Chapitre 4

MÉTHODE DES ÉLÉMENTS FINIS

4.1 Approximation variationnelle

4.1.1 Introduction

Dans ce chapitre nous présentons la méthode des **éléments finis** qui est la méthode numérique de référence pour le calcul des solutions de problèmes aux limites. Le principe de cette méthode est directement issu de **l'approche variationnelle** que nous avons étudiée en détail dans les chapitres précédents.

L'idée de base de la méthode des éléments finis est de remplacer l'espace de Hilbert V sur lequel est posée la formulation variationnelle par un sous-espace V_h de

Hilbert V sur lequel est posée la formulation variationnelle par un sous-espace V_h de dimension finie. Le problème "approché" posé sur V_h se ramène à la simple résolution d'un système linéaire, dont la matrice est appelée **matrice de rigidité**. Par ailleurs, on peut choisir le mode de construction de V_h de manière à ce que le sous-espace V_h soit une bonne approximation de V et que la solution u_h dans V_h de la formulation variationnelle soit **"proche"** de la solution exacte u dans V.

Le plan de ce chapitre est le suivant. Dans la suite de cette section nous détaillons le processus **d'approximation variationnelle interne**. La Section 4.2 présente les éléments finis en une dimension d'espace où, sans trahir les idées générales valables en dimensions supérieures, les aspects techniques sont nettement plus simples. On discute des aspects pratiques (assemblage de la matrice de rigidité, formules de quadrature, etc.) autant que théoriques (convergence de la méthode, interpolation et estimation d'erreur). La Section 4.3 est dédiée aux éléments finis en dimension supérieure ($N \geq 2$). On introduit les notions de **maillage** (triangulaire ou quadrangulaire) et de **degrés de liberté** qui permettent de construire plusieurs familles de méthodes d'éléments finis. Pour plus de détails sur la méthode des éléments finis nous renvoyons à [3], [10], [16], [18], [23], [24] (voir aussi [13], [14], [21] pour des aspects pratiques de programmation informatique).

4.1.2 Approximation interne générale

Nous considérons à nouveau le cadre général du formalisme variationnel introduit au Chapitre 1. Étant donné un espace de Hilbert V, une forme bilinéaire

continue et coercive $a(u,v)$, et une forme linéaire continue $L(v)$, on considère la formulation variationnelle :

$$\text{trouver } u \in V \text{ tel que } a(u,v) = L(v) \quad \forall\, v \in V, \tag{4.1}$$

dont on sait qu'elle admet une unique solution par le Théorème 1.3.1 de Lax-Milgram. **L'approximation interne** de (4.1) consiste à remplacer l'espace de Hilbert V par un sous-espace de dimension finie V_h, c'est-à-dire à chercher la solution de :

$$\text{trouver } u_h \in V_h \text{ tel que } a(u_h,v_h) = L(v_h) \quad \forall\, v_h \in V_h. \tag{4.2}$$

La résolution de l'approximation interne (4.2) est facile comme le montre le lemme suivant.

Lemme 4.1.1 *Soit V un espace de Hilbert réel, et V_h un sous-espace de dimension finie. Soit $a(u,v)$ une forme bilinéaire continue et coercive sur V, et $L(v)$ une forme linéaire continue sur V. Alors l'approximation interne (4.2) admet une unique solution. Par ailleurs cette solution peut s'obtenir en résolvant un système linéaire de matrice définie positive (et symétrique si $a(u,v)$ est symétrique).*

Démonstration. L'existence et l'unicité de $u_h \in V_h$, solution de (4.2), découle du Théorème 1.3.1 de Lax-Milgram appliqué à V_h. Pour mettre le problème sous une forme plus simple, on introduit une base $(\phi_j)_{1\le j\le N_h}$ de V_h. Si $u_h = \sum_{j=1}^{N_h} u_j\phi_j$, on pose $U_h = (u_1, \ldots, u_{N_h})$ le vecteur dans $\mathbb{R}^{N_h}$ des coordonnées de u_h. Le problème

poser $U_h = (u_1, ..., u_{N_h})$ le vecteur dans $\mathbb{R}^{N_h}$ des coordonnées de u_h. Le problème (4.2) est équivalent à :

