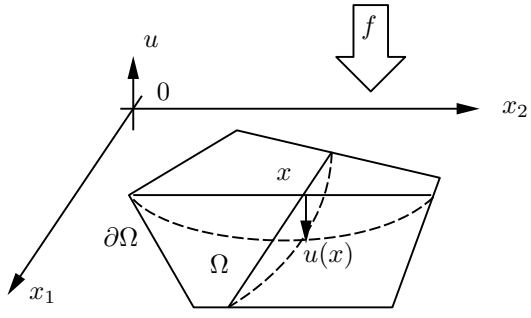

Fig. 11.1 Déformation verticale d'une membrane soumise à une force verticale.

Comme nous l'avons déjà fait dans la section 10.3, nous pouvons multiplier la première équation de (11.3) par une fonction $v : \overline{\Omega} \to \mathbb{R}$ suffisamment régulière et intégrer sur Ω. Nous obtenons

$$-\iint_{\Omega} \Delta u(x)\ v(x)dx = \iint_{\Omega} f(x)v(x)dx, \tag{11.5}$$

où $x = (x_1, x_2)$ et $dx = dx_1 dx_2$. En utilisant la formule :

$$\mathrm{div}(v\ \overrightarrow{\mathrm{grad}}u) = \overrightarrow{\mathrm{grad}}v \cdot \overrightarrow{\mathrm{grad}}u + v\Delta u, \tag{11.6}$$

nous obtenons à partir de (11.5) :

$$\iint_{\Omega} \overrightarrow{\mathrm{grad}}u(x) \cdot \overrightarrow{\mathrm{grad}}v(x)dx - \iint_{\Omega} \mathrm{div}\left(v(x)\overrightarrow{\mathrm{grad}}u(x)\right) dx = \iint_{\Omega} f(x)v(x)dx. \tag{11.7}$$

Le théorème de la divergence nous assure que :

$$\iint_{\Omega} \mathrm{div}\left(v(x)\overrightarrow{\mathrm{grad}}u(x)\right) dx = \int_{\partial\Omega} v(s)\frac{\partial u}{\partial n}(s)ds, \tag{11.8}$$

où $\partial u/\partial n$ est la dérivée de u dans la direction normale extérieure à $\partial\Omega$. Si nous imposons que v s'annule sur $\partial\Omega$, nous pouvons déduire de (11.7) et (11.8) :

$$\iint_{\Omega} \overrightarrow{\mathrm{grad}}u(x) \cdot \overrightarrow{\mathrm{grad}}v(x)dx = \iint_{\Omega} f(x)v(x)dx. \tag{11.9}$$

Soit maintenant V l'ensemble de toutes les fonctions $g : \overline{\Omega} \to \mathbb{R}$ qui sont continues sur $\overline{\Omega}$, nulles sur $\partial\Omega$, et dont les premières dérivées partielles $\partial g/\partial x_1$, $\partial g/\partial x_2$ sont continues par morceaux (ici l'expression fonction continue par morceaux n'est pas définie avec précision ; grosso modo cette expression signifie que $\overline{\Omega}$ peut être partitionné en un nombre fini de morceaux sur lesquels la fonction est continue). Nous observons que la somme de deux éléments de V reste un élément de V. De même, un élément de V multiplié par un nombre réel est encore un élément de V. Ainsi V est un espace vectoriel. Nous allons maintenant

chercher $u \in V$ qui satisfait (11.9) pour toute fonction $v \in V$.

Dans la suite, ce problème est appelé problème (11.9). Comme nous l'avions déjà fait dans la section 10.3, nous obtenons une ***formulation faible*** ou ***variationnelle*** du problème (11.3) (11.4).

La formulation variationnelle (11.9) a souvent une signification physique. Par exemple, dans le cas de la membrane, elle traduit le fait qu'une énergie est minimisée. De même, nous pourrions donner une formulation faible du problème plus général (11.1) (11.2).

Contrairement au cas unidimensionnel du chapitre 10, le problème (11.9) peut, selon la forme du domaine Ω et selon l'expression du second membre f,

ne pas avoir de solution u dans V. Cependant, nous supposerons dans la suite que ce n'est pas le cas et nous appellerons u la solution de (11.9).

Dans le chapitre 10, nous avons vu qu'une méthode d'approximation consiste à construire un sous-espace V_h de dimension finie de V et à résoudre le problème (11.9) dans V_h au lieu de V, c'est-à-dire de trouver $u_h \in V_h$ tel que

$$\iint_\Omega \overrightarrow{\text{grad}} u_h(x) \cdot \overrightarrow{\text{grad}} v_h(x) dx = \iint_\Omega f(x) v_h(x) dx \tag{11.10}$$

pour toute fonction $v_h \in V_h$. Rappelons que le problème (11.10) est appelé ***approximation de Galerkin*** de (11.9). Si $\varphi_1, \varphi_2, \ldots, \varphi_N$ est une base de V_h, nous pouvons écrire $u_h(x) = u_1\varphi_1(x) + \cdots + u_N\varphi_N(x)$ et choisir $v_h = \varphi_j$, $j = 1, 2, \ldots, N$ dans (11.10). Soit $\vec{u}$ le N-vecteur de composantes $u_1, u_2, \ldots, u_N$, soit A la $N \times N$ matrice de coefficients

$$A_{ji} = \iint_\Omega \overrightarrow{\text{grad}} \varphi_i(x) \cdot \overrightarrow{\text{grad}} \varphi_j(x) dx, \tag{11.11}$$

et $\vec{f}$ le N-vecteur de composantes $f_1, f_2, \ldots, f_N$ définies par

$$f_j = \iint_\Omega f(x)\varphi_j(x) dx. \tag{11.12}$$

Pour obtenir la solution de (11.10), il suffit donc de trouver $u_1, u_2, \ldots, u_N$ tels que

$$\sum_{i=1}^{N} A_{ji} u_i = f_j \qquad j = 1, 2, \ldots, N,$$

ou, de façon équivalente, de résoudre le système linéaire

$$A\vec{u} = \vec{f}. \tag{11.13}$$

Il est maintenant naturel de se poser la question suivante : la matrice A est-elle régulière ? Le théorème ci-dessous répond par l'affirmative.

Théorème 11.1 *A est une matrice symétrique définie positive.*

Démonstration
La symétrie est évidente. Pour montrer que A est définie positive, il suffit de constater que, pour un N-vecteur $\vec{y}$ de composantes $y_1, y_2, \ldots, y_N$, nous avons :

$$\begin{aligned}
\vec{y}^T A \vec{y} &= \sum_{i,j=1}^{N} y_i A_{ij} y_j = \sum_{i,j=1}^{N} y_i y_j \iint_\Omega \overrightarrow{\text{grad}} \varphi_i(x) \cdot \overrightarrow{\text{grad}} \varphi_j(x) dx \\
&= \iint_\Omega \left| \sum_{i=1}^{N} y_i \overrightarrow{\text{grad}} \varphi_i(x) \right|^2 dx.
\end{aligned}$$

Posons

$$\psi(x) = \sum_{i=1}^{N} y_i \varphi_i(x).$$

Nous obtenons

$$\vec{y}^T A \vec{y} = \iint_{\Omega} \left| \overrightarrow{\text{grad}} \psi(x) \right|^2 dx,$$

qui est toujours positif ou nul. Si $\vec{y}^T A \vec{y} = 0$ alors $\overrightarrow{\text{grad}} \psi(x) = 0 \ \forall x \in \overline{\Omega}$, ce qui implique $\psi(x) =$ constante. Puisque ψ est nul sur $\partial\Omega$, on aura $\psi(x) = 0 \ \forall x \in \overline{\Omega}$ et donc $\vec{y} = \vec{0}$. ■

De façon analogue à ce que nous avons déjà fait dans le cas unidimensionnel (sect. 10.4), nous proposons maintenant une construction simple d'un sous-espace V_h de V.

11.2 Eléments finis triangulaires de degré 1

Nous voulons construire des sous-espaces V_h de V de type ***éléments finis triangulaires*** (rappelons que Ω est un domaine polygonal de $\mathbb{R}^2$). Pour ce faire, nous construisons une ***triangulation*** $\mathcal{T}_h$ de $\overline{\Omega}$ en subdivisant $\overline{\Omega}$ en triangles $K_1, K_2, \ldots, K_m$ ne se recouvrant pas et tels que

- $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K = K_1 \cup K_2 \cup K_3 \cup \cdots \cup K_m$,
- 2 triangles K_i et K_j, $i \neq j$, possèdent ou bien un côté commun, ou bien un sommet P_ℓ commun, ou bien sont disjoints (fig. 11.2).

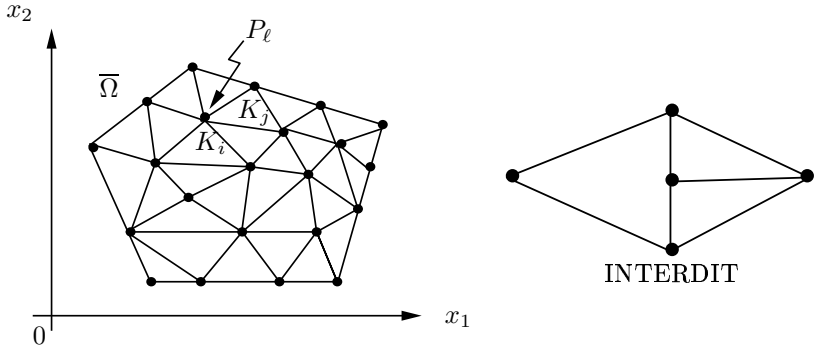

Fig. 11.2 Triangulation de $\overline{\Omega}$.

Nous introduisons encore un paramètre h mesurant le degré de finesse de la triangulation $\mathcal{T}_h$:

$$h = \max_{K \in \mathcal{T}_h} \text{diam } (K),$$

où diam (K) est le diamètre de K, c'est-à-dire le maximum des distances euclidiennes entre deux points de K. Le sous-espace V_h de dimension finie de V est défini par

$$\begin{aligned} V_h = &\{g : \overline{\Omega} \to \mathbb{R}\,;\, g \text{ est continue sur } \overline{\Omega}, \text{ s'annule sur } \partial\Omega, \\ &\text{la restriction de } g \text{ à tout triangle } K \\ &\text{de la triangulation est un polynôme de degré} \leq 1\}. \end{aligned} \tag{11.14}$$

Soit P_i, $i = 1, 2, \ldots, N$ les sommets intérieurs de la triangulation $\mathcal{T}_h$, encore appelés ***nœuds*** intérieurs. Pour décrire une fonction $g \in V_h$, nous pouvons choisir comme paramètres les valeurs $g(P_i)$ de la fonction g aux nœuds P_i, $i = 1, 2, \ldots, N$. Ces valeurs $g(P_i)$ sont appelées ***degrés de liberté***. Les fonctions de base φ_1, φ_2, ..., φ_N de V_h sont alors définies par

$$\varphi_i(P_j) = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \qquad i, j = 1, \ldots, N, \tag{11.15}$$

et sont nulles sur $\partial\Omega$ (fig. 11.3). Le support de φ_i (adhérence de l'ensemble des points où φ_i n'est pas nul) est la réunion de tous les triangles qui ont pour sommet P_i. Puisque toute fonction g de V_h peut être représentée par une combinaison linéaire des φ_i, nous avons

$$g(x) = \sum_{i=1}^{N} g_i \varphi_i(x) \qquad \text{où } g_i \in \mathbb{R} \quad i = 1, 2, \ldots, N,$$

et nous obtenons bien $g(P_i) = g_i$, $i = 1, 2, \ldots, N$, en vertu de (11.15).

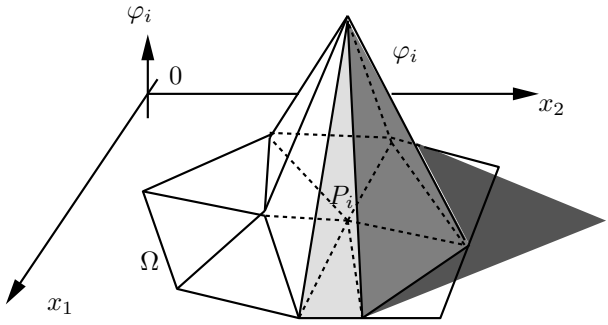

Fig. 11.3 La fonction de base φ_i.

La triangulation étant donnée, nous savons maintenant comment définir les fonctions de base $\varphi_1, \varphi_2, \ldots, \varphi_N$ et nous pouvons donc construire la matrice A et le vecteur second membre $\vec{f}$ du système linéaire (11.13). Notons que le calcul numérique de f_j, $j = 1, 2, \ldots, N$, peut nécessiter l'usage d'une formule de quadrature numérique.

Après avoir constitué la matrice A, souvent appelée ***matrice de rigidité***, ainsi que le vecteur $\vec{f}$, nous allons résoudre le système linéaire (11.13) en effectuant la décomposition de Cholesky de la matrice A, puis en résolvant deux

systèmes linéaires triangulaires (chap. 5). La solution $\vec{u}$ ainsi obtenue nous permet d'exprimer la solution approchée u_h du problème (11.3) (11.4) sous la forme $u_h(x) = u_1\varphi_1(x)+\cdots+u_N\varphi_N(x)$. En vertu de (11.15) nous aurons $u_h(P_j) = u_j$, $j = 1, 2, \ldots, N$.

Remarque 11.1 Si, en lieu et place du problème particulier (11.3) (11.4), nous considérons le problème général (11.1) (11.2), nous pouvons vérifier que la matrice de rigidité A a pour coefficients

$$A_{ij} = \iint_\Omega \left(a_{11}\frac{\partial\varphi_i}{\partial x_1}\frac{\partial\varphi_j}{\partial x_1} + a_{12}\frac{\partial\varphi_i}{\partial x_1}\frac{\partial\varphi_j}{\partial x_2} + a_{21}\frac{\partial\varphi_i}{\partial x_2}\frac{\partial\varphi_j}{\partial x_1} + a_{22}\frac{\partial\varphi_i}{\partial x_2}\frac{\partial\varphi_j}{\partial x_2}\right) dx.$$

Nous vérifions facilement que, si le problème (11.1) (11.2) est elliptique et si $a_{12} = a_{21}$, alors A reste une matrice symétrique définie positive et tout ce qui précède s'applique encore !

11.3 Un exemple particulier

Soit Ω le carré unité de $\mathbb{R}^2$ de frontière $\partial\Omega$ et soit L un entier positif. Posons $\tilde{h} = 1/(L+1)$ et notons Q_{ij} les points de coordonnées $x_1 = i\tilde{h}$ et $x_2 = j\tilde{h}$, $i, j = 0, 1, \ldots, L+1$. Considérons la triangulation $\mathcal{T}_h$ de $\overline{\Omega}$ ayant pour nœuds les points Q_{ij} (fig. 11.4). La triangulation $\mathcal{T}_h$ contient $N \equiv L^2$ nœuds intérieurs à Ω que nous numérotons comme dans la figure 11.4 ligne par ligne, c'est-à-dire $P_1 = Q_{11}$, $P_2 = Q_{21}$, $P_3 = Q_{31}$, …, $P_L = Q_{L1}$, $P_{L+1} = Q_{12}$, $P_{L+2} = Q_{22}$, …, $P_{2L} = Q_{L2}$, $P_{2L+1} = Q_{13}$, …, $P_N = Q_{LL}$.

Le support de φ_1 est représenté dans la figure 11.5 ; il est constitué des 6 triangles K_1, K_2, …, K_6. Puisque φ_1 est un polynôme de degré 1 sur chacun des triangles K_1, K_2, …, K_6 et puisque $\varphi_1(P_1) = 1$ et $\varphi_1(Q_{ij}) = 0$ si $(i,j) \neq (1,1)$, nous vérifions facilement que

$$\overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}1\\0\end{pmatrix} \text{ sur } K_1 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}-1\\0\end{pmatrix} \text{ sur } K_4,$$

$$\overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}0\\-1\end{pmatrix} \text{ sur } K_2 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}0\\1\end{pmatrix} \text{ sur } K_5,$$

$$\overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}-1\\-1\end{pmatrix} \text{ sur } K_3 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{\tilde{h}}\begin{pmatrix}1\\1\end{pmatrix} \text{ sur } K_6.$$

Un calcul simple nous permet d'affirmer que

$$A_{11} = \iint_\Omega \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx = \sum_{k=1}^{6} \iint_{K_k} \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx = 4.$$

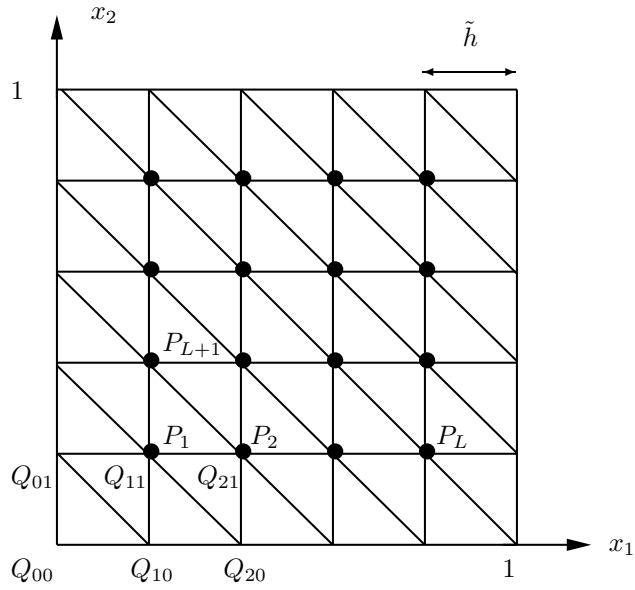

Fig. 11.4 Le carré unité et sa triangulation pour $L = 4$.

D'après la figure 11.5, l'intersection entre le support de la fonction φ_1 et celui de la fonction φ_2 est réduit aux triangles K_3 et K_4. Puisque

$$\overrightarrow{\text{grad}}\varphi_2 = \frac{1}{\tilde{h}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ sur } K_3 \quad \text{et} \quad \overrightarrow{\text{grad}}\varphi_2 = \frac{1}{\tilde{h}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ sur } K_4,$$

un calcul simple conduit à

$$\begin{aligned} A_{12} = A_{21} &= \iint_{\Omega} \overrightarrow{\text{grad}}\varphi_1 \cdot \overrightarrow{\text{grad}}\varphi_2 dx \\ &= \iint_{K_3} \overrightarrow{\text{grad}}\varphi_1 \cdot \overrightarrow{\text{grad}}\varphi_2 dx + \iint_{K_4} \overrightarrow{\text{grad}}\varphi_1 \cdot \overrightarrow{\text{grad}}\varphi_2 dx = -1. \end{aligned}$$

De façon semblable, nous montrons que

$$A_{1,L+1} = A_{L+1,1} = \iint_{\Omega} \overrightarrow{\text{grad}}\varphi_1 \cdot \overrightarrow{\text{grad}}\varphi_{L+1} dx = -1.$$

Le terme croisé $A_{2,L+1}$ est nul car $\overrightarrow{\text{grad}}\varphi_2$ et $\overrightarrow{\text{grad}}\varphi_{L+1}$ sont orthogonaux. En considérant à nouveau la figure 11.4, nous constatons que pour des raisons de symétrie

$$\begin{aligned} &A_{ii} = A_{11} = 4, && i = 1, 2, \ldots, N, \\ &A_{i,i+1} = A_{i+1,i} = A_{12} = -1, && i = 1, 2, \ldots, N-1, \quad i \neq L \bmod L, \\ &A_{i,L+i} = A_{L+i,i} = A_{1,L+1} = -1, && i = 1, 2, \ldots, N-L; \end{aligned}$$

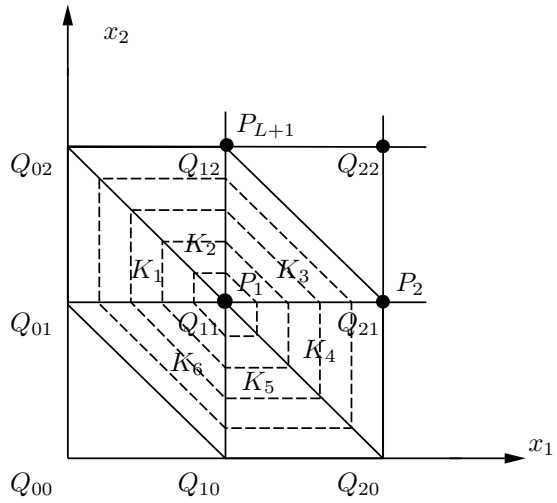

Fig. 11.5 Support et isovaleurs de la fonction de base φ_1.

tous les autres coefficients A_{ij} sont nuls. Pour $L = 4$ la matrice A a donc l'allure suivante

$$A = \begin{bmatrix} B & C & & \\ C & B & C & \\ & C & B & C \\ & & C & B \end{bmatrix}, \tag{11.16}$$

où nous avons noté

$$B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{bmatrix} \text{ et } C = \begin{bmatrix} -1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{bmatrix}.$$

Approchons le second membre $\vec{f}$ de (11.13) en utilisant la formule de quadrature

$$\iint_K g(x)dx \simeq \frac{g(S_1) + g(S_2) + g(S_3)}{3} \text{ aire}(K) \tag{11.17}$$

pour tout triangle K de sommets S_1, S_2, S_3 (nous pouvons remarquer que cette formule est exacte si g est un polynôme de degré 1, i.e. si $g(x) = \alpha x_1 + \beta x_2 + \gamma$). Nous obtenons donc :

$$f_1 = \iint_\Omega f(x)\varphi_1(x)dx = \sum_{j=1}^{6} \iint_{K_j} f(x)\varphi_1(x)dx$$

$$\simeq \frac{1}{3} f(P_1) \cdot 6 \cdot \frac{\tilde{h}^2}{2} = f(P_1)\tilde{h}^2.$$

De même, pour des raisons de symétrie (fig. 11.5), nous avons

$$f_j \simeq f(P_j)\tilde{h}^2 \qquad \text{avec } j = 1, 2, \ldots, N. \tag{11.18}$$

Résoudre le problème (11.13) revient, dans le cas particulier de cette section et en considérant le second membre intégré numériquement par la formule de quadrature (11.17), à résoudre le système linéaire

$$A\vec{u} = \vec{f}, \tag{11.19}$$

la matrice A et le vecteur $\vec{f}$ étant donnés par (11.16) et (11.18) respectivement.

Remarque 11.2 Considérons les résultats du chapitre 10, en particulier les égalités (10.2) (10.3). Nous pouvons écrire, si $u : \overline{\Omega} \to \mathbb{R}$ est assez régulière et si P est un nœud de coordonnées $(i\tilde{h}, j\tilde{h})$:

$$\begin{aligned}\Delta u(P) &= \frac{\partial^2 u}{\partial x_1^2}(P) + \frac{\partial^2 u}{\partial x_2^2}(P) \\ &= \frac{u((i+1)\tilde{h}, j\tilde{h}) - 2u(i\tilde{h}, j\tilde{h}) + u((i-1)\tilde{h}, j\tilde{h})}{\tilde{h}^2} + O(\tilde{h}^2) \\ &+ \frac{u(i\tilde{h}, (j+1)\tilde{h}) - 2u(i\tilde{h}, j\tilde{h}) + u(i\tilde{h}, (j-1)\tilde{h})}{\tilde{h}^2} + O(\tilde{h}^2).\end{aligned}$$

Notons $u_{i,j} = u(i\tilde{h}, j\tilde{h})$. Nous avons donc :

$$-\Delta u(P) \simeq \frac{4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}}{\tilde{h}^2}. \tag{11.20}$$

Soit alors $U_{i,j}$ une approximation par différences finies de $u_{i,j} = u(i\tilde{h}, j\tilde{h})$ lorsque u vérifie $-\Delta u = f$ dans Ω, $u = 0$ sur $\partial\Omega$. Le schéma numérique pour calculer $U_{i,j}$ s'écrit :

$$\frac{4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}}{\tilde{h}^2} = f(i\tilde{h}, j\tilde{h}), \tag{11.21}$$

pour $i, j = 1, \ldots, L$ et

$$U_{0,j} = U_{L+1,j} = U_{j,0} = U_{j,L+1} = 0 \qquad \text{pour } j = 0, 1, \ldots, L+1. \tag{11.22}$$

En renumérotant les variables $U_{i,j}$, $i, j = 1, \ldots, L$, comme nous l'avons déjà fait pour les nœuds de la figure 11.4, c'est-à-dire en posant $u_1 = U_{1,1}$, $u_2 = U_{2,1}$, $u_3 = U_{3,1}$, ..., $u_L = U_{L,1}$, $u_{L+1} = U_{1,2}$, $u_{L+2} = U_{2,2}$, ..., $u_N = U_{LL}$, nous constatons que le système (11.21) (11.22) est strictement équivalent au système (11.19). Donc, dans ce cas particulier, la méthode des éléments finis avec intégration numérique du second membre est équivalente à la méthode des différences finies.

