EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

DÉPARTEMENT DE MATHEMATIQUES
DMA-ECUBLENS, CH-1015 LAUSANNE
TÉLÉPHONE: 021 - [illegible]3.25.55 TÉLÉFAX: 021 - 693.43.03

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Prof. Jacques Rappaz

(introduction)

ANALYSE NUMERIQUE

Notes de cours : Leçons 1 à 10

déstiné aux étudiants :

Génie civil	2ème année
Génie rural	2ème année
Mécanique	2ème année
Physique UNIL	2ème année
Physique EPFL	1ère année

Lausanne, juin 198[illegible]4
Nouveau tira[illegible]ge

EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE LAUSANNE
POLITECNICO FEDERALE DI LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY LAUSANNE

DÉPARTEMENT DE MATHEMATIQUES
DMA-ECUBLENS, CH-1015 LAUSANNE
TELEPHONE: 021 - 693.25.55 TÉLÉFAX: 021 - 693.43.03

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Prof. Jacques Rappaz

ANALYSE NUMERIQUE

Notes de cours : Leçons 1 à 10

déstiné aux étudiants :

Génie civil	2ème année
Génie rural	2ème année
Mécanique	2ème année
Physique UNIL	2ème année
Physique EPFL	1ère année

Lausanne, juin 1994
Nouveau tirage

Avant-propos

La plupart des phénomènes physiques intervenant en technologie sont régis par des équations différentielles ou aux dérivées partielles dont la résolution numérique au moyen d'un ordinateur nécessite des connaissances en mathématiques plus ou moins importantes suivant l'approche adoptée par l'ingénieur. Ce cours de premier cycle donné aux étudiants des départements de Génie civil, Génie rural, Mécanique, Physique de l'EPFL et aux étudiants en physique de l'UNIL a pour but de fournir quelques notions de mathématiques qui sont à la base de méthodes utilisées dans le domaine de la simulation numérique pour résoudre des problèmes de technologie. Dispensé au semestre d'été sous forme de dix leçons de deux heures, ce cours n'a pas la prétention d'être un exposé exhaustif des problèmes de l'analyse numérique, ni même d'être une introduction complète à ce domaine ; il entend apporter quelques rudiments qui permettront à l'étudiant de mieux aborder et comprendre certaines difficultés numériques qu'il pourrait rencontrer dans ses activités futures.

Ces notes de cours ne sont rien d'autre que la rédaction du cours dispensé au tableau noir ; elles ne contiennent pas les exercices qui les illustrent en permettant à l'étudiant de mieux les assimiler et qui sont dispensés à raison d'une heure par leçon. Des questions de temps à disposition ne permettent pas de donner toutes les démonstrations des affirmations énoncées ; cependant, nous nous sommes efforcés de maintenir une certaine rigueur mathématique dans la manière de formuler les résultats. Les étudiants qui désirent approfondir leurs connaissances dans ce domaine pourront consulter les livres recommandés dans ce cours.

L'auteur de ces notes tient à remercier M^{me} J. Mosetti pour le soin apporté à la dactylographie du manuscrit ainsi que M. L. Gasser pour la confection des figures.

Livres recommandés

J. Baranger : Introduction à l'analyse numérique. Hermann, Paris 1977.

A. Ralston, Ph. Rabinowitz : A first course in numerical analysis. International Student Edition, McGraw-Hill 1984.

F. Scheid : Analyse numérique : cours et problèmes. Série Schaum, McGraw-Hill, Paris 1986, (édition originale : McGraw-Hill Inc., New York).

Table des matières

Leçon 1

Problèmes d'interpolation

1.1 Position du problème

Supposons vouloir chercher un polynôme p de degré $n \geq 0$ qui, pour des valeurs $t_0 < t_1 < t_2 < \cdots < t_n$ données, prenne les valeurs $p_0, p_1, p_2, \cdots, p_n$ respectivement, c'est-à-dire

$$p(t_j) = p_j \quad \text{pour } j = 0,1,2,\cdots,n. \tag{1.1}$$

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 $a_0, a_1, a_2, \cdots, a_n$ 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); on écrit ensuite les $(n+1)$ relations (1.1) comme il suit :

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

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 $a_0, a_1, a_2, \cdots, a_n$. Une manière différente d'écrire (1.3) est la suivante.

Soit T la $(n+1) \times (n+1)$-matrice dite de Vandermonde suivante :

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

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

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

on peut écrire (1.3) sous forme matricielle :

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

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

La résolution d'un système linéaire de $(n+1)$ équations à $(n+1)$ inconnues n'est pas triviale (cf. Leçon 4) et sans nul doute que la méthode que nous venons de décrire pour trouver p n'est pas une "bonne" méthode. La suite nous montrera une voie plus astucieuse pour construire p.

1.2 Base de Lagrange

Il est facile de résoudre le problème (1.1) lorsqu'on a choisi tous les p_j égaux à zéro sauf un que l'on fixe à 1. Supposons donc que pour k donné entre zéro et n on ait

$$p_k = 1 \qquad 0 \le k \le n$$

et pour $j \neq k$ on ait $p_j = 0$. Soit φ_k la fonction de t donnée par

$$\varphi_k(t) = \frac{(t-t_0)(t-t_1)(t-t_2)\cdots(t-t_{k-1})(t-t_{k+1})\cdots(t-t_n)}{(t_k-t_0)(t_k-t_1)(t_k-t_2)\cdots(t_k-t_{k-1})(t_k-t_{k+1})\cdots(t_k-t_n)} \cdot \begin{matrix} \in \mathbb{P}_n \\ \in \mathbb{K} \end{matrix} \tag{1.5}$$

$$\varphi_k(t) = \frac{\prod_{j=0, j\neq k}^{n} (t-t_j)}{\prod_{j=0, j\neq k} (t_k-t_j)}$$

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

(i) φ_k est un polynôme de degré n,

(ii) $\varphi_k(t_j) = 0$ si et $j \neq k$ $j = 0,1,2,\cdots,n$,

(iii) $\varphi_k(t_k) = 1$.

$\Big\}$ $\varphi_k(t_j) = \delta_{kj}$

En fait, au point t_k, on a associé un polynôme φ_k de degré n qui vaut 1 en t_k et 0 dans les autres points $t_j, j \neq k$. Les polynômes $\varphi_0, \varphi_1, \varphi_2, \cdots, \varphi_n$ sont linéairement indépendants, ils forment une base pour les polynômes de degré n appelée

"base de Lagrange" associée aux points t_j, $0 \leq j \leq n$.

Exemple 1.1. Prenons $n = 2, t_0 = -1, t_1 = 0, t_2 = 1$. Les polynômes $\varphi_0, \varphi_1, \varphi_2$ de degré 2 associés aux points t_0, t_1, t_2 respectivement seront donnés 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 et φ_2 sont représentés dans la figure 1.1.

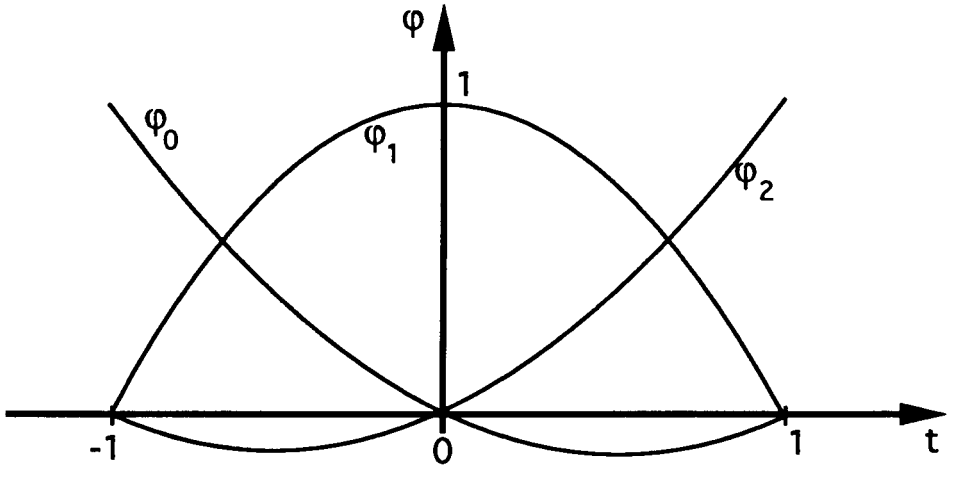


Figure 1.1 : Base de Lagrange des polynômes de degré 2

Les polynômes $\varphi_0, \varphi_1, \varphi_2$ donnés dans (1.6), (1.7) et (1.8) forment une base de Lagrange des polynômes de degré 2 associée aux points –1, 0 et 1.

1.3 Interpolation de Lagrange

Revenons au problème (1.1) où on cherche un polynôme p de degré n qui prend des valeurs données $p_0, p_1, p_2, \cdots, p_n$ en des points donnés $t_0 < t_1 < t_2 < \cdots < t_n$.

Soit $\varphi_0, \varphi_1, \varphi_2, \cdots, \varphi_n$ la base de Lagrange des polynômes de degré n associée aux points $t_j, 0 \le j \le n$. Alors la solution cherchée est donnée par

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

En effet, p étant une combinaison linéaire de $(n+1)$ polynômes $\varphi_0, \varphi_1, \cdots, \varphi_n$ de degré n, p sera lui-même un polynôme de degré n. D'autre part, si on utilise les propriétés des polynômes φ_j, on a pour $k = 0,1,2,\cdots,n$:

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

polynôme de t_k

qui est bien la relation (1.1).

Remarquons en passant que nous avons construit explicitement une solution au problème (1.1) et ceci pour n'importe quelles valeurs $p_0, p_1, p_2, \cdots, p_n$ données . Ceci montre que le problème (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) donnée au 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 $p_2 = 6$?

$t_0 = -1$, $p(t_0) = 8$
$t_1 = 0$, $p(t_1) = 3$
$t_2 = 1$, $p(t_2) = 6$
$\Big\}\ n = 2$

D'après ce qui précède, on aura $p(t) = 8\,\varphi_0(t) + 3\,\varphi_1(t) + 6\,\varphi_2(t)$ où φ_0, φ_1 et φ_2 sont donnés par (1.6), (1.7) et (1.8).

Ainsi donc on obtient

$$p(t) = 8(\tfrac{1}{2}t^2 - \tfrac{1}{2}t) + 3(1 - t^2) + 6(\tfrac{1}{2}t^2 + \tfrac{1}{2}t) = 4t^2 - t + 3. \tag{1.11}$$

✓

1.4 Interpolation d'une fonction continue par un polynôme

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

On veut interpoler f par un polynôme p de degré n aux points $t_j, 0 \le j \le n$, c'est-à-dire on veut trouver un polynôme p de degré n tel que

$$p(t_j) = f(t_j), \quad 0 \le j \le n. \tag{1.12}$$

La figure 1.2 montre un polynôme de degré 3 qui interpole $f(t) = \sin t$ dans les points $t_0 = 0, t_1 = \frac{\pi}{3}, t_2 = \frac{2\pi}{3}, t_3 = \pi$.

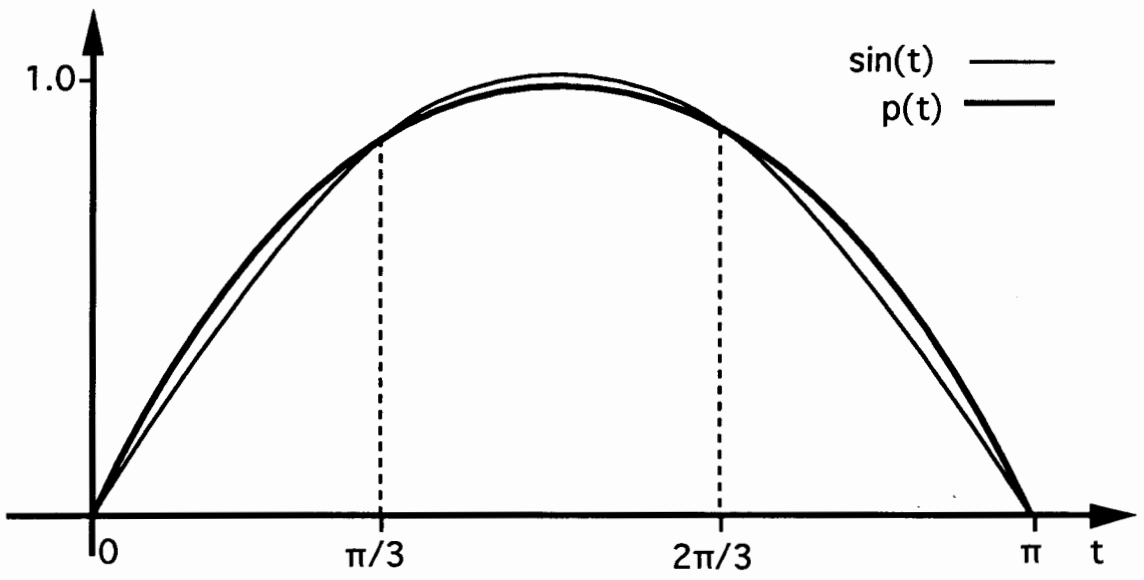


Figure 1.2 : Interpolation de la fonction sinus par un polynôme de degré 3

Clairement si $f(t)$ est donnée, alors on pose $p_j = f(t_j), 0 \le j \le n$ et en suivant ce qui est fait dans le paragraphe (1.3) on obtient $p(t) = \sum_{j=0}^{n} p_j \, \varphi_j(t)$ où les $\varphi_j, 0 \le j \le n$, forment une base de Lagrange des polynômes de degré n associée aux points $t_0, t_1, t_2, \cdots, t_n$. La solution au problème (1.12) est donc :

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

Exemple 1.3. Trouver un polynôme de degré 2 qui interpole la fonction $f(t) \equiv e^t$ dans les points -1, 0 et 1 ?

Si nous reprenons la formule (1.13), 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 on obtient

$$p(t) = \frac{1}{e}(\frac{1}{2}t^2 - \frac{1}{2}t) + (1 - t^2) + e\,(\frac{1}{2}t^2 + \frac{1}{2}t) =$$

$$= (\frac{1}{2e} - 1 + \frac{e}{2})t^2 + (\frac{e}{2} - \frac{1}{2e})t + 1.$$

Lagrange : $x_1, f(x_1)$; $x_2, f(x_2)$; ... ; $x_n, f(x_n)$ } polynôme de degré $(n-1)$

Hermite : $x_1, f(x_1)$; $x_2, f(x_2)$; $x_1, f'(x_1)$; $x_2, f'(x_2)$ } polynôme de degré (3)

1.5 Interpolation d'Hermite

avec des dérivées.

Les problèmes d'interpolation que nous venons de considérer s'appellent "problèmes d'interpolation de Lagrange" ; ils prennent en compte les valeurs de polynômes et de fonctions en certains points, mais ne tiennent pas compte des dérivées.

Il existe d'autres problèmes d'interpolation où on se donne non pas seulement la valeur de $p(t)$, mais aussi la valeur des dérivées 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.

Soit $t_0 < t_1$ deux points donnés et soit p_0, p_1, p_0', p_1' quatre nombres réels donnés. On se propose de trouver un polynôme p de degré 3 tel que

p : polynôme

p_0 : valeur du polynome avec $p(t_0)$

$$p(t_0) = p_0, \quad p(t_1) = p_1, \tag{1.14}$$

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

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

Les conditions (1.14) prescrivent la valeur de p en t_0 et t_1 ; les conditions (1.15) prescrivent la valeur de la dérivée p' de p en t_0 et t_1. Un polynôme de degré 3 s'écrit sous la forme

$$p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3; \qquad \text{(dans la base canonique)}$$

nous avons donc fixé dans (1.14), (1.15) quatre conditions pour déterminer quatre coefficients a_0, a_1, a_2 et a_3. Ici encore, nous aurons affaire à un système linéaire pour déterminer a_0, a_1, a_2 et a_3 et si nous montrons que le problème 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 appelée base d'Hermite associée aux points t_0 et t_1.

– On construit φ_0 tel que φ_0 est un polynôme de degré 3 et

$$\varphi_0(t_0) = 1, \quad \varphi_0'(t_0) = \varphi_0(t_1) = \varphi_0'(t_1) = 0. \qquad \delta_{ij}$$

On vérifie facilement que l'on a :

$$\boxed{\varphi_0(t) = -\frac{(t-t_1)^2\,(2t+t_1-3t_0)}{(t_0-t_1)^3}.} \tag{1.16}$$

– On construit φ_1 tel que φ_1 est un polynôme de degré 3 et

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

On vérifie facilement que l'on a :

$$\boxed{\varphi_1(t) = -\frac{(t-t_0)^2\,(2t+t_0-3t_1)}{(t_1-t_0)^3}.} \tag{1.17}$$

– On construit ψ_0 tel que ψ_0 est un polynôme de degré 3 et

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

On vérifie facilement que l'on a :

$$\boxed{\psi_0(t) = \frac{(t-t_1)^2\,(t-t_0)}{(t_0-t_1)^2}.} \tag{1.18}$$

– On construit ψ_1 tel que ψ_1 est un polynôme de degré 3 et

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

On vérifie facilement que l'on a :

$$\boxed{\psi_1(t) = \frac{(t-t_0)^2\,(t-t_1)}{(t_1-t_0)^2}.} \tag{1.19}$$

Dans la figure 1.3, nous avons représenté φ_0, φ_1, ψ_0 et ψ_1 sur l'intervalle $[t_0, t_1]$ uniquement.

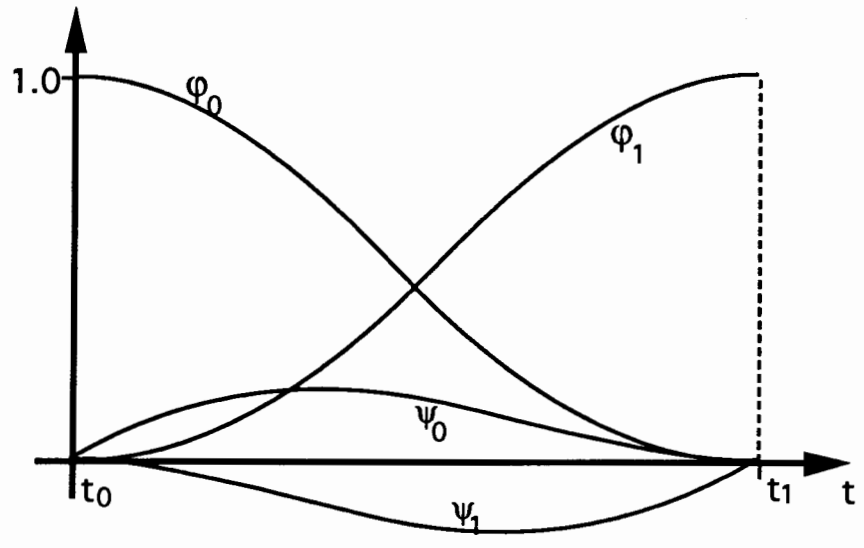


Figure 1.3 : Base d'Hermite des polynômes de degré 3

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 pouvons facilement construire la solution de (1.14), (1.15). En effet, si nous posons

$$\boxed{p(t) = p_0\,\varphi_0(t) + p_1\,\varphi_1(t) + p_0'\,\psi_0(t) + p_1'\,\psi_1(t),} \qquad (1.20)$$

nous vérifions facilement que p est un polynôme de degré 3 qui satisfait (1.14), (1.15).

1.6 Interpolation par intervalles

L'interpolation d'une fonction par des polynômes de degré élevé engendre des problèmes numériques liés aux erreurs d'arrondis (voir leçons 2 et 8 dans lesquelles ce problème est soulevé !) et à des questions de stabilité numérique. C'est la raison pour laquelle on utilise généralement l'interpolation par intervalles.

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

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

et en posant $t_0 = x_i$, $t_j = x_{i,j}$ avec $1 \le j \le n-1$, $t_n = x_{i+1}$, on peut interpoler f sur l'intervalle $[x_i, x_{i+1}]$ par un polynôme de degré n comme fait dans le paragraphe 1.4. Ainsi, on peut construire une fonction $f_h : x \in [a,b] \to f_h(x) \in \mathbb{R}$, appelée interpolant de f, telle que f_h restreinte à cet intervalle $[x_i, x_{i+1}]$ soit justement ce polynôme d'interpolation de degré n.

On peut démontrer que si f est une fonction $(n+1)$ fois continûment dérivable sur $[a,b]$, il existe une constante C (indépendante du choix des $x_i, 1 \le i \le N-1$) telle que

$$\max_{x\in[a,b]} \left| f(x) - f_h(x) \right| \le C \left(\max_{0\le i\le N-1} \left| x_{i+1} - x_i \right| \right)^{n+1}.$$

Ainsi, en prenant N de plus en plus grand, on peut faire en sorte que $\max_{x\in[a,b]} \left| f(x) - f_h(x) \right|$ devienne de plus en plus petit. Par exemple, si on pose $h = (b-a)/N$ et $x_i = a + ih$ avec $i = 0,1,2,\ldots,N$, on aura $\max_{x\in[a,b]} \left| f(x) - f_h(x) \right| \le Ch^{n+1} = C\frac{1}{N^{n+1}}$. En pratique, on prendra N grand et n petit ($n = 1$ ou 2 ou 3 ou 4).

Prenons l'exemple où $a = 0, b = 0.8, f(x) = x^{1.7} + 0.1\,e^{3x} \sin(13x), N = 4$ (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 $n = 1$ on n'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 1.4 montre le graphe de l'interpolant par intervalle.

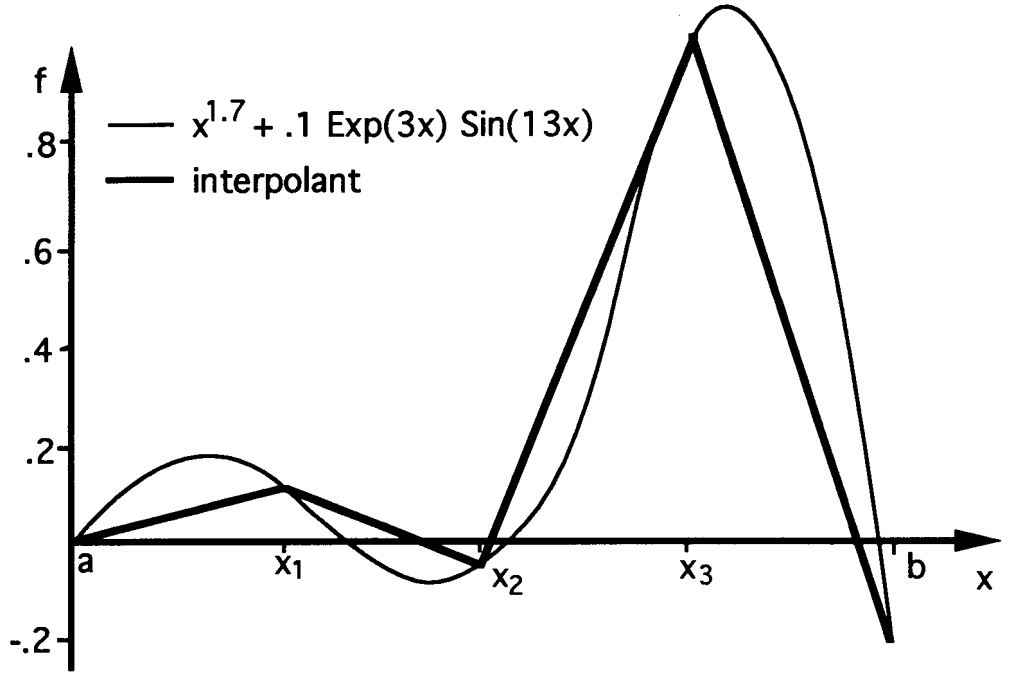


Figure 1.4 : Interpolation par intervalle de f par des polynômes de degré 1

Si $n=2$ on peut prendre 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 fait par des polynômes de degré 2. La figure 1.5 montre le graphe de l'interpolant par intervalle.

milieu : $x_{i,1} = \frac{x_i + x_{i+1}}{2}$

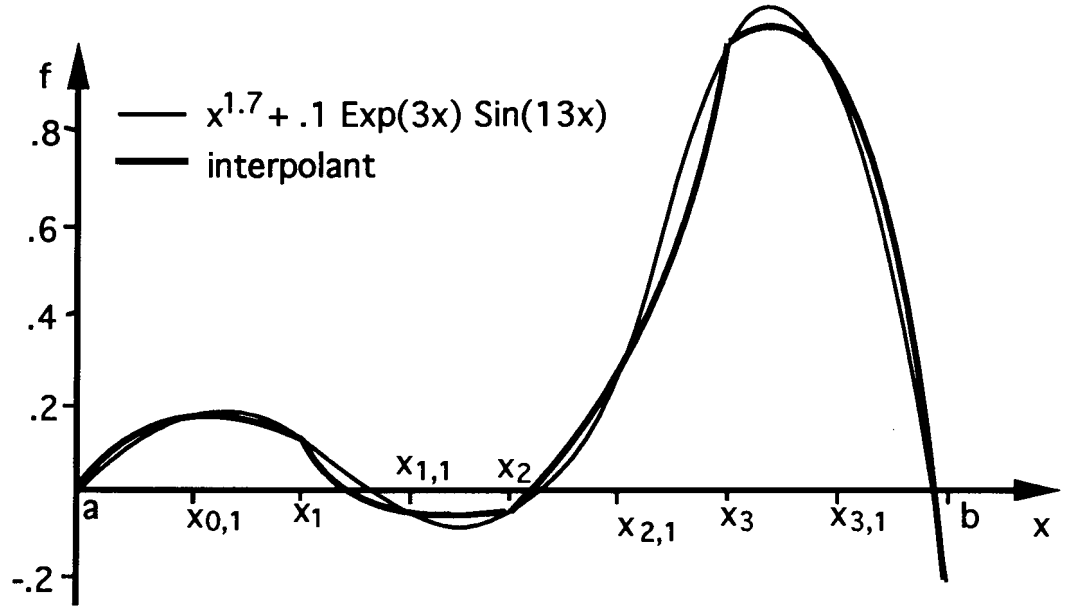


Figure 1.5 : Interpolation par intervalle de f par des polynômes de degré 2

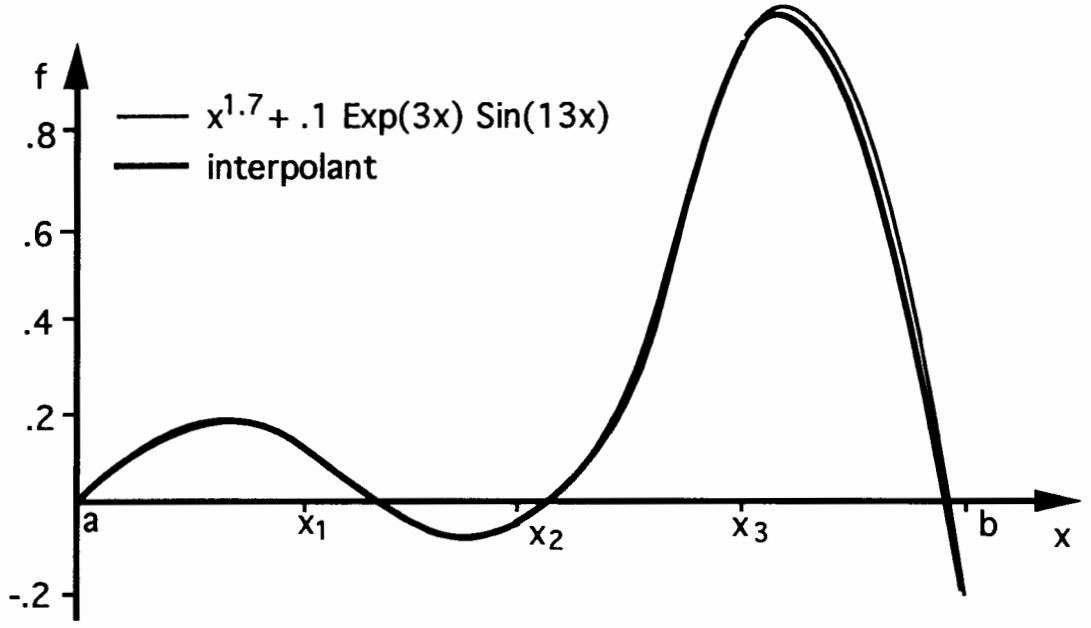


Figure 1.6 : Interpolation d'Hermite par intervalle

L'interpolation de Lagrange par intervalle met en évidence des sauts de la dérivée première en chaque point x_i comme le montre les figures 1.4 et 1.5. Une manière de construire un interpolant par intervalle plus "lisse" est d'utiliser l'interpolation d'Hermite avec des cubiques sur chaque intervalle $[x_i, x_{i+1}]$ (voir paragraphe 1.5). Clairement si f est une fonction C^1 sur l'intervalle $[a,b]$ et si, sur chaque intervalle $[x_i, x_{i+1}]$, on interpole f par un polynôme de degré 3 comme fait dans le paragraphe 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 $[a,b]$ (cf.figure 1.6).

Il existe naturellement d'autres manières de construire des interpolations par intervalles (splines, interpolation de Bézier), mais nous n'en parlerons pas ici.

Leçon 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, on a

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{h\to 0} \frac{f(x_0+h/2)-f(x_0-h/2)}{h}. \qquad (2.1)$$

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

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

après avoir défini les quantités

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

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

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

Les opérateurs Δ_h, ∇_h et δ_h sont appelés différence première respectivement progressive, rétrograde et centrée. On vérifie facilement que ces opérateurs sont linéaires, c'est-à-dire pour Δ_h : si $\alpha,\beta \in \mathbb{R}$ et si f et g sont deux fonctions continues, on a $\Delta_h(\alpha f+\beta g)(x_0) = \alpha\Delta_h f(x_0) + \beta\Delta_h g(x_0)$.

Si f est une fonction deux fois continûment dérivable, on peut écrire par développement de Taylor au 2ème ordre :

$$f(x) = \sum_{i=0}^{n} f^{(i)}(x_0) \cdot \frac{x_0^i}{i!}$$

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

où ξ est un point de l'intervalle $[x_0, x_0+h]$.

De la relation (2.6) et de la définition (2.3) on obtient clairement :

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

et on voit que si h tend vers 0, on a la convergence de $\frac{\Delta_h f(x_0)}{h}$ vers $f'(x_0)$
avec un ordre de 1 en h, c'est-à-dire qu'il existe deux constantes C et h_0 telles que :

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

(Dans (2.8) on peut poser $C = \frac{1}{2} \max_{x \in [x_0, x_0+h_0]} \left| f''(x) \right|$, et $h_0 > 0$ quelconque).

On vérifie de la même façon qu'on obtient un résultat semblable si $\Delta_h f(x_0)$ est remplacé par $\nabla_h f(x_0)$.

Par contre si on décide d'approcher $f'(x_0)$ par la valeur $\frac{\delta_h f(x_0)}{h}$, on obtient une meilleure valeur. En effet, supposons f trois fois continûment dérivable et considérons les développements de Taylor :

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

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

où ξ est un point de l'intervalle $\left[x_0, x_0+\frac{h}{2}\right]$ et η est un point de $\left[x_0-\frac{h}{2}, x_0\right]$. En soustrayant (2.10) à (2.9) et en utilisant la définition (2.5), nous avons :

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

Si h_0 est un nombre positif fixé et si $C = \frac{1}{24} \max_{x \in \left[x_0-\frac{h_0}{2}, x_0+\frac{h_0}{2}\right]} \left| f'''(x) \right|$, il est aisé de montrer à partir de (2.11) que

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

On voit donc que si h tend vers zéro, alors $\frac{\delta_h f(x_0)}{h}$ converge vers $f'(x_0)$
avec un ordre de 2 en h (erreur en h^2).

Définition. On dit que $\frac{\Delta_h f(x_0)}{h}$ et $\frac{\nabla_h f(x_0)}{h}$ sont des formules de différences finies progressive et rétrograde pour l'approximation de $f'(x_0)$. Les différences $\left| f'(x_0) - \frac{\Delta_h f(x_0)}{h} \right|$

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

La formule de différences finies centrées $\frac{\delta_h f(x_0)}{h}$ 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.

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

Définition. Nous dirons qu'un nombre a en virgule flottante est donné avec N chiffres significatifs s'il est donné avec N chiffres comptés à partir du premier chiffre non-nul.

Exemple. 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.

Un calculateur en virgule flottante ne peut que travailler avec un nombre fini de chiffres significatifs N. Clairement si nous voulons saisir le nombre $c = \frac{1}{3}$ avec un calculateur à 8 chiffres significatifs, alors nous aurons, à la place du nombre c, le nombre $a = 0{,}33333333$. L'erreur $|a - c| = \frac{1}{3}10^{-8}$ est appelée erreur d'arrondis.

Calculons maintenant $\frac{\Delta_h f(x_0)}{h}$ lorsque $f(x) = x^2$ et $x_0 = 7$.

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 :

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

Clairement nous avons $f'(x_0) = 14$ et nous observons que l'erreur obtenue par la formule aux différences $\frac{\Delta_h f(x_0)}{h}$ 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, on obtient dans ce cas les valeurs suivantes :

- $h = 0.06$: $\dfrac{\Delta_h f(x_0)}{h} = \dfrac{(7.06)^2 - (7.00)^2}{0.06} = \dfrac{49.8436 - 49.0000}{0.06} = 14.0600,$

- $h = 0.01$: $\dfrac{\Delta_h f(x_0)}{h} = \dfrac{(7.01)^2 - (7.00)^2}{0.01} = \dfrac{49.1401 - 49.0000}{0.01} = 14.0100.$

Dans la suite, nous ne ferons pas une théorie des erreurs d'arrondis. Disons simplement que si η est la précision relative du calculateur utilisé (on pose $\eta = 10^{-N}$ où N est le nombre de chiffres significatifs dont on dispose) alors l'erreur absolue obtenue sur l'évaluation d'un nombre c est $c \times \eta$. Ainsi, si on veut calculer $\frac{\Delta_h f(x_0)}{h} \underset{def}{=} \frac{f(x_0+h)-f(x_0)}{h}$ avec une valeur de h supposée donnée sans erreur d'arrondis (exemple : $h = 10^{-7}, 10^{-8}, \cdots$) on obtient :

- erreur absolue commise sur l'évaluation de $f(x_0 + h)$:
  $\eta . |f(x_0 + h)|$;

- erreur absolue commise sur l'évaluation de $f(x_0)$:
  $\eta . |f(x_0)|$;

- erreur absolue commise sur l'évaluation de $\Delta_h f(x_0)$:
  $\eta . \big(|f(x_0 + h)| + |f(x_0)|\big) \simeq 2\eta . |f(x_0)|$;

- erreur absolue commise sur l'évaluation de $\Delta_h f(x_0)/h$:
  $\sim 2\eta . \frac{|f(x_0)|}{h}$.

En conclusion, l'erreur d'arrondis commise sur l'évaluation de la quantité $\frac{\Delta_h f(x_0)}{h}$ est approximativement donnée par $2\eta . \frac{|f(x_0)|}{h}$. Si nous calculons l'erreur relative e_r due aux erreurs d'arrondis sur l'évaluation de $\frac{\Delta_h f(x_0)}{h}$, nous obtenons :

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

cette erreur relative grandit lorsque h devient petit.

En reprenant l'exemple ci-dessus, nous aurons $f(x_0) = 49$, $f'(x_0) = 14$ et

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

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

Ces considérations restent inchangées si on procède à l'évaluation de $\frac{\nabla_h f(x_0)}{h}$ ou $\frac{\delta_h f(x_0)}{h}$.

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 $f'(x_0)$, on prend la formule aux différences progressives, i.e. $f'(x_0) \simeq \frac{\Delta_h f(x_0)}{h}$. Nous avons vu que l'erreur de troncature commise est approximativement égale à $\boxed{E_t^h = \frac{1}{2}|f''(x_0)|h}$ (voir formule (2.7)). Par contre, l'erreur d'arrondis est approximativement donnée par $\boxed{E_a^h = 2\eta\frac{|f(x_0)|}{h}}$ où η est la précision relative du calculateur. Ainsi, si au lieu de calculer $f'(x_0)$, on calcule $\frac{\Delta_h f(x_0)}{h}$ avec un calculateur de précision relative η, on peut s'attendre à une erreur totale $E^h = E_a^h + E_t^h$ constituée des erreurs d'arrondis et de troncature. Pour expression de E^h on aura donc

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

Il est possible de calculer (théoriquement) h pour obtenir la plus petite erreur possible E^h. En effet, posons

$$g(x) = ax + \frac{b}{x}$$

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

Pour le faire, il suffit de calculer

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

et de vérifier que si $\bar{x} = \sqrt{b/a}$, on a $g'(\bar{x}) = 0$ et $g''(\bar{x}) > 0$. On constate que $\bar{x}$ est l'unique minimum de $g(x)$ pour $x > 0$ et la valeur de h optimale cherchée est donc

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

Il en est de même lorsqu'on prend pour approximation de $f'(x_0)$ la formule aux différences rétrogrades, i.e. $f'(x_0) \simeq \frac{\nabla_h f(x_0)}{h}$.

Par contre, si on décide de prendre pour approximation de $f'(x_0)$ la formule aux différences centrées, i.e. $f'(x_0) \simeq \frac{\delta_h f(x_0)}{h}$, l'erreur totale aura pour expression (voir (2.12)) :

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

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 itérés de la façon suivante.

Si m est 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

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

Nous vérifions que les opérateurs Δ_h^m, ∇_h^m et δ_h^m sont aussi linéaires.

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

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

sont des approximations de la $m^{\text{ième}}$ dérivée $f^{(m)}(x_0)$ de f au point x_0, d'ordre 1, 1 et 2 respectivement en h. Ainsi, il existe deux constantes C et h_0 telles que si $h \le h_0$, on a

$$\left| f^{(m)}(x_0) - \frac{\Delta_h^m f(x_0)}{h^m} \right| \le Ch, \tag{2.19}$$

$$\left| f^{(m)}(x_0) - \frac{\nabla_h^m f(x_0)}{h^m} \right| \le Ch, \tag{2.20}$$

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

Les problèmes de diffusions, déformations élastiques, propagations d'ondes, écoulements de fluides, ..., etc, etc font apparaître des dérivées deuxième ou quatrième. Ainsi, les formules très utilisées par les ingénieurs (voir leçon 9) sont des formules de différences finies centrées avec $m=2$ et $m=4$, i.e.