$$\text{trouver } U_h \in \mathbb{R}^{N_h} \text{ tel que } a\left(\sum_{j=1}^{N_h} u_j \phi_j, \phi_i\right) = L(\phi_i) \quad \forall\, 1 \leq i \leq N_h,$$

ce qui s'écrit sous la forme d'un système linéaire

$$\mathcal{K}_h U_h = b_h, \tag{4.3}$$

avec, pour $1 \leq i, j \leq N_h$,

$$(\mathcal{K}_h)_{ij} = a(\phi_j, \phi_i), \quad (b_h)_i = L(\phi_i).$$

La coercivité de la forme bilinéaire $a(u, v)$ entraîne le caractère défini positif de la matrice $\mathcal{K}_h$, et donc son inversibilité. En effet, pour tout vecteur $U_h \in \mathbb{R}^{N_h}$, on a

$$\mathcal{K}_h U_h \cdot U_h \geq \nu \left\| \sum_{j=1}^{N_h} u_j \phi_j \right\|^2 \geq C |U_h|^2 \quad \text{avec} \quad C > 0,$$

car toutes les normes sont équivalentes en dimension finie ($|\cdot|$ désigne la norme euclidienne dans $\mathbb{R}^{N_h}$). De même, la symétrie de $a(u, v)$ implique celle de $\mathcal{K}_h$. Dans les applications mécaniques la matrice $\mathcal{K}_h$ est appelée **matrice de rigidité**. $\square$

Nous allons maintenant comparer l'erreur commise en remplaçant l'espace V par son sous-espace V_h. Plus précisément, nous allons majorer la différence $\|u-u_h\|$ où u est la solution dans V de (4.1) et u_h celle dans V_h de (4.2). Précisons auparavant quelques notations : on note $\nu > 0$ la constante de coercivité et $M > 0$ la constante de continuité de la forme bilinéaire $a(u,v)$ qui vérifient

$$\begin{aligned} &a(u,u) \geq \nu \|u\|^2 \quad \forall\, u \in V, \\ &|a(u,v)| \leq M\|u\|\,\|v\| \quad \forall\, u,v \in V \end{aligned}$$

Le lemme suivant, dû à Jean Céa, montre que la distance entre la solution exacte u et la solution approchée u_h est majorée **uniformément par rapport au sous-espace** V_h par la distance entre u et V_h.

Lemme 4.1.2 (de Céa) *On se place sous les hypothèses du Lemme 4.1.1. Soit u la solution de (4.1) et u_h celle de (4.2). On a*

$$\|u-u_h\| \leq \frac{M}{\nu} \inf_{v_h \in V_h} \|u-v_h\|. \tag{4.4}$$

Démonstration. Puisque $V_h \subset V$, on déduit, par soustraction des formulations variationnelles (4.1) et (4.2), que

$$a(u-u_h, w_h) = 0 \quad \forall\, w_h \in V_h.$$

En choisissant $w_h = u_h - v_h$ on obtient

$$\nu\|u-u_h\|^2 \le a(u-u_h,u-u_h) = a(u-u_h,u-v_h) \le M\|u-u_h\|\|u-v_h\|,$$

d'où l'on déduit (4.4). □

Exercice 4.1.1 Dans le cadre du Lemme de Céa 4.1.2, démontrer que, si la forme bilinéaire $a(u,v)$ est symétrique, alors on améliore (4.4) en

$$\|u-u_h\| \le \sqrt{\frac{M}{\nu}} \inf_{v_h\in V_h} \|u-v_h\|.$$

Indication : on utilisera le fait que la solution u_h de (4.2) réalise aussi le minimum d'une énergie.

Finalement, pour démontrer la convergence de cette approximation variationnelle, nous donnons un dernier lemme général. Rappelons que dans la notation V_h le paramètre $h>0$ n'a pas encore de signification pratique. Néanmoins, nous supposerons que c'est dans la limite $h\to 0$ que l'approximation interne (4.2) "converge" vers la formulation variationnelle (4.1).