11.4 Estimations d'erreurs et méthodes de degré supérieur

Considérons à nouveau le problème de Poisson (11.3) (11.4) et soit $u : \overline{\Omega} \to \mathbb{R}$ sa solution. Notons encore u_h la solution de son approximation de Galerkin (11.10) construite avec des espaces V_h de type éléments finis triangulaires de degré 1 définis par (11.14). La question que nous posons maintenant est la suivante : que devient l'erreur entre u et u_h lorsque la triangulation devient de plus en plus fine, c'est-à-dire lorsque $h = \max_{K \in \mathcal{T}_h} \text{diam } (K)$ tend vers zéro ?

Notons

$$|u - u_h|_0 = \left(\iint_\Omega |u(x) - u_h(x)|^2 dx \right)^{1/2}$$

la norme quadratique de l'erreur $u - u_h$ et

$$|u - u_h|_1 = \left(\iint_\Omega |\overrightarrow{\text{grad}}(u(x) - u_h(x))|^2 dx \right)^{1/2}$$

la norme quadratique du gradient de l'erreur $u - u_h$. Nous avons le résultat suivant :

Théorème 11.2 *Si u est assez régulière (par exemple deux fois continûment dérivable sur $\overline{\Omega}$) et si les angles des triangles qui constituent la triangulation ne deviennent pas infiniment petits lorsque h tend vers zéro, nous avons*

$$|u - u_h|_0 \leq Ch^2, \tag{11.23}$$

$$|u - u_h|_1 \leq Ch, \tag{11.24}$$

où la constante C est indépendante de la taille du maillage h.

Ce résultat s'interprète de la manière suivante. Si les triangles K d'une triangulation $\mathcal{T}_h$ sont coupés en quatre comme sur la figure 11.6, alors le diamètre h des triangles devient deux fois plus petit, la norme quadratique de l'erreur sera réduite en principe d'un facteur quatre, alors que la norme quadratique du gradient de l'erreur sera réduite en principe d'un facteur deux.

Ce résultat peut être amélioré en utilisant des polynômes de degré $k > 1$ pour construire l'espace V_h. Dans ce cas les estimations d'erreur (11.23) (11.24) deviennent $|u - u_h|_0 \leq Ch^{k+1}$ et $|u - u_h|_1 \leq Ch^k$, pour autant que la fonction u soit très régulière.

Le cas $k = 2$ est illustré dans la figure 11.7. A chaque sommet de triangle et à chaque milieu d'arête est attachée une fonction de base polynomiale de degré 2 sur chaque triangle, nulle en tous les nœuds sauf un, où elle vaut 1. Quelques lignes du graphe de ces fonctions sont données dans la figure 11.7. Elles sont à comparer avec les graphes dessinés dans la figure 10.6.

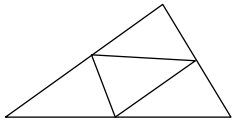

Fig. 11.6 Fragmentation d'un triangle en quatre, en utilisant le milieu des arêtes.

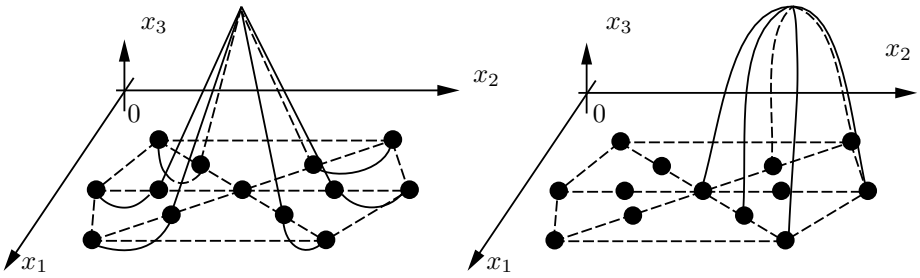

Fig. 11.7 Fonctions de base polynômiales de degré 2, associée au sommet d'un triangle (fig. de gauche), et associée au milieu d'une arête (fig. de droite).

11.5 Exercices

Exercice 11.1 (Eléments finis rectangulaires) Nous cherchons à résoudre le problème (11.3) (11.4) dans le cas où Ω est le carré unité de $\mathbb{R}^2$. Soit L un entier positif. Posons $\tilde{h} = 1/(L+1)$ et notons Q_{ij} les points de coordonnées $(i\tilde{h}, j\tilde{h})$, $i,j = 0,1,\ldots,L+1$. Considérons la subdivision de Ω en rectangles ayant pour sommets les points Q_{ij}. Cette subdivision contient $N \equiv L^2$ nœuds intérieurs à Ω. A chaque nœud intérieur Q_{ij}, $1 \leq i,j \leq L$, nous allons associer une fonction de base ϕ_{ij} construite par tensorisation de fonctions de base à une variable. Soit $s_j = j\tilde{h}$, $j = 0,1,2,\ldots,L+1$, une discrétisation de l'intervalle $[0,1]$. Nous pouvons associer aux sommets $s_1, s_2, \ldots, s_L$ les fonctions de base $\psi_1, \psi_2, \ldots, \psi_L$ à une variable s décrites dans (10.18), à savoir

$$\begin{aligned}
&\psi_i : [0,1] \to \mathbb{R} \text{ est une fonction continue,}\\
&\psi_i \text{ restreinte à l'intervalle } [s_j, s_{j+1}]\\
&\qquad\qquad \text{est un polynôme de degré 1 pour } j = 0,1,2,\ldots,L,\\
&\psi_i(s_j) = \delta_{ij} \qquad j = 0,1,2,\ldots,L+1.
\end{aligned}$$

Les fonctions de base ϕ_{ij} sont alors définies par

$$\phi_{ij}(x_1,x_2) = \psi_i(x_1)\psi_j(x_2) \qquad 1 \leq i,j \leq L.$$

En renumérotant les nœuds ligne par ligne comme dans la figure 11.4, c'est-à-dire $P_1 = Q_{11}$, $P_2 = Q_{21}$, $P_3 = Q_{31}$, ..., $P_L = Q_{L1}$, $P_{L+1} = Q_{12}$, $P_{L+2} = Q_{22}$, ..., $P_{2L} = Q_{L2}$, $P_{2L+1} = Q_{13}$, ..., $P_N = Q_{LL}$, et en renumérotant de la même

manière les fonctions de base, c'est-à-dire $\varphi_1 = \phi_{11}$, $\varphi_2 = \phi_{21}$, $\varphi_3 = \phi_{31}$, $\ldots$, $\varphi_L = \phi_{L1}$, $\varphi_{L+1} = \phi_{12}$, $\varphi_{L+2} = \phi_{22}$, $\ldots$, $\varphi_{2L} = \phi_{L2}$, $\varphi_{2L+1} = \phi_{13}$, $\ldots$, $\varphi_N = \phi_{LL}$, nous aurons associé à chaque nœud intérieur P_j une fonction de base φ_j. Finalement l'espace V_h utilisé dans l'approximation de Galerkin (11.10) est celui engendré par les fonctions $\varphi_1, \varphi_2, \ldots, \varphi_N$.

Le support de φ_1 est représenté dans la figure 11.8 ; il est constitué des 4 rectangles K_1, K_2, K_3, K_4.

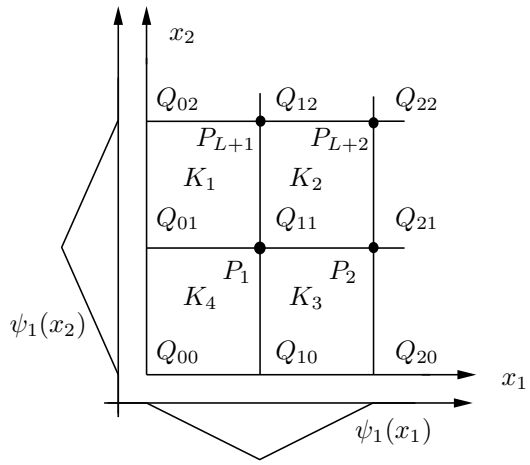

Fig. 11.8 Support de la fonction de base φ_1.

L'exercice consiste à construire la matrice de rigidité A comme nous l'avons déjà fait dans la section 11.3, mais cette fois-ci avec l'espace V_h décrit ci-dessus.

Solution
Par définition de φ_1, nous avons

$$\overrightarrow{\text{grad}}\varphi_1(x_1, x_2) = \begin{pmatrix} \psi_1'(x_1)\psi_1(x_2) \\ \psi_1(x_1)\psi_1'(x_2) \end{pmatrix},$$

et par conséquent

$$\overrightarrow{\text{grad}}\varphi_1(x_1,x_2) = \frac{1}{\tilde{h}}\begin{pmatrix} \psi_1(x_2) \\ -\psi_1(x_1) \end{pmatrix} \quad \text{sur } K_1,$$

$$\overrightarrow{\text{grad}}\varphi_1(x_1,x_2) = \frac{1}{\tilde{h}}\begin{pmatrix} -\psi_1(x_2) \\ -\psi_1(x_1) \end{pmatrix} \quad \text{sur } K_2,$$

$$\overrightarrow{\text{grad}}\varphi_1(x_1,x_2) = \frac{1}{\tilde{h}}\begin{pmatrix} -\psi_1(x_2) \\ \psi_1(x_1) \end{pmatrix} \quad \text{sur } K_3,$$

$$\overrightarrow{\text{grad}}\varphi_1(x_1,x_2) = \frac{1}{\tilde{h}}\begin{pmatrix} \psi_1(x_2) \\ \psi_1(x_1) \end{pmatrix} \quad \text{sur } K_4.$$

En utilisant la formule de Simpson qui est exacte pour l'intégration de polynômes de degré 3 (sect. 3.4), nous obtenons

$$\begin{aligned} \iint_{K_1} \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx &= \frac{1}{\tilde{h}^2}\iint_{K_1} \big(\psi_1^2(x_1)+\psi_1^2(x_2)\big)dx_1dx_2 \\ &= \frac{1}{\tilde{h}}\left(\int_0^{\tilde{h}} \psi_1^2(x_1)dx_1 + \int_{\tilde{h}}^{2\tilde{h}} \psi_1^2(x_2)dx_2\right) \\ &= \frac{1}{6}\left(1+\frac{4}{4}\right)+\frac{1}{6}\left(1+\frac{4}{4}\right) = \frac{2}{3}. \end{aligned}$$

Pour des raisons de symétrie, nous avons donc

$$A_{11} = \iint_{\Omega} \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx = 4\iint_{K_1} \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx = \frac{8}{3}.$$

Nous procédons de la même manière pour calculer A_{12} et $A_{1,L+1}$ et nous obtenons $A_{12} = A_{1,L+1} = -1/3$. Contrairement à l'exemple traité dans la section 11.3, le terme croisé $A_{1,L+2}$ n'est pas nul et vaut $-1/3$.

Dans le cas où $L = 4$, la matrice A a l'allure donnée dans (11.16). Par contre, les matrices B et C sont les suivantes :

$$B = \frac{1}{3}\begin{bmatrix} 8 & -1 & & \\ -1 & 8 & -1 & \\ & -1 & 8 & -1 \\ & & -1 & 8 \end{bmatrix} \text{ et } C = \frac{1}{3}\begin{bmatrix} -1 & -1 & & \\ -1 & -1 & -1 & \\ & -1 & -1 & -1 \\ & & -1 & -1 \end{bmatrix}.$$

Exercice 11.2 Soit Ω le carré unité de $\mathbb{R}^2$ de frontière $\partial\Omega$. Il s'agit de trouver une fonction $u : \overline{\Omega} \to \mathbb{R}^2$ telle que

$$-\frac{\partial}{\partial x_1}\left((1+x_1+x_2)\frac{\partial}{\partial x_1}u(x_1,x_2)\right) - \frac{\partial^2}{\partial x_2^2}u(x_1,x_2) = f \qquad \forall (x_1,x_2) \in \Omega, \tag{11.25}$$

$$u(x_1,x_2) = 0 \qquad \forall (x_1,x_2) \in \partial\Omega, \tag{11.26}$$

où $f : \overline{\Omega} \to \mathbb{R}$ est une fonction donnée.

1. Etablir une formulation faible du problème (11.25) (11.26).
2. Si Ω est subdivisé en 4 triangles comme dans la figure 11.9, construire explicitement la méthode d'éléments finis triangulaires de degré 1 pour approcher numériquement la solution du problème (11.25) (11.26). Utiliser la formule de quadrature (11.17) pour évaluer les intégrales. Montrer que la valeur de la solution obtenue par la méthode des éléments finis au point milieu du carré unité est égale à $1/18$ lorsque $f = 1$.

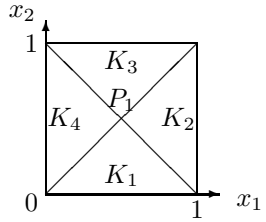

Fig. 11.9 Subdivision du carré unité en quatre triangles.

Solution

1. Pour établir une formulation faible du problème (11.25) (11.26), nous procédons comme dans la section 11.1. Multiplions (11.25) par une fonction $v : \overline{\Omega} \to \mathbb{R}$ et intégrons sur Ω. Nous obtenons

$$-\iint_\Omega \mathrm{div}\Big(\vec{w}(x)\Big)v(x)dx = \iint_\Omega f(x)v(x)dx, \tag{11.27}$$

où nous avons noté $x = (x_1, x_2)$, $dx = dx_1 dx_2$ et $\vec{w}$ le vecteur défini par

$$\vec{w}(x_1,x_2) = \begin{bmatrix} (1+x_1+x_2)\dfrac{\partial}{\partial x_1}u(x_1,x_2) \\ \dfrac{\partial}{\partial x_2}u(x_1,x_2) \end{bmatrix}.$$

En utilisant la formule

$$\mathrm{div}(v\vec{w}) = v\mathrm{div}\vec{w} + \vec{w}\cdot\overrightarrow{\mathrm{grad}}v,$$

nous obtenons

$$-\iint_\Omega \vec{w}(x)\cdot\overrightarrow{\text{grad}}v(x)dx+\iint_\Omega \text{div}\Big(v(x)\vec{w}(x)\Big)dx=\iint_\Omega v(x)\text{div}\vec{w}(x)dx.$$

En utilisant le théorème de la divergence, nous avons :

$$-\iint_\Omega \vec{w}(x)\cdot\overrightarrow{\text{grad}}v(x)dx+\iint_{\partial\Omega} v(s)\vec{w}(s)\cdot\vec{n}(s)ds=\iint_\Omega v(x)\text{div}\vec{w}(x)dx,$$

où $\vec{n}(x)$ désigne la normale extérieure à Ω. Finalement, si nous imposons que v s'annule sur $\partial\Omega$ et si nous utilisons la définition de $\vec{w}$, nous avons alors en utilisant (11.27)

$$\iint_\Omega \left((1+x_1+x_2)\frac{\partial u}{\partial x_1}(x)\frac{\partial v}{\partial x_1}(x)+\frac{\partial u}{\partial x_2}(x)\frac{\partial v}{\partial x_2}(x)\right)dx$$
$$=\iint_\Omega f(x)v(x)dx. \quad (11.28)$$

Soit V l'ensemble de toutes les fonctions $g:\overline{\Omega}\to\mathbb{R}$ qui sont continues sur $\overline{\Omega}$, nulles sur $\partial\Omega$, et dont les premières dérivées partielles $\partial g/\partial x_1$, $\partial g/\partial x_2$ sont continues par morceaux. La formulation faible du problème (11.25) (11.26) consiste donc à chercher $u\in V$ tel que (11.28) soit satisfaite pour toute fonction $v\in V$.

2. Soit V_h le sous-espace de V défini par (11.14). Le problème discrétisé consiste à chercher $u_h\in V_h$ tel que (11.28) soit satisfaite pour toute fonction $v_h\in V_h$. Dans le cas particulier de la figure 11.9, l'espace V_h est de dimension 1. Notons P_1 le sommet situé au centre du carré unité, notons encore φ_1 la fonction de base de V_h valant 1 au point P_1 et 0 aux quatre sommets du carré unité. Développons u_h dans la base φ_1, c'est-à dire écrivons

$$u_h(x_1,x_2)=u_1\varphi_1(x_1,x_2),$$

et prenons $v_h=\varphi_1$. Alors (11.28) devient

$$A_{11}u_1=f_1,$$

où

$$A_{11}=\iint_\Omega \left((1+x_1+x_2)\frac{\partial\varphi_1}{\partial x_1}(x)\frac{\partial\varphi_1}{\partial x_1}(x)+\frac{\partial\varphi_1}{\partial x_2}(x)\frac{\partial\varphi_1}{\partial x_2}(x)\right)dx$$

et

$$f_1=\iint_\Omega f(x)\varphi_1(x)dx.$$

Par définition, les dérivées partielles de φ_1 sont constantes sur chaque triangle K_1, K_2, K_3, K_4. Plus précisément nous avons

$$\overrightarrow{\text{grad}}\varphi_1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \text{ sur } K_1, \qquad \overrightarrow{\text{grad}}\varphi_1 = \begin{pmatrix} -2 \\ 0 \end{pmatrix} \text{ sur } K_2,$$

$$\overrightarrow{\text{grad}}\varphi_1 = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \text{ sur } K_3, \qquad \overrightarrow{\text{grad}}\varphi_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \text{ sur } K_4,$$

et, par conséquent

$$A_{11} = \iint_{K_1} 4dx + \iint_{K_2} 4(1 + x_1 + x_2)dx + \iint_{K_3} 4dx + \iint_{K_4} 4(1 + x_1 + x_2)dx,$$

et

$$f_1 = \sum_{i=1}^{4} \iint_{K_i} f(x)\varphi_1(x)dx.$$

Les intégrales pour calculer A_{11} ne font donc apparaître que des polynômes de degré au plus égal à 1. Nous pouvons donc, si besoin est, utiliser la formule de quadrature (11.17) pour calculer exactement ces intégrales. Nous obtenons finalement $A_{11} = 18/3$. Si nous approchons f_1 en utilisant la formule (11.17), nous avons $f_1 \simeq f(P_1)/3$. Si $f(x) = 1$ alors $f_1 = 1/3$ et nous avons $u_1 = f_1/A_{11} = 1/18$.

Exercice 11.3 (Un problème de Poisson tridimensionnel) Soit $\Omega = [0,1]^3$ le cube unité, nous cherchons une fonction $u : \overline{\Omega} \to \mathbb{R}$ telle que

$$-\Delta u(x) = f(x) \qquad \forall x \in \Omega, \tag{11.29}$$

$$u(x) = 0 \qquad \forall x \in \partial\Omega, \tag{11.30}$$

où $x = (x_1, x_2, x_3)$ et Δu est le laplacien de u, i.e. $\Delta u = \partial^2 u/\partial x_1^2 + \partial^2 u/\partial x_2^2 + \partial^2 u/\partial x_3^2$.

Soit L un entier positif. Posons $\tilde{h} = 1/(L+1)$ et notons $U_{i,j,k}$ une approximation par différences finies de $u(i\tilde{h}, j\tilde{h}, k\tilde{h})$, $i, j, k = 1, 2, \ldots, L$. Le schéma numérique pour calculer $U_{i,j,k}$ s'écrit

$$\frac{6U_{i,j,k} - U_{i+1,j,k} - U_{i-1,j,k} - U_{i,j+1,k} - U_{i,j-1,k} - U_{i,j,k+1} - U_{i,j,k-1}}{\tilde{h}^2} = f(i\tilde{h}, j\tilde{h}, k\tilde{h}), \tag{11.31}$$

pour $i, j, k = 1, \ldots, L$ et

$$U_{0,j,k} = U_{L+1,j,k} = U_{j,0,k} = U_{j,L+1,k} = U_{j,k,0} = U_{j,k,L+1} = 0, \tag{11.32}$$

pour $j, k = 1, 2, \ldots, L$. Renumérotons les variables $U_{i,j,k}$, $i, j, k = 1, 2, \ldots, L$, comme indiqué dans la figure 11.10, c'est-à-dire en posant $u_1 = U_{1,1,1}$, $u_2 = U_{2,1,1}$, $\ldots$, $u_L = U_{L,1,1}$, $u_{L+1} = U_{1,2,1}$, $u_{L+2} = U_{2,2,1}$, $\ldots$, $u_{2L} = U_{L,2,1}$, $\ldots$, $u_{L^2} = U_{L,L,1}$, $u_{L^2+1} = U_{1,1,2}$ $\ldots$, $u_{L^3} = U_{L,L,L}$. Nous écrivons les systèmes (11.31) (11.32) sous forme d'un système linéaire $A\vec{u} = \vec{f}$, de L^3 équations et L^3 inconnues.

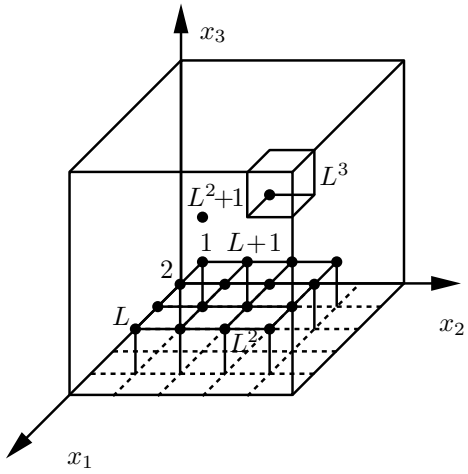

Fig. 11.10 Numérotations des inconnues dans le cube unité pour $L = 4$.

L'exercice consiste à :

1. Justifier le schéma (11.31).
2. Montrer que, pour $L = 4$, la matrice A prend l'allure suivante :

$$A = \frac{1}{\tilde{h}^2} \begin{bmatrix} B & C & & \\ C & B & C & \\ & C & B & C \\ & & C & B \end{bmatrix},$$

où les matrices B et C sont formées de blocs

$$B = \begin{bmatrix} D & E & & \\ E & D & E & \\ & E & D & E \\ & & E & D \end{bmatrix} \qquad C = \begin{bmatrix} E & & & \\ & E & & \\ & & E & \\ & & & E \end{bmatrix},$$

avec

$$D = \begin{bmatrix} 6 & -1 & & \\ -1 & 6 & -1 & \\ & -1 & 6 & -1 \\ & & -1 & 6 \end{bmatrix} \qquad E = \begin{bmatrix} -1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{bmatrix}.$$

3. Montrer que le nombre d'opérations pour résoudre le système linéaire $A\vec{u} = \vec{f}$ par une méthode Cholesky est d'ordre L^7 (utiliser le théorème 5.4).
4. Proposer une méthode d'éléments finis pour résoudre numériquement le problème (11.29) (11.30).

11.6 Notes bibliographiques et remarques

Il reste beaucoup de choses à dire sur la méthode des éléments finis. En particulier, nous avons laissé de côté ici les questions suivantes :

- Techniques de génération de maillages [9, 20], de mémorisation et construction des matrices [19, 20] ; algorithmes auto-adaptatifs [8].
- Généralisation à des éléments quadrangulaires plutôt que triangulaires (exercice 11.1).
- Généralisation à des problèmes posés dans un espace tridimensionnel (exercice 11.3).
- Généralisation à des méthodes d'éléments finis non standard [24].

Nous avons étudié dans ce chapitre des méthodes numériques pour résoudre un problème elliptique du type (11.1) (11.2). Il existe d'autres problèmes elliptiques. Citons par exemple le système de l'élasticité décrivant la déformation d'un solide élastique, le système de Stokes décrivant le mouvement stationnaire d'un fluide incompressible visqueux. Les problèmes issus de la modélisation des plaques et coques sont souvent des problèmes elliptiques. Par exemple, le problème du bi-laplacien $\Delta(\Delta u) = f$ est un problème elliptique du quatrième ordre.

Pour un exposé de la méthode des éléments finis appliquée à la mécanique du solide, nous renvoyons le lecteur à [15, 5]. Pour une approche plus mathématique nous renvoyons à [3].

Actuellement, nous trouvons sur le marché un grand nombre de logiciels de calcul scientifique que les ingénieurs utilisent pour faire des simulations numériques dans des domaines aussi divers que la mécanique des fluides, la mécanique des structures, l'électromagnétisme, la thermique, la fonderie, etc. La plupart de ces logiciels utilisent la technique des éléments finis.