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

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

qui donnent une erreur de troncature d'ordre h^2 si f est "assez régulière".

Si les erreurs de troncature dans (2.19), (2.20) et (2.21) deviennent petites lorsque h est petit, les erreurs d'arrondis par contre deviennent importantes et sont d'autant plus importantes que m est grand !

En ce qui concerne les formules de différences finies pour les dérivées d'ordre m, on peut vérifier l'affirmation suivante :

> Soient $f : \mathbb{R} \to \mathbb{R}$ une fonction continue donnée, $x_0 \in \mathbb{R}$, $h>0$ des nombres réels donnés et r un entier positif donné.
> On pose $m=2r$ et on appelle p le polynôme de degré m qui interpole f en les $(m+1)$ points $x_0-rh, x_0-(r-1)h, x_0-(r-2)h, \cdots, x_0-h, x_0, x_0+h, \cdots, x_0+(r-1)h, x_0+rh$.
> Alors on a :
>
> $$\frac{\delta_h^m f(x_0)}{h^m} = \frac{d^m p}{dx^m}(x_0).$$

Pour exemple, on considère la figure 2.1 dans laquelle on a représenté le graphe de f et celui de p dans le cas $r=1$.

Le polynôme p est de degré 2 et satisfait

$$p(x_0-h)=f(x_0-h), \quad p(x_0)=f(x_0), \quad p(x_0+h)=f(x_0+h).$$

Par la technique donnée dans la leçon 1 (voir section 1.4), on obtient

$$p(x)=\frac{1}{2h^2}\Big[\big(f(x_0-h)-2f(x_0)+f(x_0+h)\big)(x-x_0)^2 +$$

$$+h\big(f(x_0+h)-f(x_0-h)\big)\cdot(x-x_0)+2h^2 f(x_0)\Big].$$

On vérifie bien ainsi que $p''(x_0)=\frac{\delta_h^2 f(x_0)}{h^2}$.

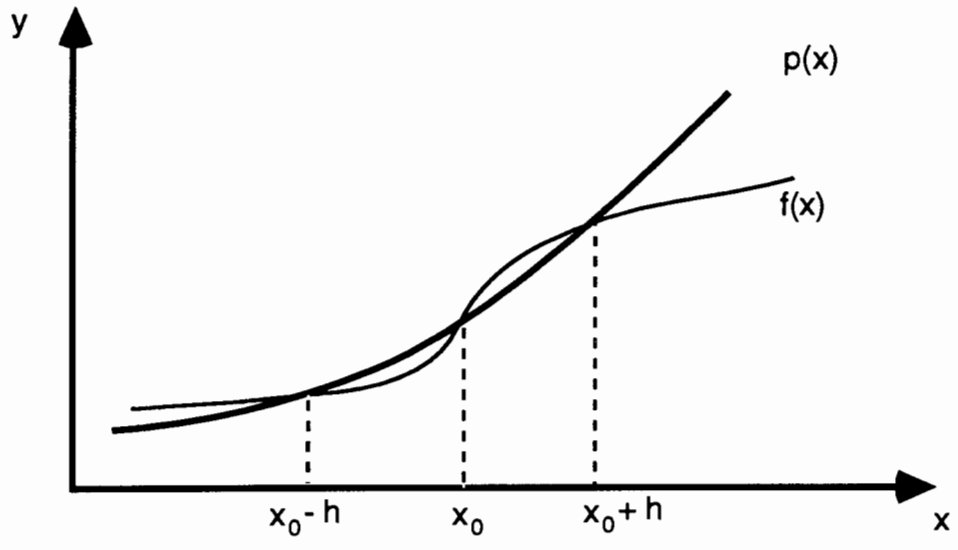


Figure 2.1 : Interpolation de f par un polynôme p de degré 2

$f(x) = \sum_{k=0}^{n} f^{(k)}(a) \frac{(x-a)^k}{k!}$, D.L. autour de a.

2.5 Extrapolation de Richardson

Il est possible de trouver des formules de dérivation numérique plus précises que celles que l'on a obtenues. Supposons par exemple la fonction f cinq fois continûment dérivable et développons de façon semblable à (2.9), (2.10) mais en allant jusqu'à l'ordre 5. On obtient :

a D.L. autour de a, avec $x = \frac{h}{2}$

$$f(x_0 + \tfrac{h}{2}) = f(x_0) + f'(x_0)\tfrac{h}{2} + \frac{f''(x_0)}{2!}(\tfrac{h}{2})^2 + \frac{f'''(x_0)}{3!}(\tfrac{h}{2})^3 + \frac{f^{IV}(x_0)}{4!}(\tfrac{h}{2})^4 + \frac{f^{V}(\xi)}{5!}(\tfrac{h}{2})^5, \quad (1)$$

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

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

$\frac{(1)-(2)}{h} = \frac{1}{h} \cdot \left(2 \cdot f'(x_0) \cdot \frac{h}{2} + 2 \cdot \frac{f'''(x_0)}{3!} \cdot \frac{h^3}{2^3} + \text{reste} \right)$

$$\Rightarrow \quad \underbrace{\frac{f(x_0+\frac{h}{2}) - f(x_0-\frac{h}{2})}{h}}_{= \delta_h f(x_0)} = f'(x_0) + \frac{f'''(x_0)}{24}h^2 + \underbrace{\frac{f^{V}(\xi)+f^{V}(\eta)}{5!\,2^5}h^4}_{\text{reste}}$$

et en prenant la définition de $\delta_h f(x_0)$:

$$\frac{\delta_h f(x_0)}{h} = f'(x_0) + \frac{f'''(x_0)}{24}h^2 + O(h^4), \qquad (2.24)$$

où ici $O(h^4)$ signifie que c'est un reste d'ordre h^4 lorsque h tend vers zéro. Dans (2.24), si on remplace h par $\frac{h}{2}$, on obtient :

$$\frac{\delta_{h/2} f(x_0)}{h/2} = f'(x_0) + \frac{f'''(x_0)}{24}\frac{h^2}{4} + O(h^4). \qquad (2.25)$$

En soustrayant quatre fois (2.25) à (2.24) on obtient :

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

et finalement

$$f'(x_0) = \frac{\frac{8}{3}\delta_{h/2} f(x_0) - \frac{1}{3}\delta_h f(x_0)}{h} + O(h^4). \qquad (2.26)$$

En observant que

$$\frac{8}{3}\delta_{h/2} f(x_0) - \frac{1}{3}\delta_h f(x_0) = \frac{1}{3}\left[8 f(x_0 + \tfrac{h}{4}) - 8 f(x_0 - \tfrac{h}{4}) - f(x_0 + \tfrac{h}{2}) + f(x_0 - \tfrac{h}{2}) \right]$$

la formule

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

pour l'approximation de $f'(x_0)$ est consistante à l'ordre 4 en h. Le procédé pour obtenir cette formule est appelé méthode d'extrapolation de Richardson . On peut généraliser cette méthode pour obtenir des formules d'ordre 6, 8, ... en h pour l'approximation de $f'(x_0)$. Il suffira de tenir compte de $\delta_{h/4} f(x_0)$, $\delta_{h/8} f'(x_0)$, ...

Leçon 3

Intégration numérique
Formules de quadrature

3.1 Généralités

Nous avons pour but maintenant 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 donner une approximation numérique de la quantité

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

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

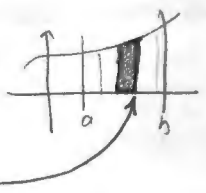
$$a = x_0 < x_1 < x_2 < x_3 < \cdots \quad \cdots < x_{N-1} < x_N = b. \tag{3.2}$$

$h = \frac{b-a}{N}$ (si les intervalles sont espacés régulièrement)

On pose $h = \max_{0 \le i \le N-1} |x_{i+1} - x_i|$; cette valeur mesure la finesse de la partition. Clairement, plus on prend de points x_i dans l'intervalle $[a,b]$ (N "grand"), plus on peut faire en sorte que h soit "petit". Lorsqu'aucune raison nous amène à prendre des intervalles de longueurs différentes, nous posons $h = \frac{b-a}{N}$ et $x_i = a + ih, i = 0,1,\cdots,N$.

Ayant défini la partition (3.2), nous pouvons écrire naturellement :

$$\int_a^b f(x)\,dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)\,dx. \tag{3.3}$$



Ce sont ainsi les intégrales $\int_{x_i}^{x_{i+1}} f(x)\,dx$ que nous allons approcher dans la suite par des formules dites de quadrature. Mentionnons encore que souvent, pour donner des formules de

<u>exemple</u> : on a : $\int_0^\infty e^{-t} f(t)\, dt$.

pour trouver les poids d'intégration ω_1, ω_2 en fct. de t_1, t_2 t.q. $J(P) = \int_0^\infty e^{-t} p(t)\, dt$. $\forall p \in \mathbb{P}^1$.

$$J(P) = \int_0^\infty e^{-t} p(t)\, dt = \omega_1 \cdot P(t_1) + \omega_2 P(t_2)$$

et : $p(t) = p(t_1)\, \varphi_1(t) + p(t_2)\, \varphi_2(t)$

ainsi :

$$J(p) = \int_0^\infty e^{-t} p(t)\, dt = p(t_1) \cdot \underbrace{\int_0^\infty e^{-t} \varphi_1(t)\, dt}_{\omega_1} + p(t_2) \cdot \underbrace{\int_0^\infty e^{-t} \varphi_2(t)\, dt}_{\omega_2}$$

$$\varphi_1 = \frac{t - t_2}{t_1 - t_2} \qquad \varphi_2 = \frac{t - t_1}{t_2 - t_1}$$

$$\omega_1 = \frac{1}{t_1 - t_2} \int_0^\infty e^{-t} (t - t_2)\, dt = \frac{1 - t_2}{t_1 - t_2}$$

$$\omega_2 = \frac{1}{t_2 - t_1} \int_0^\infty e^{-t} (t - t_1)\, dt = \frac{1 - t_1}{t_2 - t_1}$$

fin.

quadrature sur un intervalle "standard" (par exemple l'intervalle $[-1,+1]$), on exécute un changement de variable de la forme

$$t = 2\frac{x-x_i}{x_{i+1}-x_i} - 1 \tag{3.4}$$

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

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

et par suite

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{x_{i+1}-x_i}{2} \int_{-1}^{+1} g_i(t)\,dt, \tag{3.6}$$

où $g_i(t) = f(x_i + (x_{i+1} - x_i)\frac{t+1}{2}), t \in [-1,+1]$.

Dès lors, on peut définir ce qu'est une formule de quadrature pour $\int_{-1}^{+1} g(t)\,dt$ où g est une fonction continue donnée sur $[-1,+1]$.

Définition. La donnée de M points $-1 \le t_1 < t_2 < \cdots < t_M \le 1$ de l'intervalle $[-1,+1]$ et de M nombres réels $\omega_1, \omega_2, \cdots, \omega_M$ définit la formule de quadrature

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

où g est une fonction continue donnée sur $[-1,+1]$. Les points t_j sont appelés "points d'intégration" alors que les ω_j sont appelés "poids de la formule de quadrature".

Définition. 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 des polynômes de degré $r \ge 0$ si on a

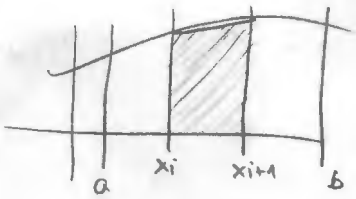
$$J(p) = \int_{-1}^{+1} p(t)\,dt$$

pour tout polynôme p de degré $\le r$.

L'introduction d'une formule de quadrature $J(\cdot)$ nous permet de remplacer une intégrale du type $\int_{-1}^{+1} g(t)\,dt$ par $J(g)$ et donc, en observant (3.6), de remplacer $\int_{x_i}^{x_{i+1}} f(x)\,dx$ par

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = \frac{x_{i+1}-x_i}{2} \sum_{j=1}^{M} \omega_j\, f\left(x_i + (x_{i+1} - x_i)\frac{t_j+1}{2}\right). \tag{3.7}$$

Par retour à (3.3), on peut remplacer $\int_a^b f(x)\,dx$ par



$$I_h = \frac{h}{2} \sum_{i=0}^{N-1} \cdot \underbrace{\sum_{j=1}^{M} \omega_j \cdot f\left(x_i + h \cdot \frac{t_j + 1}{2}\right)}_{J(G)}$$

(M → nb. de pomb d'intégration)

t_j : pts. d'intégration

$\omega_j = \int e_j$

L'idée est de dilater chaque intervalle $[x_i, x_{i+1}]$, $i = 1, .., N-1$ sur $[-1, 1]$ et d'y appliquer une formule de quadrature $J(G) = \sum_j \omega_j \cdot g(t_j)$.

Donc on divise l'intégrale sur $[a,b]$ en N intégrales qu'on dilate chacune sur l'intégrale et par laquelle on applique la formule de quadrature, p. ex. trapèzes, etc.

t_j : pts d'intégration

$$\int_a^b f(x)\,dx \approx L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1}-x_i}{2} \sum_{j=1}^{M} \omega_j\, f\left(x_i + (x_{i+1}-x_i)\frac{t_j+1}{2}\right). \qquad (3.8)$$

($x_{i+1}-x_i = h$)

Notons en passant que le calcul de $L_h(f)$ ne requiert que des évaluations de f en un certain nombre de points de $[a,b]$. Ainsi, comme nous le verrons dans les sections qui suivent, on peut choisir des points d'intégration $(t_j)_{j=1}^M$ et des poids $(\omega_j)_{j=1}^M$ de telle sorte que $L_h(f)$ soit une "bonne" approximation de $\int_a^b f(x)\,dx$. En fait, on peut démontrer le résultat suivant :

Théorème. On suppose que la formule de quadrature $J(g) = \sum_{j=1}^{M} \omega_j\, g(t_j)$ pour intégrer numériquement $\int_{-1}^{+1} g(t)\,dt$ est exacte pour des polynômes de degré r. Alors si f est une fonction assez régulière (i.e. $(r+1)$ fois continûment dérivable sur l'intervalle $[a,b]$), il existe une constante C indépendante des points x_i choisis telle que

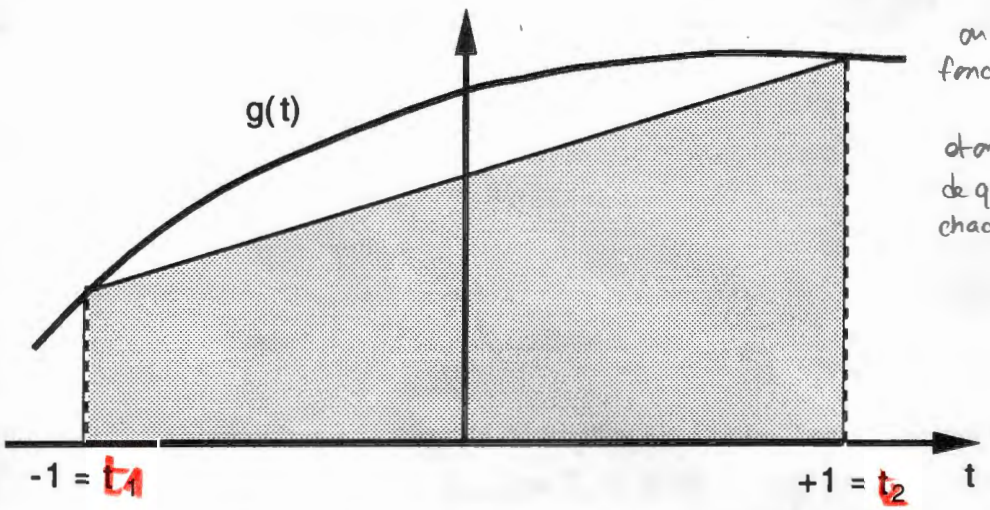
$$\left| \int_a^b f(x)\,dx - L_h(f) \right| \le C h^{r+1} = C \cdot \left(\frac{b-a}{N}\right)^{r+1} \qquad (3.9)$$

où $L_h(\cdot)$ est donnée par (3.8) et $h = \max_{0 \le i \le N-1} |x_{i+1} - x_i|$. $\quad h = \frac{b-a}{N}$

L'inégalité (3.9) montre que lorsque la partition est fine (h "petit"), l'erreur faite en calculant $L_h(f)$ au lieu de $\int_a^b f(x)\,dx$ est petite. Cette erreur est d'autant plus petite que r est grand. Il est donc légitime de chercher des points d'intégration t_j et des poids $\omega_j, 1 \le j \le M$, de telle sorte à ce que la formule $J(\cdot)$ soit exacte pour des polynômes de degré r aussi élevé que possible.

3.2 Formule du trapèze

La formule du trapèze remplace $\int_{-1}^{+1} g(t)\,dt$ par l'aire du trapèze défini par la figure 3.1.



on a donc ramené la fonction sur $[-1,1]$ de $[x_i, x_{i+1}]$ et on applique la formule de quadrature sur chacun de ces intervalles

Figure 3.1 : Formule du trapèze sur $[-1,+1]$

Ainsi on a

$$J(g) = \frac{g(-1)+g(1)}{2} \cdot 2 = g(-1) + g(1).$$

C'est une formule à 2 points $t_1 = -1, t_2 = +1$ et on a $\omega_1 = \omega_2 = 1$. On voit immédiatement qu'elle est exacte si g est un polynôme de degré 1, i.e. $\int_{-1}^{+1} g(t)\,dt = J(g)$ si g est de la forme $g(t) = \alpha t + \beta$, $\alpha, \beta \in \mathbb{R}$. En remplaçant dans (3.8) et (3.9) on aura :

$$\int_a^b f(x)\,dx = L_h(f) = \sum_{i=0}^{N-1} \underbrace{(x_{i+1} - x_i)}_{h} \frac{f(x_i)+f(x_{i+1})}{2}. \tag{3.10}$$

et

$$\left| \int_a^b f(x)\,dx - L_h(f) \right| \le C\,h^2. \tag{3.11}$$

La formule (3.10) est facile à interpréter géométriquement : la quantité $L_h(f)$ donne l'aire hachurée dans la figure 3.2.

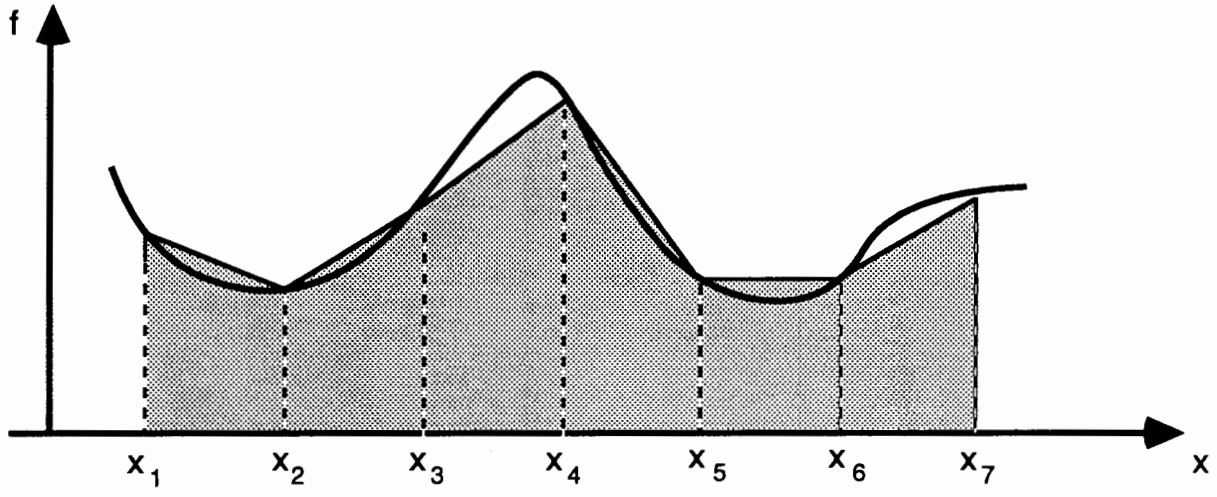


Figure 3.2 : Formule du trapèze pour $\int_a^b f(x)\,dx$

Par l'inégalité (3.11), on voit que la formule du trapèze est d'ordre 2 en h; l'erreur est majorée par h^2; en divisant chaque intervalle $[x_i, x_{i+1}]$ par 2, on divise l'erreur par un facteur 4.

3.3 Formule du rectangle

La formule du rectangle remplace $\int_{-1}^{+1} g(t)\,dt$ par l'aire du rectangle de base $[-1,+1]$ et de hauteur $g(0)$ (voir figure 3.3).

Ainsi on a

$$J(g) = 2\,g(0).$$

C'est une formule à 1 point $t_1 = 0$ et on a $\omega_1 = 2$. On voit immédiatement qu'elle est exacte pour des polynômes de degré 1. En remplaçant dans (3.8), (3.9) on aura :

normalement de degré 0, mais par symétrie on voit que l'on peut intégrer des p. de deg 2

$$L_h(f) = \sum_{i=0}^{N-1} (x_{i+1} - x_i)\, f\left(\frac{x_i + x_{i+1}}{2}\right) \qquad (3.12)$$

et

$$\left| \int_a^b f(x)\,dx - L_h(f) \right| \le C\,h^2. \qquad (3.13)$$

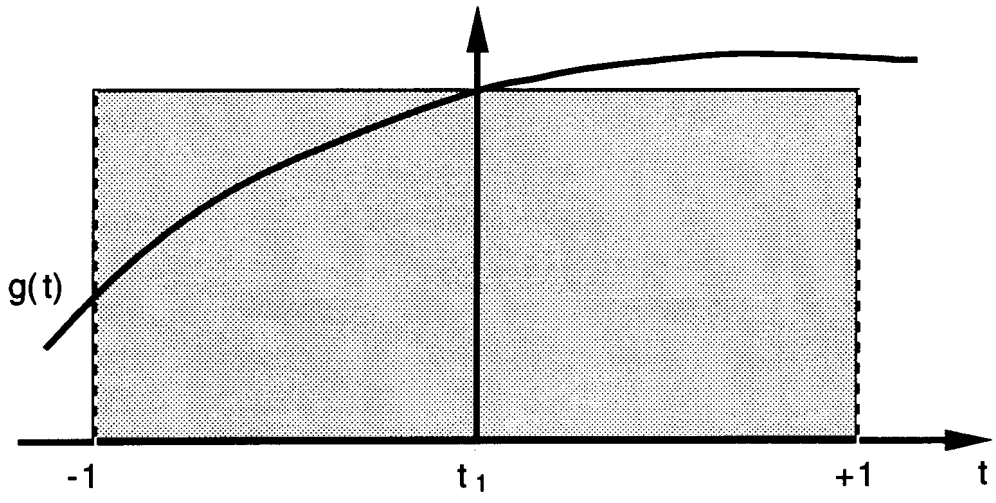


Figure 3.3 : Formule du rectangle sur $[-1,+1]$

L'interprétation géométrique de (3.12) est simple : on somme les aires des rectangles dont la base est le segment $[x_i, x_{i+1}]$ et la hauteur $f(\xi_i)$ où ξ_i est le point milieu de $[x_i, x_{i+1}]$. On obtient encore une formule d'ordre 2 en h.

3.4 Formule de Simpson

La formule de Simpson est une moyenne pondérée entre la formule du trapèze (poids $\frac{1}{3}$) et la formule du rectangle (poids $\frac{2}{3}$). Ainsi, on remplace $\int_{-1}^{+1} g(t)\,dt$ par

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

C'est une formule à 3 points $t_1 = -1, t_2 = 0, t_3 = 1$ avec les poids $\omega_1 = \frac{1}{3}, \omega_2 = \frac{4}{3}, \omega_3 = \frac{1}{3}$. On peut vérifier que $J(p) = \int_{-1}^{+1} p(t)\,dt$ lorsque $p(t) = 1, t, t^2$ et t^3. Ainsi, la formule de quadrature est exacte pour des polynômes de degré 3. En remplaçant dans (3.8), (3.9) on aura

$$L_h(f) = \sum_{i=0}^{N-1} \frac{x_{i+1}-x_i}{6}\left(f(x_i) + 4f\left(\frac{x_i+x_{i+1}}{2}\right) + f(x_{i+1})\right) \qquad (3.14)$$

et

$$\left|\int_a^b f(x)\,dx - L_h(f)\right| \le C\,h^4. \qquad (3.15)$$

La formule de Simpson est d'ordre 4 en h; 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

L'idée des formules de Gauss (ou Gauss-Legendre) est de placer au mieux les points d'intégration $t_1, t_2, \cdots, t_M$ de sorte à ce que $J(p)$ soit égal à $\int_{-1}^{+1} p(t)\,dt$ pour des polynômes p de degré r aussi grand que possible. Ainsi, pour une fonction f donnée sur $[a,b]$, l'approximation numérique de $\int_a^b f(x)\,dx$ par $L_h(f)$ sera d'autant meilleure si l'on en juge par la relation (3.9).

Considérons pour commencer que les points $-1 \le t_1 < t_2 < \cdots < t_M \le 1$ sont fixés et associons à ces points la base de Lagrange $\varphi_1, \varphi_2, \cdots, \varphi_M$ des polynômes de degré $M-1$ (voir leçon 1). Par définition, φ_j est un polynôme de degré $M-1$ et on a $\varphi_j(t_k) = 0$ si $1 \le k \le M, k \ne j$ et $\varphi_j(t_j) = 1, 1 \le j \le M$.

Si $g : t \in [-1,+1] \to g(t) \in \mathbb{R}$ est une fonction continue, on peut définir le polynôme $\tilde{g}$ de degré $M-1$ qui l'interpole aux points $t_1, t_2, \cdots, t_M$, c'est-à-dire :

$$\tilde{g}(t) = \sum_{j=1}^{M} g(t_j)\,\varphi_j(t). \qquad (3.16)$$

Il est donc naturel de remplacer $\int_{-1}^{+1} g(t)\,dt$ par

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

qui n'est rien d'autre, si on utilise (3.16), que

$$J(g) = \sum_{j=1}^{M} \omega_j\, g(t_j) \tag{3.17}$$

avec

$$\omega_j = \int_{-1}^{+1} \varphi_j(t)\, dt. \tag{3.18}$$

A ce stade, on peut résumer ce qui précède comme suit :

Si $t_1 < t_2 < \cdots < t_M$ sont M points d'intégration fixés n'importe où dans l'intervalle $[-1,+1]$ et si on définit les poids $\omega_1, \omega_2, \cdots, \omega_M$ par (3.18), alors la formule de quadrature (3.17) est exacte pour des polynômes de degré $M-1$ (affirmation évidente lorsqu'on a constaté que $g = \tilde{g}$ quand g est un polynôme de degré $M-1$). Ainsi, on sait calculer les poids d'intégration lorsque les points t_j sont donnés.

Maintenant, pour placer les points $t_1, t_2, \cdots, t_M$ de façon optimale, on est obligé de considérer les polynômes de Legendre. Dans cette leçon, nous ne rentrerons pas dans ces considérations. Nous dirons seulement que le polynôme de Legendre de degré M, noté P_M et défini par

$$P_M(t) = \frac{1}{2^M M!} \frac{d^M}{dt^M} (t^2-1)^M, \tag{3.19}$$

a exactement M zéros réels $t_1, t_2, t_3, \cdots, t_M$ tous distincts et compris dans l'intervalle ouvert $]-1,+1[$. Si, dans ce qui précède, nous choisissons ces M racines de P_M pour points d'intégration, alors la formule de quadrature (3.17) devient exacte pour des polynômes de degré $2M-1$ (et non pas seulement pour des polynômes de degré $M-1$). Dans ce cas, les points $t_1, t_2, \cdots, t_M$ sont appelés "points de Gauss" (ce sont les racines de P_M) et la formule de quadrature (3.17) est appelée "formule de Gauss-Legendre à M points" ou plus simplement "formule de Gauss à M points". Les poids de la formule de Gauss sont donnés par (3.18).

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 (3.8). Tenant compte de ce qui précède et de l'inégalité (3.9) nous avons :

$$\left| \int_a^b f(x)\, dx - L_h(f) \right| \le C\, h^{2M}, \tag{3.20}$$

où ici f est supposée "régulière" et C est une constante qui ne dépend pas des points $(x_i)_{i=0}^N$ choisis pour partitionner $[a,b]$.

Notons au passage que si $M = 1$ on a par (3.19)

$$P_1(t) = t$$

et par suite $t_1 = 0$ (seule racine de P_1). La base de Lagrange des polynômes de degré zéro associée à t_1 est donnée par

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

En calculant ω_1 par (3.18) on obtient

$$\omega_1 = 2$$

et la formule (3.17) devient dans ce cas-là :

$$J(g) = 2g(0),$$

qui n'est rien d'autre que la formule du rectangle. Clairement, nous venons de démontrer que la formule du rectangle est une formule de Gauss-Legendre à 1 point.

Dans la littérature, on trouve bien d'autres formules d'intégration suivant qu'on veuille intégrer numériquement des fonctions sur des intervalles infinis, semi-infinis ou avec des fonctions de poids sous le signe d'intégration. Ainsi, au chapitre de l'intégration gaussienne, on trouve les noms de Gauss-Laguerre, Gauss-Hermite, Gauss-Chebycheff, ...

Leçon 4

Résolution de systèmes linéaires
Elimination de Gauss
Systèmes mal conditionnés

4.1 Position du problème

Dans cette leçon, on considère un système d'équations linéaires d'ordre N de la forme

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

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

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdots & \cdots & a_{2N} \\ . & . & . & \cdots & \cdots & . \\ . & . & . & \cdots & \cdots & . \\ . & . & . & \cdots & \cdots & . \\ a_{N1} & a_{N2} & . & \cdots & \cdots & a_{NN} \end{bmatrix},$$

$$\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_N \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}.$$

Clairement, le système (4.1) peut être écrit explicitement comme un système de N équations à N inconnues $x_1, x_2, x_3, \cdots, x_N$:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1N}x_N = b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2N}x_N = b_2, \\ \vdots \\ \vdots \\ a_{N1}x_1 + a_{N2}x_2 + a_{N3}x_3 + \cdots + a_{NN}x_N = b_N. \end{cases} \tag{4.2}$$

Définitions

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

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

Supposons pour un instant que la matrice A est triangulaire supérieure. Alors on voit dans ce cas que le déterminant de A est le produit des valeurs diagonales $a_{ii} \neq 0, \forall i = 1,2,\cdots,N$. Ainsi, quitte à diviser chaque équation de (4.2) par le terme diagonal, il n'est pas restrictif de supposer que $a_{ii} = 1, \forall i = 1,2,\cdots,N$. Dans ce cas, la matrice A est triangulaire supérieure avec des valeurs 1 dans sa diagonale et, de (4.2), on tire facilement les inconnues $x_N, x_{N-1}, \cdots, x_1$ successivement car on obtient $x_N = b_N$ et pour $i = N-1, N-2, N-3, \cdots, 2,1$:

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

Si maintenant 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 dans la diagonale.

4.2 Elimination de Gauss sur un exemple

Prenons un exemple avec $N = 3$. Pour montrer comment faire la transformation d'un système linéaire en un système triangulaire supérieur équivalent, on prend :

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

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

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

La première opération consiste à diviser la 1ère équation de (4.4) par $a_{11} = 4$ (appelé 1er pivot) pour obtenir

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

Ensuite on soustrait 3 fois l'équation (4.5) à la 2ème équation de (4.4) et 2 fois l'équation (4.5) à la 3ème équation de (4.4). On obtient un système équivalent à (4.4) qui est (en répétant (4.5)) :

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1, \\ 2x_2 + 4x_3 = 2, \\ 5x_2 + 12x_3 = 9. \end{cases} \tag{4.6}$$

On observe que les 2 dernières équations de (4.6) font apparaître seulement 2 inconnues x_2 et x_3 (on a éliminé x_1) et on peut recommencer le procédé en laissant inchangée la 1ère équation. On divise donc la 2ème équation de (4.6) par 2 (2ème pivot) et on soustrait à la 3ème équation de (4.6) 5 fois cette nouvelle 2ème équation. On obtient le système équivalent :

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1, \\ x_2 + 2x_3 = 1, \\ 2x_3 = 4. \end{cases} \tag{4.7}$$

Finalement, il suffit de diviser la dernière équation de (4.7) par le 3ème pivot qui est encore ici 2 pour obtenir :

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 1, \\ x_2 + 2x_3 = 1, \\ x_3 = 2. \end{cases} \tag{4.8}$$

De (4.8), il est facile de tirer successivement comme dans (4.3) :

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

4.3 Algorithme d'élimination

L'algorithme d'élimination de Gauss pour résoudre un système de N équations à N inconnues $A\vec{x} = \vec{b}$ s'écrit comme il suit :

Algorithme d'élimination de Gauss

(entrées : $(a_{ij})_{1\le i,j\le N}$ et $(b_j)_{1\le j\le N}$)

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

(4.9)

Dans le tableau ci-dessus, le signe

$$\begin{bmatrix} \textit{Faire } i = 1 \textit{ à } N-1 \\ \cdots \end{bmatrix}$$

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

A noter qu'à la sortie de l'algorithme ci-dessus, seuls les éléments a_{ij} avec $i<j$ ont une signification pour constituer le système triangulaire supérieur à résoudre. Ainsi, on devra encore résoudre le système :

$$\begin{bmatrix} 1 & a_{12} & a_{13} & \cdots & \cdots & a_{1N} \\ 0 & 1 & a_{23} & \cdots & \cdots & a_{2N} \\ 0 & 0 & 1 & \cdots & \cdots & a_{3N} \\ . & & & . & & . \\ . & & \mathrm{O} & & . & . \\ 0 & & & & & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ b_N \end{bmatrix} \qquad (4.10)$$

en utilisant (4.3), les valeurs $(a_{ij})_{1\le i<j\le N}$ et $(b_j)_{1\le j\le N}$ étant les valeurs de sortie de l'algorithme d'élimination.

L'algorithme (4.9) ne peut être exécuté jusqu'à la fin que si les pivots successifs sont non-nuls. La question que l'on peut donc se poser est la suivante : "Quand est-ce que les pivots a_{ii} de l'algorithme (4.9) sont tous non-nuls ?" La réponse suit la définition ci-après :

Définition. B_k est la sous-matrice principale d'ordre k de A si B_k est la $k\times k$-matrice de coefficients $(a_{ij})_{1\le i,j\le k}$, $1\le k\le N$.

Théorème. Si toutes les sous-matrices principales B_k d'éléments $(a_{ij})_{1\le i,j\le k}$ sont régulières, $k=1,2,3,\cdots,N$, alors les pivots obtenus successivement dans l'élimination de Gauss (4.9) sont tous non-nuls.

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 présenté à la section 4.3. A partir de (4.9), nous obtenons :

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

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", $N_m = \frac{N^3}{3}+\mathrm{O}(N^2)$ où $\mathrm{O}(N^2)$ désigne un reste d'ordre N^2 lorsque $N\to\infty$.

Il est facile de vérifier que

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

Ainsi nous aurons

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

ce qui prouve que $N_m = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6}$.

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

> Si un système linéaire est deux fois plus grand qu'un autre (deux fois plus d'équations et d'inconnues), il faudra huit ($=2^3$) fois plus d'opérations pour le résoudre par l'algorithme décrit en section 4.3, 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. Il est évident qu'il est impossible de traiter par cet algorithme le système suivant

$$\begin{cases} 0x_1 + x_2 + 3x_3 = 1, \\ 5x_1 + 2x_2 + 3x_3 = 4, \\ 6x_1 + 8x_2 + x_3 = 1, \end{cases} \tag{4.11}$$

car dans ce cas, on ne peut pas diviser la 1ère ligne par le 1er pivot qui est nul. On voit immédiatement que les choses se présentent mieux si on échange la 1ère et la 3ème ligne pour obtenir

$$\begin{cases} 6x_1 + 8x_2 + x_3 = 1, \\ 5x_1 + 2x_2 + 3x_3 = 4, \\ 0x_1 + x_2 + 3x_3 = 1. \end{cases} \tag{4.12}$$

En effet, maintenant nous pouvons diviser la 1ère 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. On modifie ainsi l'algorithme (4.9) donné en section 4.3 en intercalant entre la 1ère ligne "Faire $i = 1$ à $N-1$" et la 2ème ligne "$p := 1/a_{ii}$" la procédure suivante :

Choix du plus grand pivot en valeur absolue

(à intercaler entre les deux premières lignes de l'algorithme d'élimination de Gauss)

Algorithme	**Commentaires**
$k := i$ $m := abs\ (a_{ii})$	On élimine l'inconnue x_i et donc i est donné. On initialise $m = \lvert a_{ii} \rvert$
Faire $j = i+1$ *à* N	Recherche du plus grand pivot en valeur absolue dans la $i^{ème}$ colonne
$s := abs\ (a_{ji})$ *Si* $m < s$ *alors* $k := j$ *et* $m := s$	Après cette séquence, on sait que le plus grand pivot en valeur absolue est m qui 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	Echange des lignes k et i
$t := a_{ij}$ $a_{ij} := a_{kj}$ $a_{kj} := t$	Echange de a_{ij} et a_{kj}
$t := b_i$ $b_i := b_k$ $b_k := t$	Echange de b_i et b_k

(4.13)

On peut 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, il convient de remarquer qu'il est toujours possible, lorsqu'on est en présence d'un système linéaire, 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 pivot, il convient de procéder à un "équilibrage" de la matrice dans le but que toutes les lignes (toutes les équations) aient un "poids semblable". L'équilibrage consiste à multiplier la $i^{\text{ème}}$ 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

$$\max_{1 \le j \le N} r_i \left| a_{ij} \right| = 1,$$

et ceci pour $i = 1, 2, \cdots, N$.