Lemme 4.1.3 *On se place sous les hypothèses du Lemme 4.1.1. On suppose qu'il existe un sous-espace $\mathcal{V}\subset V$ dense dans V et une application r_h de $\mathcal{V}$ dans V_h (appelée* **opérateur d'interpolation***) tels que*

$$\lim_{h\to 0} \|v-r_h(v)\| = 0 \quad \forall\, v\in\mathcal{V}. \tag{4.5}$$

Alors la méthode d'approximation variationnelle interne converge, c'est-à-dire que

$$\lim_{h\to 0} \|u - u_h\| = 0. \tag{4.6}$$

Démonstration. Soit $\epsilon > 0$. Par densité de $\mathcal{V}$, il existe $v \in \mathcal{V}$ tel que $\|u - v\| \leq \epsilon$. Par ailleurs, il existe un $h_0 > 0$ (dépendant de ϵ) tel que, pour cet élément $v \in \mathcal{V}$, on a

$$\|v - r_h(v)\| \leq \epsilon \quad \forall\, h \leq h_0.$$

En vertu du Lemme 4.1.2, on a

$$\|u - u_h\| \leq C\|u - r_h(v)\| \leq C\left(\|u - v\| + \|v - r_h(v)\|\right) \leq 2C\epsilon,$$

d'où l'on déduit le résultat. □

4.1.4 Méthode des éléments finis (principes généraux)

Le principe de la méthode des éléments finis est de construire des espaces d'approximation interne V_h dont la définition est basée sur la notion géométrique de **maillage** du domaine Ω. Un maillage est un pavage de l'espace en volumes élémentaires très simples : triangles, tétraèdres, parallélépipèdes (voir, par exemple, la Figure 4.7). Dans ce contexte le paramètre h de V_h correspond à la **taille maximale des mailles** ou cellules qui composent le maillage. Typiquement une base de V_h sera constituée de fonctions dont le support est **localisé** sur une ou quelques mailles. Ceci aura deux conséquences importantes : d'une part, dans la limite $h \to 0$

finies. Ceci aura deux conséquences importantes : d'une part, dans la limite $h \to 0$, l'espace V_h sera de plus en plus "gros" et approchera de mieux en mieux l'espace V tout entier, et d'autre part, la matrice de rigidité $\mathcal{K}_h$ du système linéaire (4.3) sera **creuse**, c'est-à-dire que la plupart de ses coefficients seront nuls (ce qui limitera le coût de la résolution numérique).

4.2 Éléments finis en dimension $N = 1$

Pour simplifier l'exposition nous commençons par présenter la méthode des éléments finis en une dimension d'espace. Sans perte de généralité nous choisissons le domaine $\Omega =]0,1[$. En dimension 1 un maillage est simplement constitué d'une collection de points $(x_j)_{0\leq j\leq n+1}$ (comme pour la méthode des différences finies) tels que

$$x_0 = 0 < x_1 < ... < x_n < x_{n+1} = 1.$$

Le maillage sera dit **uniforme** si les points x_j sont équidistants, c'est-à-dire que

$$x_j = jh \quad \text{avec} \quad h = \frac{1}{n+1},\ 0 \leq j \leq n+1.$$

Les points x_j sont aussi appelés les **sommets** (ou noeuds) du maillage. Par souci de simplicité nous considérons, pour l'instant, le problème modèle suivant

$$\left\{ \begin{array}{l} -u'' = f \ \text{ dans }]0,1[\\ u(0) = u(1) = 0, \end{array} \right. \tag{4.7}$$

dont nous savons qu'il admet une solution unique dans $H_0^1(\Omega)$ si $f \in L^2(\Omega)$ (voir le Chapitre 3). Dans tout ce qui suit on notera $\mathbb{P}_k$ l'ensemble des polynômes à coefficients réels d'une variable réelle de degré inférieur ou égal à k.