Chapitre 12

Approximation des problèmes paraboliques. Problème de la chaleur

12.1 Equation de la chaleur 1D et différences finies

Considérons un barreau métallique de longueur L et dont les deux extrémités sont en contact avec des réservoirs de chaleur de température constante égale à $0°C$. Supposons que ce barreau occupe l'intervalle $[0, L]$ de l'axe Ox et qu'au temps $t = 0$ sa température soit connue en tout point $x \in]0, L[$ et égale à $w(x)$, $x \in]0, L[$. Supposons en outre avoir placé sous le barreau une source de chaleur $f(x, t)$, donnée. La quantité $f(x, t)$ représente la puissance par unité de longueur fournie au point $x \in]0, L[$ et à l'instant $t > 0$. Si ρ, c_p et k sont des constantes positives données, représentant respectivement la densité volumique, la chaleur spécifique massique et la conductivité thermique, la température $u(x, t)$ du barreau au point x et à l'instant t est liée à $f(x, t)$ par l'équation :

$$\rho c_p \frac{\partial u}{\partial t}(x,t) - k\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) \qquad \forall x \in]0, L[, \quad \forall t > 0. \tag{12.1}$$

A cette équation on adjoint les ***conditions aux limites*** :

$$u(0,t) = u(L,t) = 0 \qquad \forall t > 0, \tag{12.2}$$

et la ***condition initiale*** :

$$u(x,0) = w(x) \qquad \forall x \in]0, L[. \tag{12.3}$$

L'équation (12.1) est souvent appelée ***équation de la chaleur*** et traduit le principe de conservation de l'énergie calorifique emmagasinée dans le barreau.

Dans la suite, pour alléger l'écriture, nous prendrons les constantes ρ, c_p et L égales à 1, ce qui ne modifie pas fondamentalement le problème mathématique. Ainsi, nous cherchons la fonction u satisfaisant les relations suivantes :

$$\frac{\partial u}{\partial t}(x,t) - k\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) \qquad \forall x \in]0,1[, \quad \forall t > 0, \tag{12.4}$$

$$u(0,t) = u(1,t) = 0 \qquad \forall t > 0, \tag{12.5}$$

$$u(x,0) = w(x) \qquad \forall x \in]0,1[. \tag{12.6}$$

Pour résoudre numériquement le problème (12.4) (12.5) (12.6) par la méthode des différences finies, nous commençons par le discrétiser par rapport à la variable x de façon semblable à ce qui a été fait dans la section 10.1. Si N est un entier positif, nous posons $h = \frac{1}{N+1}$ et $x_i = ih$ avec $i = 0, 1, 2, \ldots, N+1$. Soit $u_i(t)$ une approximation de $u(x,t)$ au point $x = x_i$. Nous noterons $u_i(t) \simeq u(x_i,t)$, $i = 1, 2, \ldots, N$. Au vu de la section 10.1, il est naturel de considérer le schéma :

$$\frac{d}{dt}u_i(t) + \frac{k}{h^2}\Big(-u_{i-1}(t)+2u_i(t) - u_{i+1}(t)\Big) = f(x_i,t)$$

$$i = 1, \ldots, N, \quad \forall t > 0 \tag{12.7}$$

$$u_0(t) = u_{N+1}(t) = 0 \qquad \forall t > 0, \tag{12.8}$$

$$u_i(0) = w(x_i) \qquad i = 1, \ldots, N. \tag{12.9}$$

Les fonctions $u_i(t)$, $i = 1, \ldots, N$, sont maintenant les inconnues du problème.

Nous dirons que le schéma (12.7) (12.8) (12.9) est une ***semi-discrétisation*** en espace du problème (12.4) (12.5) (12.6) par la méthode des différences finies. Si A est la $N \times N$ matrice tridiagonale définie par

$$A = \frac{k}{h^2}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \tag{12.10}$$

si $\vec{u}(t)$ est le N-vecteur de composantes $u_1(t)$, $u_2(t)$, ..., $u_N(t)$, si $\vec{f}(t)$ est le N-vecteur de composantes $f(x_1,t)$, $f(x_2,t)$, ..., $f(x_N,t)$ et si $\vec{w}$ est le N-vecteur de composantes $w(x_1)$, $w(x_2)$, ..., $w(x_N)$, alors le schéma (12.7) (12.8) (12.9) est équivalent au système différentiel :

$$\dot{\vec{u}}(t) = -A\vec{u}(t) + \vec{f}(t) \qquad \forall t > 0, \tag{12.11}$$

$$\vec{u}(0) = \vec{w}, \tag{12.12}$$

où $\dot{\vec{u}}(t)$ est la dérivée de $\vec{u}(t)$ par rapport à t au temps t, soit le N-vecteur de composantes $du_1(t)/dt$, $du_2(t)/dt$, ..., $du_N(t)/dt$.

La semi-discrétisation en espace du problème (12.4) (12.5) (12.6) conduit donc à la résolution d'un système différentiel du premier ordre avec une condition initiale. Nous pouvons donc utiliser les méthodes du chapitre 9 pour intégrer numériquement ce système différentiel. Nous allons choisir les deux schémas d'Euler progressif et rétrograde.

Schéma d'Euler progressif

Soit $\tau > 0$ un pas de temps donné, soit $t_n = n\tau$ avec $n = 0, 1, 2, \ldots$, et soit $\vec{u}^n$ une approximation de $\vec{u}(t)$ au temps $t = t_n$; nous noterons $\vec{u}^n \simeq \vec{u}(t_n)$. Considérons le schéma :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^n + \vec{f}(t_n), \qquad n = 0, 1, 2, \ldots, \tag{12.13}$$

$$\vec{u}^0 = \vec{w}. \tag{12.14}$$

Clairement, nous avons

$$\vec{u}^{n+1} = (I - \tau A)\vec{u}^n + \tau\vec{f}(t_n), \qquad n = 0, 1, 2, \ldots, \tag{12.15}$$

où I est la $N \times N$ matrice identité ; le vecteur $\vec{u}^{n+1}$ peut être calculé explicitement à partir du vecteur $\vec{u}^n$. Ainsi, à partir de $\vec{u}^0 = \vec{w}$, on peut calculer de proche en proche $\vec{u}^1$, $\vec{u}^2$, $\vec{u}^3$, ..., en utilisant (12.15) ; la j-ème composante u_j^n de $\vec{u}^n$ est une approximation de $u(x_j, t_n)$, $j = 1, \ldots, N$, $n \geq 0$. Le schéma numérique (12.13) (12.14) est ***une discrétisation*** (complète) du problème (12.4) (12.5) (12.6) par la méthode des différences finies.

Comme dans la section 9.3, une ***condition de stabilité*** limite le choix du pas temporel τ. Cette limitation est fonction du pas spatial et est exprimée par la condition :

$$\tau \leq \frac{h^2}{2k}. \tag{12.16}$$

Comme nous l'avons déjà fait dans la section 9.3, nous pourrions montrer que, si la condition (12.16) est respectée et si la fonction f est identiquement nulle, alors $\max_{1 \leq j \leq N} |u_j^n|$ décroît lorsque n croît, voir l'exercice 12.1. Cette propriété est heureuse car lorsque la source f dans (12.4) est nulle, alors nous pouvons montrer que $\lim_{t \to \infty} \max_{0 \leq x \leq 1} |u(x,t)| = 0$. Lorsque la condition (12.16) n'est pas respectée, plus n devient grand et plus les valeurs de u_i^n deviennent grandes, en changeant de signe : on dit que le schéma numérique est ***instable***, voir l'exercice 12.1.

Schéma d'Euler rétrograde

Si nous choisissons un schéma d'Euler rétrograde à la place du schéma d'Euler progressif pour discrétiser (12.11) (12.12), nous avons, à la place de (12.13) :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^{n+1} + \vec{f}(t_{n+1}), \quad n = 0, 1, 2, \ldots, \tag{12.17}$$

ou, de façon équivalente

$$(I + \tau A)\vec{u}^{n+1} = \vec{u}^n + \tau\vec{f}(t_{n+1}), \quad n = 0, 1, 2, \ldots. \tag{12.18}$$

Avec ce schéma, nous devons résoudre un système linéaire de N équations à N inconnues pour obtenir $\vec{u}^{n+1}$ à partir de $\vec{u}^n$; ce schéma est donc ***implicite***. La matrice $(I + \tau A)$ étant définie positive et tridiagonale (c'est-à-dire une matrice bande de demi-largeur de bande 2), nous pouvons résoudre ce système linéaire en

utilisant la méthode de Cholesky présentée dans la section 5.5. Contrairement au schéma explicite (12.13), ce schéma implicite est ***inconditionnellement stable*** au sens suivant.

Théorème 12.1 *Soit $\vec{u}^n$, $n = 0, 1, 2, \ldots$, la solution de (12.18) avec $f \equiv 0$. Alors, quel que soit le choix du pas de temps τ, nous avons*

$$\lim_{n \to \infty} \| \vec{u}^n \| = 0. \tag{12.19}$$

Démonstration
En posant $f \equiv 0$ dans (12.18), nous avons

$$\vec{u}^{n+1} = (I + \tau A)^{-1} \vec{u}^n. \tag{12.20}$$

Si $|||\cdot|||$ est la norme spectrale d'une matrice (sect. 4.6), alors en utilisant (12.20) nous avons :

$$\| \vec{u}^{n+1} \| \leq ||| (I + \tau A)^{-1} ||| \cdot \| \vec{u}^n \| \tag{12.21}$$

et par suite, puisque $(I + \tau A)^{-1}$ est symétrique :

$$\| \vec{u}^{n+1} \| \leq \beta \| \vec{u}^n \| \tag{12.22}$$

où β est le maximum des valeurs propres de $(I + \tau A)^{-1}$ en valeur absolue. Dans la section 4.9, nous avons vu que A est symétrique définie positive et ses valeurs propres λ_A sont donc réelles positives. Puisque les valeurs propres de $(I + \tau A)^{-1}$ sont $(1 + \tau \lambda_A)^{-1}$, on conclut facilement qu'elles sont toutes comprises entre zéro et un et donc on a $0 < \beta < 1$. De (12.22) on tire

$$\| \vec{u}^n \| \leq \beta^n \| \vec{u}^0 \| \tag{12.23}$$

et par suite la relation (12.19) est démontrée.■

Comme nous l'avons déjà dit, si f est identiquement nulle dans (12.4), alors la solution de (12.4) (12.5) (12.6) est telle que $\lim_{t \to \infty} \max_{0 \leq x \leq 1} |u(x,t)| = 0$. Le théorème 12.1 nous montre que, quel que soit le choix du pas de temps, cette propriété est maintenue sur la solution discrète $\vec{u}^n$ lorsqu'on utilise le schéma d'Euler rétrograde.

Les schémas d'Euler progressif et rétrograde sont tous les deux d'ordre 1 en temps et d'ordre 2 en espace. Donc, si nous résolvons numériquement (12.4) (12.5) (12.6) jusqu'à un temps fixé $T > 0$ en utilisant soit le schéma (12.13) sous la condition (12.16), soit le schéma (12.17), l'erreur commise est d'ordre $(\tau + h^2)$.

12.2 Equation de la chaleur 1D et éléments finis

Pour résoudre numériquement le problème (12.4) (12.5) (12.6) par la méthode des éléments finis, nous commençons par le discrétiser par rapport à la variable x de façon semblable à ce qui a été fait dans les sections 10.3 et 10.4.

Multiplions l'équation (12.4) par une fonction v, ne dépendant que de $x \in [0,1]$, supposée une fois continûment dérivable et intégrons entre $x = 0$ et $x = 1$. Nous obtenons :

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)v(x)dx - \int_0^1 k\frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = \int_0^1 f(x,t)v(x)dx. \qquad (12.24)$$

En intégrant par partie le second terme de (12.24) et en supposant $v(0) = v(1) = 0$, nous déduisons (comme nous l'avons fait pour obtenir la relation (10.9)) :

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)v(x)dx + \int_0^1 k\frac{\partial u}{\partial x}(x,t)v'(x)dx = \int_0^1 f(x,t)v(x)dx, \qquad (12.25)$$

où naturellement $v'(x)$ désigne la dérivée de v (par rapport à la variable x bien sûr !). Comme dans la section 10.3, nous introduisons l'espace V de toutes les fonctions $g : [0,1] \to \mathbb{R}$ continues, de premières dérivées g' continues par morceaux et telles que $g(0) = g(1) = 0$. Nous pouvons alors chercher, pour tout $t > 0$, une fonction $u(\cdot,t) \in V$ qui satisfait la condition initiale (12.6) ainsi que (12.25) pour toute fonction $v \in V$. Dans la suite, ce problème est appelé problème (12.25).

Le problème (12.25) est une formulation faible en espace du problème (12.4) (12.5) (12.6). Comme nous l'avons fait dans la section 10.3, nous pouvons considérer l'approximation de Galerkin suivante. Si $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont N fonctions linéairement indépendantes de V, nous construisons l'espace V_h en considérant toutes les combinaisons linéaires des fonctions φ_i. Dès lors, l'approximation de Galerkin du problème (12.25) se formule de la manière suivante : pour tout $t > 0$, trouver une fonction $u_h(\cdot,t) \in V_h$ qui satisfait

$$\int_0^1 \frac{\partial u_h}{\partial t}(x,t)v_h(x)dx + \int_0^1 k\frac{\partial u_h}{\partial x}(x,t)v_h'(x)dx = \int_0^1 f(x,t)v_h(x)dx, \qquad (12.26)$$

pour toute fonction $v_h \in V_h$. De plus nous exigeons que

$$u_h(x,0) = w_h(x) \qquad \forall x \in [0,1], \qquad (12.27)$$

où w_h est une approximation de la condition initiale w dans V_h. La détermination de w_h sera discutée ultérieurement.

Comme nous l'avons fait dans la section 10.3 pour obtenir (10.11) à partir de (10.10), nous développons $u_h(\cdot,t)$ dans la base $\varphi_1, \varphi_2, \ldots, \varphi_N$ de V_h, ce qui nous permet d'écrire :

$$u_h(\cdot,t) = \sum_{i=1}^{N} u_i(t)\varphi_i \qquad \forall t > 0.$$

Les valeurs $u_i(t)$ sont les composantes de $u_h(\cdot,t)$ dans la base des φ_i et dépendent du temps t. De façon équivalente

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \qquad \forall x \in [0,1], \quad \forall t > 0. \qquad (12.28)$$

En remplaçant (12.28) dans (12.26) et en choisissant comme fonctions test $v_h = \varphi_j$, $j = 1, 2, \ldots, N$, nous obtenons donc :

$$\sum_{i=1}^{N} \dot{u}_i(t) \int_0^1 \varphi_i(x)\varphi_j(x)dx + \sum_{i=1}^{N} u_i(t) \int_0^1 k\varphi_i'(x)\varphi_j'(x)dx$$

$$= \int_0^1 f(x,t)\varphi_j(x)dx \qquad j = 1, \ldots, N. \quad (12.29)$$

Dans (12.29), nous avons noté $\dot{u}_i(t)$ la dérivée de $u_i(t)$ par rapport à t et $\varphi_i'(x)$ la dérivée de $\varphi_i(x)$ par rapport à x.

Si A est la $N \times N$ matrice de coefficients

$$A_{ji} = \int_0^1 k\varphi_i'(x)\varphi_j'(x)dx \quad (12.30)$$

(A est appelée ***matrice de rigidité***), si M est la $N \times N$ matrice de coefficients

$$M_{ji} = \int_0^1 \varphi_i(x)\varphi_j(x)dx \quad (12.31)$$

(M est appelée ***matrice de masse***), si $\vec{u}(t)$ est le N-vecteur de composantes $u_1(t), u_2(t), \ldots, u_N(t)$ et si $\vec{f}(t)$ est le N-vecteur dont la j-ème composante est

$$f_j(t) = \int_0^1 f(x,t)\varphi_j(x)dx, \quad (12.32)$$

alors les relations (12.29) sont équivalentes à chercher $\vec{u}(t)$ tel que

$$M\dot{\vec{u}}(t) + A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0. \quad (12.33)$$

Comme pour la méthode des différences finies (voir (12.11)), l'approximation de Galerkin (12.33) conduit à un système différentiel du premier ordre. Les inconnues de ce système sont les composantes $u_j(t)$ de la solution u_h dans la base des φ_i. Pour établir la condition initiale du système, nous écrivons w_h dans la base des φ_i, c'est-à-dire

$$w_h(x) = \sum_{i=1}^{N} w_i\varphi_i(x).$$

Si $\vec{w}$ est le N-vecteur de composantes $w_1, \ldots, w_N$, alors la condition initiale du système différentiel (12.33) est définie par :

$$\vec{u}(0) = \vec{w}. \quad (12.34)$$

Nous avons donc obtenu une ***semi-discrétisation*** spatiale du problème (12.4) (12.5) (12.6). Il est facile de vérifier que les matrices M et A sont des $N \times N$ matrices symétriques définies positives. Le système différentiel (12.33) est équivalent à

$$\dot{\vec{u}}(t) = -M^{-1}A\vec{u}(t) + M^{-1}\vec{f}(t) \qquad \forall t > 0, \tag{12.35}$$

et dès lors nous pouvons calculer une approximation de la solution $u(x,t)$ du problème (12.4) (12.5) (12.6) en procédant de la façon suivante :

- on définit concrètement une base $\varphi_1, \varphi_2, \ldots, \varphi_N$ de type éléments finis comme nous l'avons fait dans la section 10.4 et on construit les matrices M, A et le vecteur $\vec{f}$;
- en supposant la condition initiale w continue sur $[0,1]$, on construit w_h en interpolant w par des polynômes de degré 1 sur chaque élément géométrique (sect. 1.6) ;
- on détermine une approximation $\vec{u}^n$ de $\vec{u}(t_n)$ en utilisant un schéma d'Euler progressif ou rétrograde comme dans (12.13) ou (12.17).

Nous terminons cette section par deux remarques.

Remarque 12.1 Le schéma d'Euler progressif s'écrit dans le cas présent :

$$M\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^n + \vec{f}(t_n)$$

ou, de façon équivalente

$$M\vec{u}^{n+1} = (M - \tau A)\vec{u}^n + \tau\vec{f}(t_n), \tag{12.36}$$

où τ est le pas de temps et $t_n = n\tau$.

Clairement, si $\vec{u}^n$ est connu dans (12.36), nous devons encore résoudre un système pour obtenir $\vec{u}^{n+1}$ car la matrice de masse M calculée par éléments finis n'est pas diagonale. Ici donc le schéma d'Euler progressif n'est pas explicite ! Pour le rendre explicite, il faut calculer concrètement la matrice de masse M en utilisant la formule de quadrature des trapèzes. Ainsi, nous obtenons en utilisant la formule (10.24) avec $c = 1$:

$$M_{ji} = \int_0^1 \varphi_i(x)\varphi_j(x)dx \simeq L_h(\varphi_i\varphi_j) = \begin{cases} h & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Ce procédé consiste à approcher la matrice de masse M par une matrice diagonale (on parle ici de ***mass lumping***) et donc à rendre explicite le schéma d'Euler progressif.

Remarque 12.2 Comme nous l'avons mentionné dans la section 9.3, les schémas d'Euler sont d'ordre 1 en τ (voir inégalité (9.22)). Pour avoir un schéma d'ordre 2, nous pouvons utiliser une moyenne des schémas d'Euler progressif et rétrograde pour obtenir :

$$M\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} + A\frac{\vec{u}^{n+1} + \vec{u}^n}{2} = \frac{\vec{f}(t_{n+1}) + \vec{f}(t_n)}{2},$$

ou, de façon équivalente

$$(M + \frac{\tau}{2}A)\vec{u}^{n+1} = (M - \frac{\tau}{2}A)\vec{u}^{n} + \frac{\tau}{2}(\vec{f}(t_{n+1}) + \vec{f}(t_n)). \tag{12.37}$$

Le schéma (12.37) est appelé ***schéma de Crank-Nicholson*** ; c'est un schéma numérique d'ordre 2, implicite, inconditionnellement stable (en norme quadratique !).

12.3 Problèmes paraboliques 2D et leurs approximations

Soit Ω un domaine polygonal dans le plan Ox_1x_2, de frontière $\partial\Omega$ et soit $\overline{\Omega} = \Omega \cup \partial\Omega$. Si $a_{11}, a_{12}, a_{21}, a_{22}$ sont quatre fonctions de $(x,t) \in \overline{\Omega} \times \mathbb{R}^+$ données, si $f : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to f(x,t) \in \mathbb{R}$ et $w : x \in \overline{\Omega} \to w(x) \in \mathbb{R}$ sont deux autres fonctions données, nous posons le problème de trouver une fonction $u : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ satisfaisant les relations suivantes :

$$\frac{\partial u}{\partial t}(x,t) - \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial}{\partial x_j}u(x,t)\right) = f(x,t)$$

$$\forall x \in \Omega, \quad \forall t > 0, \tag{12.38}$$

$$u(x,t) = 0, \qquad \forall x \in \partial\Omega, \quad \forall t > 0, \tag{12.39}$$

$$u(x,0) = w(x), \qquad \forall x \in \Omega. \tag{12.40}$$

La relation (12.39) est appelée ***condition aux limites*** alors que la relation (12.40) est appelée ***condition initiale***.

Définition 12.1 *Nous dirons que le problème (12.38) (12.39) (12.40) est parabolique si les fonctions a_{ij}, $1 \leq i,j \leq 4$, sont telles qu'il existe un nombre réel positif α qui satisfasse, pour tout $x \in \Omega$, pour tout $t > 0$ et pour tout couple de nombres réels (ξ_1, ξ_2), la relation*

$$\sum_{i,j=1}^{2} a_{ij}(x,t)\xi_i\xi_j \geq \alpha(\xi_1^2 + \xi_2^2).$$

Remarquons que si le problème (12.38) (12.39) (12.40) est parabolique au sens de la définition ci-dessus, alors pour tout $x \in \Omega$ et pour tout $t > 0$ l'équation en ξ_1, ξ_2, ξ_3 donnée par

$$\xi_3 = a_{11}(x,t)\xi_1^2 + \Big(a_{12}(x,t) + a_{21}(x,t)\Big)\xi_1\xi_2 + a_{22}(x,t)\xi_2^2$$

est l'équation d'un paraboloïde d'axe $O\xi_3$, d'où la terminologie ***problème parabolique*** (on a remplacé $\partial u/\partial t$ par ξ_3 et $\partial u/\partial x_i$ par ξ_i, $i = 1,2$).

Les problèmes paraboliques interviennent dans de nombreuses modélisations physiques telles que les phénomènes de diffusion de la chaleur, d'écoulement de

fluides, ... Dans le cas où $a_{11} = a_{22} = 1$ et $a_{12} = a_{21} = 0$ indépendamment de x et t, nous obtenons

$$\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = f(x,t) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{12.41}$$

$$u(x,t) = 0 \qquad \forall x \in \partial\Omega, \quad \forall t > 0, \tag{12.42}$$

$$u(x,0) = w(x) \qquad \forall x \in \Omega, \tag{12.43}$$

où ici $\Delta u(x,t)$ est le laplacien de u dans les variables spatiales. Le problème (12.41) (12.42) (12.43) modélise un problème de diffusion de chaleur dans une plaque Ω ; $u(x,t)$ est alors la température au point $x \in \Omega$ et à l'instant $t > 0$; $f(x,t)$ est la puissance par unité de surface introduite au point x et à l'instant $t > 0$.

Si nous superposons les résultats de la section 11.2 avec ceux de la section précédente, nous pouvons sans difficulté construire une semi-discrétisation spatiale du problème (12.41) (12.42) (12.43) par la méthode des éléments finis qui donnera lieu à un système différentiel en temps. Il suffira ensuite d'intégrer numériquement ce dernier par des méthodes que nous connaissons déjà. Dans la suite de cette section, nous exposons très brièvement cette construction.

En suivant ce qui a été fait dans la section 11.2, nous écrivons une relation semblable à (12.28) :

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{12.44}$$

où les fonctions de bases φ_j, $j = 1, \ldots, N$, sont données par (11.15) et leur graphe est illustré dans la figure 11.3. Il suffit ensuite de remplacer dans (12.29) les intégrales

$$\int_0^1 dx \qquad \text{par} \qquad \iint_\Omega dx$$

et les grandeurs φ_i' par des grandeurs $\overrightarrow{\text{grad}}\varphi_i$ pour obtenir :

$$\sum_{i=1}^{N} \dot{u}_i(t) \iint_\Omega \varphi_i(x)\varphi_j(x)dx + \sum_{i=1}^{N} u_i(t) \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx$$

$$= \iint_\Omega f(x,t)\varphi_j(x)dx, \qquad j = 1, 2, \ldots, N. \tag{12.45}$$

Pour justifier (12.45), il suffit d'opérer exactement de la même manière que pour passer de (11.5) à (11.9) et de (12.24) à (12.25).