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 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 $i^{\text{ème}}$ colonne, mais aussi dans la $i^{\text{ème}}$ ligne. On parle dans ce cas de "pivotage complet" (échange de colonnes et de lignes). Le pivotage requiert un nombre d'opérations de l'ordre de N^2 alors que le pivotage complet requiert un nombre de l'ordre de N^3 (donc coûteux !). Ce dernier n'est jamais pratiqué.

4.6 Systèmes mal conditionnés

Considérons le système de 2 équations à 2 inconnues :

$$\begin{cases} 4.218613\, x_1 + 6.327917\, x_2 = 10.546530 \\ 3.141592\, x_1 + 4.712390\, x_2 = 7.853982. \end{cases} \tag{4.14}$$

On vérifie que le système est régulier (déterminant de A non-nul) et que la solution est

$$x_1 = x_2 = 1. \tag{4.15}$$

Considérons maintenant un système d'équations voisin ($\rightarrow$ signifie "changement") :

$$\begin{cases} 4.218611 x_1 + 6.327917 x_2 = 10.546530 \\ 3.141594 x_1 + 4.712390 x_2 = 7.853980. \end{cases} \tag{4.16}$$

On vérifie encore que le système est régulier, mais cette fois-ci la solution est

$$x_1 = -5, \quad x_2 = +5. \tag{4.17}$$

On conclut que bien que les systèmes (4.14) et (4.16) soient "voisins", leurs solutions sont "très différentes". On parle dans ce cas de systèmes mal conditionnés. L'interprétation géométrique donne une explication évidente :

Le système (4.14) est formé de deux équations qui, dans le plan Ox_1,x_2, décrivent deux droites presque parallèles. Ainsi, résoudre le système (4.14) 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 (4.16)), alors le point d'intersection est fortement modifié ! $\Rightarrow$ il faut beaucoup de chiffres significatifs pour palier à ce problème

Résoudre un problème mal conditionné avec un ordinateur peut être une affaire délicate si le calculateur ne calcule pas avec suffisamment de chiffres significatifs. Dans l'exemple ci-dessus, on voit que si le calculateur ne retient que 6 chiffres significatifs, il est complètement désespéré d'obtenir une réponse significative à la résolution du système (4.14). Avec un calculateur à 10 chiffres significatifs, on peut espérer une réponse avec au plus 4 chiffres significatifs corrects. Ci-après, nous établissons une théorie grossière de ce que nous prétendons.

Nous commençons par donner deux définitions.

$\vec{y}$ est un vecteur quelconque

Définition. Soit A une $N \times N$ matrice. Si pour un N-vecteur $\vec{y}$ de composantes $(y_j)_{1 \le j \le N}$ nous notons $\|\vec{y}\| = (\sum_{j=1}^{N} y_j^2)^{1/2}$ sa norme euclidienne, alors la norme spectrale de A est définie par

$$|||A||| = \max_{\vec{y} \neq 0} \frac{\|A\vec{y}\|}{\|\vec{y}\|}. \tag{4.18}$$

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

$$\|A\vec{x}\| \le |||A||| \cdot \|\vec{x}\|. \tag{4.19}$$

En remplaçant (4.25) dans (4.24) et en utilisant la définition (4.20), on obtient :

$$\frac{\| \vec{\delta x} \|}{\| \vec{x} \|} \leq \chi(A) \frac{\| \vec{\delta b} \|}{\| \vec{b} \|} . \tag{4.26}$$

Nous avons prouvé 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 $\chi(A)$. On peut montrer que l'estimation (4.26) est pessimiste (au sens où souvent $\| \vec{\delta x} \| / \| \vec{x} \|$ est nettement plus petit que $\chi(A) \cdot \| \vec{\delta b} \| / \| \vec{b} \|$) mais non améliorable (au sens où il existe des exemples où $\| \vec{\delta x} \| / \| \vec{x} \| = \chi(A) \cdot \| \vec{\delta b} \| / \| \vec{b} \|$).

Cependant, on peut garder la règle suivante : si on calcule la solution de (4.21) avec un calculateur à p chiffres significatifs, 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). Dans l'exemple (4.14), on peut voir que $\chi(A)$ est de l'ordre de 10^7 ce qui signifie "en gros" qu'on peut avoir une perte de 7 chiffres significatifs sur le résultat !

Décomposons $LU - LL^t$:

écrire les matrices et voir pour chaque élément
i.e. a_{ik} = ligne i × colonne k.

Leçon 5

Décomposition *LU*
Décomposition de Cholesky
Matrices de bande

5.1 Décomposition *LU*

Soit A une $N \times N$ matrice dont toutes les sous-matrices principales sont régulières. Dans la leçon 4, nous avons vu alors qu'il est possible de procéder par élimination de Gauss sans choix du pivot pour résoudre un système du type $A\vec{x} = \vec{b}$. En fait, il est possible de montrer dans ce cas-là le résultat suivant :

Théorème. 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. De plus, la matrice U est celle obtenue par l'algorithme d'élimination de Gauss (4.9).

Nous ne démontrons pas ce résultat. Par contre, nous allons donner un algorithme de décomposition LU.

Pour commencer, prenons l'exemple d'une matrice $A = (a_{ij})_{1 \le i,j \le 3}$ composée de 3 lignes et 3 colonnes données. Nous devons déterminer dans ce cas $(\ell_{ij})_{1 \le j \le i \le 3}$ et les 3 nombres $(u_{ij})_{1 \le i < j \le 3}$ qui satisfont le produit matriciel suivant :

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}}_{U}. \tag{5.2}$$

En multipliant L par la 1ère colonne de U, on obtient la 1ère colonne de A. Ainsi on a immédiatement :

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = \begin{bmatrix} \ell_{11} \\ \ell_{21} \\ \ell_{31} \end{bmatrix}. \tag{5.3}$$

En multipliant la 1ère ligne de L par la 2ème et 3ème colonne de U, on obtient respectivement le 2ème et 3ème terme de la 1ère ligne de A. Ainsi on a immédiatement :

$$a_{12} = \ell_{11}\, u_{12} \qquad \text{et} \qquad a_{13} = \ell_{11}\, u_{13},$$

ou, puisque ℓ_{11} est déjà connu

$$u_{12} = a_{12}/\ell_{11}, \quad u_{13} = a_{13}/\ell_{11}. \tag{5.4}$$

Les relations (5.3) et (5.4) donnent donc la 1ère colonne de L et la 1ère ligne de U. Celles-ci étant connues, on peut déterminer la 2ème colonne de L en multipliant la 2ème et 3ème ligne de L par la 2ème colonne de U (qui est connue) pour obtenir :

$$\ell_{21}\, u_{12} + \ell_{22} = a_{22} \qquad \text{et} \qquad \ell_{31}\, u_{12} + \ell_{32} = a_{32}$$

ou, de façon équivalente,

$$\ell_{22} = a_{22} - \ell_{21}\, u_{12} \qquad \text{et} \qquad \ell_{32} = a_{32} - \ell_{31}\, u_{12}. \tag{5.5}$$

Pour déterminer la 2ème ligne de U, il suffit de multiplier la 2ème ligne de L par la 3ème colonne de U pour obtenir :

$$\ell_{21}\, u_{13} + \ell_{22}\, u_{23} = a_{23}$$

ou, de façon équivalente,

$$u_{23} = (a_{23} - \ell_{21} u_{13}) / \ell_{22}. \tag{5.6}$$

Les relations (5.5) et (5.6) déterminent la $2^{\text{ème}}$ colonne de L et la $2^{\text{ème}}$ ligne de U. Enfin ℓ_{33} est donné par

$$\ell_{33} = a_{33} - \ell_{31} u_{13} - \ell_{32} u_{23}. \tag{5.7}$$

Si A est une $N \times N$ matrice, on voit que si les k premières colonnes de L et les k premières lignes de U sont connues ($1 \leq k < N$), on peut calculer la $(k+1)^{\text{ème}}$ colonne de L et la $(k+1)^{\text{ème}}$ ligne de U. De plus, il est possible de stocker les deux matrices L et U dans le tableau A de la façon suivante :

$$\begin{bmatrix} \ell_{11} & u_{12} & u_{13} & u_{14} & \cdots & u_{1N} \\ \ell_{21} & \ell_{22} & u_{23} & u_{24} & \cdots & u_{2N} \\ \ell_{31} & \ell_{32} & \ell_{33} & u_{34} & \cdots & u_{3N} \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \ell_{N1} & \ell_{N2} & \cdots & \cdots & \cdots & \ell_{NN} \end{bmatrix}. \tag{5.8}$$

L'algorithme LU peut se faire en définissant un seul $N \times N$ tableau qui, au départ, contient la matrice A et, à la sortie, contient l'information LU sous la forme (5.8). On vérifie facilement que l'algorithme LU peut s'écrire comme suit :

Algorithme *LU*

(entrées : $(a_{ij})_{1\le i,j\le N}$ représentant la matrice A ;
sorties : $(a_{ij})_{1\le j\le i\le N}$ représentant la matrice L
et $(a_{ij})_{1\le i<j<N}$ représentant la matrice U)

Algorithme	**Commentaires**
Faire $i = 2$ *à* N $a_{1i} := a_{1i} / a_{11}$	La 1ère colonne de A est la 1ère colonne de L. Ici on forme la 1ère ligne de U
Faire $k = 2$ *à* $N-1$	Parcours des 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 $a_{ik} := a_{ik} - \sum_{j=1}^{k-1} a_{ij} * a_{jk}$	Construction de la k ème colonne de L
$a_{ki} := \frac{1}{a_{kk}}(a_{ki} - \sum_{j=1}^{k-1} a_{kj} * a_{ji})$	Construction de la k ème ligne de U
$a_{NN} := a_{NN} - \sum_{j=1}^{N-1} a_{Nj} * a_{jN}$	Construction du dernier pivot ℓ_{NN}

(5.9)

On montre que le nombre d'opérations est d'ordre N^3 comme pour l'élimination de Gauss (4.9).

5.2 Utilité de la décomposition de *LU*

Supposons que l'on doive résoudre m systèmes linéaires

$$A\vec{x}^{(\ell)} = \vec{b}^{(\ell)}, \quad \ell = 1,2,\cdots,m, \tag{5.10}$$

où, pour $\ell = 2,3,\cdots,m$, $\vec{b}^{(\ell)}$ dépend des solutions précédentes $\vec{x}^{(1)}, \vec{x}^{(2)}, \cdots, \vec{x}^{(\ell-1)}$.

- Par exemple si $m = 2$, on pourrait vouloir résoudre

$$A\vec{x}^{(1)} = \vec{b}^{(1)},$$

puis

$$A\vec{x}^{(2)} = \left\| \vec{x}^{(1)} \right\|^2 \vec{x}^{(1)},$$

où $\vec{b}^{(1)}$ est un N-vecteur donné (ici $\vec{b}^{(2)} = \left\| \vec{x}^{(1)} \right\|^2 \vec{x}^{(1)}$).

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 de faire une décomposition $A = LU$ et (5.10) peut s'écrire

$$LU\,\vec{x}^{(\ell)} = \vec{b}^{(\ell)}, \quad \ell = 1,2,\cdots,m. \tag{5.11}$$

En posant

$$\vec{y}^{(\ell)} = U\vec{x}^{(\ell)},$$

il suffit de résoudre successivement pour $\ell = 1,2,\cdots,m$:

$$L\vec{y}^{(\ell)} = \vec{b}^{(\ell)}, \tag{5.12}$$

puis

$$U\vec{x}^{(\ell)} = \vec{y}^{(\ell)}. \tag{5.13}$$

Le système (5.12) est un système triangulaire inférieur alors que (5.13) est un système triangulaire supérieur ; tous les deux sont facilement solubles.

Dans la leçon 6, nous verrons des applications de ce qui précède.

Avant de terminer cette section, il faut mentionner que si on a la décomposition $A = LU$, alors $\det(A) = \det(L) \cdot \det(U)$. Puisque U n'a que des 1 dans sa diagonale, on a $\det(U) = 1$. Le déterminant de L est donné par $\det(L) = \ell_{11} \cdot \ell_{22} \cdot \ell_{33} \cdot \cdots \cdot \ell_{NN}$ puisque L est triangulaire. Ainsi

$$\det(A) = \prod_{j=1}^{N} \ell_{jj}. \tag{5.14}$$

Il est important de mentionner ici 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 une machine du type CRAY-2, on peut faire au plus 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. Par décomposition LU de A, le calcul du déterminant d'une matrice A d'ordre 100 exigera sur un CRAY-2 un temps CPU de l'ordre de la 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 ($1 \leq k < N$), on trouve le pivot a_{kk} nul et par suite une division par zéro dans l'algorithme (5.9). Comme nous l'avons fait pour l'élimination de Gauss, il conviendra dans ce cas d'échanger la $k^{\text{ème}}$ ligne de A avec une $j^{\text{ème}}$ ($j > k$) pour assurer un pivot non-nul. Clairement si nous échangeons des lignes tout comme dans l'élimination de Gauss avec choix du pivot, 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 le faire, on définit un N-vecteur $\vec{p}$ (dit de permutation) et on pose

$$p_k = j$$

pour signifier que la ligne k a été échangée avec la $j^{\text{ème}}$.

5.4 Matrices symétriques définies positives Décomposition de Cholesky

Dans la suite, le symbole ^{T} qui accompagne une matrice ou un vecteur signifie que l'on a affaire à son transposé. Rappelons la définition suivante :

Définition. Une $N \times N$ matrice A est dite symétrique définie positive si :

(i) $A = A^{\mathrm{T}}$ (A est symétrique),

(ii) $\vec{y}^{\mathrm{T}} A\vec{y} \geq 0$ pour tout N-vecteur $\vec{y}$,

(iii) si $\vec{y}^{\mathrm{T}} A\vec{y} = 0$ alors nécessairement $\vec{y} = 0$.

Une matrice A symétrique définie positive a toutes ses sous-matrices principales symétriques définies positives, donc régulières. Ainsi, on peut toujours faire une décomposition $A = LU$ d'une matrice A symétrique définie positive. On peut même montrer que les pivots a_{kk} obtenus dans l'algorithme (5.9) vérifient la relation $\prod_{j=1}^{k} a_{jj} = \det(B_k)$, où B_k est la sous-matrice principale d'ordre k de A (cf. paragraphe 4.3), et donc ces pivots a_{kk} sont strictement positifs. Si maintenant, pour $k = 1, 2, \cdots, N$, nous multiplions la $k^{\text{ème}}$ ligne de U par la racine carrée du $k^{\text{ème}}$ pivot et nous divisons la $k^{\text{ème}}$ colonne de L par cette même valeur, nous obtenons respectivement des matrices $\tilde{U}$ et $\tilde{L}$ qui sont toujours triangulaires supérieure et inférieure et nous avons toujours $A = \tilde{L}\tilde{U}$. Puisque la diagonale de L contient les pivots et celle de U des 1, on peut voir ainsi que les diagonales de $\tilde{L}$ et $\tilde{U}$ sont les mêmes. Un raisonnement tenant compte de l'unicité de la décomposition LU et de la symétrie de A nous montre que $\tilde{L}^{\mathrm{T}} = \tilde{U}$. On obtient donc le théorème de décomposition de Cholesky :

Théorème. 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 $A = LL^{\mathrm{T}}$.

⚠ Dans le théorème énoncé ci-dessus, on a noté L au lieu de $\tilde{L}$ la matrice triangulaire inférieure qui satisfait $A = LL^{\mathrm{T}}$. Evidemment, cette matrice L n'est en principe pas la même que celle obtenue dans la décomposition LU puisqu'il s'agit en fait de $\tilde{L}$.

De façon semblable à l'algorithme (5.9), on peut écrire l'algorithme de Cholesky (5.15). Il suffit de voir dans (5.9) que la $k^{\text{ème}}$ colonne de L est divisée par $\sqrt{a_{kk}}$ et de tenir compte de la relation $a_{kj} = a_{jk}$ puisque A est symétrique.

Remarque. Si, dans l'algorithme ci-après, 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 (5.15).

Algorithme LL^T

(entrées : $(a_{ij})_{1\le j\le i\le N}$ représentant la partie triangulaire inférieure de A

(A est supposée symétrique) ;

sorties : $(a_{ij})_{1\le j\le i\le N}$ représentant L qui satisfait $A = LL^T$)

Algorithme	**Commentaires**
$a_{11} := \sqrt{a_{11}}$	Construction de ℓ_{11}
Faire $i = 2$ *à* N $a_{i1} := a_{i1} / a_{11}$	Construction de la 1ère colonne de L
Faire $k = 2$ *à* $N-1$	Parcours des colonnes de L
$a_{kk} := (a_{kk} - \sum_{j=1}^{k-1} a_{kj}^2)^{1/2}$	Construction de ℓ_{kk}
Faire $i = k+1$ *à* N $a_{ik} := \frac{1}{a_{kk}}(a_{ik} - \sum_{j=1}^{k-1} a_{ij} * a_{kj})$	Construction de la k ème colonne de L
$a_{NN} := (a_{NN} - \sum_{j=1}^{N-1} a_{Nj}^2)^{1/2}$	Construction de ℓ_{NN}

(5.15)

Lorsqu'on veut résoudre un système symétrique défini positif

$$A\vec{x} = \vec{b},$$

on fait toujours une décomposition de Cholesky $A = LL^T$ et on résout successivement

$$L\vec{y} = \vec{b},$$

puis

$$L^T\vec{x} = \vec{y}.$$

5.5 Matrices de bande

Nous commençons par donner une définition.

Définition. Soit A une $N \times N$ matrice de coefficients $(a_{ij})_{1 \le i,j \le N}$ 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 $1 \le i, j \le N$ et $|i - j| \ge \ell$.

Une matrice de bande de demi-largeur ℓ est donc de la forme

$$A = \begin{bmatrix} * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & 0 & 0 \\ 0 & * & * & * & * & * & * & * & * & * & 0 \\ 0 & 0 & * & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & * \end{bmatrix}$$

ℓ lignes

ℓ colonnes

où ici $*$ mentionne la place d'un élément non nécessairement nul et 0 mentionne la place d'un élément nécessairement nul.

Une matrice de bande de demi-largeur 1 est une matrice diagonale alors que si la demi-largeur est 2, on dit que la matrice est tridiagonale.

On vérifie facilement, en reprenant les algorithmes LU (5.9) et LL^T (5.15), qu'on a le résultat suivant :

Théorème. 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 $N\ell^2$.

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 suivant.

Exemple. Soit la $N \times N$ matrice tridiagonale A donnée par

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

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \\ a_{21} & a_{22} & a_{23} & \\ a_{31} & & \ddots & \\ & & & a_{nn} \end{pmatrix}$$

$\Rightarrow L = (l_{ij}) = 0 \quad \forall i<j$

On voit immédiatement que A est symétrique. De plus, si on calcule $\vec{y}^{\mathrm{T}} A\vec{y}$, pour un N-vecteur $\vec{y}$ de composantes $y_1, y_2, \cdots, y_N$, on obtient :

$$\begin{aligned} \vec{y}^{\mathrm{T}} A\vec{y} &= y_1(2y_1 - y_2) + y_2(-y_1 + 2y_2 - y_3) + y_3(-y_2 + 2y_3 - y_4) + \cdots + \\ &+ y_{N-1}(-y_{N-2} + 2y_{N-1} - y_N) + y_N(-y_{N-1} + 2y_N); \end{aligned}$$

et donc

$$\vec{y}^{\mathrm{T}} A\vec{y} = y_1^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + \cdots + (y_{N-1} - y_N)^2 + y_N^2.$$

Clairement, on a bien $\vec{y}^{\mathrm{T}} A\vec{y} \geq 0$ et $\vec{y}^{\mathrm{T}} A\vec{y} = 0$ implique $\vec{y} = 0$; la matrice A est donc symétrique définie positive. Il est donc possible de faire une décomposition de Cholesky $A = LL^{\mathrm{T}}$. D'après le théorème qui précède, on peut utiliser deux vecteurs $\vec{d} = (d_j)_{1 \leq j \leq N}$ et $\vec{e} = (e_j)_{1 \leq j \leq N-1}$ pour mémoriser L de la façon suivante :

$$L = \begin{bmatrix} d_1 & & & & & \\ e_1 & d_2 & & & \mathrm{O} & \\ & e_2 & d_3 & & & \\ & & e_3 & d_4 & & \\ & \mathrm{O} & & \ddots & \ddots & \\ & & & & e_{N-1} & d_N \end{bmatrix}.$$

En faisant le produit LL^{T} et en l'égalant à A, on obtient successivement :

$$d_1^2 = 2\,;\ e_1 d_1 = -1\,;\ e_1^2 + d_2^2 = 2\,;\ e_2 d_2 = -1\,;\ e_2^2 + d_3^2 = 2\,;\ e_3 d_3 = -1\,;\ \cdots\,;\ e_{N-1}^2 + d_N^2 = 2.$$

Ainsi on obtient :

$$d_1 = \sqrt{2}$$

et pour $j = 1,2,\cdots,N-1$:

$$e_j = -1/d_j \qquad \text{et} \qquad d_{j+1} = \sqrt{2 - e_j^2}\,.$$

Leçon 6

Méthode de la puissance inverse pour le calcul des valeurs propres

Méthode des moindres carrés pour les systèmes surdéterminés

6.1 Méthode de la puissance

Soit A une $N \times N$ matrice symétrique dont les valeurs propres $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_N$ (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 $\vec{\varphi}_1, \vec{\varphi}_2, \cdots, \vec{\varphi}_N$ est une base de vecteurs propres orthonormalisée correspondante, on aura

$$A\vec{\varphi}_j = \lambda_j \vec{\varphi}_j, \quad 1 \leq j \leq N, \tag{6.1}$$

et

$$\vec{\varphi}_j{}^{\mathrm{T}} \vec{\varphi}_k = \delta_{jk}, \quad 1 \leq j,k \leq N, \tag{6.2}$$

où δ_{jk} est le symbole de Kronecker qui vaut 1 si $j = k$ et 0 autrement. Si $\vec{x}^0$ est un N-vecteur donné, on peut le décomposer dans la base des $(\vec{\varphi}_j)_{1 \leq j \leq N}$ sous la forme

$$\vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j \vec{\varphi}_j \tag{6.3}$$

où α_j est donné par

$$\alpha_j = \vec{\varphi}_j{}^{\mathrm{T}} \vec{x}^{(0)}, \quad 1 \leq j \leq N. \tag{6.4}$$

Si, pour $n = 1,2,3,\cdots$, on définit

$$\vec{x}^{(n)} = A\vec{x}^{(n-1)}, \tag{6.5}$$

on obtient, en utilisant (6.1) et (6.3) :

$$\vec{x}^{(n)} = A^n \vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j \lambda_j^n \vec{\varphi}_j . \tag{6.6}$$

En supposant maintenant que

$$|\lambda_1| > |\lambda_j|, \quad \forall j = 2,3,\cdots,N, \tag{6.7}$$

et

$$\alpha_1 \equiv \vec{\varphi}_1{}^{\mathrm{T}} \vec{x}^{(0)} \neq 0, \tag{6.8}$$

on montre facilement, à partir de (6.6), que

$$\lim_{n\to\infty} \frac{\vec{x}^{(n)}}{\lambda_1^n} = \sum_{j=1}^{N} \lim_{n\to\infty} \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^n \vec{\varphi}_j = \alpha_1 \vec{\varphi}_1 . \tag{6.9}$$

En utilisant (6.9) on obtient :

$$\lim_{n\to\infty} \frac{\vec{x}^{(2n)}}{\| \vec{x}^{(2n)} \|} = \lim_{n\to\infty} \frac{\vec{x}^{(2n)}/\lambda_1^{2n}}{\| \vec{x}^{(2n)}/\lambda_1^{2n} \|} = \frac{\alpha_1}{|\alpha_1|} \vec{\varphi}_1 = \pm \vec{\varphi}_1 . \tag{6.10}$$

Ainsi, si λ_1 est la plus grande valeur propre de A en valeur absolue et si $\vec{x}^{(0)}$ n'est pas choisi orthogonalement au vecteur propre $\vec{\varphi}_1$ correspondant, alors la suite des vecteurs normalisés $\left(\frac{\vec{x}^{(2n)}}{\| \vec{x}^{(2n)} \|}\right)_{n=0}^{\infty}$ définie au moyen de (6.5) converge vers $\vec{\varphi}_1$ ou $-\vec{\varphi}_1$.

En utilisant (6.9), il est facile de voir que la suite des termes impairs $\left(\frac{\vec{x}^{(2n+1)}}{\| \vec{x}^{(2n+1)} \|}\right)_{n=0}^{\infty}$ converge aussi vers $\vec{\varphi}_1$ ou $-\vec{\varphi}_1$; la limite est la même que la précédente si $\lambda_1 > 0$ et est de signe opposé si $\lambda_1 < 0$.

Définissons maintenant la quantité (dite quotient de Rayleigh) :

$$\boxed{\mu_n = \frac{\vec{x}^{(n)\mathrm{T}} A \vec{x}^{(n)}}{\| \vec{x}^{(n)} \|^2} = \frac{\vec{x}^{(n)\mathrm{T}} \cdot \vec{x}^{(n+1)}}{\vec{x}^{(n)\mathrm{T}} \vec{x}^{(n)}}.} \tag{6.11}$$

En utilisant (6.2) et (6.6), nous vérifions les calculs suivants :

$$\mu_n = \frac{\bar{x}^{(n)T}\,\bar{x}^{(n+1)}}{\bar{x}^{(n)T}\,\bar{x}^{(n)}} = \frac{\sum_{j=1}^{N}\alpha_j^2\,\lambda_j^{2n+1}}{\sum_{j=1}^{N}\alpha_j^2\,\lambda_j^{2n}} =$$

$$= \lambda_1 \frac{\sum_{j=1}^{N}\alpha_j^2\,(\frac{\lambda_j}{\lambda_1})^{2n+1}}{\sum_{j=1}^{N}\alpha_j^2\,(\frac{\lambda_j}{\lambda_1})^{2n}} = \lambda_1 \frac{\alpha_1^2 + \sum_{j=2}^{N}\alpha_j^2\,(\frac{\lambda_j}{\lambda_1})^{2n+1}}{\alpha_1^2 + \sum_{j=2}^{N}\alpha_j^2\,(\frac{\lambda_j}{\lambda_1})^{2n}} =$$

$$\mu_n = \lambda_1 \frac{1+\sum_{j=2}^{N}(\frac{\alpha_j}{\alpha_1})^2\,(\frac{\lambda_j}{\lambda_1})^{2n+1}}{1+\sum_{j=2}^{N}(\frac{\alpha_j}{\alpha_1})^2\,(\frac{\lambda_j}{\lambda_1})^{2n}} = \lambda_1 \frac{1+\delta}{1+\varepsilon} \qquad (6.12)$$

où $\delta = \sum_{j=2}^{N}(\frac{\alpha_j}{\alpha_1})^2\,(\frac{\lambda_j}{\lambda_1})^{2n+1}$ et $\varepsilon = \sum_{j=2}^{N}(\frac{\alpha_j}{\alpha_1})^2\,(\frac{\lambda_j}{\lambda_1})^{2n}$.

En tenant compte de (6.7), nous avons :

$$\left|\frac{\lambda_N}{\lambda_1}\right| \le \left|\frac{\lambda_{N-1}}{\lambda_1}\right| \le \cdots \le \left|\frac{\lambda_2}{\lambda_1}\right| < 1,$$

et on en déduit que δ et ε tendent vers zéro lorsque n tend vers l'infini puisque :

$$|\delta| \le \left|\frac{\lambda_2}{\lambda_1}\right|^{2n+1} \sum_{j=2}^{N}\left(\frac{\alpha_j}{\alpha_1}\right)^2, \quad |\varepsilon| \le \left|\frac{\lambda_2}{\lambda_1}\right|^{2n} \sum_{j=2}^{N}\left(\frac{\alpha_j}{\alpha_1}\right)^2.$$

En utilisant la relation

$$\frac{1}{1+\varepsilon} = 1-\varepsilon+O(\varepsilon^2) \qquad \text{si } \varepsilon \to 0,$$

nous obtenons à partir de (6.12) :

$$\mu_n = \lambda_1 + O(|\varepsilon|+|\delta|)$$

d'où l'on déduit d'après les majorations de $|\delta|$ et $|\varepsilon|$ que :

$$\mu_n = \lambda_1 + O\left(\frac{\lambda_2}{\lambda_1}\right)^{2n} \qquad \text{si } n \to +\infty. \qquad (6.13)$$

La relation (6.13) signifie mathématiquement qu'il existe une constante C (indépendante de n) telle que

$$|\lambda_1 - \mu_n| \leq C\left(\frac{\lambda_2}{\lambda_1}\right)^{2n} \tag{6.14}$$

et puisque $n \to \infty$, on a $\lim_{n\to\infty} \mu_n = \lambda_1$.

Ainsi, si λ_1 est la plus grande valeur propre de A en valeur absolue et si $\vec{x}^{(0)}$ n'est pas choisi orthogonalement au vecteur propre $\vec{\varphi}_1$ correspondant, alors la suite des quotients de Rayleigh μ_n donnée par (6.11) converge vers λ_1.

Pour finir cette section, nous remarquons que (6.10) et (6.14) nous permettent de calculer une approximation numérique du vecteur propre et de la valeur propre de A lorsque cette dernière est supposée être la plus grande en valeur absolue (cette méthode porte le nom de "méthode de la puissance" à cause de (6.6)).

Si maintenant on suppose

$$|\lambda_1| = |\lambda_2| > |\lambda_3| \geq |\lambda_4| \geq \cdots \geq |\lambda_N|,$$

on peut avoir deux cas :

- ou bien λ_1 est valeur propre de multiplicité 2 ($\lambda_1 = \lambda_2$) et il lui correspond un sous-espace propre (plan) engendré par 2 vecteurs propres linéairement indépendants correspondant à λ_1 ;

- ou bien on a $\lambda_1 = -\lambda_2$ et ces deux valeurs propres λ_1 et λ_2 ont une multiplicité 1.

Dans le 1er cas, on peut remarquer que la méthode marche toujours. De façon semblable à (6.9), nous aurons $\lim_{n\to\infty} \frac{\vec{x}^{(n)}}{\lambda_1^n} = \alpha_1 \vec{\varphi}_1 + \alpha_2 \vec{\varphi}_2$ et ainsi le vecteur $\vec{x}^{(n)}$ se "couche" dans le sous-espace propre lorsque n tend vers l'infini. On vérifie que $\lim_{n\to\infty} \mu_n = \lambda_1$. Dans le 2ème cas, la méthode ne marche plus ; on devra itérer sur deux vecteurs à la fois. Une autre méthode est d'ajouter à A la matrice εI où I est la $N \times N$ matrice identité et ε est un réel (shift des valeurs propres). Clairement les valeurs propres de $A + \varepsilon I$ sont données par $\lambda_j + \varepsilon, 1 \leq j \leq N$, et ce sont ces valeurs qui vont nous intéresser lorsqu'on considérera la méthode de la puissance sur la matrice $A + \varepsilon I$ au lieu de A.

La méthode de la puissance peut être généralisée à des matrices non nécessairement symétriques.

6.2 Méthode de la puissance inverse

permet de calculer toutes les valeurs propres

Soit A une $N \times N$ matrice symétrique dont les valeurs propres sont $\lambda_1, \lambda_2, \cdots, \lambda_N$ et les vecteurs orthonormalisés correspondants sont $\vec{\varphi}_1, \vec{\varphi}_2, \cdots, \vec{\varphi}_N$ et soit μ un nombre réel tel que

$$\mu \neq \lambda_j, \quad 1 \leq j \leq N. \tag{6.15}$$

Si I est la $N \times N$ matrice identité, on vérifie par (6.15) que la matrice $(A - \mu I)^{-1}$ existe et que les valeurs propres ω_j, $1 \leq j \leq N$, de cette dernière sont

$$\omega_j = \left(\lambda_j - \mu\right)^{-1}, \quad 1 \leq j \leq N. \tag{6.16}$$

Supposons maintenant qu'il existe k tel que

$$\left|\lambda_k - \mu\right| < \left|\lambda_j - \mu\right|, \quad \forall j = 1,2,\cdots,N\,;\, j \neq k. \tag{6.17}$$

Emettre l'hypothèse (6.17) est équivalent à dire que la valeur propre λ_k la plus proche de μ est de multiplicité 1 et que $2\mu - \lambda_k$ n'est pas valeur propre de A.

En considérant (6.16) et (6.17), nous aurons

$$\left|\omega_k\right| > \left|\omega_j\right|, \quad \forall j = 1,2,\cdots,N\,;\, j \neq k; \tag{6.18}$$

clairement, la méthode de la puissance sur la matrice $(A - \mu I)^{-1}$ nous permettra de déterminer ω_k et par suite, en utilisant (6.16), nous aurons λ_k. La méthode nous donnera aussi le vecteur propre correspondant car les vecteurs propres de A sont ceux de $(A - \mu I)^{-1}$.

Remplaçons donc, dans (6.5), la matrice A par la matrice $(A - \mu I)^{-1}$. Nous obtenons si $\vec{x}^{(0)}$ est donné et pour $n = 1,2,\cdots$:

$$\vec{x}^{(n)} = (A - \mu I)^{-1}\, \vec{x}^{(n-1)}. \tag{6.19}$$

En décomposant $\vec{x}^{(0)}$ dans la base des vecteurs propres de A comme fait dans (6.3), nous obtenons

$$\vec{x}^{(n)} = (A - \mu I)^{-n}\, \vec{x}^{(0)} = \sum_{j=1}^{N} \alpha_j\, \omega_j^n\, \vec{\varphi}_j. \tag{6.20}$$

En définissant le quotient de Rayleigh

$$\mu_n = \frac{\vec{x}^{(n)\mathrm{T}} A \vec{x}^{(n)}}{\| \vec{x}^{(n)} \|^2}, \tag{6.21}$$

on montre, en utilisant (6.20) et de façon semblable à ce qui a été fait pour calculer (6.12), que

$$\mu_n - \mu = \frac{\vec{x}^{(n)\mathrm{T}}(A-\mu I)\vec{x}^{(n)}}{\| \vec{x}^{(n)} \|^2} = (\lambda_k - \mu)\frac{1+\sum\limits_{j\neq k}\left(\frac{\alpha_j}{\alpha_k}\right)^2\left(\frac{\omega_j}{\omega_k}\right)^{2n-1}}{1+\sum\limits_{j\neq k}\left(\frac{\alpha_j}{\alpha_k}\right)^2\left(\frac{\omega_j}{\omega_k}\right)^{2n}} \tag{6.22}$$

et par suite

$$\lim_{n\to\infty} \mu_n = \lambda_k . \tag{6.23}$$

Remarquons encore que $\frac{\omega_j}{\omega_k} = \frac{\lambda_k - \mu}{\lambda_j - \mu}$ et la relation (6.22) nous indique que plus μ est choisi proche de λ_k, plus rapide sera la convergence dans (6.23). La méthode de la puissance inverse consistera à changer μ au cours des itérations pour le prendre égal au quotient de Rayleigh.

Ainsi, la méthode devient, après avoir choisi un vecteur $\vec{x}^{(0)}$ (approximation du vecteur propre $\vec{\varphi}_k$) de départ :

pour $n = 1,2,3,\cdots$, calculer

- $\mu_{n-1} = \dfrac{\vec{x}^{(n-1)\mathrm{T}} A \vec{x}^{(n-1)}}{\| \vec{x}^{(n-1)} \|^2}$,
- $\vec{x}^{(n)} = (A - \mu_{n-1} I)^{-1} \vec{x}^{(n-1)}$.

Mentionnons encore qu'en pratique on calcule $\vec{x}^{(n)}$ en résolvant le système $(A - \mu_{n-1} I)\, \vec{x}^{(n)} = \vec{x}^{(n-1)}$ et de plus, on multiplie $\vec{x}^{(n)}$ par un nombre r_n tel que $\| r_n \vec{x}^{(n)} \| = 1$ pour éviter que les vecteurs grandissent au cours des itérations !

Il existe bien d'autres méthodes de calcul des valeurs propres d'une matrice (Givens-Householder, Jacobi, QR, ...), mais nous n'en parlerons pas ici !

6.3 Systèmes surdéterminés Méthode des moindres carrés

Soit A une $M \times N$ matrice dont le nombre M de lignes est plus grand que le nombre N de colonnes. 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 :

On appelle cette méthode la méthode des moindres carrés parce que si par exemple on veut résoudre

$$\begin{cases} \ln m_1 = \ln(a) - b \\ \ln m_2 = \ln(a) - 2b \\ \ln m_3 = \ln(a) - 3b \end{cases} \quad , a, b = ?$$

Cela revient à <u>minimiser</u> :

$$\sum_{i=1}^{3} \left(\ln(a) - b \cdot i - \ln m_i \right)^2 = mm.$$

$$\underbrace{\begin{pmatrix} 1 & -1 \\ 1 & -2 \\ 1 & -3 \end{pmatrix}}_{A} \cdot \underbrace{\begin{pmatrix} \ln(a) \\ b \end{pmatrix}}_{\vec{x}'} = \underbrace{\begin{pmatrix} \ln m_1 \\ \ln m_2 \\ \ln m_3 \end{pmatrix}}_{\vec{b}'}$$

$$\| A \cdot \vec{x}' - \vec{b}' \| \leq \| A \cdot \vec{y}' - \vec{b}' \| \quad \forall y \in V$$

↓

ceci s'écrit :

$$\left[(\ln(a) - b - \ln m_1)^2 + (\ln(a) - 2b - \ln m_2)^2 + (\ln(a) - 3b - \ln m_3)^2 \right]^{1/2}$$

donc on recherche bien à minimiser :

$$\sum_{i=1}^{3} \left(\ln(a) - b \cdot i - \ln m_i \right)^2 = min$$

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

$$\| A\vec{x} - \vec{b} \| \leq \| A\vec{y} - \vec{b} \|. \tag{6.24}$$

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. Supposons que A soit une $M \times N$ matrice ($M > N$) de rang N. Alors il existe un et un seul vecteur $\vec{x}$ qui satisfait (6.24). De plus, ce vecteur $\vec{x}$ est donné comme solution du système $A^T A\vec{x} = A^T \vec{b}$.