4.2.1 Éléments finis $\mathbb{P}_1$

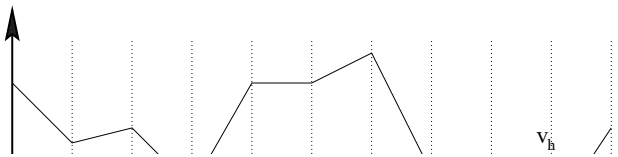
La méthode des éléments finis $\mathbb{P}_1$ repose sur l'espace discret des fonctions globalement continues et affines sur chaque maille

$$V_h = \left\{ v \in C([0,1]) \text{ tel que } v\,\big|_{[x_j, x_{j+1}]} \in \mathbb{P}_1 \text{ pour tout } 0 \leq j \leq n \right\}, \tag{4.8}$$

et sur son sous-espace

$$V_{0h} = \{ v \in V_h \text{ tel que } v(0) = v(1) = 0 \}. \tag{4.9}$$

La méthode des éléments finis $\mathbb{P}_1$ est alors simplement la méthode d'approximation variationnelle interne de la Sous-section 4.1.2 appliquée aux espaces V_h ou V_{0h} définis par (4.8) ou (4.9).



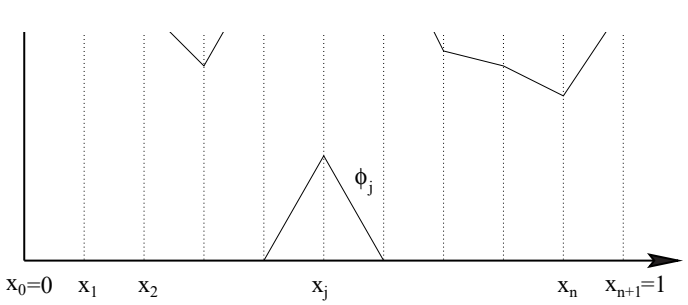


FIGURE 4.1 – Maillage de $\Omega =]0,1[$ et fonction de base en éléments finis $\mathbb{P}_1$.

On peut représenter les fonctions de V_h ou V_{0h}, affines par morceaux, à l'aide de fonctions de base très simples. Introduisons la "fonction chapeau" ϕ définie par

$$\phi(x) = \begin{cases} 1 - |x| & \text{si } |x| \leq 1, \\ 0 & \text{si } |x| > 1. \end{cases}$$

Si le maillage est uniforme, pour $0 \leq j \leq n+1$ on définit les fonctions de base (voir la Figure 4.1)

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h}\right). \tag{4.10}$$

Lemme 4.2.1 *L'espace V_h, défini par (4.8), est un sous-espace de $H^1(0,1)$ de dimension $n+2$, et toute fonction $v_h \in V_h$ est définie de manière unique par ses*

valeurs aux sommets $(x_j)_{0\leq j\leq n+1}$

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j)\phi_j(x) \quad \forall x \in [0,1].$$

De même, V_{0h}*, défini par (4.9), est un sous-espace de* $H_0^1(0,1)$ *de dimension* n*, et toute fonction* $v_h \in V_{0h}$ *est définie de manière unique par ses valeurs aux sommets* $(x_j)_{1\leq j\leq n}$

$$v_h(x) = \sum_{j=1}^{n} v_h(x_j)\phi_j(x) \quad \forall x \in [0,1].$$

Démonstration. Rappelons que les fonctions continues et de classe C^1 par morceaux appartiennent à $H^1(\Omega)$. Donc V_h et V_{0h} sont bien des sous-espaces de $H^1(0,1)$. Le reste de la preuve est immédiat en remarquant que $\phi_j(x_i) = \delta_{ij}$, où δ_{ij} est le symbole de Kronecker qui vaut 1 si $i = j$ et 0 sinon (voir la Figure 4.1). □

Remarque 4.2.2 La base (ϕ_j), définie par (4.10), permet de caractériser une fonction de V_h par ses valeurs aux noeuds du maillage. Dans ce cas on parle **d'éléments finis de Lagrange**. Par ailleurs, comme les fonctions sont localement $\mathbb{P}_1$, on dit que l'espace V_h, défini par (4.8), est l'espace des éléments finis de Lagrange d'ordre 1.