Ici encore nous pouvons construire les matrices de masse M et de rigidité A dont les éléments sont donnés par :

$$M_{ji} = \iint_\Omega \varphi_i(x)\varphi_j(x)dx$$

et

$$A_{ji} = \iint_{\Omega} \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx, \qquad i,j = 1,\ldots,N.$$

Le système différentiel résultant de (12.45) sera alors

$$M\dot{\vec{u}}(t) + A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0, \tag{12.46}$$

qui pourra être résolu numériquement par un schéma d'Euler, ou de Crank-Nicholson (ou un autre) après avoir naturellement posé

$$f_j(t) = \iint_{\Omega} f(x,t)\varphi_j(x)dx.$$

La condition initiale sera donnée par exemple par $u_i(0) = w(P_i)$, $1 \leq i \leq N$, puisque l'interpolant de w est $\sum_{j=1}^{N} w(P_j)\varphi_j$.

12.4 Un exemple particulier

Comme dans la section 11.3, nous nous plaçons dans la situation où Ω est le carré unité de frontière $\partial\Omega$ et nous considérons le problème (12.41) (12.42) (12.43) avec ce domaine Ω particulier. Nous choisissons la même triangulation $\mathcal{T}_h$ que celle décrite dans la figure 11.4 (N nœuds intérieurs, $2(L+1)^2$ triangles avec $N = L^2$ et $\tilde{h} = 1/(L+1)$), et les mêmes fonctions de base $\varphi_1, \varphi_2, \ldots, \varphi_N$ que celles données dans la section 11.3. Nous obtenons alors la matrice de rigidité A décrite dans (11.16). Il ne reste plus qu'à calculer la matrice de masse M et le second membre $\vec{f}(t)$ pour obtenir concrètement le système (12.46). Le calcul de $\vec{f}(t)$ peut se faire par le biais d'une formule de quadrature, comme nous l'avons fait pour obtenir (11.18). Nous aurons $f_j(t) \simeq f(P_j,t)\tilde{h}^2$ où P_j désigne aussi bien le nœud P_j que ses deux coordonnées dans le repère Ox_1,x_2.

Dans le cas où $L = 4$, le lecteur peut se convaincre que la matrice de masse M a l'allure suivante :

$$M = \frac{\tilde{h}^2}{12}\begin{bmatrix} \tilde{B} & \tilde{C} & & \\ \tilde{C}^T & \tilde{B} & \tilde{C} & \\ & \tilde{C}^T & \tilde{B} & \tilde{C} \\ & & \tilde{C}^T & \tilde{B} \end{bmatrix},$$

où nous avons noté

$$\tilde{B} = \begin{bmatrix} 6 & 1 & & \\ 1 & 6 & 1 & \\ & 1 & 6 & 1 \\ & & 1 & 6 \end{bmatrix} \text{ et } \tilde{C} = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \end{bmatrix}.$$

Dans cet exemple concret, nous pouvons choisir $u_i(0) = w(P_i)$, $i = 1,\ldots,N$ pour condition initiale.

12.5 Exercices

Exercice 12.1 Soit u la solution du problème (12.4) (12.5) (12.6) avec $f(x,t) = 0$, $\forall x \in]0,1[$, $\forall t > 0$. Soit N un entier positif, $h = 1/(N+1)$, $x_i = ih$, $i = 0, 1, 2, \ldots, N+1$. Soit $\tau > 0$ donné, $t_n = n\tau$, $n = 0, 1, 2, \ldots$ Soit u_i^n une approximation de $u(x_i, t_n)$, $i = 1, \ldots, N$, $n = 0, 1, 2, \ldots$, calculée en utilisant le schéma d'Euler progressif (12.13) (12.14).

1. On fixe l'entier n et on suppose que $2k\tau \leq h^2$. Montrer que si tous les u_i^n, $i = 1, 2, \ldots, N$ sont positifs, alors tous les u_i^{n+1}, $i = 1, 2, \ldots, N$ sont aussi positifs. Montrer que

$$\max_{1 \leq i \leq N} |u_i^{n+1}| \leq \max_{1 \leq i \leq N} |u_i^n|. \tag{12.47}$$

   La relation (12.47) garantit que le schéma d'Euler progressif (12.13) (12.14) est stable dès que $2k\tau \leq h^2$ (il s'agit de la condition(12.16)).

2. On considère le cas où la condition initiale est définie par $w(x) = 1$ si $0 < x < 1$. On choisit $k = 1$, $N = 4$ et $\tau = 1/25$ de sorte que $2k\tau > h^2$. Construire et représenter la solution numérique pour $n = 1, 2, \ldots, 9$. En déduire que le schéma est numériquement instable.

Solution

1. Le schéma numérique (12.13) s'écrit

$$u_i^{n+1} = \frac{k\tau}{h^2} u_{i-1}^n + \left(1 - 2\frac{k\tau}{h^2}\right) u_i^n + \frac{k\tau}{h^2} u_{i+1}^n. \tag{12.48}$$

Notons que si $2k\tau \leq h^2$ alors $1 - 2k\tau/h^2 \geq 0$. En vertu de la relation (12.48), il est clair que si u_{i-1}^n, u_i^n et u_{i+1}^n sont positifs, alors on obtient u_i^{n+1} aussi positif. Nous avons bien montré que si tous les u_i^n sont positifs, alors tous les u_i^{n+1} sont également positifs.

Montrons maintenant (12.47). Soit $1 \leq i \leq N$ fixé. L'égalité (12.48) implique

$$|u_i^{n+1}| \leq \frac{k\tau}{h^2} |u_{i-1}^n| + |1 - 2\frac{k\tau}{h^2}| \, |u_i^n| + \frac{k\tau}{h^2} |u_{i+1}^n|.$$

Puisque $1 - 2k\tau/h^2 \geq 0$, nous avons donc

$$\begin{aligned} |u_i^{n+1}| &\leq \frac{k\tau}{h^2} |u_{i-1}^n| + \left(1 - 2\frac{k\tau}{h^2}\right) |u_i^n| + \frac{k\tau}{h^2} |u_{i+1}^n| \\ &\leq \frac{k\tau}{h^2} \max_{1 \leq j \leq N} |u_j^n| + \left(1 - 2\frac{k\tau}{h^2}\right) \max_{1 \leq j \leq N} |u_j^n| + \frac{k\tau}{h^2} \max_{1 \leq j \leq N} |u_j^n| \\ &= \max_{1 \leq j \leq N} |u_j^n|. \end{aligned}$$

Il suffit ensuite de prendre le maximum sur tous les indices i pour obtenir le résultat.

2. Lorsque $k = 1$, $h = 1/5$, $\tau = 1/25$, le schéma numérique (12.48) s'écrit

$$u_i^{n+1} = u_{i-1}^n - u_i^n + u_{i+1}^n.$$

Les résultats numériques pour $n = 1, \ldots, 9$, sont présentés dans la figure 12.1. La solution numérique présente des oscillations qui ne cessent d'augmenter lorsque x est fixé et le temps t croît. Nous concluons que le schéma est dans ce cas numériquement instable.

$t_9 = 9/25$	0	−21	13	13	−21	0
t_8	0	13	−8	−8	13	0
t_7	0	−8	5	5	−8	0
t_6	0	5	−3	−3	5	0
t_5	0	−3	2	2	−3	0
t_4	0	2	−1	−1	2	0
t_3	0	−1	1	1	−1	0
$t_2 = 2/25$	0	1	0	0	1	0
$t_1 = 1/25$	0	0	1	1	0	0
$t_0 = 0$	0	1	1	1	1	0
	$x_0 = 0$	x_1	x_2	x_3	x_4	$x_5 = 1$

Fig. 12.1 Solution numérique du schéma (12.48) lorsque $k = 1$, $h = 1/5$ et $\tau = 1/25$.

Exercice 12.2 On considère un problème de la chaleur non linéaire, à savoir

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = -\Big(u(x,t)\Big)^3 \qquad \forall x \in]0,1[, \quad \forall t > 0, \tag{12.49}$$

$$u(0,t) = u(1,t) = 0 \qquad \forall t > 0, \tag{12.50}$$

$$u(x,0) = w(x) \qquad \forall x \in]0,1[, \tag{12.51}$$

où w est une fonction donnée. Pour résoudre numériquement ce problème on utilise une méthode de différence finies en espace et en temps. Soit $\tau > 0$ un pas de temps donné et soit $t_n = n\tau$, $n = 0, 1, 2, \ldots$ Soit N un entier positif, on note $h = 1/(N+1)$ le pas d'espace, $x_i = i\,h$, $i = 0, 1, 2, \ldots, N+1$. Finalement on note u_i^n une approximation de $u(x_i, t_n)$.

1. Proposer un schéma pour résoudre numériquement le problème (12.49) (12.50) (12.51) en utilisant un schéma d'Euler progressif en temps. Remarquer le caractère explicite du schéma obtenu.

2. Proposer un schéma pour résoudre numériquement le problème (12.49) (12.50) (12.51) en utilisant un schéma d'Euler rétrograde en temps. Remarquer le caractère implicite et non linéaire du schéma obtenu.
3. A partir du schéma obtenu au point précédent, exécuter à chaque pas de temps un pas de la méthode de Newton. Remarquer le caractère implicite, mais linéaire de ce nouveau schéma. Comment calcule-t-on $\vec{u}^{n+1}$ à partir de $\vec{u}^n$?

Solution

1. Etant donné la condition initiale (12.51), nous posons $u_i^0 = w(x_i)$, $i = 1, 2, \ldots, N$. Le schéma d'Euler progressif consiste, à partir de u_i^n, $i = 1, 2, \ldots, N$, à calculer u_i^{n+1}, $i = 1, 2, \ldots, N$, à l'aide du schéma suivant

$$\frac{u_i^{n+1} - u_i^n}{\tau} + \frac{-u_{i-1}^n + 2u_i^n - u_{i+1}^n}{h^2} = -(u_i^n)^3,$$

où, en vertu des conditions aux limites (12.50), nous avons posé $u_0^n = u_{N+1}^n = 0$. Le schéma ci-dessus peut s'écrire sous forme matricielle

$$\begin{aligned} &\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^n - (\vec{u}^n)^3, \qquad n = 0, 1, 2, \ldots, \\ &\vec{u}^0 = \vec{w}, \end{aligned}$$

où $\vec{u}^n$ est le N-vecteur de composantes u_i^n, $(\vec{u}^n)^3$ est le N-vecteur de composantes $(u_i^n)^3$, où A est la $N \times N$ matrice tridiagonale définie par (12.10) avec $k = 1$ et où $\vec{w}$ est le N-vecteur de composantes $w(x_i)$. Le schéma ci-dessus est explicite. Il permet de calculer explicitement $\vec{u}^{n+1}$ à partir de $\vec{u}^n$ de la façon suivante

$$\vec{u}^{n+1} = (I - \tau A)\vec{u}^n - \tau(\vec{u}^n)^3, \qquad n = 0, 1, 2, \ldots.$$

2. Le schéma d'Euler rétrograde conduit au schéma

$$\begin{aligned} &\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} + A\vec{u}^{n+1} = -(\vec{u}^{n+1})^3, \qquad n = 0, 1, 2, \ldots, \\ &\vec{u}^0 = \vec{w}. \end{aligned}$$

Ce schéma permet de calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$ de la façon suivante

$$(I + \tau A)\vec{u}^{n+1} + \tau(\vec{u}^{n+1})^3 = \vec{u}^n, \qquad n = 0, 1, 2, \ldots. \tag{12.52}$$

Ce schéma est implicite et non linéaire.

3. Nous procédons comme dans la section 10.6. Soit $\vec{F} : \mathbb{R}^N \to \mathbb{R}^N$ la fonction définie par

$$\vec{F}(\vec{y}) = (I + \tau A)\vec{y} + \tau(\vec{y})^3 - \vec{u}^n.$$

Le schéma (12.52) consiste donc à trouver $\vec{u}^{n+1}$ tel que $\vec{F}(\vec{u}^{n+1}) = \vec{0}$. La méthode de Newton s'écrit

$$\begin{aligned} &DF(\vec{y}^k)(\vec{y}^{k+1} - \vec{y}^k) = -\vec{F}(\vec{y}^k) \qquad k = 0, 1, 2, \ldots, \\ &\text{avec } \vec{y}^0 \text{ donné,} \end{aligned}$$

où DF est la matrice jacobienne de $\vec{F}$. Pour calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$ nous effectuons un seul pas de la méthode de Newton et nous choisissons $\vec{y}^0 = \vec{u}^n$. Le schéma ainsi obtenu s'écrit

$$DF(\vec{u}^n)(\vec{u}^{n+1} - \vec{u}^n) = -\vec{F}(\vec{u}^n) \qquad n = 0, 1, 2, \ldots .$$

Ce schéma est linéaire mais toujours implicite, car il nécessite la résolution d'un système linéaire avec $DF(\vec{u}^n)$ comme matrice. Calculons la matrice jacobienne $DF(\vec{u}^n)$. Nous obtenons $DF(\vec{u}^n) = I + \tau A + 3\tau B_n$, où nous avons noté B_n la $N \times N$ matrice diagonale de valeurs diagonales $(u_1^n)^2$, $(u_2^n)^2$, ..., $(u_N^n)^2$. Finalement, compte tenu de la définition de $\vec{F}$, le schéma ci-dessus devient

$$(I + \tau A + 3\tau B_n)(\vec{u}^{n+1} - \vec{u}^n) = -\tau A\vec{u}^n - \tau(\vec{u}^n)^3.$$

Ainsi, pour calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$, nous effectuons les étapes suivantes.

- On construit la matrice C définie par $C = I + \tau A + 3\tau B_n$ ainsi que le vecteur $\vec{b} = -\tau A\vec{u}^n - \tau(\vec{u}^n)^3$.
- On résout le système linéaire $C\vec{y} = \vec{b}$ par une des méthodes décrites dans les chapitres 4, 5 ou 6.
- On pose $\vec{u}^{n+1} = \vec{u}^n + \vec{y}$.

12.6 Notes bibliographiques et remarques

Dans ce chapitre nous avons résolu les sytèmes différentiels (12.11) et (12.33) en utilisant des schémas d'Euler et Crank-Nicholson. Il est bien évident que nous pouvons utiliser des méthodes d'ordre plus élevé en temps, par exemple la méthode Runge-Kutta classique (sect. 9.5).

Dans les sections 12.2 et 12.3 nous avons discrétisé un problème parabolique par des méthodes d'éléments finis en espace seulement. Il existe des méthodes d'éléments finis en espace-temps pour discrétiser les problèmes paraboliques. Il s'agit des méthodes de ***Galerkin discontinues***, voir par exemple [8, 24].

Il existe des méthodes auto-adaptives pour la résolution numérique des problèmes paraboliques, voir par exemple [8]. Ces méthodes ont pour but de calculer, de façon automatique, le pas d'espace et le pas de temps de sorte que la solution numérique soit aussi précise que possible. Ces méthodes sont aujourd'hui encore un sujet de recherche, mais il semble certain qu'à terme la plupart des codes industriels d'éléments finis en soient pourvus.

Chapitre 13

Approximation de problèmes hyperboliques. Equation de transport et équation des ondes

13.1 Equation de transport 1D et différences finies

Soit deux fonctions $c : (x,t) \in \mathbb{R}\times\mathbb{R}^+ \to c(x,t) \in \mathbb{R}$ et $f : (x,t) \in \mathbb{R}\times\mathbb{R}^+ \to f(x,t) \in \mathbb{R}$ données, continues et soit $w : x \in \mathbb{R} \to w(x) \in \mathbb{R}$ une autre fonction donnée. Nous posons le problème de trouver une fonction $u : (x,t) \in \mathbb{R}\times\mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ satisfaisant l'équation :

$$\frac{\partial u}{\partial t}(x,t) + c(x,t)\frac{\partial u}{\partial x}(x,t) = f(x,t) \quad \forall x \in \mathbb{R}, \quad \forall t > 0, \tag{13.1}$$

avec la condition initiale

$$u(x,0) = w(x) \qquad \forall x \in \mathbb{R}. \tag{13.2}$$

Les équations (13.1) (13.2) modélisent, par exemple, le transport (en fonction du temps t) d'un gaz polluant dans une colonne parallèle à l'axe Ox, remplie de charbon actif et d'air. La grandeur inconnue $u(x,t)$ représente alors la concentration par unité de volume du gaz dans l'air au point x et à l'instant t, c représente la vitesse du gaz le long de la colonne et f la quantité (par unité de temps) de gaz retenu par le charbon actif. Naturellement w est la concentration du gaz au temps initial, que nous supposons connue.

Remarquons que le problème (13.1) (13.2) est facile à résoudre lorsque $c = c_0 =$ constante et $f = 0$. En effet, la fonction u définie par

$$u(x,t) = w(x - c_0 t) \tag{13.3}$$

satisfait

$$\frac{\partial u}{\partial t}(x,t) + c_0 \frac{\partial u}{\partial x}(x,t) = 0 \qquad \forall x \in \mathbb{R}, \quad \forall t > 0, \tag{13.4}$$

et

$$u(x,0) = w(x) \qquad \forall x \in \mathbb{R}. \tag{13.5}$$

Nous avons représenté dans la figure 13.1 le graphe de la fonction u dans le cas où la condition initiale w est définie par

$$w(x) = \begin{cases} 1-x & \text{si } x \in [0,1], \\ 1+x & \text{si } x \in [-1,0], \\ 0 & \text{si } x \notin [-1,+1]. \end{cases}$$

Nous observons que la condition initiale w est transportée le long de l'axe Ox, à la vitesse c_0.

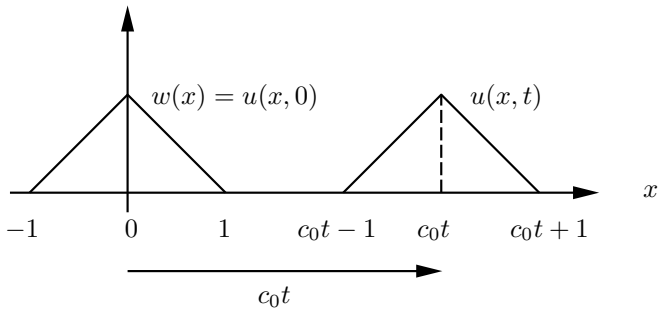

Fig. 13.1 Transport de $w(x)$ au temps $t > 0$, avec $c_0 > 0$.

Revenons au problème (13.1) (13.2) et proposons une approximation de la fonction inconnue u par une méthode de différences finies. Pour ce faire, nous introduisons un pas spatial $h > 0$, un pas temporel τ et nous posons $x_j = jh$, $j = 0, \pm 1, \pm 2, \ldots$, ainsi que $t_n = n\tau$, $n = 0, 1, 2, \ldots$, La figure 13.2 représente la grille ainsi obtenue dans l'espace temps Oxt. Si $u_j^n \simeq u(x_j, t_n)$ est une approximation de la solution u de (13.1) (13.2) au point x_j et au temps t_n, il semble naturel de choisir le schéma numérique d'ordre deux en espace ($O(h^2)$) et un en temps ($O(\tau)$) suivant :

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c(x_j, t_n) \frac{u_{j+1}^n - u_{j-1}^n}{2h} = f(x_j, t_n), \tag{13.6}$$

pour $j = 0, \pm 1, \pm 2, \ldots$, et $n = 0, 1, 2, \ldots$ L'approximation initiale est définie par

$$u_j^0 = w(x_j) \qquad \text{pour } j = 0, \pm 1, \pm 2, \ldots \tag{13.7}$$

Le schéma (13.6) (13.7) est appelé ***schéma explicite centré*** ; il permet de calculer explicitement u_j^{n+1}, $j = 0, \pm 1, \pm 2, \ldots$, à partir des valeurs de u_j^n, $j = 0, \pm 1, \pm 2, \ldots$ En effet, nous pouvons réécrire (13.6) sous la forme suivante :

$$u_j^{n+1} = u_j^n + \tau \left(f(x_j, t_n) - c(x_j, t_n) \frac{u_{j+1}^n - u_{j-1}^n}{2h} \right). \tag{13.8}$$

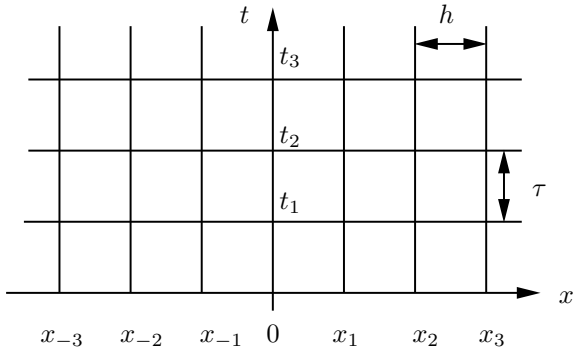

Fig. 13.2 Grille de différences finies.

Revenons au cas où $c = c_0 =$ constante et $f = 0$. Nous aurons alors

$$u_j^{n+1} = u_j^n - \frac{c_0\tau}{2h}(u_{j+1}^n - u_{j-1}^n). \tag{13.9}$$

Choisissons maintenant une condition initiale w assez régulière et 2π-périodique de telle sorte à ce que nous puissions l'identifier à sa série de Fourier complexe

$$w(x) = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imx}, \tag{13.10}$$

la grandeur i désignant naturellement l'unité imaginaire. Les coefficients de Fourier α_m sont les nombres complexes définis par

$$\alpha_m = \frac{1}{2\pi}\int_0^{2\pi} w(x)e^{-imx}dx.$$

Puisque $x_j = jh$, nous aurons donc

$$u_j^0 = w(jh) = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh}, \quad j = 0, \pm 1, \pm 2, \ldots \tag{13.11}$$

En utilisant (13.9) et (13.11), nous obtenons :

$$\begin{aligned} u_j^1 &= u_j^0 - \frac{c_0\tau}{2h}(u_{j+1}^0 - u_{j-1}^0) \\ &= \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh}\left(1 - \frac{c_0\tau}{2h}(e^{imh} - e^{-imh})\right) \\ &= \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh}\left(1 - \frac{c_0\tau}{h} i \sin mh\right). \end{aligned} \tag{13.12}$$

Nous vérifions facilement, en itérant n fois le passage de (13.11) à (13.12), qu'à l'étape n, $n = 0, 1, 2, \ldots$, nous obtenons :

$$u_j^n = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh}\left(1 - \frac{c_0\tau}{h} i \sin mh\right)^n, \tag{13.13}$$

pour $j = 0, \pm 1, \pm 2, \ldots$ Le coefficient $1 - c_0 \tau h^{-1} i \sin mh$ est un nombre complexe appelé ***coefficient d'amplification de la m-ième harmonique***. Son module

$$\sqrt{1 + \Big(\frac{c_0 \tau}{h} \sin mh\Big)^2},$$

est strictement plus grand que 1 si $c_0 \neq 0$ et si $m \neq k\pi/h$, $k = 0, \pm 1, \pm 2, \ldots$ Ainsi, les valeurs $|u_j^n|$ deviennent de plus en plus grandes lorsque n tend vers l'infini (du moins pour certains j). La solution numérique peut donc exploser alors que la solution exacte du problème (13.4) (13.5) est donné par (13.3) et satisfait donc

$$|u(x,t)| = |w(x - c_0 t)| \leq \max_{s \in [0, 2\pi]} |w(s)| \qquad \forall x \in \mathbb{R}, \quad \forall t > 0.$$

Le calcul que nous venons de faire montre que le ***schéma explicite centré est toujours instable*** ; c'est un mauvais schéma numérique qu'il ne faut surtout pas utiliser ! Comment alors établir un bon schéma numérique ?