Démonstration. Remarquons tout d'abord que la matrice $B \equiv A^T A$ 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

$$\vec{z}^T B\vec{z} = \vec{z}^T A^T A\vec{z} = \| A\vec{z} \|^2. \tag{6.25}$$

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 (6.25) garantissent que $A^T A$ est symétrique, définie positive (voir leçon 5) et donc régulière. Ainsi il existe un et un seul N-vecteur $\vec{x}$ tel que

$$A^T A\vec{x} = A^T \vec{b}. \tag{6.26}$$

Montrons que, lorsque $\vec{x}$ est solution de (6.26), alors pour tout N-vecteur $\vec{y}$ on a :

$$\| A\vec{x} - \vec{b} \| \leq \| A\vec{y} - \vec{b} \|. \tag{6.27}$$

On a, si on pose $\vec{z} = \vec{x} - \vec{y}$:

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

Comme $\vec{x}$ est solution de (6.26), on obtient :

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

et donc la relation (6.27) est bien vérifiée.

Il n'est pas très difficile de montrer, en faisant des calculs semblables, que si $\vec{x}$ satisfait (6.27), alors il satisfait nécessairement (6.26) et est donc l'unique minimum recherché.

#

$A \in M_n(K)$

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

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

en opérant au préalable une décomposition de Cholesky sur la matrice $A^T A$.

Remarquons encore que dans le cas où $M = N$ et lorsque A est une matrice régulière, non-symétrique, on peut penser a priori trouver plus favorable de résoudre $A^T A\vec{x} = A^T \vec{b}$ au lieu de $A\vec{x} = \vec{b}$ puisque ce premier système est un système symétrique équivalent au deuxième. Cependant, nous ne le faisons jamais car, entre autre, ce système symétrique est souvent mal conditionné. En effet, le nombre de condition spectral de $A^T A$ (voir leçon 4) est le carré de $\chi(A)$ comme le montre le calcul suivant :

$$\chi(A^T A) = |||A^T A||| \cdot |||A^{-1} A^{-T}|||,$$

or

$$|||A^T A||| = \sup_{\vec{x} \neq 0} \frac{\vec{x}^T A^T A\vec{x}}{\vec{x}^T \vec{x}} = \sup_{\vec{x} \neq 0} \frac{\| A\vec{x} \|^2}{\| \vec{x} \|^2} = |||A|||^2,$$

et

$$|||A^{-1} A^{-T}||| = \sup_{\vec{x} \neq 0} \frac{\vec{x}^T A^{-1} A^{-T} \vec{x}}{\vec{x}^T \vec{x}} = \sup_{\vec{x} \neq 0} \frac{\| A^{-T} \vec{x} \|^2}{\| \vec{x} \|^2} = |||A^{-T}|||^2 = |||A^{-1}|||^2,$$

d'où nous déduisons que :

$$\chi(A^T A) = |||A|||^2 \cdot |||A^{-1}|||^2 = \chi(A)^2. \qquad (6.28)$$

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

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

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

$$\|\vec{r}(\vec{x})\| \leq \|\vec{r}(\vec{y})\|, \qquad \forall \vec{y}\ N\text{-vecteur}, \tag{6.29}$$

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

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

$$\sum_{i=1}^{M} p_i r_i^2(\vec{y}) \tag{6.30}$$

au lieu de $\sum_{i=1}^{M} r_i^2(\vec{y})$ comme fait dans (6.29). Nous dirons dans ce cas que nous cherchons une solution de $A\vec{x} = \vec{b}$ au sens des moindres carrés avec poids. Si D est la $M \times M$ matrice diagonale donnée par

$$D = diag\left(\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_M}\right),$$

il est facile de voir que

$$\sum_{i=1}^{M} p_i r_i^2(\vec{y}) = \vec{r}^{\,\mathrm{T}}(\vec{y}) D^2 \vec{r}(\vec{y}) = \| D(A\vec{y} - \vec{b}) \|^2. \tag{6.31}$$

Ainsi, on cherche un N-vecteur $\vec{x}$ tel que $\| DA\vec{x} - D\vec{b} \| = \min$! 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

$$A^{\mathrm{T}} D^2 A\vec{x} = A^{\mathrm{T}} D^2 \vec{b} \tag{6.32}$$

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

Leçon 7

Equations et systèmes d'équations non linéaires

7.1 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. $f(\bar{x}) = 0$. La méthode numérique consistera alors à

(*i*) localiser grossièrement le ou les zéros de f en procédant à des évaluations qui souvent sont de type graphique ; on appellera x_0 cette solution "grossière" ;

(*ii*) construire, à partir de x_0, une suite $x_1, x_2, x_3, \cdots, x_n, \cdots$, telle que $\lim_{n\to\infty} x_n = \bar{x}$ avec $\bar{x}$ satisfaisant $f(\bar{x}) = 0$; la méthode sera qualifiée dans ce cas de convergente.

Définition. Si la méthode est convergente, on dira qu'elle est d'ordre p où p est un entier positif, s'il existe une constante C telle que

$$\underbrace{|\bar{x} - x_{n+1}|}_{\varepsilon} \leq C |\bar{x} - x_n|^p.$$

- Si $p = 1$ (et $C < 1$) on parlera de convergence linéaire.
- Si $p = 2$ on parlera de convergence quadratique.
- 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 parlera de convergence surlinéaire.

Exemple. $f(x) = x - \cos x$.

$f(x) = 0 \iff g(x) = x - f(x) = x$

par trouver un zéro de cos x :

$f(x) = x - \cos x = 0$

$\cos \bar{x} = \bar{x}$

En écrivant $f(\bar{x}) = 0$, on peut aussi écrire $\bar{x} = \cos \bar{x}$. Sur la figure 7.1, on voit immédiatement qu'il y a un seul zéro de f et qu'il est compris entre 0 et $\pi/2$. On peut prendre $x_0 = 0.75$ par exemple.

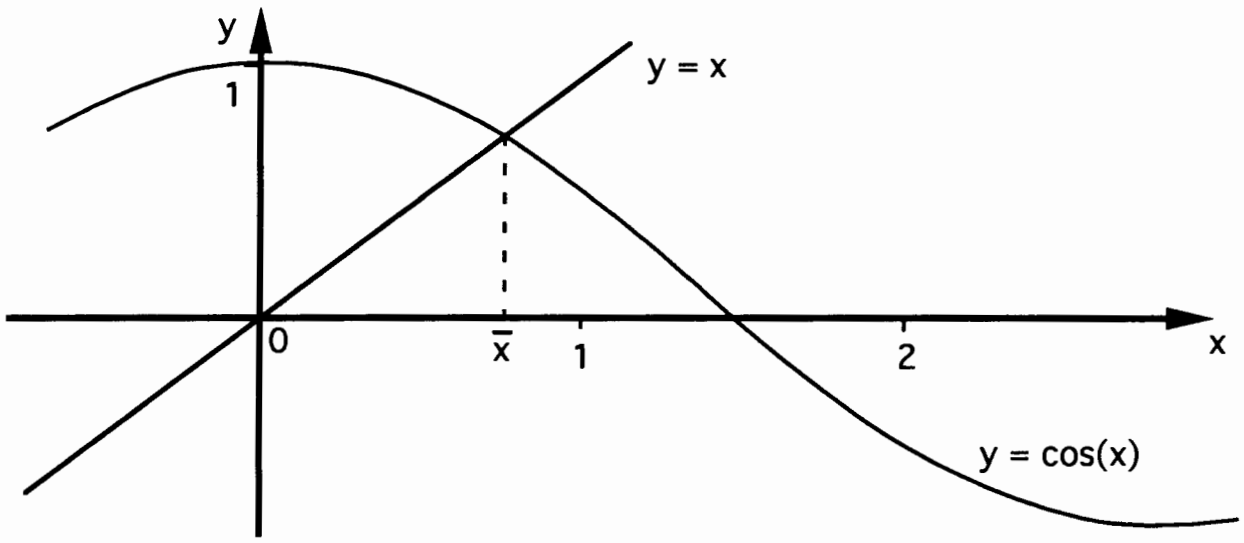


Figure 7.1 : Le point $\bar{x}$ est tel que $\bar{x} = \cos \bar{x}$

En s'inspirant de $x = \cos x$, on est tenté de poser, pour $n = 0,1,2,\cdots$,

$$x_{n+1} = \cos x_n \tag{7.1}$$

qui permet de construire successivement $x_1, x_2, \cdots$, lorsqu'on part de x_0 (itération de Picard).

Résultat. La suite donnée par (7.1) converge vers $\bar{x}$ et sa convergence est linéaire.

Démonstration. Puisque $\bar{x} = \cos \bar{x}$ on obtient par (7.1) :

$$| \bar{x} - x_{n+1} | = | \cos \bar{x} - \cos x_n | = \left| \int_{\bar{x}}^{x_n} \sin t \, dt \right|$$

et donc si $\varepsilon_n = | \bar{x} - x_n |$:

$$| \bar{x} - x_{n+1} | \leq \max_{t \in [\bar{x}-\varepsilon_n, \bar{x}+\varepsilon_n]} | \sin t | \cdot | \bar{x} - x_n |. \tag{7.2}$$

Par l'inégalité (7.2), on a clairement $| \bar{x} - x_{n+1} | \leq | \bar{x} - x_n |$ pour tout n et par suite $| \bar{x} - x_k | \leq | \bar{x} - x_0 |$ pour tout k. Si on pose $\chi = \max_{t \in [\bar{x}-\varepsilon_0, \bar{x}+\varepsilon_0]} | \sin t |$, on obtient donc $\chi < 1$ et par (7.2) :

$$| \bar{x} - x_{n+1} | \leq \chi | \bar{x} - x_n |. \tag{7.3}$$

Puisque la relation (7.3) est vraie pour tout n, on aura

$$| \bar{x} - x_n | \leq \chi^n | \bar{x} - x_0 |. \tag{7.4}$$

La relation (7.4) et le fait que $\chi < 1$ impliquent la convergence de la suite $(x_n)_{n=1}^{\infty}$ vers $\bar{x}$. De plus, (7.3) montre que cette convergence est linéaire.

#

Dans l'exemple ci-dessus, nous dirons que $\bar{x}$ est un point fixe de $\cos x$ car nous avons $\bar{x} = \cos \bar{x}$. La méthode (7.1) est appelée méthode de point fixe. Dans le paragraphe 7.2, nous présenterons quelques généralités sur les méthodes de point fixe. En attendant, nous voulons très rapidement décrire une autre méthode appelée "méthode de la bissection" qui elle, n'est pas une méthode de point fixe.

Méthode de la bissection. Revenons au cadre général où on veut calculer numériquement un zéro $\bar{x}$ d'une fonction continue f. On suppose, pour commencer, avoir deux valeurs α et β qui sont telles que $f(\alpha) f(\beta) < 0$. Ainsi f change de signe entre α et β et on pose $x_0 = \frac{\alpha+\beta}{2}$ qui est le point milieu de l'intervalle d'extrémités α et β. Si $f(x_0) = 0$, on a terminé car x_0 est un zéro de f. Ainsi, on peut supposer que $f(x_0)$ est différent de zéro et on construit 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 $x_1 = \frac{\alpha+\beta}{2}$;

- 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 $x_1 = \frac{\alpha+\beta}{2}$.

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 7.2 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, l'erreur $|\bar{x} - x_M|$ est contrôlée puisqu'à chaque pas on divise l'intervalle par 2. On aura donc

$$|\bar{x} - x_M| \leq \frac{\varepsilon}{2^{M+1}}.$$

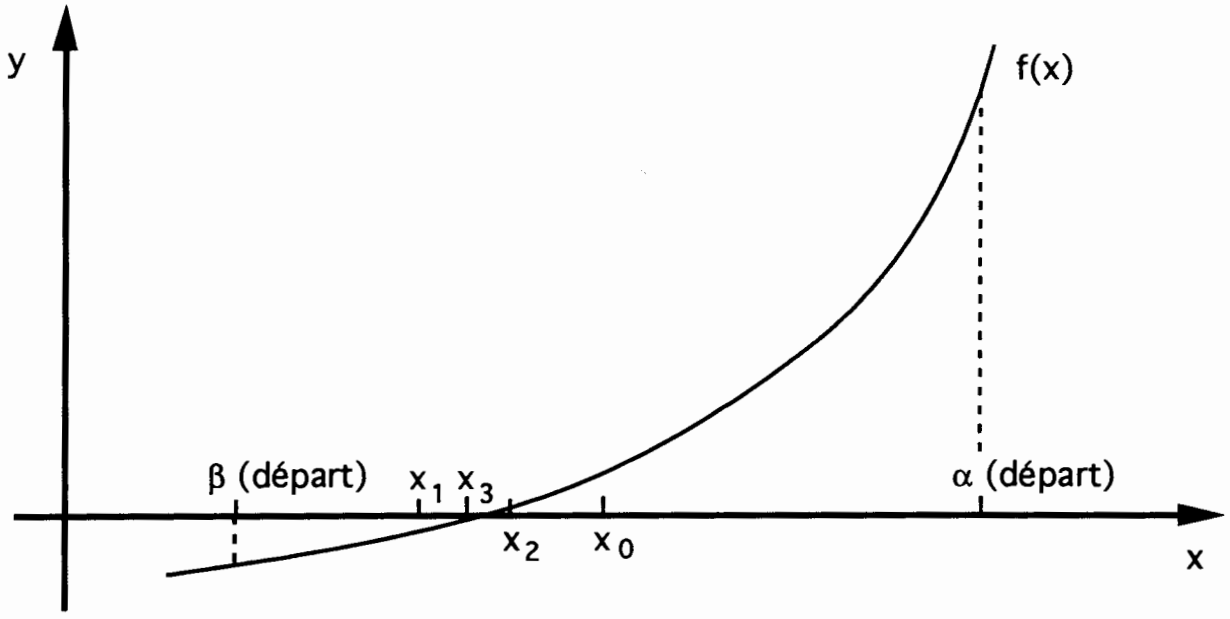


Figure 7.2 : Méthode de la bissection

Méthode de la bissection

On connaît deux valeurs α et β telles que $f(\alpha)\, f(\beta) < 0$.

La valeur de t à la sortir de l'algorithme sera une approximation d'un zéro de f.

On se donne M entier > 0 qui permet de fixer une tolérance sur l'erreur.

Algorithme	**Commentaires**
Faire $n = 0$ *à* M	
$t := \frac{\alpha+\beta}{2}$	Ici $t = x_n$
si $f(t) = 0$ *STOP*	Ici t est un zéro de f
si $f(\alpha)\, f(t) < 0$: *poser* $\beta := t$	Ici on change les valeurs de α ou β
sinon : *poser* $\alpha := t$	pour mieux "cerner" le zéro de f

7.2 Méthodes de point fixe : méthodes de Newton et de la corde

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

$$x = g(x). \tag{7.5}$$

Clairement, il existe une infinité de manière pour opérer cette transformation. Par exemple, on peut poser, comme fait dans l'exemple de la figure 7.1,

$$g(x) = x - f(x),$$

ou plus généralement

$$g(x) = x + \alpha f(x) \tag{7.6}$$

avec $\alpha \in \mathbb{R}$, $\alpha \neq 0$ quelconque. Dans (7.6), on peut même prendre pour α une fonction de x pour autant qu'elle ne s'annule pas.

Définition. Si x satisfait (7.5), on dit que x est un point fixe de g ; l'image de x par g est x lui-même.

Après avoir transformé le problème $f(x) = 0$ en un problème de point fixe (7.5), la méthode consiste à :

(i) évaluer par x_0 un point fixe $\bar{x} = g(\bar{x})$; (7.7)

(ii) poser $x_{n+1} = g(x_n)$, $n = 0,1,2,\cdots$. (7.8)

Naturellement, cette méthode ne va pas toujours converger. Cependant, nous avons le résultat suivant :

Théorème. Supposons g 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 $|\, \bar{x} - x_0 \,| \leq \varepsilon$, la suite donnée par

$$x_{n+1} = g(x_n), \quad n = 0,1,2,\cdots, \tag{7.8}$$

converge vers $\bar{x}$ lorsque n tend vers l'infini. La convergence est linéaire.

Démonstration. Si $|\, g'(\bar{x})\,| < 1$, alors par continuité de g', il existe $\varepsilon > 0$ et $\chi < 1$ tels que

$$|\, g'(x)\,| \le \chi, \qquad \text{pour tout } x \in [\bar{x}-\varepsilon, \bar{x}+\varepsilon]. \tag{7.9}$$

Supposons un instant que pour n fixé dans (7.8), on a $x_n \in [\bar{x}-\varepsilon, \bar{x}+\varepsilon]$. Le théorème des accroissements finis donne l'existence de ξ dans l'intervalle d'extrémités $x_n, \bar{x}$ tel que

$$g(\bar{x}) - g(x_n) = g'(\xi)(\bar{x} - x_n),$$

et par suite, en utilisant (7.8) et (7.9),

$$|\, \bar{x} - x_{n+1}\,| = |\, g(\bar{x}) - g(x_n)\,| = |\, g'(\xi)\,|\,|\, \bar{x} - x_n\,| \le \chi\,|\, \bar{x} - x_n\,|. \tag{7.10}$$

Puisque $\chi < 1$, la relation (7.10) implique entre autre que $x_{n+1} \in [\bar{x}-\varepsilon, \bar{x}+\varepsilon]$.

Ainsi, si $x_0 \in [\bar{x}-\varepsilon, \bar{x}+\varepsilon]$, alors toute la suite $(x_n)_{n=1}^{\infty}$ est inclue dans $[\bar{x}-\varepsilon, \bar{x}+\varepsilon]$ et en itérant (7.10), on obtient :

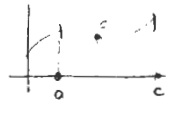
$$|\, \bar{x} - x_{n+1}\,| \le \chi^{n+1}\,|\, \bar{x} - x_0\,|, \quad n = 0,1,2,\cdots. \tag{7.11}$$

De (7.11) on tire que $\lim_{n\to\infty} x_n = \bar{x}$ et de (7.10) on observe que la convergence est linéaire.

##

Remarque. De l'inégalité (7.9) et du théorème des accroissements finis, on vérifie que pour tout $x, y \in [\bar{x}-\varepsilon, \bar{x}+\varepsilon]$ on a

$$|\, g(x) - g(y)\,| \le \chi\,|\, x - y\,|. \tag{7.12}$$



Comme $\chi < 1$, on dit que g est une contraction stricte dans un voisinage de son point fixe $\bar{x}$. En fait, g peut être une contraction stricte dans un voisinage d'un point fixe $\bar{x}$ sans qu'elle soit nécessairement une fois continûment dérivable. Dans ce cas, la relation (7.12) avec $\chi < 1$ suffit pour garantir la conclusion du théorème ci-dessus.

##

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 connaître une valeur x_n proche de $\bar{x}$. Pour calculer x_{n+1} on prendra l'intersection de l'axe Ox avec la tangente au graphe de f passant par le point $(x_n, f(x_n))$ (voir figure 7.3).

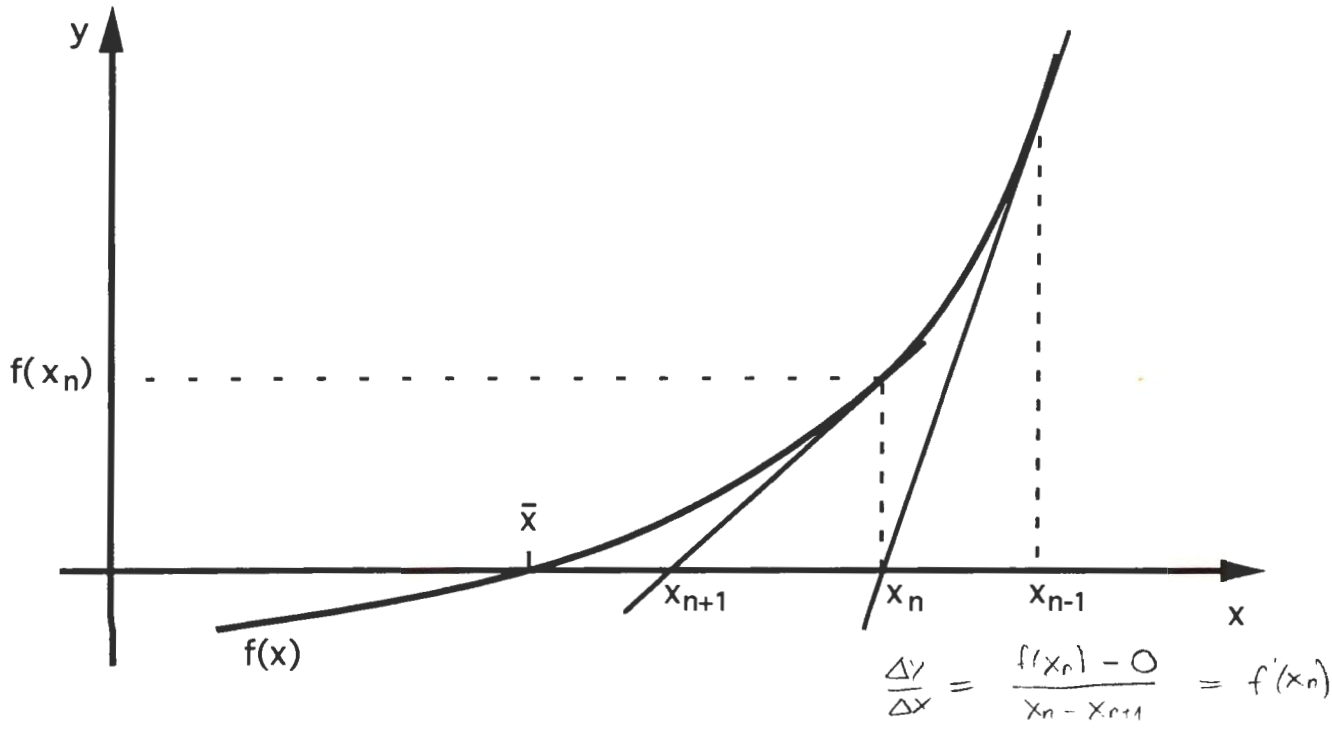


Figure 7.3 : Méthode de Newton

Clairement, nous avons la relation $\frac{f(x_n)}{x_n - x_{n+1}} = f'(x_n)$ qui donne, lorsque x_0 est choisi proche de $\bar{x}$, la méthode de Newton :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0,1,2,\cdots. \tag{7.13}$$

Nous voyons ainsi que la méthode de Newton est une méthode de point fixe pour calculer $\bar{x}$; 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 $\bar{x}$) et (7.13) est équivalent à $x_{n+1} = g(x_n)$.

En vue d'utiliser le théorème précédent, calculons $g'(x)$, puis $g'(\bar{x})$. Nous vérifions que si f est deux fois continûment dérivable :

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} \tag{7.14}$$

et par suite, puisque $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$:

$$g'(\bar{x}) = 0. \tag{7.15}$$

Ainsi, en utilisant le théorème précédent, nous concluons que si x_0 est choisi "suffisamment proche" de $\bar{x}$ alors la méthode de Newton (7.13) est convergente et on a $\lim_{n\to\infty} x_n = \bar{x}$ (du moins si f est deux fois continûment dérivable et si $f(\bar{x}) = 0$, $f'(\bar{x}) \neq 0$; hypothèses qu'il est possible d'affaiblir !). On peut montrer que la convergence est quadratique lorsqu'on se place dans le cadre des hypothèses ci-dessus ; ceci provient du fait que $g'(\bar{x}) = 0$.

Disons encore que si $f'(\bar{x}) = 0$, cette méthode converge encore (pour autant que $x_0 \neq \bar{x}$), mais la convergence devient alors linéaire.

Méthode de la corde (ou Newton-corde). C'est une méthode qui permet d'éviter qu'à chaque itération de (7.13) on évalue $f'(x_n)$. La méthode de la corde consiste donc à remplacer $f'(x_n)$ par $f'(x_0)$ dans (7.13), ce qui donne :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \quad n = 0,1,2,\cdots. \tag{7.16}$$

L'interprétation géométrique de cette méthode est immédiate lorsqu'on regarde la figure 7.4 ; le terme "méthode de la corde" est aussi facile à interpréter !

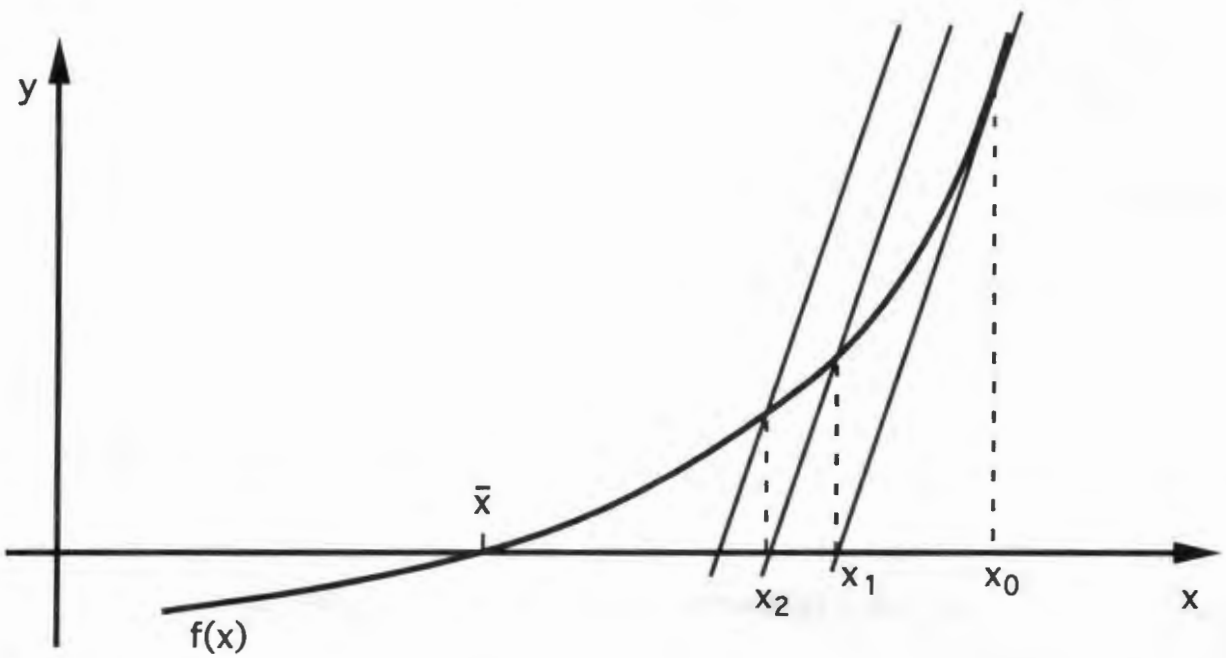


Figure 7.4 : Méthode de la corde

Ici encore, si on pose $g(x) = x - \frac{f(x)}{f'(x_0)}$, on constate que si $f(\bar{x}) = 0$ alors $\bar{x} = g(\bar{x})$ et (7.16) donne $x_{n+1} = g(x_n)$. Ainsi la méthode de la corde est une méthode de point fixe. Dans ce cas, on obtient $g'(\bar{x}) = 1 - \frac{f'(\bar{x})}{f'(x_0)}$ et on vérifie, en utilisant la continuité de f', que si x_0 est "assez proche" de $\bar{x}$, alors $|g'(\bar{x})| < 1$. En utilisant notre théorème, nous montrons donc que si x_0 est "suffisamment proche" de $\bar{x}$, la méthode de la corde converge ; sa convergence est linéaire.

7.3 Systèmes non linéaires

Soit une application régulière $f : \mathbb{R}^N \to \mathbb{R}^N$ (N entier positif) dont on veut chercher 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 $x_1, x_2, \cdots, x_N$ et x peut être vu comme un vecteur (on ne mettra pas de flèche sur x pour ne pas alourdir les notations)

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix}.$$

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 :

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \\ \vdots \\ f_N(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, x_3, \cdots, x_N) \\ f_2(x_1, x_2, x_3, \cdots, x_N) \\ f_3(x_1, x_2, x_3, \cdots, x_N) \\ \vdots \\ f_N(x_1, x_2, x_3, \cdots, x_N) \end{bmatrix}.$$

En fait, l'équation

$$f(x) = 0 \tag{7.17}$$

est un système de N équations (non linéaires en principe) à N inconnues $x_1, x_2, \cdots, x_N$:

$$\begin{cases} f_1(x_1, x_2, x_3, \cdots, x_N) = 0, \\ f_2(x_1, x_2, x_3, \cdots, x_N) = 0, \\ \vdots \\ f_N(x_1, x_2, x_3, \cdots, x_N) = 0. \end{cases} \tag{7.18}$$

Si $x \in \mathbb{R}^N$, on peut définir la $N \times N$ matrice jacobienne $Df(x)$ de f au point x de la façon suivante :

$$Df(x)=\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_N}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_N}(x) \\ \vdots & & & \\ \frac{\partial f_N}{\partial x_1}(x) & \frac{\partial f_N}{\partial x_2}(x) & \cdots & \frac{\partial f_N}{\partial x_N}(x) \end{bmatrix};$$

les coefficients de la matrice $Df(x)$ sont donc

$$Df(x)_{ij}=\frac{\partial f_i}{\partial x_j}(x). \tag{7.19}$$

La méthode de Newton pour les systèmes se généralise comme suit :

$$x^{n+1}=x^n-Df(x^n)^{-1}f(x^n),\quad n=0,1,2,\cdots, \tag{7.20}$$

après avoir choisi naturellement un vecteur de départ x^0 "suffisamment proche" du vecteur $\bar{x}$ solution de $f(\bar{x})=0$. Remarquons que dans (7.20) 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.

On peut montrer, ici encore, que si f est "assez régulière", 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 (7.20) converge vers $\bar{x}$ lorsque n tend vers l'infini, ceci pour autant que x^0 soit choisi "suffisamment proche" de $\bar{x}$ (ceci signifie que si $\| x^0-\bar{x}\| \underset{def}{=} \left(\sum_{j=1}^{N}(x_j^0-\bar{x}_j)^2\right)^{1/2}$ est "petit", alors $\lim_{n\to\infty}\| x^n-\bar{x}\|=0$). La convergence est quadratique, c'est-à-dire il existe une constante C telle que

$$\| x^{n+1}-\bar{x}\| \le C\| x^n-\bar{x}\|^2,\quad n=0,1,2,\cdots. \tag{7.21}$$

Pour calculer x^{n+1} lorsqu'on connaît x^n, on écrit (7.20) sous la forme

$$Df(x^n)(x^n-x^{n+1})=f(x^n),\quad n=0,1,2,\cdots, \tag{7.22}$$

et donc, en pratique, on procède ainsi :

1) on construit le vecteur $\vec{b}=f(x^n)$;

2) on construit la matrice $A=Df(x^n)$;

3) on résout le système $A\vec{y} = \vec{b}$ par élimination de Gauss ou décomposition LU de A (ou décomposition LL^T de A si A est symétrique définie positive) (voir leçon 5) ;

4) on pose $x^{n+1} = x^n - \vec{y}$.

L'inconvénient majeur de la méthode de Newton pour résoudre numériquement un système de N équations non linéaires à N inconnues est qu'à chaque pas de la méthode, on doit construire la matrice $Df(x^n)$ et procéder à sa décomposition LU. Pour pallier cet inconvénient, on peut utiliser la méthode de la corde qui, traduite pour les systèmes, devient :

$$x^{n+1} = x^n - Df(x^0)^{-1} f(x^n), \quad n = 0,1,2,\cdots. \tag{7.23}$$

Ainsi, une bonne fois pour toutes, on peut construire au départ la matrice $Df(x^0)$ et en faire sa décomposition LU (ou LL^T si la matrice est symétrique définie positive (voir leçon 5)). Après quoi, pour calculer x^{n+1} à partir de x^n sur un plan pratique, la méthode de la corde consistera à

1) construire le vecteur $\vec{b} = f(x^n)$;

2) résoudre le système triangulaire $L\vec{z} = \vec{b}$;

3) résoudre le système triangulaire $U\vec{y} = \vec{z}$;

4) poser $x^{n+1} = x^n - \vec{y}$.

A chaque pas de la méthode de la corde, on a deux systèmes triangulaires à résoudre dont les matrices ont été formées avant de procéder au calcul des différents itérés x^n. Cette méthode peut paraître considérablement plus économique que la méthode de Newton. En revanche, sa convergence est linéaire et donc considérablement plus lente que celle de la méthode de Newton qui est quadratique. On devra donc, avec la méthode de la corde, procéder à un plus grand nombre d'itérations pour obtenir la même précision que celle obtenue par la méthode de Newton !

Dans ces notes, nous laissons de côté les méthodes utilisées par les ingénieurs pour obtenir un "bon" vecteur de départ x^0, ainsi que les critères d'arrêt des différentes méthodes !

Leçon 8

Equations différentielles

8.1 Equations différentielles du 1er ordre : généralités

Dans ce paragraphe, on note par $\mathbb{R}^+$ l'ensemble des nombres réels non-négatifs et par $f:(x,t)\in\mathbb{R}\times\mathbb{R}^+\mapsto f(x,t)\in\mathbb{R}$ une fonction à deux variables (x,t) donnée (dans les applications, x est souvent une variable spatiale alors que t est une variable temporelle). On suppose f "assez régulière" (par exemple continue) et on veut résoudre le problème suivant :

étant donnée une valeur $u_0\in\mathbb{R}$ (u_0 est dite valeur initiale),
trouver une fonction $u: t\in\mathbb{R}^+\to u(t)\in\mathbb{R}$ qui satisfait

$$(P)\begin{cases}\dot{u}(t)=f(u(t),t) \quad si\ t>0, & (8.1)\\ u(0)=u_0, & (8.2)\end{cases}$$

où ici $\dot{u}(t)=\frac{du}{dt}(t)$.

Le problème (P) est appelé problème de Cauchy pour l'équation différentielle (8.1). La condition (8.2) est une condition de Cauchy ; une fonction u qui satisfait (8.1) est appelée intégrale de l'équation différentielle.

Exemple 1. On se donne $f(x,t)=x+t$ et $u_0=1$. Le problème de Cauchy devient

$$\begin{cases}\dot{u}(t)=u(t)+t \quad si\ t>0,\\ u(0)=1;\end{cases}$$

sa solution est donnée par $u(t)=2e^t-t-1$.

Exemple 2. On se donne $f(x,t) = \sqrt[3]{x}, u_0 = 0$. Le problème de Cauchy devient

$$\begin{cases} \dot{u}(t) = \sqrt[3]{u(t)} & si\ t > 0, \\ u(0) = 0; \end{cases}$$

dont on vérifiera que les fonctions $u(t) \equiv 0 \quad \forall t \geq 0$, et $u(t) = \pm\sqrt{\frac{8}{27}t^3} \quad \forall t \geq 0$, sont toutes trois des solutions.

L'exemple 2 nous montre que le problème (P) n'a pas nécessairement une solution unique. Les équations différentielles sont un sujet d'étude d'analyse mathématique et nous n'aborderons pas ici les problèmes d'existence et d'unicité. Cependant, nous énonçons, sans le démontrer, un théorème classique concernant cette question :

Théorème (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 $y,z \in \mathbb{R}$ et pour tout $t \in \mathbb{R}^+$, on ait

$$| f(y,t) - f(z,t) | \leq L | y - z |. \tag{8.3}$$

Alors le problème (8.1), (8.2) admet une solution et une seule.

Exemple 3. On se donne $f(x,t) = \sin x + e^{-t^2/2}, u_0 = 1$. On vérifie facilement que $| f(y,t) - f(z,t) | = | \sin y - \sin z | = | \int_z^y \cos\theta\, d\theta | \leq | y - z |$. Par le théorème de Cauchy-Lipschitz, le problème (8.1), (8.2) a une solution unique $u(t)$. Cependant, nous ne pouvons donner une expression explicite pour $u(t)$; il est donc nécessaire d'utiliser une méthode numérique si nous désirons obtenir des valeurs $u(t)$ pour différents $t \in \mathbb{R}^+$.

Soit $0 = t_0 < t_1 < t_2 < t_3 < \cdots < t_n < t_{n+1} < \cdots$ des points de $\mathbb{R}^+$ et supposons connue une approximation u^n de u en $t = t_n$, i.e. $u^n \simeq u(t_n)$. Un schéma d'intégration numérique à un pas consistera à calculer u^{n+1} (approximation de u en $t = t_{n+1}$) à partir de u^n. Pour exemple, prenons le schéma d'Euler progressif qui, en s'inspirant de (8.1), s'écrit

$$\frac{u^{n+1} - u^n}{t_{n+1} - t_n} = f(u^n, t_n)$$

ou de façon équivalente

$$u^{n+1} = u^n + (t_{n+1} - t_n) f(u^n, t_n). \tag{8.4}$$

Clairement si on pose $u^0 = u(0) = u_0$ (valeur connue), on voit immédiatement que (8.4) permet de calculer successivement u^1, puis u^2, puis u^3, ... etc.

La question qui se pose dès lors est de savoir si l'erreur

$$| u(t_n) - u^n | \qquad \text{pour } n = 1,2,\cdots,$$

est "petite" ou non. Dans la suite, nous tenterons d'y répondre (du moins partiellement !).

8.2 Problèmes numériquement mal posés

Pour illustrer notre propos, prenons pour exemple le problème (P) avec :

$$f(x,t) = 3x - 3t \qquad \text{et} \qquad u_0 = \alpha,$$

où α est un nombre réel donné. Concrètement parlant, le problème (8.1), (8.2) devient :

$$\begin{cases} \dot{u}(t) = 3u(t) - 3t \quad si\ t > 0, & (8.5) \\ u(0) = \alpha, & (8.6) \end{cases}$$

et sa solution est donnée par

$$u(t) = (\alpha - \tfrac{1}{3})\, e^{3t} + t + \tfrac{1}{3}. \qquad (8.7)$$

Si nous désirons intégrer (8.5), (8.6) jusqu'à $t = 10$ avec une valeur de $\alpha = \frac{1}{3}$, nous obtiendrons $u(10) = 10 + \frac{1}{3} = \frac{31}{3}$. Par contre, si nous faisons le calcul avec $\alpha = 0.333333$ au lieu de $\alpha = \frac{1}{3}$, nous aurons

$$\begin{aligned} u(10) &= (0.333333 - \tfrac{1}{3})\, e^{30} + 10 + \tfrac{1}{3} \\ &= -\tfrac{1}{3} 10^{-6} \cdot e^{30} + \frac{31}{3} \end{aligned}$$

qui présente une différence avec la précédente valeur de $\frac{1}{3} 10^{-6} \cdot e^{30}$ évaluée à $\frac{1}{3} 10^{7}$. 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 peut calculer qu'avec 6 chiffres significatifs (en virgule flottante), alors $\alpha = \frac{1}{3}$ deviendra $\alpha = 0.333333$ et il est donc inutile d'essayer d'inventer une méthode numérique qui calculera $u(10)$. En effet, seule l'erreur sur la condition initiale provoque une erreur inadmissible sur la solution sans même utiliser de méthodes numériques ! Nous sommes ici en présence d'un problème numériquement mal posé.

Sans définir de façon précise ce qu'est un problème numériquement mal posé, nous pouvons dire tout de même que cette notion est relative au problème, bien sûr, mais aussi au calculateur

mis à disposition. Dans l'exemple précédent, il suffit d'avoir un calculateur avec 16 chiffres significatifs pour que l'erreur relative sur la donnée initiale soit de 10^{-16} et que cette dernière induise (avec un calcul exact) une erreur relative de l'ordre de 10^{-4} sur la valeur $u(10)$.

Si, dans le problème (P), nous avions pris $f(x,t) = -3x - 3t$ au lieu de $f(x,t) = 3x - 3t$, alors nous aurions eu

$$\begin{cases} \dot{u}(t) = -3u(t) - 3t, \\ u(0) = \alpha, \end{cases}$$

dont la solution $u(t)$ est donnée par :

$$u(t) = (\alpha - \tfrac{1}{3})\, e^{-3t} - t + \tfrac{1}{3}.$$

Ce 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 $u(10) = \frac{1}{3}10^{-6} e^{-30} - 10 + \frac{1}{3} \cong -\frac{29}{3} + \frac{1}{3}10^{-19}$.

Dans la suite, nous nous intéresserons aux problèmes numériquement bien posés.

8.3 Schémas d'Euler

Pour établir un schéma d'approximation du problème (P) (8.1), (8.2), nous commençons par partitionner l'axe Ot par

$$0 = t_0 < t_1 < t_2 < \cdots < t_n < t_{n+1} < \cdots.$$

En posant $h_n = t_{n+1} - t_n$, nous pouvons approcher $\dot{u}(t_n)$ ou $\dot{u}(t_{n+1})$ par $\frac{u(t_{n+1}) - u(t_n)}{h_n}$ et si u^n est une approximation de $u(t_n)$, ces deux approches nous suggèrent les schémas suivants :

a) Schéma d'Euler progressif :

$$\begin{cases} \dfrac{u^{n+1} - u^n}{h_n} = f(u^n, t_n), \quad n = 0,1,2,\cdots, \\ u^0 = u_0 ; \end{cases} \tag{8.8}$$

b) Schéma d'Euler rétrograde :

$$\begin{cases} \dfrac{u^{n+1}-u^n}{h_n} = f(u^{n+1},t_{n+1}), & n=0,1,2,\cdots, \\ u^0 = u_0. \end{cases} \tag{8.9}$$

Ces deux schémas nous permettent de calculer u^{n+1} lorsqu'on connaît u^n et donc il est possible de déterminer successivement $u^1, u^2, u^3, \cdots$.

Le schéma d'Euler progressif est un schéma explicite car il permet d'expliciter u^{n+1} en fonction de u^n :

$$u^{n+1} = u^n + h_n f(u^n,t_n). \tag{8.8'}$$

Le schéma d'Euler rétrograde est un schéma implicite car il ne permet pas d'expliciter en principe u^{n+1} en fonction de u^n. En effet, on a dans ce cas :

$$u^{n+1} - h_n f(u^{n+1},t_{n+1}) = u^n \tag{8.9'}$$

et si on veut calculer u^{n+1}, il reste à définir la fonction

$$\ell(x) = x - h_n f(x,t_{n+1}) - u^n$$

et à chercher un zéro de $\ell(x)$ en prenant par exemple une méthode de Newton !

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 allons voir ci-dessous que le schéma progressif peut engendrer des difficultés.

Prenons un problème test où $f(x,t) = -\beta x$, où $\beta > 0$ est un nombre réel positif donné. Le problème (8.1), (8.2) devient :

$$\begin{cases} \dot{u}(t) = -\beta u(t), t > 0, \\ u(0) = u_0, \end{cases} \tag{8.10}$$

dont la solution est trivialement $u(t) = e^{-\beta t} u_0$. Puisque β est positif, ce problème est numériquement bien posé.

Pour discrétiser l'axe Ot, nous choisissons un nombre réel "petit" $h>0$ et nous posons $t_n = nh$ avec $n = 0,1,2,\cdots$; clairement, nous avons $h_n = h$ pour tout n. Nous allons regarder dans ce cadre successivement les deux schémas introduits.

- Le schéma d'Euler progressif (8.8) ou (8.8′) devient

$$u^{n+1} = (1-\beta h)u^n, \quad n = 0,1,2,\cdots \tag{8.11}$$

et par suite

$$u^n = (1-\beta h)^n u_0, \quad n = 0,1,2,\cdots. \tag{8.12}$$

Bien que la solution $u(t)$ de (8.10) tende vers zéro lorsque t tend vers l'infini, nous voyons dans (8.12) 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. Pour éviter ce phénomène, il convient donc d'imposer $-1 \leq 1-\beta h \leq 1$ ce qui aura pour effet de limiter h à :

$$h \leq \frac{2}{\beta}. \tag{8.13}$$

La condition (8.13) est appelée condition de stabilité ; elle limite le "pas d'avance" h lorsqu'on utilise le schéma d'Euler progressif.

Comme nous venons de le voir en prenant $u_0 \neq 0$ et $h > \frac{2}{\beta}$, les grandeurs $u^n, n = 0,1,2,\cdots$ tendent vers l'infini en alternant de signe et on dit dans ce cas que l'on a une instabilité numérique !

- Le schéma d'Euler rétrograde (8.9) ou (8.9′) devient sur notre exemple

$$(1+\beta h)u^{n+1} = u^n, \quad n = 0,1,2,\cdots \tag{8.14}$$

et par suite

$$u^n = \left(\frac{1}{1+\beta h}\right)^n u_0, \quad n = 0,1,2,\cdots. \tag{8.15}$$

Dans ce cas, nous voyons que pour tout $h>0$ on a

$$\lim_{n\to\infty} u^n = 0;$$

le schéma d'Euler rétrograde est toujours stable ; h n'a pas à être limité !

Nous pouvons cependant montrer que pour les deux schémas, si $T>0$ est fixé, si $h=T/N$ avec $N=1,2,3,\cdots$ alors il existe une constante C (indépendante de h ou N) telle que

$$|\,u(T)-u^N\,|\le Ch. \tag{8.16}$$

Ainsi $\lim_{N\to\infty}|\,u(T)-u^N\,|=0$ et la convergence est d'ordre $h=\frac{T}{N}$; on dit que les deux schémas sont d'ordre 1 en h !

8.4 Méthodes de Runge-Kutta d'ordre 2

Reprenons le problème (8.1), (8.2) et supposons avoir une solution $u(t)$. En intégrant (8.1) entre t_n et t_{n+1}, nous obtenons

$$u(t_{n+1})-u(t_n)=\int_{t_n}^{t_{n+1}} f(u(t),t)\,dt. \tag{8.17}$$

Si u^n est une approximation de $u(t_n)$ et u^{n+1} une approximation de $u(t_{n+1})$, nous pouvons proposer pour schéma, en intégrant le membre de droite de (8.17) par la formule des trapèzes :

$$u^{n+1}-u^n=\frac{1}{2}h_n\Big(f(u^n,t_n)+f(u^{n+1},t_{n+1})\Big),\quad n=0,1,\cdots, \tag{8.18}$$

c.f. intégration numérique : formule des rectangles

où $h_n=t_{n+1}-t_n$. Clairement ,(8.18) est un schéma implicite (c'est une moyenne des schémas d'Euler progressif et rétrograde). On peut montrer qu'il est d'ordre 2 en h.

Pour éviter le calcul implicite de u^{n+1} dans (8.18), on pourrait remplacer u^{n+1} dans le membre de droite de (8.18) par une "prédiction d'Euler progressive", i.e. par

$$\tilde{u}^{n+1}=u^n+h_n f(u^n,t_n). \tag{8.19}$$

Ainsi, nous avons construit un nouveau schéma, appelé "méthode de Heun" qui, connaissant u^n, consiste à calculer $\tilde{u}^{n+1}$ par (8.19), puis à calculer u^{n+1} par (8.18) dans laquelle on remplace u^{n+1} du membre de droite par $\tilde{u}^{n+1}$. Le schéma de Heun devient :

$$\begin{cases} p_1=f(u^n,t_n),\\ p_2=f(u^n+h_n p_1,t_{n+1}),\\ u^{n+1}=u^n+h_n(\frac{1}{2}p_1+\frac{1}{2}p_2). \end{cases} \tag{8.20}$$

rectangles

Le schéma de Heun est une méthode dite de "Runge-Kutta d'ordre 2".

On aurait pu penser poser $t_{n+1/2} = \frac{t_n + t_{n+1}}{2}$ (point milieu de $[t_n, t_{n+1}]$) et faire une prédiction d'Euler progressive, notée $u^{n+1/2}$, par la relation :

$$u^{n+1/2} = u^n + \frac{h_n}{2} f(u^n, t_n). \tag{8.21}$$

En s'inspirant de la formule des rectangles pour l'intégration du membre de droite de (8.17), nous obtenons un nouveau schéma :

$$u^{n+1} - u^n = h_n f(u^{n+1/2}, t_{n+1/2}) \tag{8.22}$$

où $u^{n+1/2}$ est donné par (8.21). Cette méthode est aussi une méthode de Runge-Kutta d'ordre 2. Connaissant u^n, elle s'écrit :

$$\begin{cases} p_1 = f(u^n, t_n), \\ p_2 = f(u^n + \frac{h_n}{2} p_1, t_n + \frac{h_n}{2}), \\ u^{n+1} = u^n + h_n p_2. \end{cases} \tag{8.23}$$

Nous ne discuterons pas ici des questions de stabilité de ces méthodes.

8.5 Méthode de Runge-Kutta classique

C'est une méthode d'ordre 4 en h qui, lorsqu'on connaît u^n (approximation de u en $t = t_n$), permet de calculer u^{n+1} (approximation de u en $t = t_{n+1}$) par la séquence suivante :

$$\begin{cases} p_1 = f(u^n, t_n) \\ p_2 = f(u^n + \frac{h_n}{2} p_1, t_n + \frac{h_n}{2}) \\ p_3 = f(u^n + \frac{h_n}{2} p_2, t_n + \frac{h_n}{2}) \\ p_4 = f(u^n + h_n p_3, t_{n+1}) \qquad ; t_{n+1} = t_n + h_n \\ u^{n+1} = u^n + \frac{h_n}{6}(p_1 + 2p_2 + 2p_3 + p_4). \end{cases} \tag{8.24}$$

C'est une méthode explicite. Si nous voulons intégrer numériquement (8.1), (8.2) jusqu'à $t = T$ où T est un nombre positif donné, si nous posons $h = \frac{T}{N}$ (où N est un entier positif donné), $t_j = jh$ avec $j = 0,1,2,\cdots,N$, et si nous utilisons le schéma (8.24), nous aurons l'estimation d'erreur :

Simpson : $J(g) = \frac{1}{3}\left(g(-1) + g(1)\right) + \frac{4}{3} g(0)$

Sous certaines conditions de régularité,

$$| u(T) - u^N | \leq C h^4 = C \frac{T^4}{N^4}, \tag{8.25}$$

lorsque N tend vers l'infini. Ici C est une constante indépendante de N. La relation (8.25) montre que la méthode de Runge-Kutta classique est d'ordre 4 en h. Si f ne dépend que de t, on peut voir que la méthode de Runge-Kutta est la méthode de Simpson pour intégrer $f(t)$ entre $t = 0$ et $t = T$.

8.6 Systèmes différentiels du 1er ordre

Soit $\vec{f} : (\vec{x}, t) \in \mathbb{R}^M \times \mathbb{R} \rightarrow \vec{f}(\vec{x}, t) \in \mathbb{R}^M$ une fonction vectorielle donnée et supposée continue (ici M est un entier positif). Si $\vec{u}_0$ est un vecteur à M composantes donné, on pose le problème de trouver une solution réelle à valeurs vectorielles

$$\vec{u} : t \in \mathbb{R}^+ \rightarrow \vec{u}(t) \in \mathbb{R}^M$$

telle que

$$(Q) \begin{cases} \dot{\vec{u}}(t) = \vec{f}(\vec{u}(t), t), \quad t > 0, & (8.26) \\ \vec{u}(0) = \vec{u}_0. & (8.27) \end{cases}$$

Clairement, on a affaire à un système différentiel de M équations à M inconnues qui sont les composantes $u_1(t), u_2(t), \cdots, u_M(t)$ de $\vec{u}(t)$.

Les schémas d'Euler progressif et rétrograde, la méthode de Heun, les méthodes de Runge-Kutta présentés dans les paragraphes 8.3 à 8.5 sont généralisables au système différentiel (Q). Le schéma d'Euler progressif par exemple devient :

$$\vec{u}^{n+1} = \vec{u}^n + h_n \vec{f}(\vec{u}^n, t_n), \tag{8.28}$$

où $\vec{u}^n$ est une approximation connue de $\vec{u}(t_n)$ et $\vec{u}^{n+1}$ est une approximation de $\vec{u}(t_{n+1})$ calculée à partir de $\vec{u}^n$, $(h_n = t_{n+1} - t_n)$.

8.7 Equations différentielles d'ordre supérieur

Pour commencer, prenons une fonction à 3 variables

$$f : (x,y,t) \in \mathbb{R}^3 \rightarrow f(x,y,t) \in \mathbb{R}$$

que nous supposons continue et soit deux nombres donnés u_0 et v_0. On pose le problème de trouver une fonction

$$u : t \in \mathbb{R}^+ \rightarrow u(t) \in \mathbb{R}$$

deux fois continûment dérivable telle que

$$(R) \begin{cases} \ddot{u}(t) = f(u(t),\dot{u}(t),t), \quad t > 0, & (8.29) \\ u(0) = u_0, \quad \dot{u}(0) = v_0, & (8.30) \end{cases}$$

où ici $\ddot{u}(t) = \frac{d^2u}{dt^2}$, $\dot{u}(t) = \frac{du}{dt}(t)$. Clairement, (R) est un problème différentiel du $2^{\text{ème}}$ ordre ; on a deux conditions initiales (8.30) qui, en dynamique par exemple, sont souvent la position et la vitesse initiales.

Une première idée pour résoudre numériquement le problème (R) est d'introduire une nouvelle inconnue $v(t) \underset{def}{=} \dot{u}(t)$; ainsi (8.29), (8.30) est ramené à un système du 1^{er} ordre pour les inconnues $u(t)$ et $v(t)$:

$$\begin{cases} \dot{u}(t) = v(t), \\ \dot{v}(t) = f(u(t),v(t),t), \quad t > 0 \\ u(0) = u_0 \quad et \quad v(0) = v_0. \end{cases} \tag{8.31}$$

On peut ainsi appliquer les méthodes décrites plus haut (cf. paragraphe 8.6) au système (8.31).

On peut en faire de même si l'équation différentielle de départ est d'ordre plus élevé que 2. Cependant, il existe des méthodes spécifiquement adaptées aux problèmes d'ordre 2 dont une catégorie est appelée "méthodes de Newmark". Par exemple si f est indépendante de y, i.e. $f(x,y,t) = g(x,t)$, nous aurons

$$\begin{cases} \ddot{u}(t) = g(u(t),t), \quad t > 0, & (8.32) \\ u(0) = u_0, \quad \dot{u}(0) = v_0. & (8.33) \end{cases}$$

En nous inspirant de la leçon 2 pour l'approximation des dérivées secondes au moyen de différences finies, nous prendrons $h>0, t_n = nh$ avec $n = 0,1,2,\cdots$ et si u^n est une approximation de $u(t_n)$, nous écrirons le schéma :

$$\frac{u^{n+1}-2u^n+u^{n-1}}{h^2} = g(u^n,t_n), \quad n=1,2,3,\cdots, \tag{8.34}$$

avec

$$u^0 = u_0 \tag{8.35}$$

et

$$u^1 = u_0 + hv_0 + \frac{1}{2}h^2 g(u_0,0). \tag{8.36}$$

Remarquons que (8.34) permet de calculer u^{n+1} lorsqu'on connaît u^n et u^{n-1}. Pour interpréter (8.36), il suffit d'utiliser (8.32), (8.33) et d'écrire :

$$u^1 = u(0) + h\dot{u}(0) + \frac{h^2}{2}\ddot{u}(0); \tag{8.37}$$

le membre de droite de (8.37) est la somme des trois premiers termes du développement limité à l'ordre 2 de $u(t)$ en $t=0$.

La méthode (8.34)-(8.36) est d'ordre 2; elle est utilisée lorsqu'on veut résoudre numériquement l'équation des ondes.

Leçon 9

Différences finies et éléments finis pour des problèmes aux limites unidimensionnels

9.1 Approximation par différences finies d'un problème aux limites

Considérons le problème suivant : étant données deux fonctions c et f continues sur l'intervalle $[0,1]$, trouver une fonction u deux fois continûment dérivable sur $[0,1]$ telle que

$$\begin{cases} -u''(x) + c(x)u(x) = f(x) & si\ 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{9.1}$$

Un exemple de situation physique où ce problème est rencontré est celui du fléchissement d'une poutre de longueur 1, étirée selon son axe par une force P, soumise à une densité linéaire de charge $f(x)$ et simplement appuyée à ses extrémités 0 et 1. Alors le moment fléchissant $u(x)$ au point d'abscisse x est solution du problème (9.1) 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 à l'endroit x.

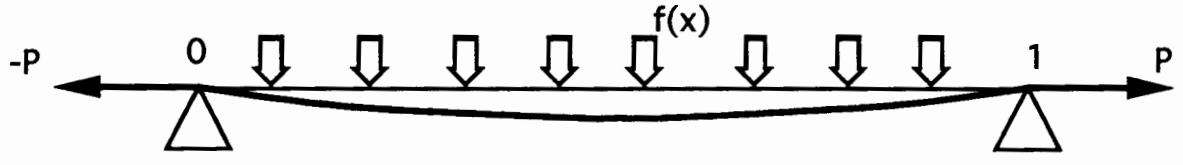


Figure 9.1 : Fléchissement d'une poutre

Un autre exemple est celui du déplacement vertical $u(x)$ à l'endroit x d'une corde tendue entre les points 0 et 1 avec une tension unité, et soumise à une densité de charge verticale $f(x)$; dans ce cas, on a $c(x) = 0, \forall x \in [0,1]$.

Remarquons que l'équation différentielle dans (9.1) est du deuxième ordre et nous avons deux conditions $u(0) = 0, u(1) = 0$, qui sont appelées conditions aux limites.

Si l'on suppose $c \geq 0$ sur l'intervalle $[0,1]$, on peut montrer que le problème (9.1) a une et une seule solution. Sauf dans quelques cas très rares, on ne connaît pas de "formule" qui permettrait d'obtenir une valeur numérique pour $u(x)$, $x \in (0,1)$. Il convient donc de trouver un moyen "d'approcher" les valeurs de la solution du problème (P) "d'aussi près que l'on veut". Une des méthodes pour atteindre ce but est la méthode des différences finies que nous décrivons ci-dessous.

Soit N un entier positif ; on pose $h = \frac{1}{N+1}$ et les points $x_j = jh$ avec $j = 0,1,2,\cdots,N+1$, seront appelés les points de discrétisation (cf. figure 9.2).

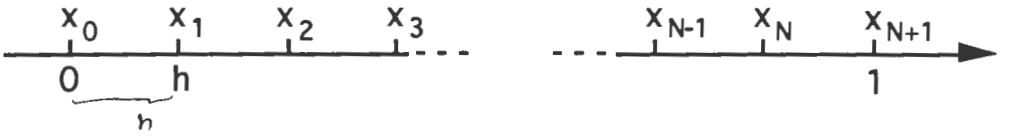


Figure 9.2 : Points de discrétisation

Dans la leçon 2, nous avons montré que si u est quatre fois continûment dérivable, alors

$$u''(x) = \frac{\delta_h^2 u(x)}{h^2} + O(h^2) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2), \qquad (9.2)$$

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 (9.1), nous nous inspirerons de (9.2) et nous calculerons des valeurs u_j sensées être proches de $u(x_j)$ et satisfaisant pour $1 \leq j \leq N$:

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} + c(x_j)u_j = f(x_j) \qquad (9.3)$$

avec

$$u_0 = u_{N+1} = 0. \qquad (9.4)$$

Le problème (9.3) est appelé "problème approché" ou "problème discret". C'est un schéma d'approximation par différences finies du problème (9.1) dit par opposition "continu". Résoudre (9.3) avec (9.4) revient à chercher un nombre fini de valeurs u_j qui doivent 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 $u_1, u_2, \cdots, u_N$, si $\vec{f}$ est le N-vecteur de composantes $f(x_1), f(x_2), \cdots, f(x_N)$ et si A est la $N \times N$ matrice définie par

$$A = \frac{1}{h^2} \begin{bmatrix} 2+c_1 h^2 & -1 & & & & \\ -1 & 2+c_2 h^2 & -1 & & \mathrm{O} & \\ & -1 & 2+c_3 h^2 & -1 & & \\ & & \cdot & \cdot & \cdot & \\ & \mathrm{O} & & \cdot & \cdot & -1 \\ & & & & -1 & 2+c_N h^2 \end{bmatrix} \tag{9.5}$$

où $c_i = c(x_i)$, alors le problème approché (9.3) avec (9.4) est clairement équivalent à chercher $\vec{u}$ tel que

$$A\vec{u} = \vec{f}. \tag{9.6}$$

Si $c(x) \geq 0$ pour tout $x \geq 0$, on montre, comme fait pour l'exemple du paragraphe 5.5 de la leçon 5, que A est une matrice symétrique définie positive. Elle est donc régulière et soit $\vec{u}$ la solution de (9.6) que l'on peut calculer après avoir fait une décomposition de Cholesky de la matrice A (voir paragraphe 5.5). Si la solution u de (9.1) est quatre fois continûment dérivable, il est possible de démontrer (en utilisant des notions de stabilité et consistance que nous n'introduirons pas ici) qu'il existe une constante C telle que

$$\max_{1 \leq j \leq N} | u(x_j) - u_j | \leq Ch^2. \tag{9.7}$$

On constate donc que si $(u_j)_{1 \leq j \leq N}$ est solution de (9.3) et si u est solution de (9.1), on a

$$\lim_{N \to \infty} \max_{1 \leq j \leq N} | u(x_j) - u_j | = 0; \tag{9.8}$$

l'erreur est en principe quatre fois plus petite lorsqu'on double le nombre de points de discrétisation.

bien expliqué !
pedagogie = O

9.2 Approximation par la méthode de Galerkin d'un problème aux limites

Considérons le problème (9.1) 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 aurons :

$$\int_0^1 u'(x)v'(x)\,dx-u'(1)v(1)+u'(0)v(0)+\int_0^1 c(x)u(x)v(x)\,dx=\int_0^1 f(x)v(x)\,dx.$$

Si nous imposons à la fonction v d'être nulle en $x=0$ et $x=1$ alors nous déduisons l'égalité :

$$\int_0^1 u'(x)v'(x)\,dx+\int_0^1 c(x)u(x)v(x)\,dx=\int_0^1 f(x)v(x)\,dx. \qquad (9.9)$$

Soit maintenant V l'ensemble de toutes les fonctions g continues, de première dérivée g' continue par morceaux et telles que $g(0)=g(1)=0$. Nous pouvons alors chercher $u\in V$ qui satisfait (9.9) pour toute fonction $v\in V$. Dans la suite ce problème est appelé "problème (9.9)".

Le problème (9.9) est dit "problème faible" ou "problème variationnel" ; on cherche a priori des fonctions u moins régulières que dans le problème différentiel (9.1).

De par notre façon de déduire le problème (9.9) du problème (9.1), il est évident que toute solution u de (9.1) est solution de (9.9). En fait, on peut montrer que si $c(x)\geq 0, \forall x\in[0,1]$, alors le problème (9.9) a une et une seule solution u qui est celle du problème (9.1).

Alors que la méthode des différences finies est une méthode d'approximation basée sur la formulation différentielle (9.1) du problème, la méthode des éléments finis a pour point de départ la formulation faible (9.9).

Si $\varphi_1,\varphi_2,\cdots,\varphi_N$ 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 par

$$g(x)=\sum_{i=1}^{N} g_i\,\varphi_i(x),$$

où les g_i sont N nombres réels.

Dès lors, une approximation du problème (9.9) se formulera de la manière suivante : trouver une fonction $u_h\in V_h$ telle que

$$\int_0^1 u_h'(x)v_h'(x)\,dx+\int_0^1 c(x)u_h(x)v_h(x)\,dx=\int_0^1 f(x)v_h(x)\,dx \qquad (9.10)$$

pour toute fonction $v_h\in V_h$. On dira que (9.10) est une approximation de Galerkin de (9.9).

Puisque u_h est cherché dans V_h, on peut écrire :

$$u_h(x) = \sum_{i=1}^{N} u_i \varphi_i(x),$$

où $u_1, u_2, \cdots, u_N$ sont N nombres réels à déterminer.

Le problème (9.10) est alors équivalent à chercher $u_1, u_2, \cdots, u_N$ tels que

$$\sum_{i=1}^{N} u_i \left(\int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)\,dx \right) = \int_0^1 f(x)\varphi_j(x)\,dx \qquad (9.11)$$

pour tout $j = 1,2,3,\cdots,N$.

Si A est la $N \times N$ matrice de coefficients

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)\,dx, \qquad (9.12)$$

(A est dite "matrice de rigidité" lorsque $c = 0$)
si $\vec{u}$ est le N-vecteur de composantes $u_1, u_2, \cdots, u_N$ et si $\vec{f}$ est le N-vecteur dont la $j^{\text{ème}}$ composante est

$$f_j = \int_0^1 f(x)\varphi_j(x)\,dx, \qquad (9.13)$$

alors le problème (9.10) ou (9.11) est équivalent à chercher $\vec{u}$ tel que

$$A\vec{u} = \vec{f}. \qquad (9.14)$$

Comme pour la méthode des différences finies, cette méthode d'approximation conduit à la résolution d'un système linéaire après la construction de la matrice A et du vecteur $\vec{f}$.

La méthode des éléments finis revient à faire un choix judicieux des fonctions $\varphi_1, \varphi_2, \cdots, \varphi_N$ qui définissent V_h de telle sorte que :

1) La matrice A soit de bande pour résoudre (9.14) par la technique donnée dans le paragraphe 5.5 (leçon 5).

2) La solution u_h du problème (9.10) converge dans un certain sens vers la solution u du problème (9.9) lorsque N croît.

9.3 Méthode d'éléments finis sous la forme la plus simple

Divisons l'intervalle $[0,1]$ en $(N+1)$ parties (N étant un entier positif) et posons $h = \frac{1}{N+1}$, $x_i = ih$ avec $i = 0,1,2,\cdots,N+1$, comme dans la figure 9.2.

Pour $i = 1,2,3,\cdots,N$ on définit les fonctions

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} = h & si\ x_{i-1} \le x \le x_i, \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}} = h & si\ x_i \le x \le x_{i+1}, \\ 0 & si\ x \le x_{i-1}\ ou\ x \ge x_{i+1}. \end{cases}$$

Le graphe de la fonction φ_i est représenté en figure 9.3.

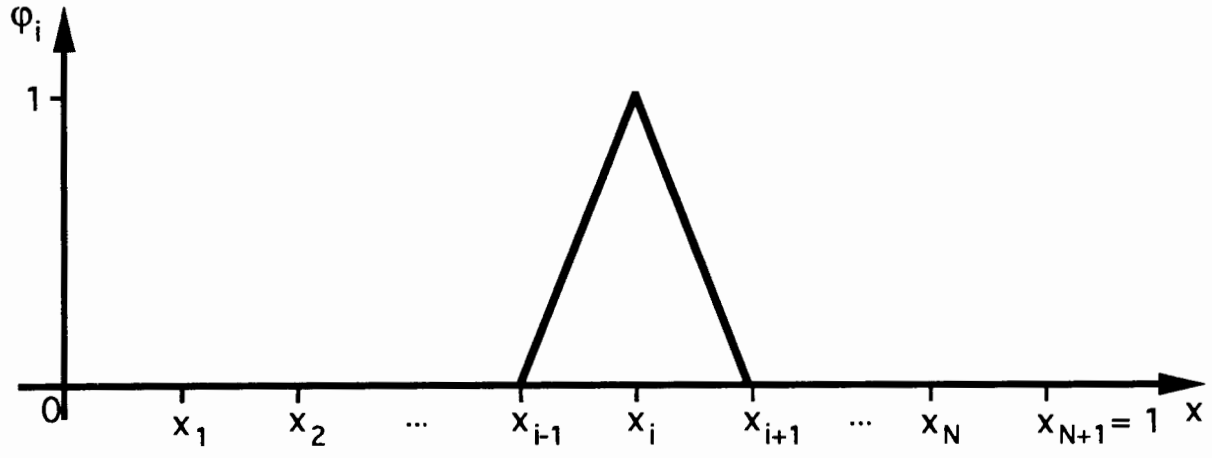


Figure 9.3 : Fonction de base de V_h

Remarquons dans ce cas-là que 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 donné dans la figure 9.4.

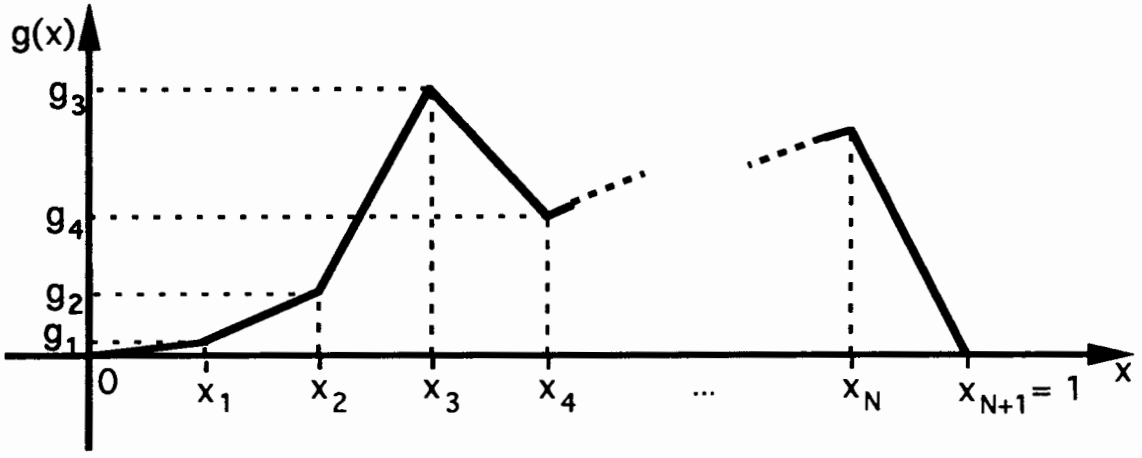


Figure 9.4 : Graphe d'un élément de V_h

On dira que :

- $x_0, x_1, x_2, \cdots, x_{N+1}$ sont les noeuds de la discrétisation,
- $[x_0, x_1], [x_1, x_2], \cdots, [x_N, x_{N+1}]$ sont les éléments géométriques,
- $\varphi_1, \varphi_2, \cdots, \varphi_N$ sont des fonctions de base du sous-espace V_h de type "éléments finis".

Nous avons vu que pour résoudre numériquement le problème (9.9), i.e. pour résoudre le problème (9.10), on doit construire la matrice A et le vecteur $\vec{f}$, puis résoudre le système $A\vec{u} = \vec{f}$. On est donc conduit à calculer les expressions :

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx + \int_0^1 c(x)\varphi_i(x)\varphi_j(x)\,dx$$

et

$$f_j = \int_0^1 f(x)\varphi_j(x)\,dx.$$

On vérifiera que

$$\int_0^1 \varphi_i'(x)\varphi_j'(x)\,dx = \begin{cases} 2/h & si\ i = j, \\ -1/h & si\ i \neq j,\ \ |i-j| = 1, \\ 0 & autrement. \end{cases}$$

D'autre part, pour obtenir les valeurs de $\int_0^1 c(x)\varphi_i(x)\varphi_j(x)\,dx$ et $\int_0^1 f(x)\varphi_j(x)\,dx$, on est conduit à prendre une formule d'intégration numérique. Choisissons la formule des trapèzes, i.e. on approche $\int_0^1 \ell(x)\,dx$ par

on intègre sur chaque élément géométrique.

$$L_h(\ell) = h\big(\tfrac{1}{2}\ell(x_0) + \ell(x_1) + \ell(x_2) + \cdots + \ell(x_N) + \tfrac{1}{2}\ell(x_{N+1})\big).$$

$$\ell_h = \frac{h}{2}\big(L(x_0) + L(x_1)\big) + \frac{h}{2}\big(L(x_1) + L(x_2)\big) + \cdots + \frac{h}{2}\big(L(x_N) + L(x_{N+1})\big)$$

On vérifiera aisément que

$$L_h(c\,\varphi_i\,\varphi_j) = \begin{cases} h\,c(x_j) & si\ i = j, \\ 0 & si\ i \neq j, \end{cases} \qquad (9.15)$$

et

$$L_h(f\,\varphi_j) = h\,f(x_j). \qquad (9.16)$$

Si, dans (9.12) on remplace $\int_0^1 c(x)\varphi_i(x)\varphi_j(x)\,dx$ par (9.15) et si on remplace (9.13) par (9.16), alors on vérifie que le système (9.14) est exactement égal à h fois le système obtenu

dans (9.6). 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 se laisse facilement généraliser aux situations suivantes :

- distribution non uniforme des points de discrétisation $(x_j)_{1\le j\le N}$ ce qui permet de concentrer les noeuds aux endroits de "forte variation de la solution" ;

- fonctions de base de V_h définies par des polynômes de degré plus élevé que 1 sur chaque élément géométrique $[x_j, x_{j+1}]$ ce qui permet d'augmenter la précision de la méthode. (Ici les notions d'interpolation mentionnés en leçon 1 interviennent de façon capitale).

De plus, la théorie des éléments finis est complètement "mathématisée" et des études de convergence avec estimation des erreurs existent dans des cadres tout à fait généraux qui comprennent bon nombre d'applications physiques.

9.4 Approximation par différences finies d'un problème aux limites non linéaire

Revenons au problème (9.1) et supposons maintenant que c ne dépende pas seulement de x mais aussi de u. Dans ce cas, nous considérons non seulement la fonction f connue mais aussi nous nous donnons une fonction à deux variables $\tilde{c} : (x,v) \in [0,1] \times \mathbb{R} \to \tilde{c}(x,v) \in \mathbb{R}$ pour poser le problème aux limites

$$\begin{cases} -u''(x) + \tilde{c}(x,u(x)) = f(x) \quad si\ 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{9.17}$$

où u est naturellement la fonction inconnue. Si $\tilde{c}(x,v) = c(x)v$, où $c(x)$ est une fonction donnée, nous retrouvons le problème (9.1).

Pour résoudre numériquement (9.17), nous reprenons la discrétisation de la figure (9.2) et de façon semblable à (9.3) nous aurons, si u_j est une approximation de $u(x_j)$:

$$\frac{-u_{j-1}+2u_j-u_{j+1}}{h^2} + \tilde{c}(x_j,u_j) = f(x_j) \quad si\ 1 \le j \le N, \tag{9.18}$$

$$u_0 = u_{N+1} = 0. \tag{9.19}$$

Clairement, (9.18) avec (9.19) est une approximation par différences finies de (9.17) ; nous sommes en présence d'un système en principe non linéaire de N équations pour les N inconnues $u_1, u_2, \cdots, u_N$.

Si nous voulons maintenant résoudre le système (9.18) par la méthode de Newton, nous posons, si $\vec{u}$ est le N-vecteur de composantes $u_1, u_2, \cdots, u_N$:

$$F(\vec{u}) \underset{def}{\equiv} \begin{bmatrix} \frac{2u_1 - u_2}{h^2} & + \tilde{c}(x_1, u_1) - f(x_1) \\ \frac{-u_1 + 2u_2 - u_3}{h^2} & + \tilde{c}(x_2, u_2) - f(x_2) \\ \vdots & \\ \frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} & + \tilde{c}(x_j, u_j) - f(x_j) \\ \vdots & \\ \vdots & \\ \frac{-u_{N-1} + 2u_N}{h^2} & + \tilde{c}(x_N, u_N) - f(x_N) \end{bmatrix}. \tag{9.20}$$

Clairement F est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}^N$ et le problème (9.18)-(9.19) est équivalent à chercher $\vec{u}$ tel que :

$$F(\vec{u}) = 0. \tag{9.21}$$

Si nous connaissons une solution approchée $\vec{u}^0$, la méthode de Newton pour résoudre (9.21) devient (voir leçon 7) :

$$\vec{u}^{n+1} = \vec{u}^n - DF(\vec{u}^n)^{-1} F(\vec{u}^n), \quad n = 0,1,2,\cdots. \tag{9.22}$$

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 dans la suite, nous définissons encore

$$d_j^n \underset{def}{=} d(x_j, u_j^n), \quad 1 \le j \le N.$$

Il est alors facile de vérifier que la matrice $DF(\vec{u}^n)$ est donnée par :

$$DF(\vec{u}^n) = \frac{1}{h^2}\begin{bmatrix} 2+d_1^n h^2 & -1 & & & & \\ -1 & 2+d_2^n h^2 & -1 & & \mathrm{O} & \\ & -1 & 2+d_3^n h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & \mathrm{O} & & \ddots & \ddots & -1 \\ & & & & -1 & 2+d_N^n h^2 \end{bmatrix}. \quad (9.23)$$

Ainsi, pour calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$ dans (9.22), on calculera (cf. leçon 7) :

- le vecteur $F(\vec{u}^n)$ en remplaçant $u_1, u_2, \cdots, u_N$ dans (9.20) par $u_1^n, u_2^n, \cdots, u_N^n$;
- la matrice $DF(\vec{u}^n)$ par l'expression (9.23) ;
- le vecteur $\vec{y}$ solution de $DF(\vec{u}^n)\vec{y} = F(\vec{u}^n)$ (élimination de Gauss) ;
- le vecteur $\vec{u}^{n+1} = \vec{u}^n - \vec{y}$.