Considérons à nouveau la fonction u définie par (13.3) et solution du problème (13.4) (13.5). Nous constatons que la condition initiale $w(x)$ est transportée à la vitesse c_0 dans le sens des x positifs lorsque c_0 est positif et dans le sens des x négatifs lorsque c_0 est négatif. Il semble dès lors naturel que, si c_0 est positif, il faille tenir compte de u_{j-1}^n et u_j^n (au lieu de u_{j+1}^n) pour calculer u_j^{n+1} et que, si c_0 est négatif, il faille tenir compte de u_j^n (au lieu de u_{j-1}^n) et u_{j+1}^n pour calculer u_j^{n+1}. Nous proposons donc le ***schéma décentré*** suivant :

$$\frac{u_j^{n+1} - u_j^n}{\tau} + (c_j^n)^+ \, \frac{u_j^n - u_{j-1}^n}{h} + (c_j^n)^- \, \frac{u_{j+1}^n - u_j^n}{h} = f(x_j, t_n), \qquad (13.14)$$

pour $j = 0, \pm 1, \pm 2, \ldots$, $n = 0, 1, 2, \ldots$; les coefficients $(c_j^n)^+$ et $(c_j^n)^-$ étant définis par

$$(c_j^n)^+ = \begin{cases} c(x_j, t_n) & \text{si } c(x_j, t_n) > 0, \\ 0 & \text{si } c(x_j, t_n) \leq 0, \end{cases}$$

et

$$(c_j^n)^- = \begin{cases} c(x_j, t_n) & \text{si } c(x_j, t_n) < 0, \\ 0 & \text{si } c(x_j, t_n) \geq 0. \end{cases}$$

La relation (13.14) peut encore s'écrire

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c(x_j, t_n) \frac{u_j^n - u_{j-1}^n}{h} = f(x_j, t_n) \qquad \text{si } c(x_j, t_n) > 0$$

(on dit que le schéma est décentré en arrière) et

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c(x_j, t_n) \frac{u_{j+1}^n - u_j^n}{h} = f(x_j, t_n) \qquad \text{si } c(x_j, t_n) < 0$$

(on dit que le schéma est décentré en avant). Remarquons encore que

$$(c_j^n)^+ = \frac{1}{2}\Big(c(x_j,t_n) + |c(x_j,t_n)|\Big)$$
$$(c_j^n)^- = \frac{1}{2}\Big(c(x_j,t_n) - |c(x_j,t_n)|\Big).$$

Le schéma décentré (13.14) est explicite ; il permet de calculer explicitement les valeurs u_j^{n+1} à partir des valeurs u_j^n. Dans le cas où $c = c_0$ et $f = 0$, nous pouvons faire une analyse de stabilité similaire à celle que nous avons déjà faite pour le schéma centré. Nous obtenons alors que le coefficient d'amplification de la m-ième harmonique est égal à $1 - c_0\tau h^{-1}(1 - e^{-imh})$ si $c_0 > 0$ et $1 - c_0\tau h^{-1}(e^{imh} - 1)$ si $c_0 < 0$. Ainsi son module est plus petit ou égal à 1 (indépendamment de m) lorsque la condition

$$\frac{\tau}{h} \leq \frac{1}{|c_0|}$$

est satisfaite. Cette condition est appelée ***condition de stabilité***. Dans le cas où c n'est pas constant et si nous utilisons le schéma explicite décentré (13.14), la condition de stabilité devient

$$\frac{\tau}{h} \leq \frac{1}{\sup\limits_{x\in\mathbb{R},t>0} |c(x,t)|}. \tag{13.15}$$

En pratique, le pas spatial h et le pas temporel τ devront être choisis de sorte à ce que la condition (13.15) soit satisfaite. Nous dirons que *le schéma explicite décentré est conditionnellement stable*. La condition de stabilité (13.15) est appelée ***condition de Courant-Friedrichs-Lewy*** ou plus simplement ***condition CFL***. Ainsi, si nous fixons le pas spatial $h > 0$, nous devons choisir un pas temporel τ plus petit que $h/\sup|c(x,t)|$. Dans le cas contraire, le schéma (13.14) produit à un moment ou à un autre des valeurs $|u_j^n|$ qui augmentent indéfiniment lorsque n augmente !

Il existe bien d'autres schémas de différences finies pour résoudre numériquement le problème de transport (13.1) (13.2) (Lax, Lax-Wendroff, saute-mouton, ...) que nous n'aborderons pas ici.

13.2 Equation des ondes 1D et différences finies

Soit $f : (x,t) \in [0,1] \times \mathbb{R}^+ \to f(x,t) \in \mathbb{R}$ une fonction continue donnée, et soit $w : x \in [0,1] \to w(x) \in \mathbb{R}$ et $v : x \in [0,1] \to v(x) \in \mathbb{R}$ deux autres fonctions. Etant donné un nombre positif c, nous posons le problème de trouver une fonction $u : (x,t) \in [0,1] \times \mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ telle que

$$\frac{\partial^2 u}{\partial t^2}(x,t) - c^2\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) \qquad \forall x \in]0,1[, \quad \forall t > 0, \tag{13.16}$$
$$u(0,t) = u(1,t) = 0 \qquad \forall t > 0, \tag{13.17}$$
$$u(x,0) = w(x) \text{ et } \frac{\partial u}{\partial t}(x,0) = v(x) \qquad \forall x \in]0,1[. \tag{13.18}$$

Le problème (13.16) (13.17) (13.18) est appelé ***problème hyperbolique*** d'ordre deux ; l'équation (13.16) est une équation aux dérivées partielles d'ordre deux en temps et deux en espace. A cette équation nous ajoutons les deux conditions aux limites (13.17) ainsi que les deux conditions initiales (13.18). Remarquons que si nous remplaçons symboliquement $\partial^2 u/\partial t^2$ par t^2, $\partial^2 u/\partial x^2$ par x^2 et f par 1 alors l'équation (13.16) se réduit à $t^2 - c^2 x^2 = 1$ qui est l'équation d'une hyperbole dans le plan Oxt, d'où le nom de ***problème hyperbolique***.

Le problème de la ***corde vibrante*** est l'exemple d'une situation physique régie par les équations (13.16) (13.17) (13.18). Considérons une corde élastique, tendue entre les points $x = 0$ et $x = 1$ et soumise à une densité de force verticale f (c'est-à-dire $f(x,t)$ est la force par unité de longueur exercée sur la corde au point x et à l'instant t). Alors $u(x,t)$ représente la déformation verticale de la corde au point x et à l'instant t et satisfait l'équation (13.16). Le nombre c dépend de la masse spécifique de la corde et de sa tension. Les conditions aux limites (13.17) traduisent le fait que la corde est tendue entre les points $x = 0$ et $x = 1$. La déformation initiale w et la vitesse de déformation initiale v sont spécifiées par le biais des deux conditions (13.18).

Considérons le cas où $f = 0$, $v = 0$, $w(0) = w(1) = 0$ et introduisons la fonction 2-périodique ω définie par $\omega(x) = w(x)$ si $x \in [0,1]$, $\omega(x) = -w(-x)$ si $x \in [-1,0]$. Nous vérifions facilement que la fonction u définie par

$$u(x,t) = \frac{1}{2}\Big(\omega(x-ct) + \omega(x+ct)\Big) \quad \forall x \in [0,1], \quad \forall t \geq 0, \tag{13.19}$$

est solution du problème (13.16) (13.17) (13.18). Du point de vue physique, le déplacement vertical u de la corde vibrante est la somme de deux ondes se propageant de droite à gauche et de gauche à droite à la vitesse c. Pour cette raison, l'équation (13.16) est appelée ***équation des ondes***.

Nous allons maintenant proposer une méthode de différences finies couramment utilisée pour résoudre numériquement (13.16) (13.17) (13.18). Soit N un entier positif, $h = \frac{1}{N+1}$, $x_j = jh$ avec $j = 0, 1, 2, \ldots, N+1$. De façon semblable à ce qui a été fait pour le problème parabolique (sect. 12.1), nous commençons par établir une ***semi-discrétisation*** en espace du problème (13.16) (13.17) (13.18) par différences finies, à savoir

$$\frac{d^2}{dt^2}u_j(t) + c^2\frac{-u_{j-1}(t) + 2u_j(t) - u_{j+1}(t)}{h^2}$$
$$= f(x_j,t) \qquad j = 1, ..., N, \forall t > 0, \tag{13.20}$$
$$u_0(t) = u_{N+1}(t) = 0 \qquad \forall t > 0, \tag{13.21}$$
$$u_j(0) = w(x_j) \text{ et } \frac{d}{dt}u_j(0) = v(x_j) \qquad j = 1, \ldots, N. \tag{13.22}$$

Ici $u_j(t)$ est une approximation de $u(x_j,t)$ pour $j = 1, \ldots, N$. De façon similaire à ce que nous avons fait dans le cadre du problème de la chaleur (sect. 12.1),

nous introduisons la $N \times N$ matrice A définie par

$$A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \tag{13.23}$$

le N-vecteur $\vec{u}(t)$ de composantes $u_1(t)$, ..., $u_N(t)$, le N-vecteur $\vec{f}(t)$ de composantes $f(x_1,t)$, ..., $f(x_N,t)$, le N-vecteur $\vec{w}$ de composantes $w(x_1)$, ..., $w(x_N)$ et le N-vecteur $\vec{v}$ de composantes $v(x_1)$, ..., $v(x_N)$. Le système (13.20) (13.21) (13.22) peut alors s'écrire sous la forme condensée

$$\ddot{\vec{u}}(t) + c^2 A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0, \tag{13.24}$$

$$\vec{u}(0) = \vec{w}, \qquad \dot{\vec{u}}(0) = \vec{v}, \tag{13.25}$$

où nous avons noté $\dot{\vec{u}}(t)$ le N-vecteur de composantes $du_1(t)/dt$, ..., $du_N(t)/dt$ et $\ddot{\vec{u}}(t)$ le N-vecteur de composantes $d^2u_1(t)/dt^2$, ..., $d^2u_N(t)/dt^2$. Le système différentiel (13.24) (13.25) est d'ordre deux et nous pouvons utiliser la méthode de Newmark décrite dans la section 9.7 pour le résoudre numériquement, voir (9.37) (9.38) (9.39). Si $\tau > 0$ est un pas de temps donné, si $t_n = n\tau$ avec $n = 0, 1, 2, \ldots$, et si $\vec{u}^n$ est une approximation de $\vec{u}(t_n)$ (i.e. $u_j^n \simeq u_j(t_n) \simeq u(x_j, t_n)$, $j = 1, \ldots, N$), alors une discrétisation en temps du schéma (13.24) (13.25) est la suivante :

$$\frac{\vec{u}^{n+1} - 2\vec{u}^n + \vec{u}^{n-1}}{\tau^2} + c^2 A\vec{u}^n = \vec{f}(t_n), \quad n = 1, 2, \ldots, \tag{13.26}$$

$$\vec{u}^0 = \vec{w}, \qquad \vec{u}^1 = \vec{w} + \tau\vec{v} + \frac{1}{2}\tau^2\Big(\vec{f}(0) - c^2 A\vec{w}\Big). \tag{13.27}$$

Le schéma (13.26) (13.27) est un schéma explicite. Connaissant $\vec{u}^0$ et $\vec{u}^1$ nous pouvons calculer pour $n = 1, 2, \ldots$

$$\vec{u}^{n+1} = (2I - \tau^2 c^2 A)\vec{u}^n - \vec{u}^{n-1} + \tau^2 \vec{f}(t_n), \tag{13.28}$$

où I désigne la $N \times N$ matrice identité. Posons $\lambda = \tau^2 c^2/h^2$ et utilisons la convention $u_0^n = u_{N+1}^n = 0$. La relation (13.28) s'écrit composante par composante :

$$u_j^{n+1} = 2\,(1-\lambda)\,u_j^n + \lambda\Big(u_{j-1}^n + u_{j+1}^n\Big) - u_j^{n-1} + \tau^2 f(x_j, t_n), \tag{13.29}$$

pour $j = 1, \ldots, N$.

Posons à nouveau $f = 0$ et $v = 0$ et comparons la solution u du problème (13.16) (13.17) (13.18) définie par (13.19) à son approximation numérique définie par (13.29). Supposons, pour simplifier, que

$$w(x) = \sin m\pi x \tag{13.30}$$

où m est un entier positif (si ce n'est pas le cas nous pouvons développer $w(x)$ en série de Fourier et les calculs sont similaires). En utilisant la formule trigonométrique

$$\sin(\alpha+\beta)+\sin(\alpha-\beta)=2\sin\alpha\cos\beta, \quad \alpha,\beta\in\mathbb{R}, \tag{13.31}$$

et (13.19) nous obtenons :

$$u(x,t)=\sin(m\pi x)\cos(m\pi ct).$$

Puisque $x_j=jh$ et $t_n=n\tau$, nous avons donc

$$u(x_j,t_n)=\sin(m\pi jh)\cos(m\pi cn\tau). \tag{13.32}$$

D'autre part, le schéma numérique (13.26) (13.27) s'écrit avec $f=0$ et $v=0$:

$$\begin{aligned}
u_j^0 &= w(x_j),\\
u_j^1 &= (1-\lambda)w(x_j)+\frac{1}{2}\lambda(w(x_{j-1})+w(x_{j+1})),\\
u_j^2 &= 2(1-\lambda)u_j^1+\lambda(u_{j-1}^1+u_{j+1}^1)-u_j^0,\\
&\vdots\\
u_j^{n+1} &= 2(1-\lambda)u_j^n+\lambda(u_{j-1}^n+u_{j+1}^n)-u_j^{n-1},
\end{aligned} \tag{13.33}$$

pour $j=1,\ldots,N$. En utilisant la condition initiale (13.30) et la formule trigonométrique (13.31) nous obtenons

$$u_j^n=\alpha_n\sin(m\pi jh), \tag{13.34}$$

où les coefficients α_n dépendent de m et sont donnés par les formules de récurrence :

$$\begin{aligned}
\alpha_0 &= 1,\\
\alpha_1 &= 1-\lambda(1-\cos(m\pi h)),\\
\alpha_2 &= 2\alpha_1\alpha_1-\alpha_0,\\
&\vdots\\
\alpha_n &= 2\alpha_1\alpha_{n-1}-\alpha_{n-2}.
\end{aligned} \tag{13.35}$$

Par conséquent, compte tenu de (13.32) et (13.35), u_j^n est une bonne approximation de $u(x_j,t_n)$, si et seulement si α_n est une bonne approximation de $\cos(m\pi cn\tau)$. Une condition nécessaire pour que ce soit le cas est que $|\alpha_n|$ reste borné indépendamment de m et n. Nous adoptons donc la définition suivante :

Définition 13.1 *Le schéma (13.26) (13.27) est stable s'il existe une constante C telle que les valeurs $(\alpha_n)_{n=0}^{\infty}$ définies par (13.35) satisfassent*

$$|\alpha_n|\le C, \qquad n=0,1,2,\ldots, \qquad m=1,2,\ldots$$

Nous sommes maintenant en mesure de montrer le résultat suivant.

Théorème 13.1 *Le schéma (13.26) (13.27) est stable si la condition CFL suivante est satisfaite :*

$$\tau \leq \frac{h}{c}.$$

Démonstration
Soit α_1 le coefficient défini en (13.35) et soit p le polynôme de degré 2 en s défini par :

$$p(s) = s^2 - 2\alpha_1 s + 1.$$

Clairement les zéros de p sont donnés par les 2 valeurs

$$s_+ = \alpha_1 + \sqrt{\alpha_1^2 - 1} \quad \text{et} \quad s_- = \alpha_1 - \sqrt{\alpha_1^2 - 1}, \tag{13.36}$$

où nous avons noté $\sqrt{\alpha_1^2 - 1}$ la racine positive de $\alpha_1^2 - 1$ si $|\alpha_1| \geq 1$ et $\sqrt{\alpha_1^2 - 1} = i\sqrt{1 - \alpha_1^2}$ si $|\alpha_1| < 1$, i étant l'unité imaginaire. Vérifions que le coefficient α_n défini en (13.35) est tel que

$$\alpha_n = \frac{1}{2}(s_+^n + s_-^n), \qquad n = 0, 1, 2, \ldots \tag{13.37}$$

En effet, nous constatons immédiatement que l'égalité (13.37) est vraie pour $n = 0$ et $n = 1$. Supposons que (13.37) soit vraie pour $n \leq k$ et montrons qu'elle reste vraie pour $n = k + 1$. L'hypothèse de récurrence et les relations (13.35), nous assurent que

$$\begin{aligned}
\alpha_{k+1} &= 2\alpha_1\alpha_k - \alpha_{k-1} \\
&= 2\alpha_1 \frac{1}{2}(s_+^k + s_-^k) - \frac{1}{2}(s_+^{k-1} + s_-^{k-1}) \\
&= \frac{1}{2}s_+^{k-1}(2\alpha_1 s_+ - 1) + \frac{1}{2}s_-^{k-1}(2\alpha_1 s_- - 1).
\end{aligned}$$

En utilisant (13.36), nous avons $s_\pm^2 = 2\alpha_1 s_\pm - 1$ et donc

$$\alpha_{k+1} = \frac{1}{2}(s_+^{k+1} + s_-^{k+1})$$

qui est bien la formule (13.37) pour $n = k + 1$.

Revenons à la question de la stabilité du schéma (13.26) (13.27). Pour obtenir $|\alpha_n| \leq C$ pour tout $n = 0, 1, 2, \ldots$, et pour tout $m = 1, 2, \ldots$, il suffit, en vertu de (13.37), d'assurer que

$$|s_\pm| \leq 1, \qquad m = 1, 2, \ldots \tag{13.38}$$

Nous constatons que si $|\alpha_1| \leq 1$ alors le critère (13.38) est satisfait. Ainsi $|\alpha_n| \leq C$ si

$$-1 \leq \alpha_1 \leq 1 \qquad m = 1, 2, \ldots,$$

soit, en utilisant (13.35) :

$$-1 \leq 1 - \lambda(1 - \cos m\pi h) \leq 1, \quad m = 1, 2, \ldots \tag{13.39}$$

L'inégalité de droite dans (13.39) est toujours satisfaite. L'inégalité de gauche est satisfaite pour autant que

$$\lambda \leq \frac{2}{1 - \cos mh}, \qquad m = 1, 2, \ldots, \tag{13.40}$$

ce qui est le cas si

$$\lambda \leq 1.$$

Puisque $\lambda = \tau^2 c^2 / h^2$, nous obtenons bien le résultat de stabilité annoncé dans notre théorème. ■

Remarque 13.1 Il est possible de montrer que si la condition CFL est satisfaite, le schéma numérique (13.26) (13.27) est d'ordre 2. Plus précisément, nous voulons pour un temps $T > 0$, approcher numériquement $u(x, T)$, $0 < x < 1$, au moyen du schéma (13.26) (13.27) en prenant M pas de temps. Nous posons donc $\tau = T/M$, $t_n = n\tau$, pour $n = 0, 1, \ldots, M$, et nous choisissons un entier N tel que $(N+1)cT \leq M$ de sorte que le pas d'espace h défini par $h = 1/(N+1)$ satisfasse $h \geq \tau c$. Nous calculons ensuite les valeurs $u_1^M, \ldots, u_N^M$ et l'erreur maximale satisfait

$$\max_{j=1,\ldots,N} |u_j^M - u(x_j, T)| \leq Ch^2, \qquad \text{si } h \to 0, \tag{13.41}$$

la constante C étant indépendante de M et N.

13.3 Equations des ondes 2D et éléments finis

Soit Ω un domaine polygonal de $\mathbb{R}^2$, de frontière $\partial\Omega$ et soit $\overline{\Omega} = \Omega \cup \partial\Omega$. Donnons-nous trois fonctions continues

$$\begin{aligned} &f : (x, t) \in \overline{\Omega} \times \mathbb{R}^+ \to f(x, t) \in \mathbb{R}, \\ &w : x \in \overline{\Omega} \to w(x) \in \mathbb{R}, \\ &v : x \in \overline{\Omega} \to v(x) \in \mathbb{R}, \end{aligned}$$

où le point $x \in \overline{\Omega}$ a naturellement deux composantes x_1 et x_2 ; nous noterons $x = (x_1, x_2)$. Si c est un nombre positif donné, nous posons le problème de trouver

$$u : (x, t) \in \overline{\Omega} \times \mathbb{R}^+ \to u(x, t) \in \mathbb{R}$$

tel que

$$\frac{\partial^2 u}{\partial t^2}(x, t) - c^2 \Delta u(x, t) = f(x, t) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{13.42}$$

$$u(x, t) = 0 \qquad \forall x \in \partial\Omega, \quad \forall t > 0, \tag{13.43}$$

$$u(x, 0) = w(x) \text{ et } \frac{\partial u}{\partial t}(x, 0) = v(x) \qquad \forall x \in \Omega. \tag{13.44}$$

Les équations (13.42) (13.43) (13.44) modélisent par exemple les vibrations d'une membrane élastique. Considérons une membrane tendue dans le plan horizontal Ox_1x_2, attachée en son bord $\partial\Omega$ et soumise à un champ de force vertical de densité $f(x,t)$ au point $x \in \Omega$ et à l'instant t. La déformation verticale de cette membrane au point x et à l'instant t satisfait alors, en première approximation, les équations (13.42) (13.43) (13.44), c étant un nombre dépendant de la masse spécifique et de la tension de la membrane. Les égalités (13.44) décrivent la déformation verticale initiale de la membrane ainsi que la vitesse initiale de déformation. Notons encore que le cas $f \equiv 0$ et $v \equiv 0$ correspond à la situation où nous lâchons la membrane après l'avoir déformée. Les vibrations de la membrane se traduiront dans ce cas par des propagations d'ondes comme dans le cas de la corde vibrante.

Pour discrétiser spatialement les équations (13.42) (13.43) (13.44) par la méthode des éléments finis, nous procédons comme dans la section 11.1. Pour ce faire, nous multiplions (13.42) par une fonction test $\varphi : x \in \overline{\Omega} \to \varphi(x) \in \mathbb{R}$ appartenant à l'espace V introduit dans la section 11.1, s'annulant sur le bord $\partial\Omega$, et nous intégrons par partie comme en (11.5)-(11.9). Nous obtenons :

$$\begin{aligned} \iint_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\varphi(x)dx + c^2 \iint_\Omega \overrightarrow{\text{grad}}u(x,t) \cdot \overrightarrow{\text{grad}}\varphi(x)dx \\ = \iint_\Omega f(x,t)\varphi(x)dx \qquad \forall t > 0. \end{aligned} \tag{13.45}$$

Si $\varphi_1, \varphi_2, \ldots, \varphi_N$ sont N fonctions linéairement indépendantes de V, nous construisons l'espace V_h en considérant toutes les combinaisons linéaires des fonctions φ_i, comme nous l'avons déjà fait dans les sections 11.1 et 12.3. Soit u_h l'approximation de u définie par

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \qquad \forall x \in \overline{\Omega}.$$

Remplaçons u par u_h dans (13.45) et choisissons pour fonctions test $\varphi = \varphi_j$, $j = 1, \ldots, N$. Nous obtenons :

$$\begin{aligned} \sum_{i=1}^{N} \ddot{u}_i(t) \iint_\Omega \varphi_i(x)\varphi_j(x)dx + c^2 \sum_{i=1}^{N} u_i(t) \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx \\ = \iint_\Omega f(x,t)\varphi_j(x)dx, \qquad j = 1, \ldots, N, \quad \forall t > 0. \end{aligned} \tag{13.46}$$

Utilisons à nouveau les notations de la section 12.3. Soit M la matrice de masse de coefficients

$$M_{ji} = \iint_\Omega \varphi_i(x)\varphi_j(x)dx, \qquad i,j = 1, \ldots, N,$$

soit A la matrice de rigidité de coefficients

$$A_{ji} = \iint_{\Omega} \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx, \qquad i,j = 1,\ldots,N,$$

soit $\vec{u}(t)$ le N-vecteur de composantes $u_1(t),\ldots,u_N(t)$ et $\vec{f}(t)$ le N-vecteur de composantes $f_1(t),\ldots,f_N(t)$ définies par :

$$f_j(t) = \iint_{\Omega} f(x,t)\varphi_j(x)dx, \qquad j = 1,\ldots,N.$$

Nous pouvons alors écrire les relations (13.46) sous forme d'un système différentiel, à savoir :

$$M\ddot{\vec{u}}(t) + c^2 A\vec{u}(t) = \vec{f}(t), \qquad \forall t > 0. \tag{13.47}$$

Puisque ce système différentiel est du deuxième ordre, nous devons ajouter les conditions initiales

$$\vec{u}(0) = \vec{w} \text{ et } \dot{\vec{u}}(0) = \vec{v}. \tag{13.48}$$

Les N-vecteurs $\vec{w}$ et $\vec{v}$ ont pour composantes les coefficients $w_1,\ldots,w_N$ et $v_1,\ldots,v_N$ qui sont tels que les quantités

$$\sum_{j=1}^{N} w_j\varphi_j(x) \text{ et } \sum_{j=1}^{N} v_j\varphi_j(x)$$

soient des approximations des conditions initiales $w(x)$ et $v(x)$, respectivement (par exemple les interpolants aux nœuds de la triangulation $\mathcal{T}_h$). Nous utiliserons à nouveau la méthode de Newmark décrite dans la section 9.7 pour résoudre numériquement (13.47) (13.48). Si nous voulons, comme nous l'avons fait dans la section précédente, que cette méthode soit explicite, nous devrons utiliser une méthode d'intégration numérique de sorte que la matrice de masse M soit diagonale (*mass lumping*, voir aussi la remarque 12.1).