La méthode de Newton pour résoudre (9.21) est très rapidement convergente (convergence quadratique). Si $\vec{u}^0$ est choisi "relativement proche" d'une solution $\vec{u}$ du problème (9.21) (ou (9.18)-(9.19)), l'algorithme proposé ci-dessus permettra d'obtenir en peu de pas une solution numérique par différences finies du problème (9.17).

Remarquons enfin que si $\tilde{c}(x,u) = c(x)u$ comme dans (9.1), 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 (9.5) et devient indépendante de n. Ainsi, (9.22) nous fournit pour $n = 1$ (1 seul pas) la solution de (9.21).

Leçon 10

Une méthode d'éléments finis pour le problème de Poisson

Une méthode de différences finies pour le problème de la chaleur

10.1 Problème de Poisson et formulation variationnelle

Soit Ω un domaine polygonal dans le plan Ox_1x_2 et soit $\partial\Omega$ sa frontière. Si $f : \Omega \to \mathbb{R}$ est une fonction continue donnée, on pose le problème de trouver $u : \Omega \to \mathbb{R}$ telle que

$$\begin{aligned} &-\Delta u(x_1,x_2) = f(x_1,x_2), \quad \forall (x_1,x_2) \in \Omega, \\ &u(x_1,x_2) = 0, \quad \forall (x_1,x_2) \in \partial\Omega, \end{aligned} \tag{10.1}$$

où Δu est le Laplacien de u, i.e. $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$. Le problème (10.1) est appelé "problème de Poisson" ; c'est un problème différentiel aux limites d'ordre 2, la condition limite étant la $2^{\text{ème}}$ relation de (10.1).

Mentionnons que le problème (10.1) apparaît dans de nombreuses modélisations physiques telles que des problèmes de potentiel, de déformations de membranes, d'écoulements de fluides, ... L'exemple traditionnel de situation physique est celui du déplacement vertical $u(x_1,x_2)$ au point (x_1,x_2) d'une membrane tendue, attachée à $\partial\Omega$, et soumise à une densité de force verticale "normalisée" f (voir figure 10.1).

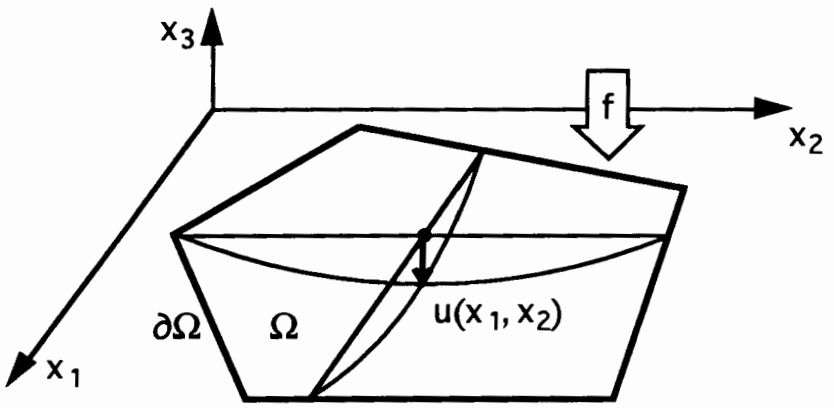


Figure 10.1 : Déformation verticale d'une membrane soumise à une force verticale

Comme fait dans le paragraphe 9.2, nous pouvons multiplier la première équation de (10.1) par une fonction $v : \Omega \to \mathbb{R}$ "suffisamment régulière" et intégrer sur Ω. Nous obtenons

$$-\iint_{\Omega} \Delta u(x)\, v(x)\, dx = \iint_{\Omega} f(x)\, v(x)\, dx, \tag{10.2}$$

où $x = (x_1, x_2)$ et $dx = dx_1\, dx_2$.

Connaissant la formule (voir cours Analyse III) :

$$div\,(v\, g\vec{r}ad\, u) = g\vec{r}ad\, v \cdot g\vec{r}ad\, u + v\, \Delta u, \tag{10.3}$$

nous obtenons à partir de (10.2) :

$$\iint_{\Omega} g\vec{r}ad\, u(x) \cdot g\vec{r}ad\, v(x)\, dx - \iint_{\Omega} div\,(v(x)\, g\vec{r}ad\, u(x))\, dx = \iint_{\Omega} f(x)\, v(x)\, dx. \tag{10.4}$$

Le théorème de la divergence permet d'écrire :

$$\iint_{\Omega} div\,(v(x)\, g\vec{r}ad\, u(x))\, dx = \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, ds, \tag{10.5}$$

où $\frac{\partial u}{\partial n}$ est la dérivée de u dans la direction normale extérieure à $\partial\Omega$. Si nous assurons que v s'annule sur $\partial\Omega$, nous pouvons déduire de (10.4) et (10.5) :

$$\iint_{\Omega} g\vec{r}ad\, u(x) \cdot g\vec{r}ad\, v(x)\, dx = \iint_{\Omega} f(x)\, v(x)\, dx. \tag{10.6}$$

Soit maintenant V l'ensemble de toutes les fonctions $g : \overline{\Omega} \to \mathbb{R}$ $(\overline{\Omega} = \Omega \cup \partial\Omega)$ qui sont continues sur $\overline{\Omega}$, nulles sur $\partial\Omega$, et dont les premières dérivées partielles $\frac{\partial g}{\partial x_1}$ et $\frac{\partial g}{\partial x_2}$ sont continues par morceaux. Nous pouvons chercher $u \in V$ qui satisfait (10.6) pour toute fonction $v \in V$. Dans la suite, ce problème est appelé "problème (10.6)". Ici encore, nous obtenons une formulation faible ou variationnelle du problème (10.1), (cf. leçon 9).

La formulation variationnelle (10.6) a souvent une signification physique (par exemple signifie qu'une énergie est minimisée) alors que la formulation différentielle (10.1) est formelle.

Contrairement au cas unidimensionnel vu dans la leçon 9, suivant la forme du domaine Ω et suivant le second membre f, le problème (10.6) peut ne pas avoir de solution u dans V. Cependant, nous supposerons dans la suite que ce n'est pas le cas et nous appelerons u la solution de (10.6).

Dans la leçon 9, nous avons vu qu'une méthode d'approximation consiste à construire un sous-espace V_h de dimension finie de V et de résoudre le problème (10.6) dans V_h au lieu de V, c'est-à-dire de trouver $u_h \in V_h$ tel que

$$\iint_\Omega \vec{grad}\, u_h(x) \cdot \vec{grad}\, v_h(x)\, dx = \iint_\Omega f(x)\, v_h(x)\, dx, \quad \forall v_h \in V_h. \tag{10.7}$$

Rappelons que le problème (10.7) est dit "approximation de Galerkin" de (10.6). Si $\varphi_1, \varphi_2, \cdots, \varphi_N$ est une base de V_h, il suffit de construire la matrice A de coefficients $A_{ji} = \iint_\Omega \vec{grad}\, \varphi_i(x) \cdot \vec{grad}\, \varphi_j(x)\, dx$, le vecteur $\vec{f}$ de composantes $f_j = \iint_\Omega f(x)\, \varphi_j(x)\, dx$ et de résoudre le système linéaire

$$A\vec{u} = \vec{f}, \tag{10.8}$$

pour obtenir la solution de (10.7) (on aura en effet $u_h(x) = \sum_{i=1}^{N} u_i\, \varphi_i(x)$ où u_i est la $i^{\text{ème}}$ composante de $\vec{u}$; la démarche est analogue à celle du problème unidimensionnel considéré à la leçon 9).

Ci-après nous donnons une construction simple d'un sous-espace V_h de V.

10.2 Méthode d'éléments finis triangulaires de degré 1

Nous voulons construire des sous-espaces V_h de V de type "éléments finis triangulaires" dans le cas où Ω est un domaine polygonal de $\mathbb{R}^2$. Pour réaliser ce but, nous prenons une triangulation τ_h de $\overline{\Omega}$ en subdivisant $\overline{\Omega}$ en triangles $K_1, K_2, \cdots, K_m$ qui ne se recouvrent pas et tels que

- $\overline{\Omega} = \bigcup_{K \in \tau_h} K = K_1 \cup K_2 \cup K_3 \cup \cdots \cup K_m$,
- 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 (voir figure 10.2).

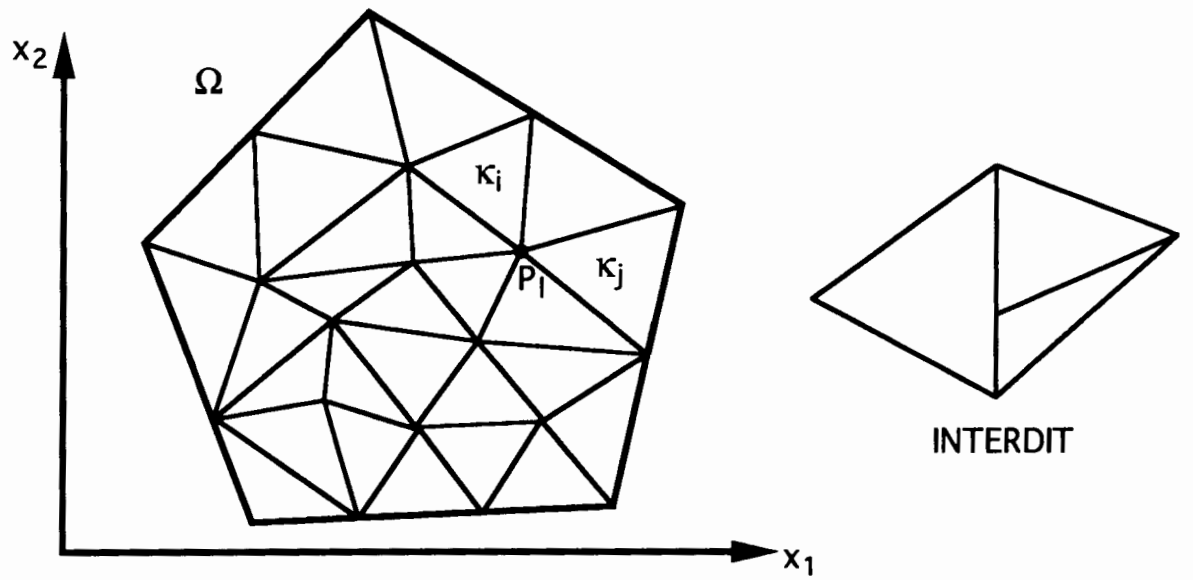


Figure 10.2 : Triangulation de Ω

Nous introduisons encore un paramètre h qui mesure le "degré de finesse" de la triangulation τ_h :

$$h = \max_{K \in \tau_h} \operatorname{diam}(K), \tag{10.9}$$

où $\operatorname{diam}(K)$ 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 sera défini par

$$V_h = \{\, g : \overline{\Omega} \to \mathbb{R} : g \text{ est continue sur } \overline{\Omega}, \text{ s'annule sur } \partial\Omega, \\ \text{et la restriction de } g \text{ à } K \text{ est un polynôme de degré} \le 1, \quad \forall K \in \tau_h \,\}. \tag{10.10}$$

Pour décrire une fonction $g \in V_h$, nous pouvons choisir comme paramètres les valeurs $g(P_i)$ de la fonction g aux noeuds $P_i, i = 1,2,\cdots,N$, intérieurs à la triangulation τ_h ; ces valeurs $g(P_i)$ sont appelées "degrés de liberté". Les fonctions de base $\varphi_1, \varphi_2, \cdots, \varphi_N$ de V_h sont définies alors par

$$\varphi_i(P_j) = \delta_{ij} = \begin{cases} 1 & si\ i = j \\ 0 & si\ i \neq j \end{cases} \quad 1 \le i, j \le N, \tag{10.11}$$

elles sont prises nulles sur $\partial\Omega$. Le support de φ_i (fermeture de l'ensemble des points où φ_i n'est pas nul) est composé de tous les triangles qui ont pour sommet P_i. Ainsi, à chaque noeud intérieur P_i, nous y associons la fonction de base φ_i (voir figure 10.3).

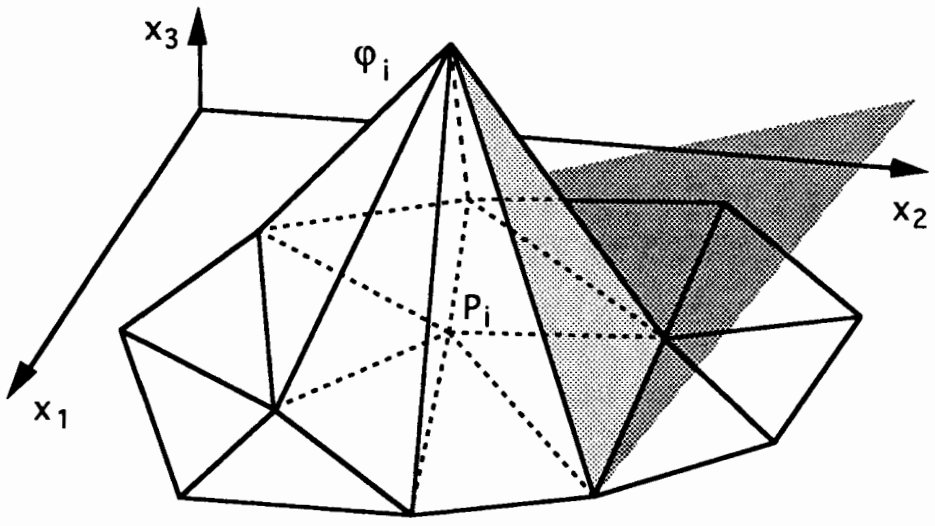


Figure 10.3 : La fonction de base φ_i

Une fonction g de V_h peut être représentée par une combinaison linéaire des φ_i et donc :

$$g(x) = \sum_{i=1}^{N} g_i \, \varphi_i(x) \quad où \; g_i \in \mathbb{R}, \; i = 1, \cdots, N;$$

on a $g(P_i) = g_i, 1 \le i \le N$; ici $x = (x_1, x_2)$.

Clairement, la triangulation étant donnée, nous pouvons avoir une connaissance explicite des fonctions de base $\varphi_1, \varphi_2, \cdots, \varphi_N$ et ainsi il est possible de construire la matrice A et le vecteur second membre $\vec{f}$ apparaissant dans (10.8). A noter que pour calculer numériquement $f_j = \iint_\Omega f(x) \, \varphi_j(x) \, dx, j = 1,2, \cdots, N$, il conviendra peut-être de faire usage d'une formule de quadrature numérique.

Il est facile de vérifier le résultat suivant :

Théorème. A est une matrice symétrique définie positive.

Démonstration. La symétrie est évidente ; quant à l'aspect "défini-positif", il suffit de voir que pour un N-vecteur $\vec{y}$ de composantes $y_1, y_2, \cdots, y_N$ on a :

$$\vec{y}^{\,T} A \vec{y} = \sum_{i,j=1}^{N} y_i \, A_{ij} \, y_j = \sum_{i,j=1}^{N} y_i \, y_j \iint_\Omega \vec{grad} \, \varphi_i(x) \cdot \vec{grad} \, \varphi_j \, dx = \iint_\Omega \Big| \sum_{i=1}^{N} y_i \, \vec{grad} \, \varphi_i(x) \Big|^2 \, dx;$$

en posant

$$\psi(x) = \sum_{i=1}^{N} y_i \, \varphi_i(x),$$

nous obtenons

$$\vec{y}^{\,T} A\vec{y} = \iint_{\Omega} \left| \vec{grad}\ \psi(x) \right|^2 dx$$

qui est toujours positif ou nul. Si $\vec{y}^{\,T} A\vec{y} = 0$ alors $\vec{grad}\ \psi(x) = 0$ ce qui implique $\psi(x) =$ constante. Puisque ψ est nul sur $\partial\Omega$, on aura $\psi \equiv 0$ et donc $\vec{y} = 0$.

#

La matrice A, dite de rigidité, et le vecteur $\vec{f}$ étant constitués, on résoudra le système (10.8) en faisant une décomposition de Cholesky de la matrice A, puis deux résolutions de systèmes triangulaires (voir leçon 5). La solution $\vec{u}$ obtenue ainsi permettra d'exprimer une solution approchée u_h du problème (10.1) sous la forme $u_h(x) = \sum_{i=1}^{N} u_i\, \varphi_i(x)$; en outre, on aura $u_h(P_j) = u_j, 1 \le j \le N$.

Il reste beaucoup de choses à dire sur la méthode des éléments finis. En particulier, nous avons laissé de côté ici les questions importantes suivantes :

- techniques de maillage, de mémorisation et de construction des matrices de rigidité ;
- généralisation à des polynômes de degré plus élevé que 1 sur chaque triangle de la triangulation τ_h ;
- estimation de l'erreur entre u et u_h en fonction du paramètre h défini par (10.9) ;
- généralisation à des éléments quadrangulaires plutôt que triangulaires ;
- généralisation à des problèmes posés dans un espace tridimensionnel ;
- généralisation à des méthodes d'éléments finis non-standard ... .

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, les structures solides, l'électromagnétisme, la thermique, ... La plupart de ces logiciels utilisent la technique des éléments finis.

10.3 Une méthode de différences finies pour le problème de la chaleur

Considérons le problème modèle donné par la diffusion de la chaleur à une dimension d'espace dans un barreau occupant l'intervalle (0,1) (voir cours d'Analyse III) et à température nulle à chaque extrémité. On se donnera donc deux fonctions :

$$w : x \in (0,1) \to w(x) \in \mathbb{R}$$

et

$$f : (x,t) \in (0,1) \times (0,\infty) \to f(x,t) \in \mathbb{R} ;$$

il s'agira alors de trouver une fonction $u : (x,t) \in (0,1) \times (0,\infty) \to u(x,t) \in \mathbb{R}$ telle que

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad \forall x \in (0,1), \quad \forall t > 0, \tag{10.12}$$

et satisfaisant les conditions aux limites

$$u(0,t) = u(1,t) = 0, \quad \forall t > 0, \tag{10.13}$$

et la condition initiale

$$u(x,0) = w(x), \quad \forall x \in (0,1). \tag{10.14}$$

Rappelons que dans le problème physique, l'équation (10.12) est en fait $\rho\, c_p \frac{\partial u}{\partial t} - k \frac{\partial^2 u}{\partial x^2} = f$ où $u(x,t)$ est la température à l'endroit x et à l'instant t, $f(x,t)$ est la puissance par unité de longueur fournie à l'endroit x et à l'instant t, ρ est la densité, c_p la chaleur spécifique et k la conductivité thermique du barreau supposée constante ; naturellement $w(x)$ est la température initiale du barreau à l'endroit x.

Pour résoudre numériquement le problème (10.12), (10.13), (10.14), nous commençons par le discrétiser par rapport à la variable x de façon semblable à ce qui a été fait dans le paragraphe 9.1. Si N est un entier positif, nous posons $h = \frac{1}{N+1}$ et $x_j = jh$ avec $j = 0,1,2,\cdots,N+1$. Au vu du paragraphe 9.1, il est naturel de prendre le schéma :

$$\frac{d}{dt}u_i(t) + \frac{1}{h^2}(-u_{i-1}(t) + 2u_i(t) - u_{i+1}(t)) = f(x_i,t), \quad i = 1,2,\cdots,N, \tag{10.15}$$

$$u_0(t) = u_{N+1}(t) = 0, \quad \forall t \in (0,T), \tag{10.16}$$

$$u_i(0) = w(x_i), \quad i = 1,2,\cdots,N, \tag{10.17}$$

où $u_i(t)$ est une approximation de $u(x,t)$ au point $x = x_i$. On notera $u_i(t) \simeq u(x_i,t)$; les fonctions $u_i(t), 1 \le i \le N$ sont maintenant les inconnues du problème.

On dit que le schéma (10.15), (10.16), (10.17) est une semi-discrétisation en espace du problème (10.12), (10.13), (10.14) par la méthode des différences finies. Si A est la $N \times N$ matrice donnée par

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & \mathrm{O} & \\ & -1 & 2 & -1 & & \\ & & \cdot & \cdot & \cdot & \\ & \mathrm{O} & & \cdot & \cdot & -1 \\ & & & & -1 & 2 \end{bmatrix},$$

si $\vec{u}(t)$ est le N-vecteur de composantes $u_1(t), u_2(t), \cdots, u_N(t)$, si $\vec{f}(t)$ est le N-vecteur de composantes $f(x_1,t), f(x_2,t), \cdots, f(x_N,t)$ et si $\vec{w}$ est le N-vecteur de composantes $w(x_1), w(x_2), \cdots, w(x_N)$, alors le schéma (10.15), (10.16), (10.17) est équivalent au système différentiel :

$$\dot{\vec{u}}(t) = -A\vec{u}(t) + \vec{f}(t), \quad t \in (0,\infty), \tag{10.18}$$

$$\vec{u}(0) = \vec{w}, \tag{10.19}$$

où $\dot{\vec{u}}(t)$ est la dérivée de $\vec{u}(t)$ par rapport à t au temps t.

Dès lors, une semi-discrétisation en espace du problème (10.12), (10.13), (10.14) nous conduit à résoudre un système différentiel du 1er ordre avec une condition initiale qui pourrait être intégré numériquement par une méthode vue dans la leçon 8.

Pour illustrer ceci, choisissons deux méthodes d'Euler.

1) **Méthode d'Euler progressive :** si $\tau > 0$ est un "pas de temps" choisi et si $t_n = n\tau$ avec $n = 0,1,2,\cdots$, nous notons par $\vec{u}^n$ une approximation de $\vec{u}$ au temps t_n, i.e. $\vec{u}^n \simeq u(t_n)$, et nous écrivons le schéma :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^n + \vec{f}(t_n), \quad n = 0,1,2,\cdots, \tag{10.20}$$

$$\vec{u}^0 = \vec{w}. \tag{10.21}$$

Clairement nous avons

$$\vec{u}^{n+1} = (I - \tau A)\vec{u}^n + \tau \vec{f}(t_n), \quad n = 0,1,2,\cdots, \tag{10.22}$$

où I est la $N \times N$ matrice identité ; le vecteur $\vec{u}^{n+1}$ peut être calculé explicitement dès que l'on connaît $\vec{u}^n$. Ainsi, à partir de $\vec{u}^0 = \vec{w}$, on peut calculer de proche en proche $\vec{u}^1, \vec{u}^2, \vec{u}^3, \cdots$ en utilisant (10.22) ; la $j^{\text{ème}}$ composante u_j^n de $\vec{u}^n$ devient une approximation de $u(x_j, t_n), 1 \le j \le N, n \ge 0$. Le schéma numérique (10.20), (10.21) est une discrétisation (complète) du problème (10.12), (10.13), (10.14) par la méthode des différences finies en utilisant une méthode d'Euler explicite pour la discrétisation en temps.

Tout comme dans le paragraphe 8.3, il intervient ici une condition de stabilité qui limite le pas temporel τ. Cette limitation est fonction du pas spatial et est exprimée par la condition :

$$\tau \le \frac{h^2}{2}. \tag{10.23}$$

Nous pourrions montrer de façon semblable à ce qui a été fait dans le paragraphe 8.2, que si f est identiquement nul alors $\vec{u}(t)$ converge vers zéro lorsque t tend vers l'infini. La condition de stabilité (10.23) permet d'éviter les instabilités numériques qui se manifestent par des grandeurs u_i^n qui deviennent de plus en plus grandes en alternant le signe lorsque n tend vers l'infini.

2) **Méthode d'Euler rétrograde :** si nous avions choisi un schéma d'Euler rétrograde à la place du schéma d'Euler progressif pour discrétiser (10.18), (10.19), nous aurions eu à la place de (10.20) :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^{n+1} + \vec{f}(t_{n+1}), \quad n = 0,1,2,\cdots, \tag{10.24}$$

ou de façon équivalente

$$(I + \tau A)\,\vec{u}^{n+1} = \vec{u}^n + \tau \vec{f}(t_{n+1}), \quad n = 0,1,2,\cdots. \tag{10.25}$$

Avec ce schéma, nous devons résoudre un système de N équations à N inconnues pour calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$; ce schéma est donc implicite. La matrice $(I + \tau A)$ étant définie positive et de bande (demi-largeur 2), nous pouvons le résoudre par la méthode donnée dans le paragraphe 5.5. Contrairement au schéma explicite donné précédemment, ce schéma implicite est inconditionnellement stable, ce qui signifie que pour n'importe quel pas temporel τ choisi, si f est identiquement nul, les valeurs u_i^n restent bornées lorsque n tend vers l'infini. Nous allons même montrer dans ce cas-là le résultat suivant :

Théorème. Si $f \equiv 0$ alors on a pour tout pas de temps τ :

$$\lim_{n \to \infty} \| \vec{u}^n \| = 0, \tag{10.26}$$

lorsque $\vec{u}^n, n = 0,1,2,\cdots$ satisfait (10.25).

Démonstration. En utilisant (10.25) nous aurons

$$\vec{u}^{n+1} = (I + \tau A)^{-1} \vec{u}^n. \tag{10.27}$$

Si $|||\cdot|||$ est la norme spectrale d'une matrice (cf. leçon 4, paragraphe 4.6), alors en utilisant (10.27) nous aurons :

$$\|\vec{u}^{n+1}\| \leq |||(I + \tau A)^{-1}||| \cdot \|\vec{u}^n\| \tag{10.28}$$

et par suite, puisque $(I + \tau A)^{-1}$ est symétrique :

$$\| \vec{u}^{n+1} \| \leq \beta \| \vec{u}^n \| \tag{10.29}$$

où β est le maximum des valeurs propres de $(I + \tau A)^{-1}$ en valeur absolue. Dans la leçon 5 (paragraphe 5.5), 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 $(I + \tau \lambda_A)^{-1}$, on conclut facilement qu'elles sont toutes comprises entre zéro et un et donc on a $\beta < 1$. De (10.29) on tire

$$\| \vec{u}^n \| \leq \beta^n \| \vec{u}^0 \| \tag{10.30}$$

et par suite la relation (10.26) est démontrée.

#

Les méthodes d'Euler ci-dessus sont toutes les deux d'ordre 1 en temps et d'ordre 2 en espace, c'est-à-dire lorsqu'on veut intégrer numériquement (10.12) jusqu'à un temps $T > 0$ par le schéma (10.20) sous la condition $\tau < h^2/2$ ou par le schéma (10.24), on obtient une erreur de l'ordre $O(\tau + h^2)$.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Département de Mathématiques

Prof. Jacques Rappaz

ANALYSE NUMÉRIQUE

Complément du polycopié
Leçons 10 à 13

Table des matières

Chapitre 10

Une méthode d'éléments finis pour l'approximation de problèmes elliptiques

10.1 Problèmes elliptiques et formulation variationnelle

Soit Ω un domaine polygonal dans le plan Ox_1x_2 de frontière $\partial\Omega$ et soit $\overline{\Omega} = \Omega \cup \partial\Omega$ (voir figure 10.1). Si a_{11}, a_{12}, a_{21}, a_{22} sont quatre nombres réels donnés et si $f : \overline{\Omega} \to \mathbb{R}$ est une fonction continue donnée, nous posons le problème de trouver la fonction $u : \overline{\Omega} \to \mathbb{R}$ satisfaisant les relations :

$$-\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}u(x_1,x_2)\right) = f(x_1,x_2) \qquad \forall (x_1,x_2) \in \Omega, \quad (10.1)$$

$$u(x_1,x_2) = 0 \qquad \forall (x_1,x_2) \in \partial\Omega, \quad (10.2)$$

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 (10.1) (10.2) est un problème différentiel aux limites d'ordre 2, la condition limite étant l'équation (10.2).

Définition 10.1 *Nous dirons que le problème (10.1) (10.2) est elliptique si les coefficients a_{ij}, $1 \leq i,j \leq 2$, sont choisis de sorte à ce que l'équation $a_{11}x_1^2 + (a_{12}+a_{21})x_1x_2 + a_{22}x_2^2 = 1$ soit l'équation d'une ellipse (ou d'un cercle).*

Si $a_{12} = a_{21}$, une condition nécessaire et suffisante pour que cette propriété soit vraie, est que la matrice

$$S = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

soit symétrique définie positive.

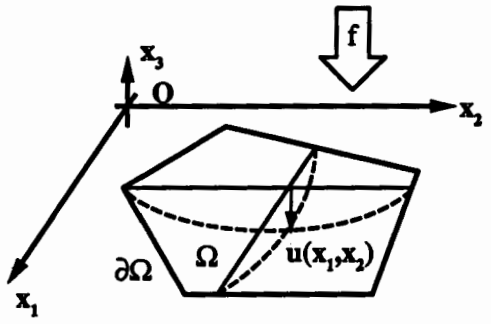


FIG. 10.1 - *Déformation verticale d'une membrane soumise à une force verticale.*

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. La méthode des éléments finis, que nous décrivons ci-dessous, est une méthode souvent utilisée pour résoudre numériquement ce type de problèmes.

Mentionnons qu'un exemple typique de problème 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

$$-\Delta u(x_1, x_2) = f(x_1, x_2) \qquad \forall (x_1, x_2) \in \Omega, \tag{10.3}$$

$$u(x_1, x_2) = 0 \qquad \forall (x_1, x_2) \in \partial\Omega, \tag{10.4}$$

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 (10.3) (10.4) est appelé "problème de Poisson".

La solution u du problème de Poisson modélise le déplacement vertical $u(x_1, x_2)$ au point (x_1, x_2) d'une membrane Ω tendue, attachée à $\partial\Omega$, et soumise à une densité de force verticale et proportionnelle à f (voir figure 10.1).

Comme nous l'avons déjà fait dans le paragraphe 9.2, nous pouvons multiplier la première équation de (10.3) par une fonction $v : \overline{\Omega} \to \mathbb{R}$ "suffisamment régulière" et intégrer sur Ω. Nous obtenons

$$-\iint_\Omega \Delta u(x)\, v(x) dx = \iint_\Omega f(x) v(x) dx, \tag{10.5}$$

où $x = (x_1, x_2)$ et $dx = dx_1 dx_2$. En utilisant la formule (voir le cours Analyse III) :

$$\operatorname{div}(v\, \overrightarrow{\operatorname{grad} u}) = \overrightarrow{\operatorname{grad} v} \cdot \overrightarrow{\operatorname{grad} u} + v\Delta u, \tag{10.6}$$

nous obtenons à partir de (10.5) :

$$\iint_\Omega \overrightarrow{\operatorname{grad} u}(x) \cdot \overrightarrow{\operatorname{grad} v}(x) dx - \iint_\Omega \operatorname{div}\left(v(x)\overrightarrow{\operatorname{grad} u}(x)\right) dx = \iint_\Omega f(x) v(x) dx. \tag{10.7}$$

Le théorème de la divergence nous assure que :

$$\iint_\Omega \operatorname{div}\left(v(x)\overrightarrow{\operatorname{grad}}u(x)\right)dx = \int_{\partial\Omega} v(s)\frac{\partial u}{\partial n}(s)ds, \tag{10.8}$$

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 (10.7) et (10.8) :

$$\iint_\Omega \overrightarrow{\operatorname{grad}}u(x)\cdot\overrightarrow{\operatorname{grad}}v(x)dx = \iint_\Omega f(x)v(x)dx. \tag{10.9}$$

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. Nous pouvons

chercher $u \in V$ qui satisfait (10.9) pour toute fonction $v \in V$.

Dans la suite, ce problème est appelé "problème (10.9)". Ici encore, nous obtenons une formulation faible ou variationnelle du problème (10.3) (10.4), (cf. leçon 9).

La formulation variationnelle (10.9) 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 variationnelle du problème plus général (10.1) (10.2).

Contrairement au cas unidimensionnel de la leçon 9, le problème (10.9) 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 (10.9).

Dans la leçon 9, 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 (10.9) dans V_h au lieu de V, c'est-à-dire de trouver $u_h \in V_h$ tel que

$$\iint_\Omega \overrightarrow{\operatorname{grad}}u_h(x)\cdot\overrightarrow{\operatorname{grad}}v_h(x)dx = \iint_\Omega f(x)v_h(x)dx \tag{10.10}$$

pour toute fonction $v_h \in V_h$. Rappelons que le problème (10.10) est appelé "approximation de Galerkin" de (10.9). 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 $v_h = \varphi_j$, $j = 1, \ldots, N$ dans (10.10). Soit $\vec{u}$ le N-vecteur de composantes $u_1, \ldots, u_N$, soit A la $N \times N$ matrice de coefficients

$$A_{ji} = \iint_\Omega \overrightarrow{\operatorname{grad}}\varphi_i(x)\cdot\overrightarrow{\operatorname{grad}}\varphi_j(x)dx, \tag{10.11}$$

et $\vec{f}$ le N-vecteur de composantes $f_1, \ldots, f_N$ définies par

$$f_j = \iint_\Omega f(x)\varphi_j(x)dx. \tag{10.12}$$

Pour obtenir la solution de (10.10) il suffit donc de trouver $u_1, \ldots, u_N$ tels que

$$\sum_{i=1}^{N} A_{ji}u_i = f_j \qquad j = 1, \ldots, N,$$

ou, de façon équivalente, de résoudre le système linéaire

$$A\vec{u} = \vec{f}. \tag{10.13}$$

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 10.1 *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 :

$$\vec{y}^T A\vec{y} = \sum_{i,j=1}^{N} y_i A_{ij} y_j = \sum_{i,j=1}^{N} y_i y_j \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx$$

$$= \iint_\Omega \left| \sum_{i=1}^{N} y_i \overrightarrow{\text{grad}}\varphi_i(x) \right|^2 dx.$$

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 $\overrightarrow{\text{grad}}\psi(x) = 0$ ce qui implique $\psi(x) =$ constante. Puisque ψ est nul sur $\partial\Omega$, on aura $\psi \equiv 0$ et donc $\vec{y} = \vec{0}$. ■

De façon analogue à ce que nous avons déjà fait dans le cas unidimensionnel, voir chap. 9, nous proposons maintenant une construction simple d'un sous-espace V_h de V.

10.2 Méthode d'éléments 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 (voir figure 10.2).

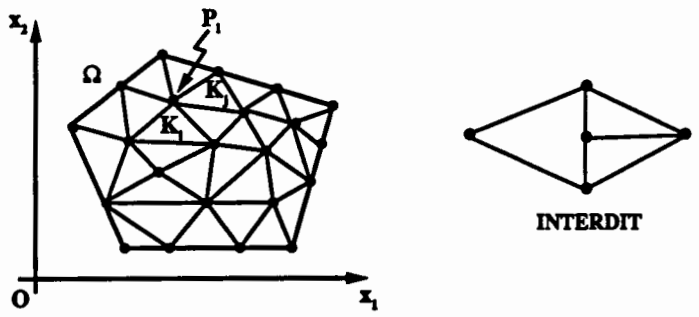


FIG. 10.2 - *Triangulation de* Ω.

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

$$\begin{aligned} V_h = \{ & g : \overline{\Omega} \to \mathbb{R} \,;\, g \text{ est continue sur } \overline{\Omega}, \text{ s'annule sur } \partial\Omega, \\ & \text{la restriction de } g \text{ à tout triangle } K \\ & \text{de la triangulation est un polynôme de degré} \leq 1 \}. \end{aligned} \tag{10.14}$$

Soit P_i, $i = 1, 2, \ldots, N$ les sommets intérieurs de la triangulation $\mathcal{T}_h$, encore appelés "noeuds". Pour décrire une fonction $g \in V_h$, nous pouvons choisir comme paramètres les valeurs $g(P_i)$ de la fonction g aux noeuds P_i, $i = 1, 2, \ldots, N$. Ces valeurs $g(P_i)$ sont appelées "degrés de liberté". Les fonctions de base φ_1, $\varphi_2, \ldots, \varphi_N$ de V_h sont alors définies par

$$\varphi_i(P_j) = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \qquad i, j = 1, \ldots, N, \tag{10.15}$$

et sont nulles sur $\partial\Omega$ (voir figure 10.3). Le support de φ_i (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

$$g(x) = \sum_{i=1}^{N} g_i \varphi_i(x) \qquad \text{où } g_i \in \mathbb{R} \quad i = 1, \ldots, N,$$

et nous obtenons bien $g(P_i) = g_i$, $i = 1, \ldots, N$, en vertu de (10.15).

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 f du système linéaire (10.13). Notons que le calcul numérique de f_j, $j = 1, 2, \ldots, N$, peut nécessiter l'usage d'une formule de quadrature numérique.

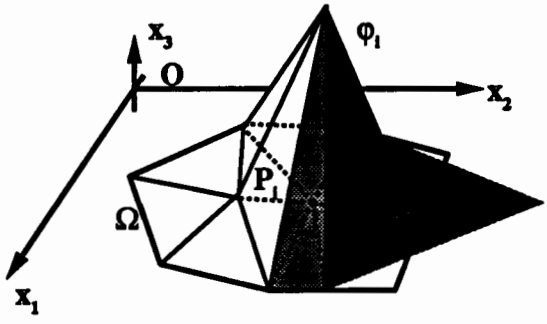


FIG. 10.3 - *La fonction de base φ_i.*

Après avoir constitué la matrice A, souvent appelée **matrice de rigidité**, ainsi que le vecteur $\vec{f}$, nous allons résoudre le système linéaire (10.13) en effectuant la décomposition de Cholesky de la matrice A, puis en résolvant deux systèmes linéaires triangulaires (voir chap. 5). La solution $\vec{u}$ ainsi obtenue nous permet d'exprimer la solution approchée u_h du problème (10.3) sous la forme $u_h(x) = u_1\varphi_1(x) + \cdots + u_N\varphi_N(x)$. En vertu de (10.15) nous aurons $u_h(P_j) = u_j$, $j = 1, \ldots, N$.