13.4 Equation de transport 1D non linéaire

Supposons que l'on ait un continuum unidimensionnel de particules, réparties sur la droite réelle Ox et sans interactions entre elles. Notons $u(x,t)$ la vitesse de la particule se trouvant au point $x \in \mathbb{R}$ et à l'instant $t > 0$ (description eulérienne). Si nous désignons par $x = g_{\overline{x}}(t)$, $t > 0$, la trajectoire horaire de la particule se trouvant en $x = \overline{x}$ au temps $t = 0$ (description lagrangienne), alors sa vitesse au temps $t > 0$ est donnée par $\dot{g}_{\overline{x}}(t)$ et nous avons par définition

$$\dot{g}_{\overline{x}}(t) = u\big(g_{\overline{x}}(t),t\big), \qquad t > 0, \tag{13.49}$$

$$g_{\overline{x}}(0) = \overline{x}. \tag{13.50}$$

En dérivant (13.49) par rapport au temps, nous obtenons

$$\ddot{g}_{\overline{x}}(t) = \frac{\partial u}{\partial x}\big(g_{\overline{x}}(t),t\big)\dot{g}_{\overline{x}}(t) + \frac{\partial u}{\partial t}\big(g_{\overline{x}}(t),t\big). \tag{13.51}$$

Puisque les particules n'interagissent pas entre elles, l'accélération $\ddot{g}_{\overline{x}}(t)$ est nulle. En utilisant (13.49), l'équation (13.51) s'écrit :

$$\frac{\partial u}{\partial t}\big(g_{\overline{x}}(t),t\big)+u\big(g_{\overline{x}}(t),t\big)\frac{\partial u}{\partial x}\big(g_{\overline{x}}(t),t\big)=0, \qquad t>0, \tag{13.52}$$

$$u\big(g_{\overline{x}}(0),0\big)=u(\overline{x},0). \tag{13.53}$$

Les équations (13.52) (13.53) justifient l'étude du problème de transport non linéaire suivant (appelé ***problème de Burgers***) :

Etant donnée une fonction $w : x \in \mathbb{R} \to w(x) \in \mathbb{R}$, trouver une fonction de deux variables $u : (x,t) \in \mathbb{R} \times \mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ telle que

$$\frac{\partial u}{\partial t}(x,t)+u(x,t)\frac{\partial u}{\partial x}(x,t)=0 \qquad \forall x\in\mathbb{R}, \qquad \forall t>0, \tag{13.54}$$

$$u(x,0)=w(x) \qquad \forall x\in\mathbb{R}. \tag{13.55}$$

Le problème (13.54) (13.55) peut présenter des difficultés que nous décrivons très brièvement. Supposons connaître une solution $u(x,t)$ du problème ci-dessus et résolvons le problème de Cauchy (sect. 9.1) suivant : trouver $\beta : t \in \mathbb{R}^+ \to \beta(t) \in \mathbb{R}$ tel que

$$\dot{\beta}(t)=u\big(\beta(t),t\big), \qquad t>0, \tag{13.56}$$

$$\beta(0)=\overline{x} \tag{13.57}$$

où $\overline{x} \in \mathbb{R}$ est un nombre donné. Si nous supposons que u vérifie la condition (9.4) du théorème 9.2, le problème (13.56) (13.57) a une solution globale unique β. Posons maintenant

$$\gamma(t)=u\big(\beta(t),t\big).$$

Nous avons

$$\dot{\gamma}(t)=\frac{\partial u}{\partial x}\big(\beta(t),t\big)u\big(\beta(t),t\big)+\frac{\partial u}{\partial t}\big(\beta(t),t\big).$$

Puisque u satisfait (13.54), nous obtenons

$$\dot{\gamma}(t)=0,$$

et par conséquent, compte tenu de (13.55) et (13.57),

$$\begin{aligned}\gamma(t)&=\text{constante}=u\big(\beta(t),t\big)=u\big(\beta(0),0\big)\\&=u(\overline{x},0)=w(\overline{x}) \qquad \forall t>0.\end{aligned} \tag{13.58}$$

En utilisant (13.56) et (13.58), nous avons

$$\dot{\beta}(t)=\gamma(t)=w(\overline{x}) \qquad \forall t>0,$$

soit, en intégrant de 0 à t et en utilisant (13.57)

$$\beta(t)=w(\overline{x})t+\overline{x} \qquad \forall t>0. \tag{13.59}$$

Nous avons donc montré que si $u(x,t)$ est une fonction suffisamment régulière qui satisfait (13.54) (13.55), alors

$$u\big(w(\overline{x})t+\overline{x},t\big)=w(\overline{x}) \qquad \forall \overline{x}\in\mathbb{R}, \quad \forall t>0. \tag{13.60}$$

La relation (13.60) traduit le fait que la solution u reste constante sur la droite d'équation $x = w(\overline{x})t + \overline{x}$ dans le plan Oxt appelée ***courbe caractéristique***. Dans la figure 13.3, nous représentons ces courbes caractéristiques (qui sont les lignes de niveau de la solution $u(x,t)$) dans le cas où la fonction w est définie par $w(x) = x$ et par $w(x) = -x$, respectivement.

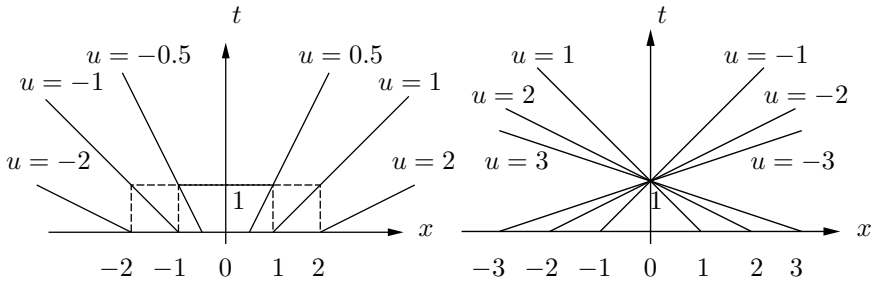

Fig. 13.3 Courbes caractéristiques lorsque $w(x) = x$, i.e. $x = \overline{x}t + \overline{x}$ (fig. de gauche) et lorsque $w(x) = -x$, i.e. $x = -\overline{x}t + \overline{x}$ (fig. de droite).

Lorsque la fonction w est définie par $w(x) = x$, nous avons $u(xt + x, t) = x$ et nous sommes en présence d'une ***onde de détente***. En revenant au modèle particulaire du début de cette section, nous affirmons que les trajectoires dans l'espace-temps des particules sont des droites et s'éloignent les unes des autres au cours du temps.

Lorsque la fonction w est définie par $w(x) = -x$, nous constatons que les courbes caractéristiques se coupent au point $(0, 1)$ dans l'espace-temps. Ainsi, lorsque le temps t atteint la valeur 1, la solution u devient discontinue et, puisque u n'est plus régulière, l'égalité (13.60) n'est plus valable. Au temps $t = 1$ une ***onde de choc*** est engendrée et il faudra dès lors élaborer une théorie plus complexe pour trouver une solution physique du problème.

Ce dernier exemple nous montre une des difficultés inhérentes aux problèmes hyperboliques non linéaires tels que les équations (13.54) (13.55) ou plus généralement les équations qui régissent la dynamique des gaz compressibles. Une question importante est de proposer des schémas numériques qui permettent de décrire correctement les chocs. Cette question fait l'objet de nombreux articles et est trop complexe pour être abordée dans ce chapitre !

13.5 Exercices

Exercice 13.1 Soit u la solution du problème (13.4) (13.5), avec $c_0 > 0$. Soit h le pas spatial, τ le pas temporel et posons $x_j = jh$, $j \in \mathbb{Z}$ et $t_n = n\tau$,

$n = 0, 1, 2, \ldots$ Soit u_j^n l'approximation de $u(x_j, t_n)$ définie par le schéma décentré

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c_0 \frac{u_j^n - u_{j-1}^n}{h} = 0. \tag{13.61}$$

1. On fixe l'entier n et on suppose que $c_0\tau \leq h$. Montrer que si tous les u_j^n, $j \in \mathbb{Z}$, sont positifs, alors tous les u_j^{n+1}, $j \in \mathbb{Z}$, sont aussi positifs. Montrer que
$$\sup_{j\in\mathbb{Z}} |u_j^{n+1}| \leq \sup_{j\in\mathbb{Z}} |u_j^n|. \tag{13.62}$$
La relation (13.62) garantit que le schéma décentré (13.61) est stable dès que $c_0\tau \leq h$.
2. Pourquoi est-il souhaitable que le schéma numérique satisfasse le point 1 ?
3. On considère le cas où la condition initiale est définie par $w(x) = 1$ si $x \leq 0$, $w(x) = 0$ si $x > 0$. On choisit $c_0 = 2$, $h = \tau$ de sorte que $c_0\tau > h$. Construire et représenter la solution numérique pour $n = 1, 2, 3, 4$. Vérifier sur le graphique que $u_n^n = 2^n$ et en déduire que le schéma est numériquement instable (en norme du maximum).

Solution

1. Le schéma numérique (13.61) s'écrit

$$u_j^{n+1} = (1 - \alpha)u_j^n + \alpha u_{j-1}^n, \tag{13.63}$$

où nous avons noté $\alpha = c_0\tau/h$. Puisque c_0 est positif α est aussi positif. De plus, si la condition $c_0\tau \leq h$ est satisfaite, alors $1 - \alpha \geq 0$. En vertu de la relation (13.63), il est clair que si u_{j-1}^n et u_j^n sont positifs, alors u_j^{n+1} est aussi positif. Nous avons bien montré que si tous les u_j^n sont positifs, alors tous les u_j^{n+1} sont également positifs.

Montrons maintenant (13.62). Soit $j \in \mathbb{Z}$ fixé. L'égalité (13.63) implique

$$|u_j^{n+1}| \leq |1 - \alpha||u_j^n| + |\alpha||u_{j-1}^n|.$$

Puisque $\alpha \geq 0$ et $1 - \alpha \geq 0$, nous avons donc

$$\begin{aligned} |u_j^{n+1}| &\leq (1 - \alpha)|u_j^n| + \alpha|u_{j-1}^n| \\ &\leq (1 - \alpha) \sup_{k\in\mathbb{Z}} |u_k^n| + \alpha \sup_{k\in\mathbb{Z}} |u_k^n| \\ &= \sup_{k\in\mathbb{Z}} |u_k^n|, \end{aligned}$$

et nous avons donc démontré (13.62).

2. Nous avons vu au début de ce chapitre que la solution exacte du problème (13.4) (13.5) est donnée par (13.3). Par conséquent, si la condition initiale w satisfait $w(x) \geq 0$, $\forall x \in \mathbb{R}$, alors on a aussi $u(x, t) \geq 0$, $\forall x \in \mathbb{R}$, $\forall t \geq 0$. D'autre part, à partir de (13.3), on a bien évidemment

$$\sup_{x\in\mathbb{R}} |u(x, t)| \leq \sup_{x\in\mathbb{R}} |w(x)|.$$

Les propriétés que nous avons démontrées au point 1 ne sont donc que l'analogue discret des propriétés ci-dessus.

3. Lorsque $c_0 = 2$, $h = \tau$, le schéma (13.61) ou (13.63) s'écrit

$$u_j^{n+1} = -u_j^n + 2u_{j-1}^n.$$

Les résultats numériques pour $n = 1, 2, 3, 4$ sont présentés dans la figure 13.4. La solution numérique présente des oscillations et on peut vérifier que $u_n^n = 2^n$, ce qui implique que

$$\lim_{n\to\infty} \sup_{j\in\mathbb{Z}} u_j^n = +\infty.$$

Nous avons donc vérifié que le schéma devient dans ce cas numériquement instable (en norme du maximum).

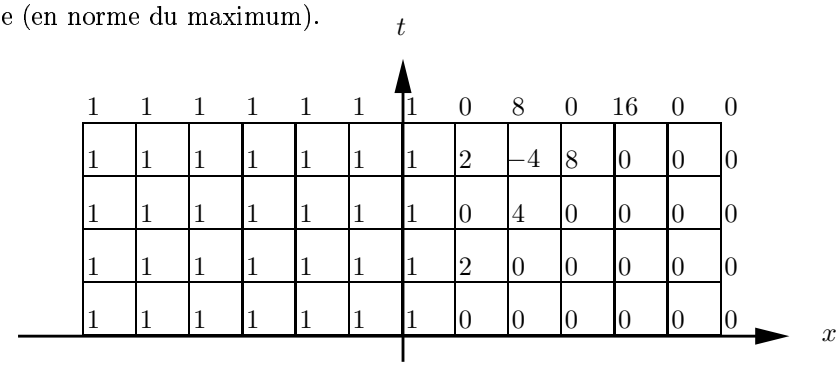

Fig. 13.4 Solution numérique du schéma (13.63) lorsque $c_0 = 2$, $h = \tau$.

Exercice 13.2 On considère un problème de transport dans un intervalle borné. Soit $T > 0$, $c_0 > 0$ donnés et soit u la solution de

$$\frac{\partial u}{\partial t}(x,t) + c_0 \frac{\partial u}{\partial x}(x,t) = 0 \qquad \forall x \in]0,1[, \quad \forall t \in]0,T[. \tag{13.64}$$

$$u(0,t) = 0 \qquad \forall t \in]0,T[, \tag{13.65}$$

$$u(x,0) = w(x) \qquad \forall x \in]0,1[, \tag{13.66}$$

où w est une fonction donnée. Si N est un entier positif, on note $h = 1/(N+1)$ le pas d'espace, $x_j = j\,h$, $j = 0, 1, \ldots, N+1$. Si M est un entier positif, on note $\tau = T/M$ le pas de temps, $t_n = n\tau$, $n = 0, 1, \ldots, M$. Finalement, on note u_j^n une approximation de $u(x_j, t_n)$. Pour résoudre numériquement (13.64) (13.65) (13.66), on utilise le schéma décentré suivant. On pose d'abord $u_0^n = 0$, $n = 0, 1, \ldots, M$. Soit $1 \le j \le N+1$ fixé. Etant donné u_{j-1}^n, $n = 0, 1, \ldots, M$, on calcule u_j^n, $n = 1, \ldots, M$, à l'aide du schéma suivant :

$$\frac{u_{j-1}^{n+1} - u_{j-1}^n}{\tau} + c_0 \frac{u_j^{n+1} - u_{j-1}^{n+1}}{h} = 0, \tag{13.67}$$

où nous avons posé $u_j^0 = w(x_j)$.

1. On fixe l'entier j, on suppose que $c_0\tau \geq h$ et que $w(x) \geq 0$, $\forall x \in [0,1]$. Montrer que si tous les u_{j-1}^n, $n = 0, 1, \ldots, M$ sont positifs, alors tous les u_j^n, $n = 0, 1, \ldots, M$ sont aussi positifs. Montrer que

$$\max_{0\leq n\leq M} |u_j^n| \leq \max_{0\leq n\leq M} |u_{j-1}^n|. \tag{13.68}$$

2. Comparer le schéma (13.67) avec le schéma décentré (13.14).
3. Que se passe-t-il lorsque $c_0 < 0$?

Solution

1. Si nous comparons le schéma (13.67) avec les schémas décrits dans la section 13.1, alors nous constatons que les rôles du temps et de l'espace ont été inversés. En effet, nous calculons une approximation de u aux points x_j à partir de l'approximation trouvée aux points x_{j-1}. Le schéma (13.67) s'écrit ainsi :

$$u_j^{n+1} = \left(1 - \frac{h}{c_0\tau}\right) u_{j-1}^{n+1} + \frac{h}{c_0\tau} u_{j-1}^n. \tag{13.69}$$

Nous pouvons donc, étant donné u_{j-1}^0, u_{j-1}^1, ..., u_{j-1}^M, calculer u_j^1, u_j^2, ..., u_j^M à l'aide de (13.69). Supposons maintenant $c_0\tau \geq h$, fixons $1 \leq j \leq M$ et supposons que les u_{j-1}^0, u_{j-1}^1, ..., u_{j-1}^M sont tous positifs. Puisque $u_j^0 = w(x_j) \geq 0$, il est clair, en vertu de (13.69), que tous les u_j^1, u_j^2, ..., u_j^M sont aussi positifs.

Montrons maintenant (13.68). Puisque $c_0\tau \geq h$, nous avons :

$$\begin{aligned}
|u_j^{n+1}| &\leq \left|1 - \frac{h}{c_0\tau}\right| |u_{j-1}^{n+1}| + \frac{h}{c_0\tau}|u_{j-1}^n| \\
&= \left(1 - \frac{h}{c_0\tau}\right) |u_{j-1}^{n+1}| + \frac{h}{c_0\tau}|u_{j-1}^n| \\
&\leq \left(1 - \frac{h}{c_0\tau}\right) \max_{0\leq m\leq M} |u_{j-1}^m| + \frac{h}{c_0\tau} \max_{0\leq m\leq M} |u_{j-1}^m|.
\end{aligned}$$

Il suffit de prendre le maximum sur l'indice n pour obtenir (13.68).

Nous concluons donc en affirmant que, si nous cherchons à résoudre un problème de transport dans un domaine borné en x et tel que la vitesse de transport c_0 soit positive, alors nous pouvons progresser selon les x positifs (de gauche à droite sur la figure 13.5).

2. Le schéma décentré (13.14) permet de calculer les valeurs u_j^{n+1}, $1 \leq j \leq N+1$, en fonction des valeurs u_j^n, $0 \leq j \leq N+1$, lorsque n est fixé. En effet le schéma (13.14) s'écrit dans le cadre de cet exercice :

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c_0 \frac{u_j^n - u_{j-1}^n}{h} = 0, \tag{13.70}$$

où nous posons $u_0^n = 0$ en vertu de la condition limite (13.65). Comme nous l'avons vu dans la section 13.1 et dans l'exercice 13.1, le schéma (13.70) est stable sous la condition $c_0\tau \leq h$, ce qui limite le pas de temps en fonction du pas d'espace. Bien évidemment, puisque nous avons permuté les rôles de l'espace

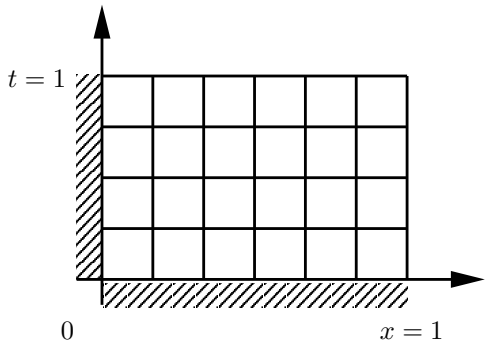

Fig. 13.5 Le domaine de calcul dans le cadre de l'exercice 13.2. Le bord hachuré indique la partie du bord sur laquelle il faut prescrire la valeur de u, dans le cas où $c_0 > 0$.

et du temps dans le schéma (13.67), il est normal que le pas d'espace soit limité en fonction du pas de temps pour assurer la stabilité de (13.67). Ainsi le schéma (13.67) est stable sous la condition $c_0\tau \geq h$.

3. Lorsque $c_0 < 0$, il faut imposer la valeur de u non pas en $x = 0$ mais en $x = 1$. Il faut donc remplacer la condition (13.65) par

$$u(1,t) = 0 \qquad \forall t \in]0,T[.$$

Du point de vue physique, cela revient à imposer la valeur de la fonction u dans la direction de transport de l'information. Lorsque $c_0 > 0$, l'information se propage vers les x croissants et il faut imposer u à l'entrée, c'est-à-dire en $x = 0$. Inversement, lorsque $c_0 < 0$, l'information se propage vers les x décroissants et il faut encore imposer u à l'entrée qui cette fois-ci est en $x = 1$. Du point de vue numérique, nous pouvons dans ce cas déterminer les valeurs u_j^n en progressant de la droite vers la gauche. Il faut donc remplacer le schéma (13.67) par le schéma suivant :

$$\frac{u_{j+1}^{n+1} - u_{j+1}^n}{\tau} + c_0 \frac{u_{j+1}^{n+1} - u_j^{n+1}}{h} = 0.$$

13.6 Notes bibliographiques et remarques

Dans la section 13.1, nous avons décrit une méthode de différences finies pour la résolution d'un problème de transport. Il existe d'autres méthodes numériques pour résoudre ce genre de problèmes. Citons par exemple la méthode des caractéristiques, les méthodes particulaires, les méthodes de volumes finis. Plusieurs méthodes d'éléments finis sont également envisageables : les méthodes d'éléments finis continus de type SUPG (Streamline Upwind Petrov Galerkin) ou GLS (Galerkin Least Squares) ainsi que les méthodes d'éléments finis discontinus. La plupart de ces méthodes sont décrites dans [8, 24].

Les problèmes de déformation dynamique des structures, les problèmes d'acoustique sont modélisés par des équations semblables à celle des ondes 2D (sect. 13.3), voir par exemple [5].

La dynamique des gaz est souvent modélisée par des systèmes hyperboliques de lois de conservation. Les équations correspondantes sont non linéaires et peuvent générer des ondes de chocs, même si les données initiales du gaz (température, densité, quantité de mouvement) sont très régulières ! La résolution numérique de tels problèmes est difficile, voir par exemple [14] pour une description des schémas numériques, [10, 24] pour une approche plus mathématique.

Chapitre 14

Approximation de problèmes de convection-diffusion

14.1 Un problème de convection-diffusion stationnaire et différences finies

Soit $f : x \in [0,1] \to f(x) \in \mathbb{R}$ et $c : x \in [0,1] \to c(x) \in \mathbb{R}$ deux fonctions continues données et soit $\varepsilon > 0$ fixé. Nous cherchons une fonction $u : x \in [0,1] \to u(x) \in \mathbb{R}$ satisfaisant

$$\begin{aligned} &-\varepsilon u''(x) + c(x)u'(x) = f(x) \qquad \forall x \in]0,1[, \\ &u(0) = u(1) = 0. \end{aligned} \tag{14.1}$$

Le problème (14.1) est appelé problème de ***convection-diffusion stationnaire***. Le terme de diffusion est $-\varepsilon u''(x)$ alors que le terme de convection est $c(x)u'(x)$. Si $c(x) = 0 \ \forall x \in [0,1]$, le problème (14.1) est le problème de diffusion que nous avons traité dans le chapitre 10 (attention : remarquer que le problème (10.1) et le problème (14.1) sont différents car dans le premier apparaît un terme de type $c(x)u(x)$ alors que dans le deuxième nous avons $c(x)u'(x)$). Le problème (14.1) avec $c(x) = 0$, $\forall x \in [0,1]$, décrit aussi le problème (12.1) de diffusion de la chaleur stationnaire (c'est-à-dire le problème (12.1) avec $\partial u/\partial t = 0$ et f indépendant de t). Si nous choisissons $\varepsilon = 0$ dans (14.1), nous obtenons alors le problème de transport stationnaire (c'est-à-dire le problème (13.1) avec $\partial u/\partial t = 0$). L'exemple typique d'une situation physique régie par les équations (14.1) est le problème de la propagation de chaleur stationnaire dans un fluide (unidimensionnel) soumis à des mouvements de convection stationnaire. Dans ce cas, u représente la température du fluide et c la vitesse du fluide.

Considérons maintenant le cas simple où $c = c_0 =$ constante $\neq 0$ et $f = f_0 =$ constante. Dans ce cas, la première équation de (14.1) est une équation différentielle linéaire à coefficients constants que nous pouvons résoudre

explicitement. En tenant compte des conditions aux limites, nous obtenons :

$$u(x) = \frac{f_0}{c_0}\left(x - \frac{1-\exp\left(\frac{c_0}{\varepsilon}x\right)}{1-\exp\left(\frac{c_0}{\varepsilon}\right)}\right) \qquad \forall x \in [0,1]. \tag{14.2}$$

Si c_0 est positif et si c_0/ε est très grand, i.e. $\varepsilon << c_0$, alors $u(x)$ se comporte comme $f_0 x/c_0$ excepté dans un voisinage d'ordre ε/c_0 du point limite $x = 1$, voisinage dans lequel la solution u présente *une couche limite* (fig. 14.1).