Remarque 10.1 Si, en lieu et place du problème particulier (10.3) (10.4), nous considérons le problème général (10.1) (10.2), 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 (10.1) (10.2) est elliptique et si $a_{12} = a_{21}$, alors A reste une matrice symétrique définie positive et tout ce qui précède s'applique encore !

10.3 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 Ω ayant pour noeuds les points Q_{ij}, voir figure 10.4. La triangulation $\mathcal{T}_h$ contient $N \equiv L^2$ noeuds intérieurs à Ω que nous numérotons comme dans la figure 10.4 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 10.5; il est constitué des 6 triangles $K_1, K_2, \ldots, K_6$. Puisque φ_1 est un polynôme de degré 1 sur chacun des triangles $K_1, K_2, \ldots, K_6$ et puisque $\varphi_1(P_1) = 1$ et $\varphi_1(Q_{ij}) = 0$ si $(i,j) \neq (1,1)$,

$$\varphi_i(P_j) = \delta_{ij}$$

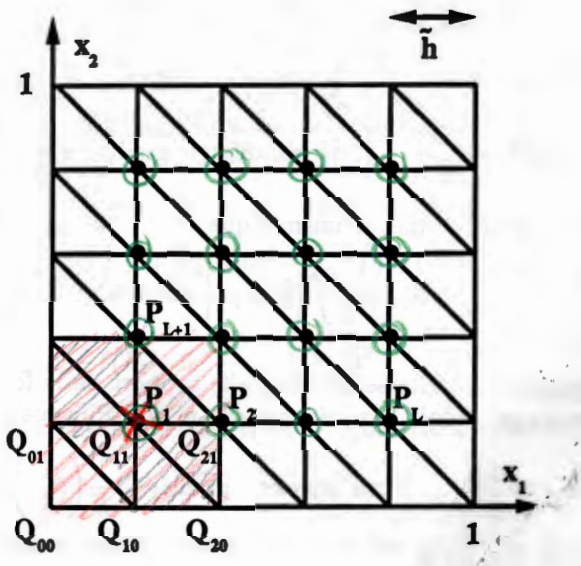
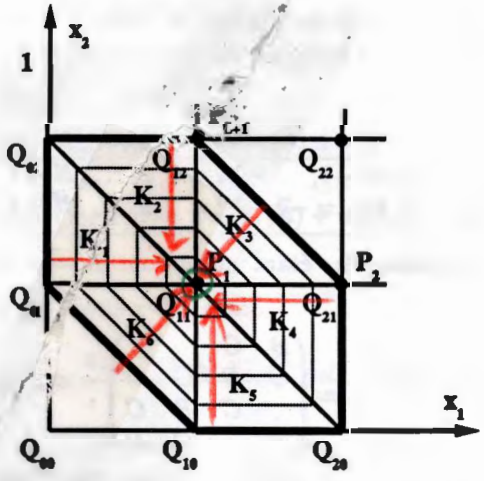


FIG. 10.4 - *Le carré unité et sa triangulation pour $L = 4$.*



le gradient est perpendiculaire aux courbes de niveau

FIG. 10.5 - *Support et isovaleurs de la fonction de base φ_1.*

nous vérifions facilement que

$$\begin{cases} \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{sur } K_1 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} -1 \\ 0 \end{pmatrix} \quad \text{sur } K_4, \\ \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} 0 \\ -1 \end{pmatrix} & \text{sur } K_2 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{sur } K_5, \\ \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} -1 \\ -1 \end{pmatrix} & \text{sur } K_3 \qquad \overrightarrow{\text{grad}}\varphi_1 = \frac{1}{h}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{sur } K_6. \end{cases}$$

Un calcul simple nous permet d'affirmer que

$$A_{11} = \iint_\Omega \left|\overrightarrow{\text{grad}}\varphi_1\right|^2 dx = 4.$$

D'après la figure 10.5, 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

$$\overrightarrow{\text{grad}}\varphi_2 = \frac{1}{h}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ sur } K_3 \quad \text{et} \quad \overrightarrow{\text{grad}}\varphi_2 = \frac{1}{h}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ sur } K_4,$$

un calcul simple conduit à

$$A_{12} = A_{21} = \iint_\Omega \overrightarrow{\text{grad}}\varphi_1 \cdot \overrightarrow{\text{grad}}\varphi_2 dx = -1.$$

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 10.4 nous constatons que pour des raisons de symétrie

$$\begin{aligned} &A_{ii} = A_{11} = 4, && \forall i = 1, 2, \ldots, N, \\ &A_{i,i+1} = A_{i+1,i} = A_{12} = -1, && \forall i = 1, 2, \ldots, N-1, \quad i \neq L \bmod L, \\ &A_{i,L+i} = A_{L+i,i} = A_{1,L+1} = -1, && \forall i = 1, 2, \ldots, N-L, \end{aligned}$$

tous les autres coefficients A_{ij} étant nuls. Pour $L = 4$ la matrice A a donc l'allure suivante

$$A = \begin{pmatrix} B & C & & \\ C & B & C & \\ & C & B & C \\ & & C & B \end{pmatrix}, \tag{10.16}$$

où nous avons noté

$$B = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{pmatrix} \text{ et } C = \begin{pmatrix} -1 & & & \\ & -1 & & \\ & & -1 & \\ & & & -1 \end{pmatrix}.$$

Approchons le second membre $\vec{f}$ de (10.13) en utilisant la formule de quadrature

$$\iint_K g(x)dx \simeq \frac{g(S_1)+g(S_2)+g(S_3)}{3} \text{ aire}(K) \qquad (10.17)$$

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 (cf. figure 10.5), nous avons

$$f_j \simeq f(P_j)\tilde{h}^2 \qquad \text{avec } j = 1, 2, \ldots, N. \qquad (10.18)$$

Résoudre le problème (10.13) revient, dans le cas particulier de ce paragraphe et en considérant le second membre intégré numériquement par la formule de quadrature (10.17), à résoudre le système linéaire

$$A\vec{u} = \vec{f}, \qquad (10.19)$$

la matrice A et le vecteur $\vec{f}$ étant donnés par (10.16) et (10.18) respectivement.

Remarque 10.2 Considérons les résultats du chap. 9, en particulier les égalités (9.2) et (9.3). Nous pouvons écrire, si $u : \overline{\Omega} \to \mathbb{R}$ est "assez régulière" et si P est un noeud de coordonnées $(i\tilde{h}, j\tilde{h})$:

$$\Delta u(P) = \frac{\partial^2 u}{\partial x_1^2}(P) + \frac{\partial^2 u}{\partial x_2^2}(P) =$$

$$= \frac{u((i+1)\tilde{h}, j\tilde{h}) - 2u(i\tilde{h}, j\tilde{h}) + u((i-1)\tilde{h}, j\tilde{h})}{\tilde{h}^2} + O(h^2)$$

$$+ \frac{u(i\tilde{h}, (j+1)\tilde{h}) - 2u(i\tilde{h}, j\tilde{h}) + u(i\tilde{h}, (j-1)\tilde{h})}{\tilde{h}^2} + O(h^2).$$

Notons $u_{i,j} = u(i\tilde{h}, j\tilde{h})$, nous avons donc :

$$-\Delta u(P) \simeq \frac{4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}}{\tilde{h}^2}. \qquad (10.20)$$

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 :

$$\frac{4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}}{\tilde{h}^2} = f(i\tilde{h}, j\tilde{h}), \qquad (10.21)$$

pour $i, j = 1, \ldots, N$ et

$$U_{0,j} = U_{L+1,j} = U_{j,0} = U_{j,L+1} = 0 \qquad \text{pour } j = 0, 1, \ldots, L+1. \tag{10.22}$$

En renumérotant les variables $U_{i,j}$, $i, j = 1, \ldots, N$, comme nous l'avons déjà fait pour les noeuds de la figure 10.4, 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 (10.21) (10.22) est strictement équivalent au système (10.19). Donc, dans ce cas particulier, la méthode des éléments finis avec intégration numérique du second membre est équivalente à la méthode des différences finies.

10.4 Estimations d'erreurs et méthodes d'ordre supérieur

Considérons à nouveau le problème de Poisson (10.3) (10.4) et soit $u : \overline{\Omega} \to \mathbb{R}$ sa solution. Notons encore u_h la solution de son approximation de Galerkin (10.10) construite avec des espaces V_h de type "éléments finis triangulaires de degré 1" définis par (10.14). 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 10.2 *Si u est "assez régulière" et si les angles des triangles qui constituent la triangulation ne deviennent pas infiniment petits lorsque h tend vers zéro, nous avons*

$$|u - u_h|_0 \leq Ch^2, \tag{10.23}$$

$$|u - u_h|_1 \leq Ch, \tag{10.24}$$

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 10.6, 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.

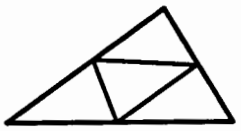


FIG. 10.6 - *Fragmentation d'un triangle en quatre, en utilisant le milieu des arêtes.*

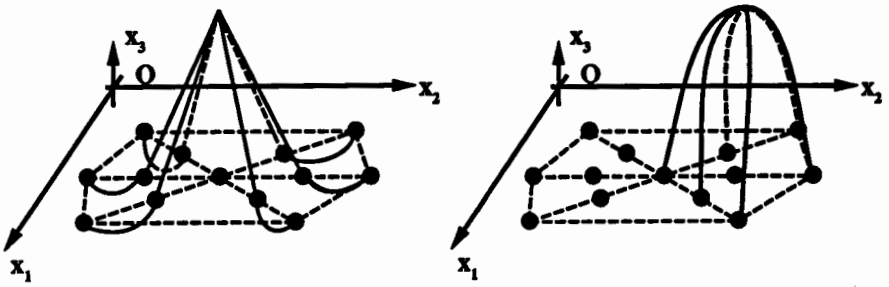


FIG. 10.7 - *Fonctions de base pour les polynômes de degré 2, sommet d'un triangle (fig. de gauche), milieu d'une arête (fig. de droite).*

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 (10.23) (10.24) deviennent $|u - u_h|_0 \leq Ch^{k+1}$ et $|u - u_h|_1 \leq Ch^k$. Le cas $k = 2$ est illustré dans la figure 10.7. A chaque sommet d'un triangle et à chaque milieu d'une arête est attachée une fonction de base polynomiale de degré 2 sur chaque triangle, nulle en tous les noeuds sauf un, où elle vaut 1. Quelques lignes du graphe de ces fonctions sont données dans la figure 10.7.

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 maillage, de mémorisation et de construction des matrices de rigidité;
- généralisation à des éléments quadrangulaires plutôt que triangulaires;
- généralisation à des problèmes posés dans un espace tridimensionnel;
- généralisation à des méthodes d'éléments finis non-standard ....

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 plupart de ces logiciels utilisent la technique des éléments finis.

Chapitre 11

Approximation des problèmes paraboliques. Problème de la chaleur

11.1 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 est connue 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 :

$$\dot{u} - \frac{\partial^2}{\partial x^2} u = f(x,t)$$

$$\rho c_p \frac{\partial u}{\partial t}(x,t) - k \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) \qquad \forall x \in]0, L[, \quad \forall t > 0. \quad (11.1)$$

A cette équation on adjoint des conditions aux limites :

$$u(0,t) = u(L,t) = 0 \qquad \forall t > 0, \quad (11.2)$$

et une condition initiale :

$$u(x,0) = w(x) \qquad \forall x \in]0, L[. \quad (11.3)$$

L'équation (11.1) 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 toutes les constantes ρ, c_p, k et L égales à 1, ce qui ne modifie pas fondamentalement le problème mathématique. (Suffit de multiplier par les cts.)

Ainsi, nous cherchons la fonction u satisfaisant les relations suivantes :

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) - \dfrac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) & \forall x \in]0,1[, \quad \forall t > 0, \qquad (11.4) \\ u(0,t) = u(1,t) = 0 & \forall t > 0, \qquad (11.5) \\ u(x,0) = w(x) & \forall x \in]0,1[. \qquad (11.6) \end{cases}$$

discrétisation spatiale

Pour résoudre numériquement le problème (11.4), (11.5), (11.6) 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 le paragraphe 9.1. 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 du paragraphe 9.1, il est naturel de considérer le schéma :

déjà vu au §

$$\frac{d}{dt}u_i(t) + \frac{1}{h^2}\Big(-u_{i-1}(t) + 2u_i(t) - u_{i+1}(t)\Big) = f(x_i,t) \qquad i = 1, \ldots, N, \quad \forall t > 0 \qquad (11.7)$$

$$u_0(t) = u_{N+1}(t) = 0 \qquad \forall t > 0, \qquad (11.8)$$

$$u_i(0) = w(x_i) \qquad i = 1, \ldots, N. \qquad (11.9)$$

Les fonctions $u_i(t)$, $i = 1, \ldots, N$, sont maintenant les inconnues du problème.

Nous dirons que le schéma (11.7), (11.8), (11.9) est une semi-discrétisation en espace du problème (11.4), (11.5), (11.6) par la méthode des différences finies. Si A est la $N \times N$ matrice tridiagonale définie par

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix},$$

si $\vec{u}(t)$ est le N-vecteur de composantes $u_1(t), u_2(t), \ldots, u_N(t)$, si $\vec{f}(t)$ est le N-vecteur de composantes $f(x_1,t), f(x_2,t), \ldots, f(x_N,t)$ et si $\vec{w}$ est le N-vecteur de composantes $w(x_1), w(x_2), \ldots, w(x_N)$, alors le schéma (11.7), (11.8), (11.9) est équivalent au système différentiel :

$$\begin{cases} \dot{\vec{u}}(t) = -A\vec{u}(t) + \vec{f}(t) & \forall t > 0, \qquad (11.10) \\ \vec{u}(0) = \vec{w}, \qquad (11.11) \end{cases}$$

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 $du_1(t)/dt, du_2(t)/dt, \ldots, du_N(t)/dt$.

La semi-discrétisation en espace du problème (11.4) (11.5) (11.6) conduit donc à la résolution d'un système différentiel du premier ordre avec une condition initiale. Nous pouvons donc utiliser les méthodes du chap. 8 pour intégrer numériquement ce système différentiel. Nous allons choisir les deux méthodes d'Euler progressive et rétrograde.

11.1.1 Méthode d'Euler progressive

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 :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^n + \vec{f}(t_n), \qquad n = 0, 1, 2, \ldots, \tag{11.12}$$

$$\vec{u}^0 = \vec{w}. \quad \text{↘ euler progressif} \tag{11.13}$$

Clairement, nous avons

$$\vec{u}^{n+1} = (I - \tau A)\vec{u}^n + \tau\vec{f}(t_n), \qquad n = 0, 1, 2, \ldots, \tag{11.14}$$

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 $\vec{u}^1, \vec{u}^2, \vec{u}^3, \ldots$ en utilisant (11.14); 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 (11.12) (11.13) est une discrétisation (complète) du problème (11.4) (11.5) (11.6) par la méthode des différences finies, en utilisant une méthode d'Euler explicite pour la discrétisation en temps.

Comme dans le paragraphe 8.3, une condition de stabilité limite le choix du pas temporel τ. Cette limitation est fonction du pas spatial et est exprimée par la condition :

$$\tau \leq \frac{h^2}{2}. \tag{11.15}$$

Comme nous l'avons déjà fait dans le paragraphe 8.2, nous pourrions montrer que, si la condition (11.15) est respectée et si la fonction f est identiquement nulle, alors $\vec{u}^n$ converge vers zéro lorsque n tend vers l'infini. Lorsque la condition (11.15) n'est pas respectée, plus n devient grand et plus les valeurs de u_i^n deviennent grandes, en changeant de signe.

11.1.2 Méthode d'Euler rétrograde

Si nous choisissons un schéma d'Euler rétrograde à la place du schéma d'Euler progressif pour discrétiser (11.10) (11.11), nous avons, à la place de (11.12) :

$$\frac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = -A\vec{u}^{n+1} + \vec{f}(t_{n+1}), \quad n = 0, 1, 2, \ldots, \tag{11.16}$$

↘ euler rétrograde

ou, de façon équivalente

$$(I + \tau A)\vec{u}^{n+1} = \vec{u}^n + \tau\vec{f}(t_{n+1}), \quad n = 0, 1, 2, \ldots. \tag{11.17}$$

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 présentée dans le paragraphe 5.5. Contrairement au

schéma explicite (11.12), ce schéma implicite est inconditionnellement stable au sens suivant.

Théorème 11.1 *Soit $\vec{u}^n$, $n = 0, 1, 2, \ldots$ la solution de (11.17) avec $f \equiv 0$. Alors, quel que soit le choix du pas de temps τ, nous avons*

$$\lim_{n\to\infty} \| \vec{u}^n \| = 0. \tag{11.18}$$

Démonstration
En posant $f \equiv 0$ dans (11.17) nous avons

$$\vec{u}^{n+1} = (I + \tau A)^{-1} \vec{u}^n. \tag{11.19}$$

Si $|||\cdot|||$ est la norme spectrale d'une matrice (cf. paragraphe 4.6), alors en utilisant (11.19) nous avons:

$$\| \vec{u}^{n+1} \| \leq |||(I + \tau A)^{-1}||| \cdot \| \vec{u}^n \| \tag{11.20}$$

et par suite, puisque $(I + \tau A)^{-1}$ est symétrique:

$$\| \vec{u}^{n+1} \| \leq \beta \| \vec{u}^n \| \tag{11.21}$$

où β est le maximum des valeurs propres de $(I + \tau A)^{-1}$ en valeur absolue. Dans le paragraphe 5.5, 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 $(I+\tau\lambda_A)^{-1}$, on conclut facilement qu'elles sont toutes comprises entre zéro et un et donc on a $0 < \beta < 1$. De (11.21) on tire

$$\| \vec{u}^n \| \leq \beta^n \| \vec{u}^0 \| \tag{11.22}$$

et par suite la relation (11.18) est démontrée. ■

Les méthodes d'Euler progressive et rétrograde sont toutes les deux d'ordre 1 en temps et d'ordre 2 en espace. Donc, si nous résolvons numériquement (11.4) jusqu'à un temps fixé $T > 0$ en utilisant soit le schéma (11.12) sous la condition (11.15), soit le schéma (11.16), l'erreur est d'ordre $O(\tau + h^2)$.

11.2 Equation de la chaleur 1D et éléments finis

Pour résoudre numériquement le problème (11.4) (11.5) (11.6) 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 paragraphes 9.2 et 9.3. Multiplions l'équation (11.4) 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:

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)v(x)dx - \int_0^1 \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = \int_0^1 f(x,t)v(x)dx. \tag{11.23}$$

En intégrant par partie le second terme de (11.23) et en supposant $v(0) = v(1) = 0$, nous déduisons (comme nous l'avons fait pour obtenir la relation (9.9)):

$$\int_0^1 \frac{\partial u}{\partial t}(x,t)v(x)dx + \int_0^1 \frac{\partial u}{\partial x}(x,t)v'(x)dx = \int_0^1 f(x,t)v(x)dx, \tag{11.24}$$

où naturellement $v'(x)$ désigne la dérivée de v (par rapport à la variable x bien sûr !). Comme dans le paragraphe 9.2, 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 (11.6) ainsi que (11.24) pour toute fonction $v \in V$. Dans la suite, ce problème est appelé "problème (11.24)".

Le problème (11.24) est une formulation faible en espace du problème (11.4)–(11.6). Comme nous l'avons fait dans le paragraphe 9.2, 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 (11.24) se formule de la manière suivante : pour tout $t > 0$, trouver une fonction $u_h(\cdot,t) \in V_h$ qui satisfait

$$\int_0^1 \frac{\partial u_h}{\partial t}(x,t)v_h(x)dx + \int_0^1 \frac{\partial u_h}{\partial x}(x,t)v_h'(x)dx = \int_0^1 f(x,t)v_h(x)dx, \tag{11.25}$$

pour toute fonction $v_h \in V_h$ et, de plus,

$$u_h(x,0) = w_h(x) \qquad \forall x \in [0,1], \tag{11.26}$$

où w_h est une approximation de la condition initiale u dans V_h ; la détermination de w_h sera discutée ultérieurement.

Comme nous l'avons fait dans le paragraphe 9.2 pour obtenir (9.11) à partir de (9.10), 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

$$u_h(x,t) = \sum_{i=1}^N u_i(t)\varphi_i(x) \qquad \forall x \in [0,1], \quad \forall t > 0. \tag{11.27}$$

En remplaçant (11.27) dans (11.25) et en choisissant comme fonctions test $v_h = \varphi_j$, $j = 1,2,\ldots,N$, nous obtenons donc :

$$\sum_{i=1}^N \dot{u}_i(t) \int_0^1 \varphi_i(x)\varphi_j(x)dx + \sum_{i=1}^N u_i(t) \int_0^1 \varphi_i'(x)\varphi_j'(x)dx = \int_0^1 f(x,t)\varphi_j(x)dx \qquad j = 1,\ldots,N. \tag{11.28}$$

Dans (11.28), 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

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx \tag{11.29}$$

(A est dite "matrice de rigidité"), si M est la $N \times N$ de coefficients

$$M_{ji} = \int_0^1 \varphi_i(x)\varphi_j(x)dx \tag{11.30}$$

(M est dite "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

$$f_j(t) = \int_0^1 f(x,t)\varphi_j(x)dx, \tag{11.31}$$

alors les relations (11.28) sont équivalentes à chercher $\vec{u}(t)$ tel que

$$M\dot{\vec{u}}(t) + A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0. \tag{11.32}$$

Comme pour la méthode des différences finies (voir (11.10)), l'approximation de Galerkin (11.32) 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 (11.32) est définie par :

$$\vec{u}(0) = \vec{w}. \tag{11.33}$$

Nous avons donc obtenu une semi-discrétisation spatiale du problème (11.4)–(11.6). Il est facile de vérifier que les matrices M et A sont des $N \times N$–matrices symétriques définies positives. Le système différentiel (11.32) est équivalent à

$$\dot{\vec{u}}(t) = -M^{-1}A\vec{u}(t) + M^{-1}\vec{f}(t) \qquad \forall t > 0, \tag{11.34}$$

et dès lors nous pouvons calculer une approximation de la solution $u(x,t)$ du problème (11.4)–(11.6) en procédant de la façon suivante :

i) on définit concrètement une base $\varphi_1, \varphi_2, \ldots, \varphi_N$ de type éléments finis comme nous l'avons fait dans le paragraphe 9.3 et on construit les matrices M, A et le vecteur $\vec{f}$;

ii) 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 (voir paragraphe 1.6);

iii) on détermine une approximation $\vec{u}^n$ de $\vec{u}(t_n)$ par une méthode d'Euler progressive ou rétrograde comme dans (11.12) ou (11.16).

Nous terminons ce paragraphe par deux remarques.

Remarque 11.1 La méthode d'Euler progressive 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

$$M\vec{u}^{n+1} = (M-\tau A)\vec{u}^n + \tau\vec{f}(t_n), \tag{11.35}$$

où τ est le pas de temps et $t_n = n\tau$.

Clairement, si $\vec{u}^n$ est connu dans (11.35), 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 la méthode d'Euler progressive n'est pas explicite ! Pour la rendre explicite, il faut calculer concrètement la matrice de masse M en utilisant la formule de quadrature des trapèzes. Ainsi, nous obtenons par (11.30) :

$$M_{ji} = \int_0^1 \varphi_i(x)\varphi_j(x)dx \simeq L_h(\varphi_i\varphi_j)$$

où $L_h(\varphi_i\varphi_j)$ est définie en (9.15). Il suffit de remarquer que $L_h(\varphi_i\varphi_j) = 0$ si $i \neq j$, pour voir que 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 la méthode d'Euler progressive.

Remarque 11.2 Comme nous l'avons mentionné dans le chap. 8, les schémas d'Euler sont d'ordre 1 en τ (voir inégalité (8.16)). Pour avoir un schéma d'ordre 2, nous pouvons utiliser une moyenne des schémas d'Euler progressif et rétrograde (voir relation (8.18)) que nous pouvons écrire comme suit :

$$M\frac{\vec{u}^{n+1}-\vec{u}^n}{\tau} + A\frac{\vec{u}^{n+1}+\vec{u}^n}{2} = \frac{\vec{f}(t_{n+1})+\vec{f}(t_n)}{2}$$

ou, de façon équivalente

$$(M+\frac{\tau}{2}A)\vec{u}^{n+1} = (M-\frac{\tau}{2}A)\vec{u}^n + \frac{\tau}{2}(\vec{f}(t_{n+1})+\vec{f}(t_n)). \tag{11.36}$$

Le schéma (11.36) est appelé "schéma de Crank-Nicholson" ; c'est un schéma numérique d'ordre 2, implicite, inconditionnellement stable (en norme quadratique !).

11.3 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 nombres réels donnés, si $f : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to f(x,t) \in \mathbb{R}$ est une fonction continue donnée et si $w : x \in \overline{\Omega} \to w(x) \in \mathbb{R}$ est une fonction donnée, 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 :

$$\frac{\partial u}{\partial t}(x,t) - \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}u(x,t)\right) = f(x,t)$$

$$\forall x = (x_1,x_2) \in \Omega, \quad \forall t > 0, \tag{11.37}$$

$$u(x,t) = 0, \qquad \forall x = (x_1,x_2) \in \partial\Omega, \quad \forall t > 0, \tag{11.38}$$

$$u(x,0) = w(x), \qquad \forall x = (x_1,x_2) \in \Omega. \tag{11.39}$$

La relation (11.38) est appelée "condition aux limites" alors que la relation (11.39) est appelée "condition initiale".

Définition 11.1 *Nous dirons que le problème (11.37)-(11.39) est parabolique si les coefficients a_{ij}, $1 \le i,j \le 4$, sont tels que l'équation $a_{11}x_1^2 + (a_{12} + a_{21})x_1x_2 + a_{22}x_2^2 = 1$ soit celle d'une ellipse (ou d'un cercle).*

Notons que si le problème est parabolique, alors l'équation

$$t - a_{11}x_1^2 - (a_{12} + a_{21})x_1x_2 - a_{22}x_2^2 = 0$$

est l'équation d'un paraboloïde dans l'espace Ox_1x_2t.

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 $a_{12} = a_{21} = 0$ nous obtenons

$$\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = f(x,t) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{11.40}$$

$$u(x,t) = 0 \qquad \forall x \in \partial\Omega, \quad \forall t > 0, \tag{11.41}$$

$$u(x,0) = w(x) \qquad \forall x \in \Omega, \tag{11.42}$$

où ici $\Delta u(x,t)$ est le Laplacien de u dans les variables spatiales. Le problème (11.40)-(11.42) 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 du paragraphe 10.2 avec ceux du paragraphe précédent, nous pouvons sans difficulté construire une semi-discrétisation spatiale du problème (11.40)-(11.42) 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 ce paragraphe, nous exposons très brièvement cette construction.

En suivant ce qui a été fait au paragraphe 10.2, nous écrivons une relation semblable à (11.27) :

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{11.43}$$

où les fonctions de bases φ_j, $j = 1, \ldots, N$, sont données par (10.15). Il suffit ensuite de remplacer dans (11.28) les intégrales

$$\text{“}\int_0^1 dx\text{”} \qquad \text{par} \qquad \text{“}\iint_\Omega dx\text{”}$$

et les grandeurs “φ_i'” par des grandeurs “$\overrightarrow{\text{grad}}\varphi_i$” pour obtenir :

$$\sum_{i=1}^N \dot{u}_i(t) \iint_\Omega \varphi_i(x)\varphi_j(x)dx + \sum_{i=1}^N u_i(t) \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx$$
$$= \iint_\Omega f(x,t)\varphi_j(x)dx, \qquad j = 1,2,\ldots,N. \quad (11.44)$$

Pour justifier (11.44), il suffit d'opérer exactement de la même manière que pour passer de (10.5) à (10.9) et de (11.23) à (11.24).

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 (11.44) sera alors

$$M\dot{\vec{u}}(t) + A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0, \qquad (11.45)$$

qui pourra être résolu numériquement par une méthode d'Euler, ou de Crank–Nicholson (ou une autre) après avoir naturellement posé

$$f_j(t) = \iint_\Omega f(x,t)\varphi_j(x)dx$$

et tenu compte d'une approximation w_h de la condition initiale (11.42).

11.4 Un exemple particulier

Comme dans le paragraphe 10.3, nous nous plaçons dans le cas où Ω est le carré unité de frontière $\partial\Omega$ et nous considérons le problème (11.40)–(11.42) avec ce domaine Ω particulier. Nous choisissons la même triangulation $\mathcal{T}_h$ que celle décrite dans la figure 10.4 (N noeuds intérieurs, $2L^2$ 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 le paragraphe 10.3. Nous obtenons alors la matrice de rigidité A décrite dans (10.16). Il ne reste plus qu'à calculer la matrice de masse M et le second membre $\vec{f}(t)$ pour obtenir concrètement le système (11.45). Le calcul de $\vec{f}(t)$ peut se faire par le biais d'une formule de quadrature, comme nous l'avons fait

pour obtenir (10.18). Nous aurons $f_j(t) \simeq f(P_j,t)\tilde{h}^2$ où P_j désigne aussi bien le noeud 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: très pédagogique BRAVO !

$$M = \frac{\tilde{h}^2}{12}\begin{pmatrix} B & C & & \\ C^T & B & C & \\ & C^T & B & C \\ & & C^T & B \end{pmatrix},$$

où nous avons noté

$$B = \begin{pmatrix} 6 & 1 & & \\ 1 & 6 & 1 & \\ & 1 & 6 & 1 \\ & & 1 & 6 \end{pmatrix} \text{ et } C = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \end{pmatrix}.$$

Dans cet exemple concret, nous pouvons choisir pour condition initiale w_h l'interpolant de w, c'est-à-dire $w_h(x) = \sum_{i=1}^{N} w(P_i)\varphi_i(x)$, ce qui imposera donc $u_i(0) = w(P_i)$, $i = 1, \ldots, N$.

Chapitre 12

Approximation de problèmes hyperboliques. Equation de transport et équation des ondes

12.1 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 :

$$\frac{\partial u}{\partial t}(x,t) + c(x,t)\frac{\partial u}{\partial x}(x,t) = f(x,t) \quad \forall x \in \mathbb{R}, \quad \forall t > 0, \tag{12.1}$$

avec la condition initiale

$$u(x,0) = w(x) \qquad \forall x \in \mathbb{R}. \tag{12.2}$$

Les équations (12.1) (12.2) 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 (12.1) (12.2) est facile à résoudre lorsque $c = c_0$ = constante et $f = 0$. En effet, la fonction u définie par

$$u(x,t) = w(x - c_0 t) \tag{12.3}$$

satisfait

$$\frac{\partial u}{\partial t}(x,t) + c_0\frac{\partial u}{\partial x}(x,t) = 0 \qquad \forall x \in \mathbb{R}, \quad \forall t > 0, \tag{12.4}$$

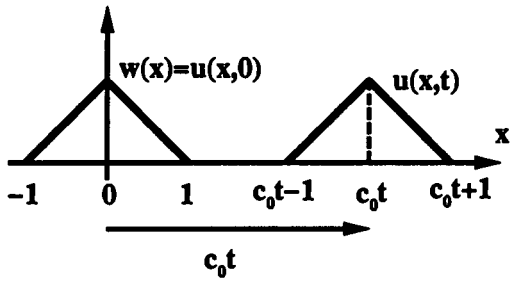


FIG. 12.1 - *Transport de $w(x)$ au temps $t > 0$, avec $c_0 > 0$.*

et

$$u(x,0) = w(x) \qquad \forall x \in \mathbb{R}. \tag{12.5}$$

Nous avons représenté figure 12.1 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.

Revenons au problème (12.1) (12.2) 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 12.2, 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 (12.1) (12.2) au point x_j et au temps t_n, il semble naturel de choisir le schéma numérique suivant :

$$\frac{u_j^{n+1} - u_j^n}{\tau} + c(x_j, t_n)\frac{u_{j+1}^n - u_{j-1}^n}{2h} = f(x_j, t_n), \tag{12.6}$$

pour $j = 0, \pm 1, \pm 2, \ldots$ et $n = 0, 1, 2, \ldots$. L'approximation initiale est définie par

$$u_j^0 = w(x_j) \qquad \text{pour } j = 0, \pm 1, \pm 2, \ldots. \tag{12.7}$$

Le schéma (12.6) (12.7) 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 (12.6) sous la forme suivante :

$$u_j^{n+1} = u_j^n + \tau\left(f(x_j, t_n) - c(x_j, t_n)\frac{u_{j+1}^n - u_{j-1}^n}{2h}\right). \tag{12.8}$$

Revenons au cas où $c = c_0 =$ constante et $f = 0$. Nous aurons alors

$$u_j^{n+1} = u_j^n - \frac{c_0\tau}{2h}(u_{j+1}^n - u_{j-1}^n). \tag{12.9}$$

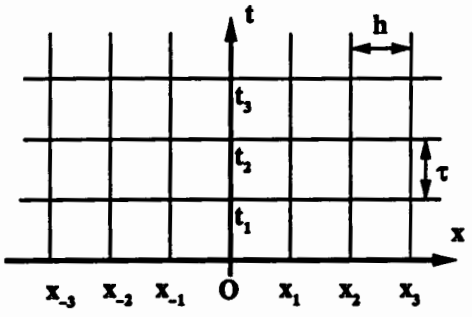


FIG. 12.2 - *Grille de différences finies.*

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

$$w(x) = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imx}, \tag{12.10}$$

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

$$u_j^0 = w(jh) = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh}, \quad j = 0, \pm 1, \pm 2, \ldots. \tag{12.11}$$

En utilisant (12.9) et (12.11) nous obtenons :

$$\begin{aligned} u_j^1 &= u_j^0 - \frac{c_0\tau}{2h}(u_{j+1}^0 - u_{j-1}^0) \\ &= \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh} \left(1 - \frac{c_0\tau}{2h}(e^{imh} - e^{-imh})\right) \\ &= \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh} \left(1 - \frac{c_0\tau}{h} i \sin mh\right). \end{aligned} \tag{12.12}$$

Nous vérifions facilement, en itérant n fois le passage de (12.11) à (12.12), qu'à l'étape n, $n = 0, 1, 2, \ldots$, nous obtenons :

$$u_j^n = \sum_{m=-\infty}^{+\infty} \alpha_m e^{imjh} \left(1 - \frac{c_0\tau}{h} i \sin mh\right)^n, \tag{12.13}$$

pour $j = 0, \pm 1, \pm 2, \ldots$. Le coefficient $1 - c_0\tau/h\, i \sin mh$ est un nombre complexe appelé "coefficient d'amplification de la m-ième harmonique". Son module

$$\sqrt{1 + \left(\frac{c_0\tau}{h} \sin mh\right)^2},$$

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 (12.4) (12.5) est donné par (12.3) 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?

On prend les conditions juste avant pour établir le transport.

Considérons à nouveau la fonction u définie par (12.3) et solution du problème (12.4) (12.5). Nous constatons que la condition initiale $w(x)$ est transportée à la vitesse c_0 dans le sens des $x > 0$ lorsque $c_0 > 0$ et dans le sens des $x < 0$ lorsque $c_0 < 0$. Il semble dès lors naturel que, si $c_0 > 0$, 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 $c_0 < 0$, 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 :

$$\frac{u_j^{n+1} - u_j^n}{\tau} + (c_j^n)^+ \frac{u_j^n - u_{j-1}^n}{h} + (c_j^n)^- \frac{u_{j+1}^n - u_j^n}{h} = f(x_j, t_n), \tag{12.14}$$

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 (12.14) 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 (12.14) 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 a un module plus petit ou égal à 1 (indépendemment 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 (12.14), la condition de stabilité devient

$$\frac{\tau}{h} \leq \frac{1}{\max\limits_{x\in\mathbb{R},t>0}|c(x,t)|}. \tag{12.15}$$

En pratique le pas spatial h et le pas temporel τ devront être choisis de sorte à ce que la condition (12.15) soit satisfaite. Nous dirons que **le schéma explicite décentré est conditionnellement stable**. La condition de stabilité (12.15) 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 $h/\max|c(x,t)|$. Dans le cas contraire le schéma (12.14) produit 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 (12.1) (12.2) (Lax, Lax–Wendroff, saute–mouton, ...). Il existe aussi bien d'autres méthodes (éléments finis discontinus, méthode des caractéristiques, méthodes particulaires, ...) que nous n'abordons pas ici.

12.2 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 donnée, continue, 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 données. 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

$$\frac{\partial^2 u}{\partial t^2}(x,t) - c^2\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t) \quad \forall x \in]0,1[, \quad \forall t > 0, \tag{12.16}$$

$$u(0,t) = u(1,t) = 0 \quad \forall t > 0, \tag{12.17}$$

$$u(x,0) = w(x) \text{ et } \frac{\partial u}{\partial t}(x,0) = v(x) \quad \forall x \in]0,1[. \tag{12.18}$$

Le problème (12.16)–(12.18) est appelé "problème hyperbolique" d'ordre deux; l'équation (12.16) 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 (12.17) ainsi que les deux conditions initiales (12.18). 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 (12.16), se réduit à $t^2 - x^2 = 1$ 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 (12.16)–(12.18). 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 (12.16). Le nombre c dépend de la masse spécifique de la corde et de sa tension. Les conditions aux limites (12.17) 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 (12.18).

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

$$u(x,t) = \frac{1}{2}\Big(\omega(x-ct) + \omega(x+ct)\Big) \quad \forall x \in [0,1], \quad \forall t \geq 0, \tag{12.19}$$

est solution du problème (12.16)-(12.18). 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 (12.16) est appelée "équation des ondes".

Nous allons maintenant proposer une méthode de différences finies couramment utilisée pour résoudre numériquement (12.16)-(12.18). 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 (voir paragraphe 11.1), nous commençons par établir une semi-discrétisation en espace du problème (12.16)-(12.18) par différences finies, à savoir

$$\frac{d^2}{dt^2}u_j(t) + c^2\frac{-u_{j-1}(t) + 2u_j(t) - u_{j+1}(t)}{h^2} = f(x_j,t) \qquad j = 1,\ldots,N,\ \forall t > 0, \tag{12.20}$$

$$u_0(t) = u_{N+1}(t) = 0 \qquad \forall t > 0, \tag{12.21}$$

$$u_j(0) = w(x_j) \text{ et } \frac{d}{dt}u_j(0) = v(x_j) \qquad j = 1,\ldots,N. \tag{12.22}$$

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 (voir le chap. 11), nous introduisons la $N \times N$ matrice A définie par

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \tag{12.23}$$

le N-vecteur $\vec{u}(t)$ de composantes $u_1(t), \ldots, u_N(t)$, le N-vecteur $\vec{f}(t)$ de composantes $f(x_1,t), \ldots, f(x_N,t)$, le N-vecteur $\vec{w}$ de composantes $w(x_1), \ldots, w(x_N)$ et le N-vecteur $\vec{v}$ de composantes $v(x_1), \ldots, v(x_N)$. Le système (12.20)-(12.22) peut alors s'écrire sous la forme condensée

$$\ddot{\vec{u}}(t) + c^2 A\vec{u}(t) = \vec{f}(t) \qquad \forall t > 0, \tag{12.24}$$

$$\vec{u}(0) = \vec{w}, \qquad \dot{\vec{u}}(0) = \vec{v}, \tag{12.25}$$

où nous avons noté $\dot{\vec{u}}(t)$ le N–vecteur de composantes $du_1(t)/dt, \ldots, du_N(t)/dt$ et $\ddot{\vec{u}}(t)$ le N–vecteur de composantes $d^2u_1(t)/dt^2, \ldots, d^2u_N(t)/dt^2$. Le système différentiel (12.24) (12.25) est d'ordre deux et nous pouvons utiliser la méthode de Newmark décrite dans le paragraphe 8.7 pour le résoudre numériquement. 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 (12.24) (12.25) est la suivante :

$$\frac{\vec{u}^{n+1} - 2\vec{u}^n + \vec{u}^{n-1}}{\tau^2} + c^2 A\vec{u}^n = \vec{f}(t_n), \quad n = 1, 2, \ldots, \tag{12.26}$$

$$\vec{u}^0 = \vec{w}, \qquad \vec{u}^1 = \vec{w} + \tau\vec{v} + \frac{1}{2}\tau^2 c^2\Big(\vec{f}(0) - A\vec{w}\Big). \tag{12.27}$$

Le schéma (12.26) (12.27) est un schéma explicite. Connaissant $\vec{u}^0$ et $\vec{u}^1$ nous pouvons calculer pour $n = 1, 2, \ldots$:

$$\vec{u}^{n+1} = (2I - \tau^2 c^2 A)\vec{u}^n - \vec{u}^{n-1} + \tau^2\vec{f}(t_n), \tag{12.28}$$

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 (12.28) s'écrit composante par composante :

$$u_j^{n+1} = 2(1-\lambda)u_j^n + \lambda\Big(u_{j-1}^n + u_{j+1}^n\Big) - u_j^{n-1} + \tau^2 f(x_j, t_n), \tag{12.29}$$

pour $j = 1, \ldots, N$.

Posons à nouveau $f = 0$ et $v = 0$ et comparons la solution u du problème (12.16)-(12.18), définie par (12.19), à son approximation numérique, définie par (12.29). Supposons, pour simplifier, que

$$w(x) = \sin m\pi x \tag{12.30}$$

où m est un entier positif (si ce n'est pas le cas nous pouvons développer $w(x)$ en série de Fourier, les calculs sont similaires). En utilisant la formule trigonométrique

$$\sin(\alpha + \beta) + \sin(\alpha - \beta) = 2\sin\alpha\cos\beta, \quad \alpha, \beta \in \mathbb{R}, \tag{12.31}$$

et (12.19) 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

$$u(x_j, t_n) = \sin(m\pi jh)\cos(m\pi cn\tau). \tag{12.32}$$

D'autre part le schéma numérique (12.26) (12.27) s'écrit avec $f = 0$ et $v = 0$:

$$\begin{aligned}
u_j^0 &= w(x_j), \\
u_j^1 &= (1-\lambda)w(x_j) + \frac{1}{2}\lambda(w(x_{j-1}) + w(x_{j+1})), \\
u_j^2 &= 2(1-\lambda)u_j^1 + \lambda(u_{j-1}^1 + u_{j+1}^1) - u_j^0, \\
&\vdots \\
u_j^{n+1} &= 2(1-\lambda)u_j^n + \lambda(u_{j-1}^n + u_{j+1}^n) - u_j^{n-1},
\end{aligned} \tag{12.33}$$

pour $j = 1, \ldots, N$. En utilisant la condition initiale (12.30) et la formule trigonométrique (12.31), nous obtenons

$$u_j^n = \alpha_n \sin(m\pi jh), \tag{12.34}$$

les coefficients α_n étant donnés par les formules de récurrence :

$$\begin{aligned} &\alpha_0 = 1, \\ &\alpha_1 = 1 - \lambda(1 - \cos(m\pi h)), \\ &\alpha_2 = 2\alpha_1\alpha_1 - \alpha_0, \\ &\quad\vdots \\ &\alpha_n = 2\alpha_1\alpha_{n-1} - \alpha_{n-2}. \end{aligned} \tag{12.35}$$

Par conséquent, compte tenu de (12.32) et (12.35), 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 12.1 *Le schéma (12.26) (12.27) est stable s'il existe une constante C telle que les valeurs $(\alpha_n)_{n=0}^{\infty}$ définies par (12.35) satisfassent*

$$|\alpha_n| \leq C, \qquad n = 0, 1, 2, \ldots, \qquad m = 1, 2, \ldots.$$

Nous sommes maintenant en mesure de montrer le résultat suivant :

Théorème 12.1 *Le schéma (12.26) (12.27) est stable si la condition CFL suivante est satisfaite :*

$$\tau \leq \frac{h}{|c|}.$$

Démonstration
Soit α_1 le coefficient défini en (12.35) 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

$$s_+ = \alpha_1 + \sqrt{\alpha_1^2 - 1} \quad \text{et} \quad s_- = \alpha_1 - \sqrt{\alpha_1^2 - 1}, \tag{12.36}$$

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 (12.35) est tel que

$$\alpha_n = \frac{1}{2}(s_+^n + s_-^n), \qquad n = 0, 1, 2, \ldots. \tag{12.37}$$

En effet, nous constatons immédiatement que l'égalité (12.37) est vraie pour $n = 0$ et $n = 1$. Supposons que (12.37) soit vraie pour $n \leq k$ et montrons qu'elle

reste vraie pour $n = k + 1$. L'hypothèse de récurrence et (12.35), nous assure 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 (12.36) 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 (12.37) pour $n = k + 1$.

Revenons à la question de la stabilité du schéma (12.26) (12.27). 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 (12.37), d'assurer que

$$|s_\pm| \leq 1, \qquad m = 1, 2, \ldots . \tag{12.38}$$

Nous constatons que si $|\alpha_1| \leq 1$ alors le critère (12.38) est satisfait. Ainsi $|\alpha_n| \leq C$ si

$$-1 \leq \alpha_1 \leq 1 \qquad m = 1, 2, \ldots ,$$

soit, en utilisant (12.35):

$$-1 \leq 1 - \lambda(1 - \cos m\pi h) \leq 1, \quad m = 1, 2, \ldots . \tag{12.39}$$

L'inégalité de droite dans (12.39) est toujours satisfaite. L'inégalité de gauche est satisfaite pour autant que

$$\lambda \leq \frac{2}{1 - \cos mh}, \qquad m = 1, 2, \ldots , \tag{12.40}$$

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 12.1 Il est possible de montrer que si la condition CFL est satisfaite, le schéma numérique (12.26) (12.27) 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 (12.26) (12.27) 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 $(N + 1)|c|T \leq M$ de telle sorte à ce 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

$$\max_{j=1,\ldots,N} |u_j^M - u(x_j, T)| \leq Ch^2, \qquad \text{si } h \to 0, \tag{12.41}$$

la constante C étant indépendante de M et N.

12.3 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 les trois fonctions continues

$$f : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to f(x,t) \in \mathbb{R},$$
$$w : x \in \overline{\Omega} \to w(x) \in \mathbb{R},$$
$$v : x \in \overline{\Omega} \to v(x) \in \mathbb{R},$$

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

$$\frac{\partial^2 u}{\partial t^2}(x,t) - c^2 \Delta u(x,t) = f(x,t) \qquad \forall x \in \Omega, \quad \forall t > 0, \tag{12.42}$$
$$u(x,0) = 0 \qquad \forall x \in \partial\Omega, \quad \forall t > 0, \tag{12.43}$$
$$u(x,t) = w(x) \text{ et } \frac{\partial u}{\partial t}(x,t) = v(x) \qquad \forall x \in \Omega. \tag{12.44}$$

Les équations (12.42)–(12.44) 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 (12.42)–(12.44), c étant un nombre dépendant de la masse spécifique et de la tension de la membrane. Les égalités (12.44) 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 lachons 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 (12.42)–(12.44) par la méthode des éléments finis, nous procédons comme dans le paragraphe 10.1. Pour ce faire, nous multiplions (12.42) par une fonction test $\varphi : x \in \overline{\Omega} \to \varphi(x) \in \mathbb{R}$ de classe C^1, s'annulant sur le bord $\partial\Omega$, et nous intégrons par partie comme en (10.5). Nous obtenons :

$$\iint_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\varphi(x)dx + c^2 \iint_\Omega \overrightarrow{\text{grad}}u(x,t) \cdot \overrightarrow{\text{grad}}\varphi(x)dx$$
$$= \iint_\Omega f(x,t)\varphi(x)dx \qquad \forall t > 0. \tag{12.45}$$

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 paragraphes 10.1 et 11.3. 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 (12.45) et choisissons pour fonctions test $\varphi = \varphi_j$, $j = 1, \ldots, N$ nous obtenons :

$$\sum_{i=1}^{N} \ddot{u}_i(t) \iint_{\Omega} \varphi_i(x)\varphi_j(x)dx + c^2 \sum_{i=1}^{N} u_i(t) \iint_{\Omega} \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx$$

$$= \iint_{\Omega} f(x,t)\varphi_j(x)dx, \qquad j = 1, \ldots, N, \quad \forall t > 0. \quad (12.46)$$

Utilisons à nouveau les notations du paragraphe 11.3. 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 (12.46) sous forme d'un système différentiel, à savoir :

$$M\ddot{\vec{u}}(t) + c^2 A\vec{u}(t) = \vec{f}(t), \qquad \forall t > 0. \quad (12.47)$$

Puisque ce système différentiel est du deuxième ordre, nous devons ajouter les conditions initiales

$$\vec{u}(0) = \vec{w} \text{ et } \dot{\vec{u}}(0) = \vec{v}. \quad (12.48)$$

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 noeuds de la triangulation $\mathcal{T}_h$). Nous utiliserons à nouveau la méthode de Newmark décrite dans le paragraphe 8.7 pour résoudre numériquement(12.47) (12.48). Si nous voulons, comme nous l'avons fait dans le paragraphe précédent, que cette méthode soit explicite, nous devrons utiliser une méthode d'intégration numérique de sorte à ce que la matrice de masse M soit diagonale ("mass lumping", voir aussi la remarque 11.1).

12.4 Equation de transport 1D non linéaire

Supposons avoir un continum 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

$$\dot{g}_{\overline{x}}(t) = u(g_{\overline{x}}(t), t), \qquad t > 0, \tag{12.49}$$

$$g_{\overline{x}}(0) = \overline{x}. \tag{12.50}$$

En dérivant (12.49) par rapport au temps, nous obtenons

$$\ddot{g}_{\overline{x}}(t) = \frac{\partial u}{\partial x}(g_{\overline{x}}(t), t)\dot{g}_{\overline{x}}(t) + \frac{\partial u}{\partial t}(g_{\overline{x}}(t), t). \tag{12.51}$$

Puisque les particules n'intéragissent pas entre elles, l'accélération $\ddot{g}_{\overline{x}}(t)$ est nulle. En utilisant (12.49), l'équation (12.51) s'écrit :

$$\frac{\partial u}{\partial x}(g_{\overline{x}}(t), t) + u(g_{\overline{x}}(t), t)\frac{\partial u}{\partial x}(g_{\overline{x}}(t), t) = 0, \qquad t > 0, \tag{12.52}$$

$$u(g_{\overline{x}}(0), 0) = u(\overline{x}, 0). \tag{12.53}$$

Les équations (12.52) (12.53) justifient l'étude du problème de transport non-linéaire suivant (dit "problème de Burger") :

Etant donné 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

$$\frac{\partial u}{\partial t}(x,t) + u(x,t)\frac{\partial u}{\partial x}(x,t) = 0 \qquad \forall x \in \mathbb{R} \qquad \forall t > 0, \tag{12.54}$$

$$u(x,0) = w(x) \qquad \forall x \in \mathbb{R}. \tag{12.55}$$

Le problème (12.54) (12.55) 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 (voir chapitre 8) suivant : trouver $\beta : t \in \mathbb{R}^+ \to \beta(t) \in \mathbb{R}$ tel que

$$\dot{\beta}(t) = u(\beta(t), t), \qquad t > 0, \tag{12.56}$$

$$\beta(0) = \overline{x} \tag{12.57}$$

où $\overline{x} \in \mathbb{R}$ est un nombre donné. Si nous supposons que u vérifie la condition (8.3) du théorème de Cauchy–Lipschitz du paragraphe 8.1, le problème (12.56) (12.57) a une et une seule solution β. Posons maintenant

$$\gamma(t) = u(\beta(t), t),$$

nous avons

$$\dot{\gamma}(t) = \frac{\partial u}{\partial x}(\beta(t), t)u(\beta(t), t) + \frac{\partial u}{\partial t}(\beta(t), t).$$

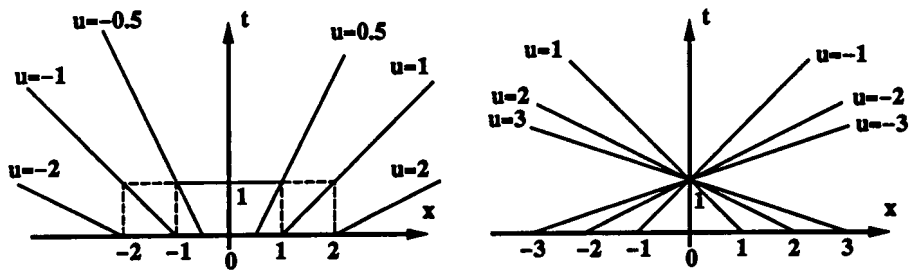


FIG. 12.3 - *Courbes caractéristiques lorsque $w(x) = x$, i.e. $x = \overline{x}t + \overline{x}$ (figure de gauche) et lorsque $w(x) = -x$, i.e. $x = -\overline{x}t + \overline{x}$ (figure de groite).*

Puisque u satisfait (12.54), nous obtenons

$$\dot{\gamma}(t) = 0,$$

et par conséquent, compte tenu de (12.55) et (12.57),

$$\begin{aligned} \gamma(t) &= \text{constante} = u(\beta(t), t) = u(\beta(0), 0) \\ &= u(\overline{x}, 0) = w(\overline{x}) \qquad \forall t > 0. \end{aligned} \tag{12.58}$$

En utilisant (12.56) et (12.58) nous avons

$$\dot{\beta}(t) = \gamma(t) = w(\overline{x}) \qquad \forall t > 0,$$

soit, en intégrant de 0 à t et en utilisant (12.57)

$$\beta(t) = w(\overline{x})t + \overline{x} \qquad \forall t > 0. \tag{12.59}$$

Nous avons donc montré que si $u(x,t)$ est une fonction suffisamment régulière qui satisfait (12.54) (12.55), alors

$$u(w(\overline{x})t + \overline{x}, t) = w(\overline{x}) \qquad \forall \overline{x} \in \mathbb{R}, \quad \forall t > 0. \tag{12.60}$$

La relation (12.60) traduit le fait que la solution u reste constante sur la droite d'équation $x = w(\overline{x})t + \overline{x}$ appelée "courbe caractéristique". Dans la figure 12.3, 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.

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 ce paragraphe, 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 (12.60) 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 inhérente aux problèmes hyperboliques non–linéaires tels que les équations (12.54) (12.55) 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 cette leçon !

Chapitre 13

Approximation de problèmes de convection–diffusion

13.1 Un problème de convection–diffusion stationnaire et différences finies

Soit $f : x \in [0,1] \to f(x)$ et $c : x \in [0,1] \to c(x) \in \mathbb{R}$ deux fonctions données, continues, et soit $\varepsilon > 0$ fixé. Nous cherchons une fonction $u : x \in [0,1] \to u(x) \in \mathbb{R}$ satisfaisant

$$\begin{aligned} &-\varepsilon u''(x) + c(x)u'(x) = f(x) \qquad \forall x \in]0,1[, \\ &u(0) = u(1) = 0. \end{aligned} \tag{13.1}$$

Le problème (13.1) est appelé un 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 (13.1) est le problème de diffusion que nous avons traité dans le chap. 9 (attention : remarquez que le problème (9.1) et le problème (13.1) sont différents car dans le premier apparaît un terme de type $c(x)u(x)$ alors que dans le deuxième nous avons $c(x)u'(x)$). Si nous choisissons $\varepsilon = 0$ dans (13.1), nous obtenons alors le problème de transport stationnaire (c'est-à-dire le problème (12.1) avec $\partial u/\partial t = 0$). L'exemple typique d'une situation physique régie par les équations (13.1) est le problème de la propagation de chaleur dans un fluide (unidimensionnel) soumis à des mouvements de convection. 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 (13.1) est une équation différentielle linéaire à coefficients constants que nous pouvons résoudre explicitement. En tenant compte des conditions aux limites nous obtenons :

$$u(x) = \frac{f_0}{c_0}\left(x - \frac{1 - \exp\left(\frac{c_0}{\varepsilon}x\right)}{1 - \exp\left(\frac{c_0}{\varepsilon}\right)}\right) \qquad \forall x \in [0,1]. \tag{13.2}$$

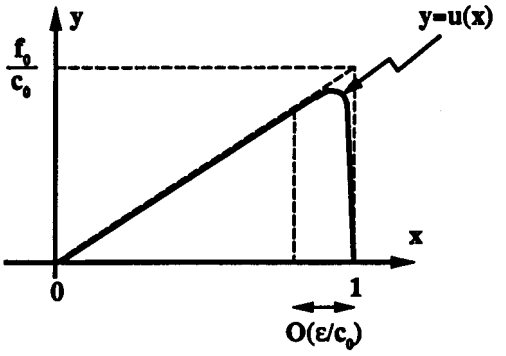


FIG. 13.1 - *Solution $u(x)$ lorsque $\varepsilon << c_0$.*

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 où la solution u présente une couche limite (voir figure 13.1).

Considérons maintenant une approximation par différences finies des équations (13.1) 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:

$$\begin{aligned} \varepsilon \frac{2u_j - u_{j-1} - u_{j+1}}{h^2} + c(x_j)\frac{u_{j+1} - u_{j-1}}{2h} \\ = f(x_j), \qquad j = 1, \ldots, N, \\ u_0 = u_{N+1} = 0. \end{aligned} \tag{13.3}$$

Clairement (13.3) est un système linéaire de N équations et N inconnues $u_1, \ldots, u_N$. Dans le cas simple où $c(x) = c_0 =$ constante $\neq 0$ et $f(x) = f_0 =$ constante et si $h = 2\varepsilon/c_0$, 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 13.2 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 figure 13.3. 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 (13.3) 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 entre la différence rétrograde et la différence progressive (voir chap. 2). Le schéma numérique pour discrétiser (13.1) devient:

$$\begin{aligned} \varepsilon \frac{2u_j - u_{j+1} - u_{j-1}}{h^2} + \frac{c(x_j)}{h}\Big(\alpha_j(u_j - u_{j-1}) + (1-\alpha_j)(u_{j+1} - u_j)\Big) \\ = f(x_j), \qquad j = 1, \ldots, N, \\ u_0 = u_{N+1} = 0. \end{aligned} \tag{13.4}$$

Il reste donc à choisir les nombres $\alpha_j \in [0, 1]$ pour obtenir la "meilleure approximation possible". Remarquons que si nous choisissons $\alpha_j = 1/2$, $j = 1, \ldots, N$,

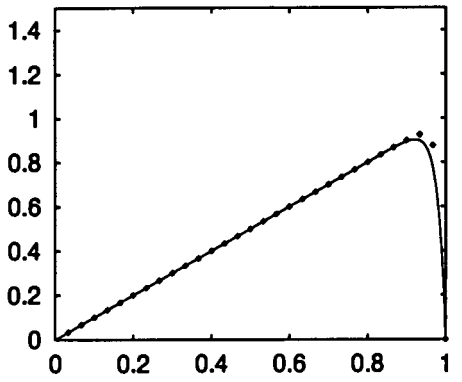


FIG. 13.2 - *Les valeurs u_j pour $h < 2\varepsilon/c_0$ (ici $f_0 = 1$, $\varepsilon = 0.02$, $c_0 = 1$, $h = 1/30$).*

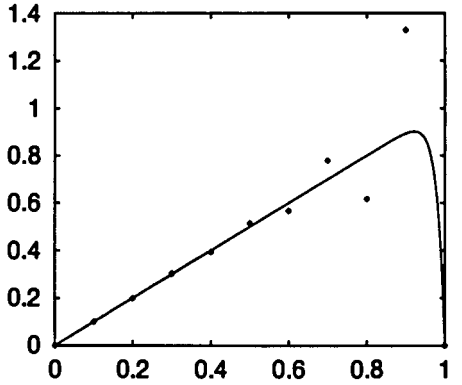


FIG. 13.3 - *Les valeurs u_j pour $h > 2\varepsilon/c_0$ (ici $f_0 = 1$, $\varepsilon = 0.02$, $c_0 = 1$, $h = 1/10$).*

le schéma (13.4) coincide avec (13.3) qui, nous l'avons vu, peut être oscillant lorsque h n'est pas "suffisamment petit". Nous allons, sur la base de (13.2), donner une formule de calcul pour les valeurs α_j.

Pour ce faire, supposons que $c(x) = c_0 =$ constante > 0, $f(x) = f_0 =$ constante et $\alpha_j = \alpha \in [0,1]$, $j = 1, \ldots, N$. Dans ce cas, le schéma (13.4) s'écrit

$$\begin{aligned} &2u_j - u_{j+1} - u_{j-1} + \gamma\Big(\alpha(u_j - u_{j-1}) + (1-\alpha)(u_{j+1} - u_j)\Big) \\ &\qquad = \frac{h^2 f_0}{\varepsilon}, \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0, \end{aligned} \tag{13.5}$$

où nous avons noté

$$\gamma = \frac{c_0}{\varepsilon} h. \tag{13.6}$$

D'autre part nous savons que la solution $u(x)$ de (13.1) est donnée par (13.2) et si nous posons $w_j = u(x_j)$ nous avons

$$w_j = \frac{f_0}{c_0}\left(h - \frac{1 - \exp(j\gamma)}{1 - \exp\left(\frac{c_0}{\varepsilon}\right)}\right), \qquad j = 0, 1, \ldots, N+1. \tag{13.7}$$

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

$$(1+\gamma\alpha)(2 - e^{-\gamma} - e^{\gamma}) + \gamma(e^{\gamma} - 1) = 0, \tag{13.8}$$

alors nous avons

$$\begin{aligned} &2w_j - w_{j-1} - w_{j+1} + \gamma\Big(\alpha(w_j - w_{j-1}) + (1-\alpha)(w_{j+1} - w_j)\Big) \\ &\qquad = \frac{f_0}{c_0}\gamma h = \frac{h^2 f_0}{\varepsilon}, \end{aligned}$$

et ainsi les valeurs $w_1, \ldots, w_N$ satisfont (13.5). Autrement dit, si l'égalité (13.8) est satisfaite, nous avons

$$u_j = u(x_j) = w_j, \qquad j = 1, \ldots, N,$$

c'est-à-dire que la solution exacte du problème (13.1) coincide aux points $x_1 \ldots, x_N$

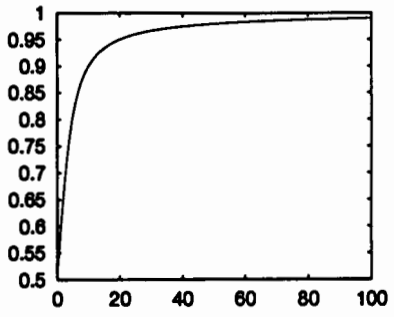


FIG. 13.4 - *Tracé de la fonction* $\gamma \to \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}$.

avec la solution du schéma (13.5). La condition (13.8) s'exprime encore

$$\begin{aligned}\alpha &= \frac{1-e^{\gamma}}{(2-e^{-\gamma}-e^{\gamma})} - \frac{1}{\gamma} \\ &= \frac{1-\frac{1}{2}e^{\gamma}-\frac{1}{2}e^{-\gamma}+\frac{1}{2}(e^{-\gamma}-e^{\gamma})}{2-e^{-\gamma}-e^{\gamma}} - \frac{1}{\gamma} \\ &= \frac{1}{2} + \frac{1}{2}\frac{e^{\gamma}-e^{-\gamma}}{e^{\gamma}+e^{-\gamma}-2} - \frac{1}{\gamma} \\ &= \frac{1}{2} + \frac{1}{2}\frac{(e^{\gamma/2}+e^{-\gamma/2})(e^{\gamma/2}-e^{-\gamma/2})}{(e^{\gamma/2}-e^{-\gamma/2})^2} - \frac{1}{\gamma},\end{aligned}$$

et donc

$$\alpha = \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}. \tag{13.9}$$

Dans la figure 13.4, nous avons représenté α comme fonction de γ, pour $\gamma \in \mathbb{R}$.

Dans le cas général où c et f ne sont pas constants, nous choisirons donc, si $c(x_j) \neq 0$:

$$\alpha_j = \frac{1}{2} + \frac{1}{2}\coth\frac{\gamma_j}{2} - \frac{1}{\gamma_j} \qquad \text{avec } \gamma_j = \frac{c(x_j)}{\varepsilon}h. \tag{13.10}$$

Le schéma numérique (13.4) avec les coefficients α_j donnés par (13.10) sera appelé "schéma upwind". Il produit la solution exacte lorsque c et f sont constants.

Remarque 13.1 Un calcul simple nous permet de vérifier que le schéma (13.4) peut se mettre sous la forme suivante :

$$\begin{aligned}&\varepsilon_j^* \frac{2u_j - u_{j+1} - u_{j-1}}{h^2} + c(x_j)\frac{u_{j+1}-u_{j-1}}{2h} \\ &\qquad = f(x_j), \qquad j = 1, \ldots, N, \\ &u_0 = u_{N+1} = 0,\end{aligned} \tag{13.11}$$

avec

$$\varepsilon_j^* = \varepsilon + c(x_j)h\left(\alpha_j - \frac{1}{2}\right). \tag{13.12}$$

Ainsi le schéma upwind est souvent interprèté comme un schéma de différences centrées (comparer (13.3) et (13.12)) dans lequel le terme de diffusion numérique $c(x_j)h(\alpha_j - 1/2)$ a été ajouté.

13.2 Un problème de convection–diffusion stationnaire et éléments finis

Approchons maintenant la solution u du problème (13.1) par une méthode d'éléments finis (voir chap. 9). Pour établir le problème faible correspondant au problème (13.1), nous pratiquons comme dans le paragraphe 9.2. 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 :

$$\varepsilon \int_0^1 u'(x)v'(x)dx + \int_0^1 c(x)u'(x)v(x)dx = \int_0^1 f(x)v(x)dx \qquad \forall v \in V. \quad (13.13)$$

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 par éléments finis de (13.13) consiste à chercher $u_h \in V_h$ tel que

$$\varepsilon \int_0^1 u_h'(x)v_h'(x)dx + \int_0^1 c(x)u_h'(x)v_h(x)dx = \int_0^1 f(x)v_h(x)dx \qquad \forall v_h \in V_h. \quad (13.14)$$

Exprimons u_h comme combinaison linéaire de $\varphi_1, \ldots, \varphi_N$, c'est-à-dire

$$u_h(x) = \sum_{i=1}^{N} u_i \varphi_i(x) \qquad \forall x \in [0,1], \quad (13.15)$$

et choisissons $v_h = \varphi_1, \ldots, \varphi_N$. Le problème (13.14) est donc équivalent à chercher $u_1, u_2, \ldots, u_N$ tels que

$$\sum_{i=1}^{N} u_i \int_0^1 \left(\varepsilon \varphi_i'(x)\varphi_j'(x) + c(x)\varphi_i'(x)\varphi_j(x)\right)dx = \int_0^1 f(x)\varphi_j(x)dx, \qquad j = 1, \ldots, N. \quad (13.16)$$

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 si les points $x_1, \ldots, x_N$, sont uniformément répartis ($h = 1/(N+1)$ et $x_j = jh$, $j = 1, \ldots, N$), alors nous vérifions facilement que A est la matrice tridiagonale définie par

$$A = \begin{pmatrix} \frac{2\varepsilon}{h} & -\frac{\varepsilon}{h} + \frac{c_0}{2} & & \\ -\frac{\varepsilon}{h} - \frac{c_0}{2} & \ddots & \ddots & \\ & \ddots & \ddots & -\frac{\varepsilon}{h} + \frac{c_0}{2} \\ & & -\frac{\varepsilon}{h} - \frac{c_0}{2} & \frac{2\varepsilon}{h} \end{pmatrix}, \tag{13.17}$$

et $f_j = f_0 h$, $j = 1, \ldots, N$. Le système linéaire (13.16) coincide donc, dans ce cas, avec le schéma centré (13.3). Nous avons vu qu'il convenait de le modifier, surtout si $\varepsilon << c_0$.

Pour ce faire, nous écrivons (13.14) de la manière suivante :

$$\sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} \Big(\varepsilon u_h'(x)v_h'(x) + c(x)u_h'(x)v_h(x)\Big)dx = \sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} f(x)v_h(x)dx \qquad \forall v_h \in V_h. \tag{13.18}$$

Ecrivons la fonction u_h sous la forme (13.15) et choisissons, sur chaque intervalle $]x_k, x_{k+1}[$, des fonctions test de la forme :

$$v_h(x) = \varphi_j(x) + \beta_k c(x_{k+1/2})\varphi_j'(x), \tag{13.19}$$

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(13.18) devient :

$$\sum_{i=1}^{N} u_i \sum_{k=0}^{N} \left(\Big(\varepsilon + \beta_k c(x_{k+1/2})^2\Big) \int_{x_k}^{x_{k+1}} \varphi_i'(x)\varphi_j'(x)dx + c(x_{k+1/2}) \int_{x_k}^{x_{k+1}} \varphi_i'(x)\varphi_j(x)dx \right)$$

$$= \sum_{k=0}^{N} \int_{x_k}^{x_{k+1}} f(x)\Big(\varphi_j(x) + \beta_k c(x_{k+1/2})\varphi_j'(x)\Big)dx, \qquad j = 1, \ldots, N. \tag{13.20}$$

Remarquons que, si $\beta_k c(x_{k+1/2}) \neq 0$, la fonction test v_h définie par (13.19) est discontinue 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 > 0, $f(x) = f_0 =$ 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 (13.20) s'écrit

$$\sum_{i=1}^{N} (A_{ji} + \beta c_0^2 B_{ji})\, u_i = f_0 h \qquad j = 1, \ldots, N, \tag{13.21}$$

où A est la $N \times N$ matrice tridiagonale définie par (13.17) et B est la $N \times N$ matrice tridiagonale définie par

$$B = \frac{1}{h} \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}.$$

En fait le schéma (13.21) 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 13.1 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 (13.9). Ainsi

$$\beta = \left(\frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}\right)\frac{h}{c_0} \tag{13.22}$$

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

$$\lim_{\gamma \to \infty} \left(\frac{1}{2}\coth\frac{\gamma}{2} - \frac{1}{\gamma}\right) = \frac{1}{2}.$$

Par conséquent, au lieu d'utiliser la formule (13.22) nous utiliserons une formule plus simple, à savoir

$$\beta = \frac{\gamma}{12} \cdot \frac{h}{c_0} \quad \text{si } \gamma < 6,$$
$$\beta = \frac{1}{2} \cdot \frac{h}{c_0} \quad \text{si } \gamma \geq 6.$$

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 (13.20):

$$\begin{aligned} \beta_k &= \frac{\gamma_k}{12} \cdot \frac{h_k}{|c(x_{k+1/2})|} \quad \text{si } \gamma_k < 6, \\ \beta_k &= \frac{1}{2}\frac{h_k}{|c(x_{k+1/2})|} \quad \text{si } \gamma_k \geq 6. \end{aligned} \tag{13.23}$$

Le schéma (13.20) avec les valeurs de β_k définies par (13.23) est appelé dans la littérature "schéma SUPG" (Streamline Upwind Petrov–Galerkin). Contrairement au schéma (13.16) (qui coincide avec (13.20) lorsque $\beta_k = 0$ pour tous les k), ce schéma ne produit pas une solution oscillante lorsque le nombre de Péclet de la maille k, γ_k, est grand. De plus, l'ordre de convergence du schéma SUPG et le même que l'ordre de convergence du schéma (13.16), lorsque $h = \max(h_0, \ldots, h_N)$ tend vers zéro.

13.3 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 $f : (x,t) \in \overline{\Omega} \times \mathbb{R}^+ \to f(x,t) \in \mathbb{R}$, 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

$$\begin{aligned} &\frac{\partial u}{\partial t}(x,t) - \varepsilon\Delta u(x,t) + \vec{c}(x,t) \cdot \overrightarrow{\text{grad}}u(x,t) \\ &\qquad = f(x,t) && \forall x \in \Omega, \quad \forall t > 0, \\ &u(x,t) = 0, && \forall x \in \partial\Omega \quad \forall t > 0, \\ &u(x,0) = w(x) && \forall x \in \Omega, \end{aligned} \tag{13.24}$$

où Δ est l'opérateur "Laplacien". Le problème (13.24) 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 (13.24) coincide avec le problème parabolique traité dans le paragraphe 11.3.

Pour obtenir une approximation par éléments finis de (13.24) en utilisant un schéma SUPG, il suffit de pratiquer comme dans le paragraphe 11.3, mais en modifiant les fonctions test comme nous l'avons fait dans le paragraphe précédent. Soit donc $\mathcal{T}_h$ une triangulation de Ω en triangles $K \in \mathcal{T}_h$. Soit $\varphi_1, \ldots, \varphi_N$

les fonctions de base définies par (10.15), affines sur chaque triangle $K \in \mathcal{T}_h$, valant 1 en un des noeuds intérieurs de la triangulation et zéro aux autres noeuds. Une discrétisation standard (voir paragraphe 11.3) en espace de (13.24) 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 semi-faible correspondante. Le problème se ramène donc à chercher les fonctions $u_1(t), \ldots, u_N(t)$ satisfaisant des conditions initiales (sur ce point, il suffit de procéder de façon identique à ce qui a été fait dans le paragraphe 11.3) ainsi que

$$\sum_{i=1}^{N} \dot{u}_i(t) \iint_\Omega \varphi_i(x)\varphi_j(x)dx + \varepsilon \sum_{i=1}^{N} u_i(t) \iint_\Omega \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx$$
$$+ \sum_{i=1}^{N} u_i(t) \iint_\Omega \left(\vec{c}(x,t) \cdot \overrightarrow{\text{grad}}\varphi_i(x)\right)\varphi_j(x)dx = \iint_\Omega f(x,t)\varphi_j(x)dx, \quad (13.25)$$

pour $j = 1, \ldots, N$. Procédons maintenant comme dans le paragraphe précédent 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 le schéma SUPG suivant:

$$\sum_{i=1}^{N} \dot{u}_i(t) \sum_{K\in\mathcal{T}_h} \iint_K \varphi_i(x) \left(\varphi_j(x) + \beta_K \vec{c}(Q_K) \cdot \overrightarrow{\text{grad}}\varphi_j(x)\right) dx$$
$$+ \sum_{i=1}^{N} u_i(t) \sum_{K\in\mathcal{T}_h} \left(\varepsilon \iint_K \overrightarrow{\text{grad}}\varphi_i(x) \cdot \overrightarrow{\text{grad}}\varphi_j(x)dx \right.$$
$$\left. + \iint_K \vec{c}(Q_K) \cdot \overrightarrow{\text{grad}}\varphi_i(x) \left(\varphi_j(x) + \beta_K \vec{c}(Q_K) \cdot \overrightarrow{\text{grad}}\varphi_j(x)\right) dx \right)$$
$$= \sum_{K\in\mathcal{T}_h} \iint_K f(x,t) \left(\varphi_j(x) + \beta_K \vec{c}(Q_K) \cdot \overrightarrow{\text{grad}}\varphi_j(x)\right) dx, \quad (13.26)$$

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 le paragraphe précédent, nous définissons les nombres β_K dans (13.26) par

$$\begin{aligned} \beta_K &= \frac{\gamma_K}{12} \cdot \frac{h_K}{|\vec{c}(Q_K)|} \quad \text{si } \gamma_K < 6, \\ \beta_K &= \frac{1}{2} \frac{h_K}{|\vec{c}(Q_K)|} \quad \text{si } \gamma_K \geq 6. \end{aligned} \quad (13.27)$$

Il reste ensuite à discrétiser (13.26) par rapport à la variable t en utilisant un schéma d'Euler rétrograde ou Crank–Nicholson, comme nous l'avons fait dans le chap. 11. Ainsi une méthode SUPG pour la résolution numérique du problème de diffusion-convection (13.24) sera donnée par une discrétisation temporelle standard (Euler rétrograde ou Crank–Nicholson par exemple) du système différentiel (13.26).