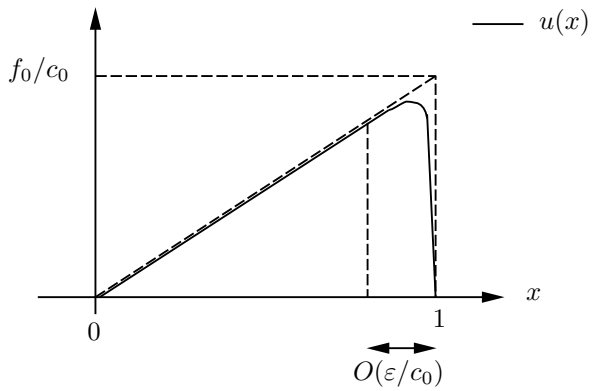

Fig. 14.1 Solution $u(x)$ lorsque $\varepsilon << c_0$.

Considérons maintenant une approximation par différences finies des équations (14.1) en prenant des différences centrées pour le terme de convection. Soit N un entier positif, $h = \frac{1}{N+1}$, $x_j = jh$, $j = 0, 1, \ldots, N+1$ et u_j une approximation de $u(x_j)$ au point x_j. Nous étudions le schéma suivant, appelé ***schéma centré*** :

$$\begin{aligned} &\varepsilon\frac{2u_j - u_{j-1} - u_{j+1}}{h^2} + c(x_j)\frac{u_{j+1} - u_{j-1}}{2h} \\ &\qquad\qquad = f(x_j), \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0. \end{aligned} \tag{14.3}$$

Clairement (14.3) est un système linéaire de N équations et N inconnues u_1, ..., u_N. Dans le cas simple où $c(x) = c_0 =$ constante $\neq 0$, $f(x) = f_0 =$ constante et $h = 2\varepsilon/c_0$, un calcul immédiat donne $u_j = f_0 x_j/c_0$, $j = 1, \ldots, N$. Si $h < 2\varepsilon/c_0$, les résultats numériques de la figure 14.2 montrent que les valeurs u_j approchent correctement la solution du problème. Par contre, si $h > 2\varepsilon/c_0$, alors les valeurs u_j présentent des oscillations au voisinage de la couche limite, voir la figure 14.3. Dans ce cas, le pas d'espace h est trop grand en regard de l'épaisseur de la couche limite (qui est de l'ordre de ε/c_0 lorsque $\varepsilon << c_0$). Nous allons proposer un autre schéma que le schéma (14.3) que nous appellerons ***schéma décentré***.

Si α_j est un nombre compris entre zéro et un, alors nous pouvons approcher $u'(x_j)$ par $\alpha_j(u_j - u_{j-1})/h + (1-\alpha_j)(u_{j+1} - u_j)/h$ qui est une moyenne pondérée

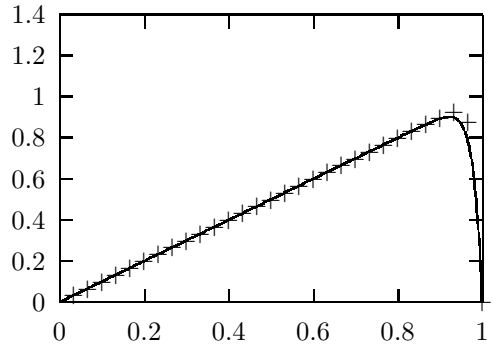

Fig. 14.2 Les valeurs u_j pour $h < 2\varepsilon/c_0$ (ici $f_0 = 1, \varepsilon = 0.02, c_0 = 1, h = 1/30$).

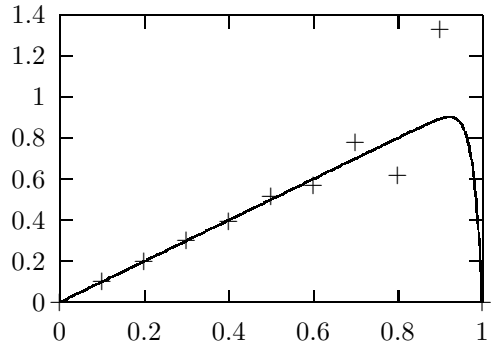

Fig. 14.3 Les valeurs u_j pour $h > 2\varepsilon/c_0$ (ici $f_0 = 1, \varepsilon = 0.02, c_0 = 1, h = 1/10$).

entre les formules de différences finies rétrograde et progressive (chap. 2). Le schéma numérique pour discrétiser (14.1) devient :

$$\begin{aligned} &\varepsilon\frac{2u_j - u_{j+1} - u_{j-1}}{h^2} + \frac{c(x_j)}{h}\Big(\alpha_j(u_j - u_{j-1}) + (1-\alpha_j)(u_{j+1} - u_j)\Big) \\ &\qquad = f(x_j), \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0. \end{aligned} \tag{14.4}$$

Il reste donc à choisir les nombres $\alpha_j \in [0,1]$ pour obtenir la meilleure approximation possible. Remarquons que si nous choisissons $\alpha_j = 1/2$, $j = 1, \ldots, N$, le schéma (14.4) coïncide avec (14.3) qui, nous l'avons vu, peut être oscillant lorsque h n'est pas suffisamment petit. Nous allons, sur la base de (14.2), donner une formule de calcul pour les valeurs α_j.

Pour ce faire, supposons que $c(x) = c_0 =$ constante positive, $f(x) = f_0 =$

constante et $\alpha_j = \alpha \in [0,1]$, $j = 1, \ldots, N$. Dans ce cas, le schéma (14.4) s'écrit

$$\begin{aligned} &2u_j - u_{j+1} - u_{j-1} + \gamma\Big(\alpha(u_j - u_{j-1}) + (1-\alpha)(u_{j+1} - u_j)\Big) \\ &\qquad = \frac{h^2 f_0}{\varepsilon}, \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0, \end{aligned} \tag{14.5}$$

où nous avons noté

$$\gamma = \frac{c_0}{\varepsilon} h. \tag{14.6}$$

D'autre part, nous savons que la solution $u(x)$ de (14.1) est donnée par (14.2) et si nous posons $w_j = u(x_j)$, nous avons

$$w_j = \frac{f_0}{c_0}\left(jh - \frac{1 - \exp(j\gamma)}{1 - \exp\left(\frac{c_0}{\varepsilon}\right)}\right), \qquad j = 0, 1, \ldots, N+1. \tag{14.7}$$

Nous vérifions facilement que

$$\begin{aligned} &2w_j - w_{j-1} - w_{j+1} + \gamma\Big(\alpha(w_j - w_{j-1}) + (1-\alpha)(w_{j+1} - w_j)\Big) \\ &\qquad = \frac{f_0}{c_0}\left(\frac{\Big((1+\gamma\alpha)(2 - e^{-\gamma} - e^{\gamma}) + \gamma(e^{\gamma} - 1)\Big)e^{j\gamma}}{1 - \exp\left(\frac{c_0}{\varepsilon}\right)} + \gamma h\right). \end{aligned}$$

Si nous choisissons α tel que

$$(1 + \gamma\alpha)(2 - e^{-\gamma} - e^{\gamma}) + \gamma(e^{\gamma} - 1) = 0, \tag{14.8}$$

alors nous avons

$$\begin{aligned} &2w_j - w_{j-1} - w_{j+1} + \gamma\Big(\alpha(w_j - w_{j-1}) + (1-\alpha)(w_{j+1} - w_j)\Big) \\ &\qquad = \frac{f_0}{c_0}\gamma h = \frac{h^2 f_0}{\varepsilon}. \end{aligned}$$

Dans ce cas les valeurs $w_1, \ldots, w_N$ satisfont le schéma (14.5) et nous avons

$$u_j = w_j = u(x_j), \qquad j = 1, \ldots, N.$$

Ainsi, lorsque l'égalité (14.8) est satisfaite, alors la solution exacte du problème (14.1) coïncide aux points $x_1 \ldots, x_N$ avec la solution du schéma (14.5). La condition (14.8) s'exprime encore de la manière suivante :

$$\begin{aligned} \alpha &= \frac{1 - e^{\gamma}}{(2 - e^{-\gamma} - e^{\gamma})} - \frac{1}{\gamma} \\ &= \frac{1 - \frac{1}{2}e^{\gamma} - \frac{1}{2}e^{-\gamma} + \frac{1}{2}(e^{-\gamma} - e^{\gamma})}{2 - e^{-\gamma} - e^{\gamma}} - \frac{1}{\gamma} \\ &= \frac{1}{2} + \frac{1}{2}\frac{e^{\gamma} - e^{-\gamma}}{e^{\gamma} + e^{-\gamma} - 2} - \frac{1}{\gamma} \\ &= \frac{1}{2} + \frac{1}{2}\frac{(e^{\gamma/2} + e^{-\gamma/2})(e^{\gamma/2} - e^{-\gamma/2})}{(e^{\gamma/2} - e^{-\gamma/2})^2} - \frac{1}{\gamma}, \end{aligned}$$

et donc

$$\alpha = \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}. \tag{14.9}$$

Dans la figure 14.4, nous avons représenté α comme fonction de γ, pour $\gamma \in \mathbb{R}$. Lorsque γ est proche de zéro, c'est-à-dire lorsque h est petit par rapport à ε/c_0, nous pouvons choisir $\alpha \simeq 0.5$ ce qui, dans (14.5), conduit au schéma centré. Lorsque γ est grand, c'est-à-dire lorsque h est grand par rapport à l'épaisseur de la couche limite ε/c_0, nous avons intérêt à choisir α proche de 1 ce qui, dans (14.5), conduit au schéma décentré arrière (le fait que le schéma soit décentré en arrière provient de l'hypothèse c_0 positif).

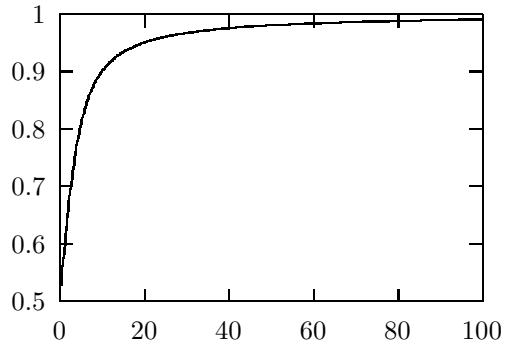

Fig. 14.4 Tracé de la fonction $\gamma \to \alpha = \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}$.

Dans le cas général où c et f ne sont pas constants, nous choisirons donc, si $c(x_j) \neq 0$:

$$\alpha_j = \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma_j}{2} - \frac{1}{\gamma_j} \qquad \text{avec } \gamma_j = \frac{c(x_j)}{\varepsilon}h. \tag{14.10}$$

Le schéma numérique (14.4) dont les coefficients α_j sont donnés par (14.10) sera appelé ***schéma upwind***. Il produit la solution exacte lorsque c et f sont constants.

Remarque 14.1 Un calcul simple nous permet de vérifier que le schéma (14.4) peut se mettre sous la forme suivante :

$$\begin{aligned} &\varepsilon_j^* \frac{2u_j - u_{j+1} - u_{j-1}}{h^2} + c(x_j)\frac{u_{j+1} - u_{j-1}}{2h} \\ &\qquad = f(x_j), \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0, \end{aligned} \tag{14.11}$$

avec

$$\varepsilon_j^* = \varepsilon + c(x_j)h\left(\alpha_j - \frac{1}{2}\right). \tag{14.12}$$

Ainsi le schéma upwind est souvent interprété comme un schéma centré (comparer (14.3) et (14.11)) où le coefficient de diffusion ε a été augmenté de la quantité $c(x_j)h(\alpha_j - 1/2)$, appelée ***diffusion numérique***.

14.2 Un problème de convection-diffusion stationnaire et éléments finis

Approchons maintenant la solution u du problème (14.1) par une méthode d'éléments finis (sect. 10.3 et 10.4). Pour établir le problème faible correspondant au problème (14.1), nous pratiquons comme dans la section 10.3 pour déduire (10.9) de (10.1). Soit V l'ensemble de toutes les fonctions g continues sur l'intervalle $[0,1]$, de premières dérivées g' continues par morceaux et telles que $g(0) = g(1) = 0$. Nous cherchons une fonction $u \in V$ telle que :

$$\begin{aligned} \varepsilon \int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u'(x)v(x)dx \\ = \int_0^1 f(x)v(x)dx \qquad \forall v \in V. \end{aligned} \tag{14.13}$$

Soit N points $x_1, \ldots, x_N$ situés à l'intérieur de l'intervalle $[0,1]$ tels que $x_0 = 0 < x_1 < x_2 < \ldots < x_N < 1 = x_{N+1}$. Considérons les N fonctions $\varphi_1, \ldots, \varphi_N$ définies de la façon suivante :

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & \text{si } x_{j-1} \le x \le x_j, \\ \dfrac{x - x_{j+1}}{x_j - x_{j+1}} & \text{si } x_j \le x \le x_{j+1}, \\ 0 & \text{si } x \notin [x_{j-1}, x_{j+1}]. \end{cases}$$

Les fonctions $\varphi_1, \ldots, \varphi_N$ appartiennent à l'espace V et sont linéairement indépendantes. Soit V_h l'espace engendré par les combinaisons linéaires de $\varphi_1, \ldots, \varphi_N$. Une approximation standard de (14.13) par éléments finis de degré 1 consiste à chercher $u_h \in V_h$ tel que

$$\begin{aligned} \varepsilon \int_0^1 u_h'(x)v_h'(x)dx + \int_0^1 c(x)u_h'(x)v_h(x)dx \\ = \int_0^1 f(x)v_h(x)dx \qquad \forall v_h \in V_h. \end{aligned} \tag{14.14}$$

Exprimons u_h comme combinaison linéaire de $\varphi_1, \ldots, \varphi_N$, c'est-à-dire

$$u_h(x) = \sum_{i=1}^{N} u_i \varphi_i(x) \qquad \forall x \in [0,1], \tag{14.15}$$

et choisissons $v_h = \varphi_j$, $j = 1, 2, \ldots, N$. Le problème (14.14) est donc équivalent à chercher $u_1, u_2, \ldots, u_N$ tels que

$$\begin{aligned} \sum_{i=1}^{N} u_i \int_0^1 \Big(\varepsilon \varphi_i'(x)\varphi_j'(x) + c(x)\varphi_i'(x)\varphi_j(x)\Big)dx \\ = \int_0^1 f(x)\varphi_j(x)dx, \qquad j = 1, \ldots, N. \end{aligned} \tag{14.16}$$

Ici encore, nous obtenons un système linéaire de N équations à N inconnues $u_1, \ldots, u_N$. Soit A la $N \times N$ matrice de coefficients

$$A_{ji} = \int_0^1 \Big(\varepsilon \varphi_i'(x)\varphi_j'(x) + c(x)\varphi_i'(x)\varphi_j(x)\Big)dx, \qquad i,j = 1, \ldots, N,$$

et soit $\vec{f}$ le N-vecteur de composantes

$$f_j = \int_0^1 f(x)\varphi_j(x)dx, \qquad j = 1, \ldots, N.$$

Dans le cas où $c(x) = c_0 =$ constante, $f(x) = f_0 =$ constante et lorsque les points $x_1, \ldots, x_N$ sont uniformément répartis ($h = 1/(N+1)$ et $x_j = jh$, $j = 1, \ldots, N$), nous vérifions facilement que A est la matrice tridiagonale définie par

$$A = \begin{bmatrix} \frac{2\varepsilon}{h} & -\frac{\varepsilon}{h} + \frac{c_0}{2} & & \\ -\frac{\varepsilon}{h} - \frac{c_0}{2} & \ddots & \ddots & \\ & \ddots & \ddots & -\frac{\varepsilon}{h} + \frac{c_0}{2} \\ & & -\frac{\varepsilon}{h} - \frac{c_0}{2} & \frac{2\varepsilon}{h} \end{bmatrix} \tag{14.17}$$

et $f_j = f_0 h$, $j = 1, \ldots, N$. Le système linéaire (14.16) coïncide donc, dans ce cas, avec le schéma centré (14.3). Nous avons vu qu'il convenait de le modifier, surtout si $\varepsilon << c_0$.

Pour ce faire, nous écrivons (14.14) de la manière suivante :

$$\sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} \Big(\varepsilon u_h'(x)v_h'(x) + c(x)u_h'(x)v_h(x)\Big)dx = \sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} f(x)v_h(x)dx \qquad \forall v_h \in V_h. \tag{14.18}$$

Ecrivons la fonction u_h sous la forme (14.15) et choisissons, sur chaque intervalle $]x_k, x_{k+1}[$, des fonctions test autres que celles que nous avons choisies jusqu'à présent, c'est-à-dire :

$$v_h(x) = \varphi_j(x) + \beta_k c(x_{k+1/2})\varphi_j'(x) \qquad \forall x \in]x_k, x_{k+1}[, \tag{14.19}$$

où β_k est un nombre réel positif à déterminer et $x_{k+1/2}$ est le point milieu de $]x_k, x_{k+1}[$. Si nous approchons

$$\int_{x_k}^{x_{k+1}} c(x)u_h'(x)v_h(x)dx \qquad \text{par} \qquad c(x_{k+1/2}) \int_{x_k}^{x_{k+1}} u_h'(x)v_h(x)dx,$$

l'équation(14.18) devient, avec de telles fonctions test :

$$\sum_{i=1}^{N} u_i \sum_{k=0}^{N} \Bigg(\Big(\varepsilon + \beta_k c(x_{k+1/2})^2\Big) \int_{x_k}^{x_{k+1}} \varphi_i'(x)\varphi_j'(x)dx + c(x_{k+1/2}) \int_{x_k}^{x_{k+1}} \varphi_i'(x)\varphi_j(x)dx \Bigg)$$
$$= \sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} f(x)\Big(\varphi_j(x) + \beta_k c(x_{k+1/2})\varphi_j'(x)\Big)dx, \qquad j = 1, \ldots, N. \quad (14.20)$$

Remarquons que, si $\beta_k c(x_{k+1/2}) \neq 0$, la fonction test v_h définie par (14.19) est discontinue aux points x_k, $1 \leq k \leq N$, et par conséquent $v_h \notin V_h$. Cependant, nous allons voir que β_k sera choisi petit de sorte à ce que v_h soit presque dans V_h.

Pour définir les valeurs β_k, $k = 1, \ldots, N$, revenons au cas simple où $c(x) = c_0 =$ constante positive, $f(x) = f_0 =$ constante et $x_j = jh$, $j = 1, \ldots, N$, avec $h = 1/(N+1)$. Dans ce cas $\beta_k = \beta$ sur chaque intervalle et

$$\sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} f(x)\varphi_j'(x)dx = f_0 \int_0^1 \varphi_j'(x)dx$$
$$= f_0(\varphi_j(1) - \varphi_j(0)) = 0.$$

Ainsi (14.20) s'écrit

$$\sum_{i=1}^{N} \left(A_{ji} + \beta c_0^2 B_{ji}\right) u_i = f_0 h \qquad j = 1, \ldots, N, \quad (14.21)$$

où A est la $N \times N$ matrice tridiagonale définie par (14.17) et B est la $N \times N$ matrice tridiagonale définie par

$$B = \frac{1}{h} \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

En fait, le schéma (14.21) est un schéma de différences finies décentrées multiplié par un facteur h. Il correspond à un schéma centré dans lequel nous avons ajouté un terme de diffusion numérique $\beta c_0^2 B\vec{u}$. Suite à la remarque 14.1, il convient de choisir β de sorte à ce que

$$\beta c_0^2 = \left(\alpha - \frac{1}{2}\right) c_0 h,$$

où α est donné par (14.9). Ainsi

$$\beta = \left(\frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}\right)\frac{h}{c_0} \quad (14.22)$$

où $\gamma = \dfrac{c_0}{\varepsilon} h$. Un développement limité autour de zéro nous assure que

$$\frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma} = \frac{\gamma}{12} + O(\gamma^3).$$

D'autre part, lorsque γ tend vers l'infini, nous avons

$$\lim_{\gamma \longrightarrow \infty} \left(\frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma} \right) = \frac{1}{2}.$$

Par conséquent, au lieu d'utiliser la formule (14.22), nous utiliserons une formule plus simple, à savoir

$$\begin{aligned} \beta &= \frac{\gamma}{12} \cdot \frac{h}{c_0} \quad \text{si } \gamma < 6, \\ \beta &= \frac{1}{2} \cdot \frac{h}{c_0} \quad \text{si } \gamma \geq 6. \end{aligned}$$

Considérons à nouveau les cas où les fonctions c et f ne sont pas constantes et les points x_j non nécessairement équidistribués. Nous posons

$$\gamma_k = |c(x_{k+1/2})| \frac{h_k}{\varepsilon}, \qquad \text{avec } h_k = x_{k+1} - x_k,$$

et γ_k est appelé ***nombre de Péclet de la maille*** k. Nous utiliserons la règle suivante pour définir les valeurs β_k intervenant dans (14.20) :

$$\begin{aligned} \beta_k &= \frac{\gamma_k}{12} \cdot \frac{h_k}{|c(x_{k+1/2})|} \quad \text{si } \gamma_k < 6, \\ \beta_k &= \frac{1}{2} \cdot \frac{h_k}{|c(x_{k+1/2})|} \quad \text{si } \gamma_k \geq 6. \end{aligned} \tag{14.23}$$

Le schéma (14.20) avec les valeurs de β_k définies par (14.23) est appelé dans la littérature ***schéma SUPG*** (Streamline Upwind Petrov-Galerkin). Contrairement au schéma (14.16) (qui coïncide avec (14.20) lorsque $\beta_k = 0$ pour tous les k), ce schéma produit une solution non oscillante (ou presque) lorsque le nombre de Péclet γ_k de la maille k est grand. De plus, l'ordre de convergence du schéma SUPG est le même que l'ordre de convergence du schéma (14.16), lorsque $h = \max_{1 \leq k \leq N} h_k$ tend vers zéro.

Remarque 14.2 Les éléments géométriques $[x_k, x_{k+1}]$ étant fixés une fois pour toutes, le schéma SUPG introduit une diffusion numérique adaptée à chacun d'eux : plus l'élément $[x_k, x_{k+1}]$ est petit (h_k petit), plus son nombre de Peclet γ_k est petit et moins nous introduisons de diffusion dans (14.20) (β_k petit).

14.3 Problèmes bidimensionnels de convection-diffusion

Soit Ω un domaine polygonal dans le plan Ox_1x_2, de frontière $\partial\Omega$ et soit $\overline{\Omega} = \Omega \cup \partial\Omega$. Nous nous donnons une fonction $f : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \longrightarrow f(x,t) \in \mathbb{R}$,

une fonction vectorielle $\vec{c}\,: (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to \vec{c}(x,t) \in \mathbb{R}^2$, un nombre positif ε et une fonction $w : x \in \overline{\Omega} \to w(x) \in \mathbb{R}$. Dès lors, nous posons le problème de chercher une fonction $u : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ telle que

$$\begin{aligned} &\frac{\partial u}{\partial t}(x,t) - \varepsilon \Delta u(x,t) + \vec{c}(x,t) \cdot \overrightarrow{\text{grad}} u(x,t) \\ &\qquad\qquad = f(x,t) \qquad\qquad \forall x \in \Omega, \quad \forall t > 0, \\ &u(x,t) = 0 \qquad\qquad \forall x \in \partial\Omega \quad \forall t > 0, \\ &u(x,0) = w(x) \qquad\qquad \forall x \in \Omega, \end{aligned} \tag{14.24}$$

où Δ est l'opérateur laplacien. Le problème (14.24) modélise par exemple un problème de propagation de la chaleur dans un fluide contenu dans le domaine Ω. La grandeur $u(x,t)$ représente alors la température du fluide au point x et à l'instant t, ε sa conductibilité thermique, $\vec{c}(x,t)$ sa vitesse et $f(x,t)$ la source de puissance par unité de surface au point x et à l'instant t. La diffusion de la chaleur est modélisée par le terme $-\varepsilon\Delta u$ alors que la convection l'est par le terme $\vec{c} \cdot \overrightarrow{\text{grad}} u$. Dans le cas où $c \equiv 0$, le problème (14.24) coïncide avec le problème parabolique traité dans la section 12.3.

Pour obtenir une approximation par éléments finis de (14.24) en utilisant un schéma SUPG, il suffit de pratiquer comme dans la section 12.3, mais en modifiant les fonctions test comme nous l'avons fait dans la section précédente. Soit donc $\mathcal{T}_h$ une triangulation de $\overline{\Omega}$ en triangles $K \in \mathcal{T}_h$. Soit $\varphi_1, \ldots, \varphi_N$ les fonctions de base définies par (11.15), affines sur chaque triangle $K \in \mathcal{T}_h$, valant 1 en un des nœuds intérieurs de la triangulation et zéro aux autres nœuds. Une discrétisation standard (sect. 12.3) en espace de (14.24) par la méthode des éléments finis consiste à décomposer la solution approchée u_h dans la base $\varphi_1, \ldots, \varphi_N$

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x),$$

et à choisir $\varphi_1, \ldots, \varphi_N$ comme fonctions test dans la formulation faible correspondante. Le problème se ramène donc à chercher les fonctions $u_1(t), \ldots, u_N(t)$ satisfaisant des conditions initiales ainsi que

$$\begin{aligned} &\sum_{i=1}^{N} \dot{u}_i(t) \iint_\Omega \varphi_i(x)\varphi_j(x)dx + \varepsilon \sum_{i=1}^{N} u_i(t) \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx \\ &\quad + \sum_{i=1}^{N} u_i(t) \iint_\Omega \Big(\vec{c}(x,t) \cdot \overrightarrow{\text{grad}}\varphi_i(x)\Big)\varphi_j(x)dx = \iint_\Omega f(x,t)\varphi_j(x)dx, \end{aligned} \tag{14.25}$$

pour $j = 1, \ldots, N$ (sur ce point, il suffit de procéder de façon identique à ce qui a été fait dans la section 12.3). Opérons maintenant comme dans la section précédente et remplaçons, sur chaque triangle K de la triangulation, la fonction test φ_j par

$$\varphi_j + \beta_K \vec{c}(Q_K) \cdot \overrightarrow{\text{grad}}\varphi_j,$$

où Q_K est le centre de gravité du triangle K et où β_K est un nombre positif à définir. Nous obtenons ainsi la semi-discrétisation spatiale SUPG suivante :

$$\sum_{i=1}^{N} \dot{u}_i(t) \sum_{K\in\mathcal{T}_h} \iint_K \varphi_i(x) \left(\varphi_j(x) + \beta_K \vec{c}(Q_K)\cdot\overrightarrow{\text{grad}}\varphi_j(x)\right) dx$$
$$+ \sum_{i=1}^{N} u_i(t) \sum_{K\in\mathcal{T}_h} \left(\varepsilon \iint_K \overrightarrow{\text{grad}}\varphi_i(x)\cdot\overrightarrow{\text{grad}}\varphi_j(x)dx \right.$$
$$\left. + \iint_K \vec{c}(Q_K)\cdot\overrightarrow{\text{grad}}\varphi_i(x)\left(\varphi_j(x) + \beta_K\vec{c}(Q_K)\cdot\overrightarrow{\text{grad}}\varphi_j(x)\right) dx \right)$$
$$= \sum_{K\in\mathcal{T}_h} \iint_K f(x,t)\left(\varphi_j(x) + \beta_K\vec{c}(Q_K)\cdot\overrightarrow{\text{grad}}\varphi_j(x)\right) dx, \quad (14.26)$$

pour $j = 1,\ldots,N$. Pour tout triangle K de la triangulation, notons h_K le diamètre de K et soit

$$\gamma_K = |\vec{c}(Q_K)|\frac{h_K}{\varepsilon}$$

le ***nombre de Péclet de la maille*** K. De façon similaire à ce que nous avons fait dans la section précédente, nous définissons les nombres β_K dans (14.26) par

$$\begin{aligned} \beta_K &= \frac{\gamma_K}{12}\cdot\frac{h_K}{|\vec{c}(Q_K)|} \quad \text{si } \gamma_K < 6, \\ \beta_K &= \frac{1}{2}\frac{h_K}{|\vec{c}(Q_K)|} \quad \text{si } \gamma_K \geq 6. \end{aligned} \quad (14.27)$$

Il reste ensuite à discrétiser (14.26) par rapport à la variable t en utilisant un schéma d'Euler rétrograde ou Crank-Nicholson, comme nous l'avons fait dans le chapitre 12. Ainsi un schéma SUPG pour la résolution numérique du problème de convection-diffusion (14.24) sera donné par une discrétisation temporelle standard (Euler rétrograde ou Crank-Nicholson par exemple) du système différentiel (14.26).

14.4 Exercices

Exercice 14.1 Soit $\varepsilon > 0$ et c_0 deux nombres réels et soit $w : x \in [0,1] \to w(x) \in \mathbb{R}$ une fonction donnée. Nous cherchons une fonction $u : (x,t) \in [0,1]\times\mathbb{R}^+ \to u(x,t) \in \mathbb{R}$ satisfaisant

$$\frac{\partial u}{\partial t}(x,t) - \varepsilon\frac{\partial^2 u}{\partial x^2}(x,t) + c_0\frac{\partial u}{\partial x}(x,t) = 0 \qquad \forall x \in]0,1[, \quad \forall t > 0, \quad (14.28)$$
$$u(0,t) = u(1,t) = 0 \qquad \forall t > 0, \quad (14.29)$$
$$u(x,0) = w(x) \qquad \forall x \in]0,1[. \quad (14.30)$$

Le problème ci-dessus est un problème de ***convection-diffusion évolutif***. Lorsque $\varepsilon = 1$ et $c_0 = 0$, ce problème coïncide avec le problème de la chaleur

étudié dans la section 12.1. Lorsque $\varepsilon = 0$, ce problème coïncide avec le problème de transport étudié dans la section 13.1.

Pour résoudre numériquement ce problème, nous utilisons une méthode de différences finies. Si N est un entier positif, nous notons $h = 1/(N+1)$, $x_i = ih$, $i = 0, 1, 2, \ldots, N+1$. Si τ est un nombre réel positif donné, nous notons $t_n = n\tau$, $n = 0, 1, 2, \ldots$ Pour n fixé, nous notons u_i^n l'approximation de $u(x_i, t_n)$, calculée à partir du schéma suivant :

$$\begin{aligned} &\frac{u_i^{n+1} - u_i^n}{\tau} + \varepsilon \frac{-u_{i-1}^n + 2u_i^n - u_{i+1}^n}{h^2} + c_0 \frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0, \qquad 1 \leq i \leq N, \\ &u_0^n = u_{N+1}^n = 0, \\ &u_i^0 = w(x_i), \qquad 1 \leq i \leq N. \end{aligned} \tag{14.31}$$

Noter que le schéma ci-dessus est explicite et que le terme de convection est centré.

1. On fixe l'entier n et on suppose que

$$\tau \leq \frac{h^2}{2\varepsilon} \qquad \text{et} \qquad h \leq \frac{2\varepsilon}{|c_0|}. \tag{14.32}$$

Montrer que si tous les u_i^n, $i = 1, 2, \ldots, N$ sont positifs ou nuls, alors tous les u_i^{n+1}, $i = 1, 2, \ldots, N$ sont aussi positifs ou nuls. Montrer que

$$\max_{1 \leq i \leq N} |u_i^{n+1}| \leq \max_{1 \leq i \leq N} |u_i^n|. \tag{14.33}$$

La relation (14.33) garantit que le schéma (14.31) est stable sous les conditions (14.32).

2. On considère le schéma (14.31) dans lequel nous prenons $w(x) = 1$, si $0 < x < 1$, $N = 4$ (et donc $h = 1/5$) et $\tau = 1/50$. Calculer et représenter $(u_i^n)_{i=1}^4$ pour $n = 1, 2, 3, 4, 5$ dans les cas suivants :
   (a) $\varepsilon = 1$, $c_0 = 10$,
   (b) $\varepsilon = 1$, $c_0 = 30$
   (c) $\varepsilon = 5$, $c_0 = 30$.

   Comment conclure ?

Solution

1. Le schéma numérique (14.31) s'écrit

$$u_i^{n+1} = \left(\varepsilon \frac{\tau}{h^2} + c_0 \frac{\tau}{2h}\right) u_{i-1}^n + \left(1 - 2\varepsilon \frac{\tau}{h^2}\right) u_i^n + \left(\varepsilon \frac{\tau}{h^2} - c_0 \frac{\tau}{2h}\right) u_{i+1}^n. \tag{14.34}$$

Par hypothèse

$$1 - 2\varepsilon \frac{\tau}{h^2} \geq 0 \qquad \text{et} \qquad \varepsilon \frac{\tau}{h^2} \pm c_0 \frac{\tau}{2h} \geq 0$$

et en vertu de la relation (14.34), il est clair que si les quantités u_{i-1}^n, u_i^n et u_{i+1}^n sont positives ou nulles, alors la quantité u_i^{n+1} est aussi positive ou nulle. Nous avons bien montré le résultat énoncé.

Montrons maintenant (14.33). Soit $1 \leq i \leq N$ fixé. Par hypothèse nous avons

$$1 - 2\varepsilon\frac{\tau}{h^2} \geq 0 \qquad \text{ainsi que} \qquad \varepsilon\frac{\tau}{h^2} \pm c_0\frac{\tau}{2h} \geq 0.$$

L'égalité (14.34) implique ainsi :

$$\begin{aligned}
|u_i^{n+1}| &\leq \left(\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{2h}\right)|u_{i-1}^n| + \left(1 - 2\varepsilon\frac{\tau}{h^2}\right)|u_i^n| + \left(\varepsilon\frac{\tau}{h^2} - c_0\frac{\tau}{2h}\right)|u_{i+1}^n| \\
&\leq \left(\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{2h}\right)\max_{1\leq j\leq N}|u_j^n| \\
&\quad + \left(1 - 2\varepsilon\frac{\tau}{h^2}\right)\max_{1\leq j\leq N}|u_j^n| + \left(\varepsilon\frac{\tau}{h^2} - c_0\frac{\tau}{2h}\right)\max_{1\leq j\leq N}|u_j^n| \\
&= \max_{1\leq j\leq N}|u_j^n|.
\end{aligned}$$

Il suffit ensuite de prendre le maximum sur tous les indices i pour obtenir le résultat.

2. Dans le cas (a) nous avons $\varepsilon = 1$, $c_0 = 10$, $\varepsilon\tau/h^2 = 1/2$, $c_0\tau/(2h) = 1/2$ et le schéma (14.31) s'écrit simplement

$$u_i^{n+1} = u_{i-1}^n.$$

Les résultats numériques pour $n = 1, 2, 3, 4, 5$ sont présentés dans la figure 14.5. Puisques les conditions (14.32) sont satisfaites, nous savons que le schéma est stable. Cependant, il produit des erreurs assez grandes car il conduit très rapidement à la solution triviale. Pour obtenir des résultats plus précis, il conviendrait de choisir h et τ plus petits !

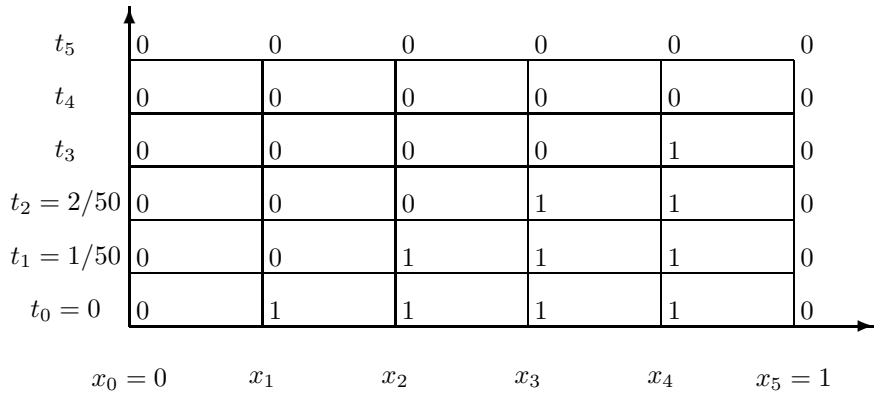

Fig. 14.5 Solution numérique du schéma (14.31) lorsque $\varepsilon = 1$, $c_0 = 10$, $h = 1/5$ et $\tau = 1/50$.

Dans le cas (b) nous avons $\varepsilon = 1$, $c_0 = 30$, $\varepsilon\tau/h^2 = 1/2$, $c_0\tau/(2h) = 3/2$ et le schéma (14.31) s'écrit

$$u_i^{n+1} = 2u_{i-1}^n - u_{i+1}^n.$$

Les résultats numériques pour $n = 1, 2, 3, 4, 5$ sont présentés dans la figure 14.6. La condition $\tau \leq h^2/(2\varepsilon)$ est toujours satisfaite, par contre la condition $h \leq 2\varepsilon/|c_0|$ ne l'est plus. Nous constatons que le schéma engendre des oscillations liées au fait que nous avons choisi une approximation centrée pour l'approximation du terme $\partial u/\partial x$.

t_5	0	−14	8	44	−8	0
t_4	0	2	14	−4	−16	0
t_3	0	3	−2	−8	0	0
$t_2 = 2/50$	0	−1	−3	0	2	0
$t_1 = 1/50$	0	−1	1	1	2	0
$t_0 = 0$	0	1	1	1	1	0
	$x_0 = 0$	x_1	x_2	x_3	x_4	$x_5 = 1$

Fig. 14.6 Solution numérique du schéma (14.31) lorsque $\varepsilon = 1$, $c_0 = 30$, $h = 1/5$ et $\tau = 1/50$.

Finalement dans le cas (c) nous avons $\varepsilon = 5$, $c_0 = 30$, $\varepsilon\tau/h^2 = 5/2$, $c_0\tau/(2h) = 3/2$ et le schéma (14.31) s'écrit

$$u_i^{n+1} = 4u_{i-1}^n - 4u_i^n + u_{i+1}^n.$$

Les résultats numériques pour $n = 1, 2, 3, 4, 5$ sont présentés dans la figure 14.7. La condition $\tau \leq h^2/(2\varepsilon)$ n'est plus satisfaite alors que la condition $h \leq 2\varepsilon/|c_0|$ l'est à nouveau. Nous constatons que le schéma engendre des oscillations considérablement plus grandes que dans le cas précédent ($\varepsilon = 1$). Contrairement au cas précédent, les instabilités numériques ne sont pas dues au fait que nous avons choisi une différence centrée pour l'approximation du terme $\partial u/\partial x$, mais au fait que notre schéma n'est pas stable pour la résolution numérique du problème de la chaleur (sect. 12.1).

Exercice 14.2 Nous considérons à nouveau le problème de convection-diffusion (14.28) (14.29) (14.30) dans lequel nous supposons que c_0 est positif. Pour résoudre numériquement ce problème nous considérons, au lieu du schéma (14.31), le schéma suivant, appelé ***schéma implicite décentré*** :

$$\begin{aligned}
&\frac{u_i^{n+1} - u_i^n}{\tau} + \varepsilon \frac{-u_{i-1}^{n+1} + 2u_i^{n+1} - u_{i+1}^{n+1}}{h^2} + c_0 \frac{u_i^{n+1} - u_{i-1}^{n+1}}{h} = 0, \qquad 1 \leq i \leq N, \\
&u_0^{n+1} = u_{N+1}^{n+1} = 0, \\
&u_i^0 = w(x_i), \qquad 1 \leq i \leq N.
\end{aligned} \tag{14.35}$$

t_5	0	-2292	5264	-5872	3264	0
t_4	0	380	-772	656	-160	0
t_3	0	-67	112	-56	-16	0
$t_2 = 2/50$	0	13	-15	0	4	0
$t_1 = 1/50$	0	-3	1	1	0	0
$t_0 = 0$	0	1	1	1	1	0
	$x_0 = 0$	x_1	x_2	x_3	x_4	$x_5 = 1$

Fig. 14.7 Solution numérique du schéma (14.31) lorsque $\varepsilon = 5$, $c_0 = 30$, $h = 1/5$ et $\tau = 1/50$.

1. Montrer que si tous les u_i^n, $i = 1, 2, \ldots, N$ sont positifs ou nuls, alors tous les u_i^{n+1}, $i = 1, 2, \ldots, N$ sont aussi positifs ou nuls.
2. Montrer que
$$\max_{0 \le i \le N+1} |u_i^{n+1}| \le \max_{0 \le i \le N+1} |u_i^n|. \tag{14.36}$$
La relation (14.36) garantit que le schéma (14.35) est stable quelles que soient les valeurs de h et τ. On dit dans ce cas que le schéma (14.35) est ***inconditionnellement stable***.

Solution

1. le schéma numérique (14.35) s'écrit
$$\left(1 + 2\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_i^{n+1} - \left(\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_{i-1}^{n+1} - \varepsilon\frac{\tau}{h^2} u_{i+1}^{n+1} = u_i^n. \tag{14.37}$$
Pour n fixé, notons k un entier tel que
$$u_k^{n+1} \le u_j^{n+1} \qquad j = 0, 1, 2, \ldots, N+1. \tag{14.38}$$
Supposons pour commencer que k soit différent de 0 et que k soit différent de $N+1$. En prenant $i = k$ dans (14.37), nous obtenons
$$\left(1 + 2\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_k^{n+1} = \left(\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_{k-1}^{n+1} + \varepsilon\frac{\tau}{h^2} u_{k+1}^{n+1} + u_k^n.$$
Puisque $u_{k-1}^{n+1} \ge u_k^{n+1}$ et $u_{k+1}^{n+1} \ge u_k^{n+1}$, nous avons donc
$$\left(1 + 2\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_k^{n+1} \ge \left(\varepsilon\frac{\tau}{h^2} + c_0\frac{\tau}{h}\right) u_k^{n+1} + \varepsilon\frac{\tau}{h^2} u_k^{n+1} + u_k^n,$$
et par suite
$$u_k^{n+1} \ge u_k^n. \tag{14.39}$$
Si $k = 0$ ou $k = N+1$, alors la relation (14.39) est trivialement vraie. Si nous supposons que tous les u_j^n sont positifs ou nuls, nous obtenons donc, en utilisant (14.38) et (14.39) :
$$0 \le u_k^{n+1} \le u_j^{n+1} \qquad j = 0, 1, 2, \ldots, N+1,$$

ce qui prouve que tous les u_j^{n+1} sont aussi positifs ou nuls.

2. La relation (14.39) étant vraie même lorsque $k = 0$ ou $k = N + 1$, nous en déduisons que

$$\min_{0\le j\le N+1} u_j^{n+1} = u_k^{n+1} \ge u_k^n \ge \min_{0\le j\le N+1} u_j^n. \tag{14.40}$$

Si nous avions choisi un entier k tel que

$$u_k^{n+1} \ge u_j^{n+1} \qquad j = 0, 1, 2, \ldots, N+1,$$

nous aurions obtenu $u_k^{n+1} \le u_k^n$ et par suite

$$\max_{0\le j\le N+1} u_j^{n+1} \le \max_{0\le j\le N+1} u_j^n. \tag{14.41}$$

Les inégalités (14.40) et (14.41) sont appelées ***principe du minimum et du maximum discrets***. Elles impliquent nécessairement l'inégalité (14.36).

Nous avons donc montré que le schéma (14.35) est inconditionnellement stable. Par contre, ce schéma est implicite et il nécessite à chaque pas de temps la résolution du système linéaire

$$A\vec{u}^{n+1} = \vec{u}^n.$$

Ici $\vec{u}^{n+1}$ est le N-vecteur de composantes u_i^{n+1}, $1 \le i \le N$, $\vec{u}^n$ est le N-vecteur de composantes u_i^n, $1 \le i \le N$ et A est la $N \times N$ matrice tridiagonale définie par

$$A = \begin{bmatrix} 1 + \dfrac{2\varepsilon\tau}{h^2} + \dfrac{c_0\tau}{h} & -\dfrac{\varepsilon\tau}{h^2} & & \\ -\dfrac{\varepsilon\tau}{h^2} - \dfrac{c_0\tau}{h} & \ddots & \ddots & \\ & \ddots & \ddots & -\dfrac{\varepsilon\tau}{h^2} \\ & & -\dfrac{\varepsilon\tau}{h^2} - \dfrac{c_0\tau}{h} & 1 + \dfrac{2\varepsilon\tau}{h^2} + \dfrac{c_0\tau}{h} \end{bmatrix}.$$

14.5 Notes bibliographiques et remarques

Nous avons présenté dans ce chapitre une méthode d'éléments finis continus permettant de résoudre un problème de convection-diffusion, en l'occurence le schéma SUPG [8]. Il existe d'autres méthodes du même type, à savoir les méthodes GLS (Galerkin Least Squares) ainsi que la stabilisation par bulles (*bubbles* en anglais), voir par exemple [22, 24]. Toutes ces méthodes ont pour but d'éliminer les oscillations numériques lorsque le maillage du domaine de

calcul est choisi une fois pour toutes et lorsque le nombre de Peclet devient grand dans certaines mailles. Une méthode moderne pour effectuer ce genre de calculs consiste à placer, dans le cadre d'un processus itératif, des mailles fines aux endroits de forte variation de la solution (par exemple aux endroits où se trouvent les couches limites). Une telle méthode est appelée méthode ***adaptative*** ; le critère permettant de détecter les endroits où il faut placer des mailles fines est basé sur des ***estimations d'erreurs a posteriori***, voir par exemple [8].

Les équations de Navier-Stokes modélisent l'écoulement d'un fluide visqueux, newtonien, incompressible et sont un exemple concret de problème de convection-diffusion non linéaire. Les méthodes que nous avons décrites dans ce chapitre s'appliquent à ces équations, voir par exemple [24].

Il existe d'autres méthodes pour la résolution numérique des problèmes de convection-diffusion. Citons par exemple la méthode des volumes finis, très utilisée par les ingénieurs, voir par exemple [22]. Mentionnons enfin l'existence de méthodes d'origine probabiliste, particulièrement efficaces lorsque la dimension de l'espace est grande, ce qui est le cas lorsqu'on aborde des modèles mathématiques issus de la finance [18].

Bibliographie

[1] E. L. Allgower and K. Georg. *Numerical Continuation Methods.* Springer-Verlag, Berlin, 1990.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK User's Guide.* SIAM, 3600 University City Science Center, Philadelphia, Pennsylvania 19104-2688, 1992.

[3] Ph. Ciarlet. Basic error estimates for elliptic problems. In Ph. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis*, volume 2, pages 17–352. North-Holland, Amsterdam, 1991.

[4] M. Crouzeix and F. Mignot. *Analyse numérique des équations différentielles.* Masson, Paris, 1992.

[5] A. Curnier. *Méthodes numériques en mécanique des solides.* Presses polytechniques et universitaires romandes, Lausanne, 1993.

[6] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration, 2nd ed.* Academic Press, Orlando, 1984.

[7] D. de Werra. *Eléments de programmation linéaire avec application aux graphes.* Presses polytechniques et universitaires romandes, Lausanne, 1990.

[8] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations.* Cambridge University Press, Cambridge, U. K., 1996.

[9] P.-L. Georges. *Génération automatique de maillages, RMA 16.* Masson, Paris, 1991.

[10] E. Godlewski and P.-A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws.* Springer, Berlin, 1996.

[11] G. Golub and J. M. Ortega. *Scientific Computing - an Introduction with Parallel Computing.* Academic Press, San Diego, CA, USA, 1993.

[12] E. Hairer, S. P. Norsett, and G. Wanner. *Solving Ordinary Differential Equations.* Springer-Verlag, Berlin, 1993.

[13] E. Hairer and G. Wanner. *Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin, 1996.

[14] C. Hirsch. *Computational Methods for Inviscid and Viscous Flows.* Wiley, Chichester, 1990.

[15] T.J.R. Hughes. *The Finite Element Method.* Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[16] E. Isaacson and H. B. Keller. *Analysis of Numerical Methods.* John Wiley and Sons, New York, 1966.

[17] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations.* SIAM, Philadelphia, 1995.

[18] D. Lamberton and B. Lapeyre. *Introduction to Stochastic Calculus Applied to Finance.* Chapman & Hall, London, 1996.

[19] P. Lascaux and J. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur, 2 tomes,* 2^e *édition.* Masson, Paris, 1994.

[20] B. Lucquin and O. Pironneau. *Introduction au calcul scientifique.* Masson, Paris, 1996.

[21] Y. Meyer. *Wavelets : Algorithms and Applications.* SIAM, Philadelphia, 1993.

[22] K.W. Morton. *Numerical Solution of Convection-Diffusion Problems.* Chapman & Hall, London, 1996.

[23] J. M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[24] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations.* Springer, Berlin, 1994.

[25] A. Ralston and P. Rabinowitz. *A First Course in Numerical Analysis, 2nd ed.* McGraw-Hill, New York, 1978.

[26] J.-J. Risler. *Méthodes mathématiques pour la CAO, RMA 18.* Masson, Paris, 1991.

[27] Y. Saad. *Numerical Methods for Large Eigenvalue Problems.* Manchester University Press, Manchester, 1992.

[28] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis, 2nd ed.* Springer-Verlag, Berlin, 1993.

[29] M. Vetterli and Jelena Kovacevic. *Wavelets and Subband Coding.* Prentice Hall, New Jersey, 1995.

Index