Cet exemple des éléments finis $\mathbb{P}_1$ permet à nouveau de comprendre l'intérêt de la formulation variationnelle. En effet, les fonctions de V_h ne sont pas deux fois

de la formulation variationnelle. En effet, les fonctions de V_h ne sont pas deux fois dérivables sur le segment $[0,1]$ et cela n'a pas de sens de résoudre, même de manière approchée, l'équation (4.7) (en fait la dérivée seconde d'une fonction de V_h est une somme de masses de Dirac aux noeuds du maillage !). Au contraire, il est parfaitement légitime d'utiliser des fonctions de V_h dans la formulation variationnelle (4.2) qui ne requiert qu'une seule dérivée. •

Décrivons la **résolution pratique** du problème de Dirichlet (4.7) par la méthode des éléments finis $\mathbb{P}_1$. La formulation variationnelle (4.2) de l'approximation interne devient ici :

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_0^1 u_h'(x)v_h'(x)\,dx = \int_0^1 f(x)v_h(x)\,dx \;\forall\, v_h \in V_{0h}. \quad (4.11)$$

On décompose u_h sur la base des $(\phi_j)_{1\le j\le n}$ et on prend $v_h = \phi_i$ ce qui donne

$$\sum_{j=1}^{n} u_h(x_j) \int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \int_0^1 f(x)\phi_i(x)\,dx.$$

En notant $U_h = (u_h(x_j))_{1\le j\le n}$, $b_h = \left(\int_0^1 f(x)\phi_i(x)\,dx\right)_{1\le i\le n}$, et en introduisant la **matrice de rigidité**

$$\mathcal{K}_h = \left(\int_0^1 \phi_j'(x)\phi_i'(x)\,dx\right)_{1\le i,j\le n},$$

la formulation variationnelle dans V_{0h} revient à résoudre dans $\mathbb{R}^n$ le système linéaire

$$\mathcal{K}_h U_h = b_h.$$

Comme les fonctions de base ϕ_j ont un "petit" support, l'intersection des supports de ϕ_j et ϕ_i est souvent vide et la plupart des coefficients de $\mathcal{K}_h$ sont nuls. Un calcul simple montre que

$$\int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \begin{cases} -h^{-1} & \text{si } j = i-1 \\ 2h^{-1} & \text{si } j = i \\ -h^{-1} & \text{si } j = i+1 \\ 0 & \text{sinon} \end{cases}$$

et la matrice $\mathcal{K}_h$ est tridiagonale

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}. \tag{4.12}$$

Pour obtenir le second membre b_h il faut calculer les intégrales

$$(b_h)_i = \int^{x_{i+1}} f(x)\phi_i(x)\,dx \quad \text{pour tout } 1 \le i \le n.$$

$J_{x_{i-1}}$

L'évaluation exacte du second membre b_h peut être difficile ou impossible si la fonction f est compliquée. En pratique on a recours à des **formules de quadrature** (ou formules d'intégration numérique) qui donnent une approximation des intégrales définissant b_h. Par exemple, on peut utiliser la formule du "point milieu"

$$\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \psi(x)\,dx \approx \psi\left(\frac{x_{i+1}+x_i}{2}\right),$$

ou la formule des "trapèzes"

$$\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \psi(x)\,dx \approx \frac{1}{2}\left(\psi(x_{i+1})+\psi(x_i)\right).$$

Ces deux formules sont exactes pour les fonctions ψ affines. Si la fonction ψ est régulière quelconque, alors ces formules sont simplement approchées avec un reste de l'ordre de $\mathcal{O}(h^2)$.

La résolution du système linéaire $\mathcal{K}_h U_h = b_h$ est la partie la plus coûteuse de la méthode en terme de temps de calcul. C'est pourquoi nous présentons dans la Section 4.4 des méthodes performantes de résolution. Rappelons que la matrice $\mathcal{K}_h$ est nécessairement inversible par application du Lemme 4.1.1.

Remarque 4.2.3 La matrice de rigidité $\mathcal{K}_h$ est très similaire à des matrices déjà rencontrées lors de l'étude des méthodes de différences finies. •

Problème de Neumann. La mise en oeuvre de la méthode des éléments finis $\mathbb{P}_1$ pour le problème de Neumann suivant est très similaire

$$\begin{cases} -u'' + au = f \text{ dans }]0,1[\\ u'(0) = \alpha, u'(1) = \beta. \end{cases} \tag{4.13}$$

Rappelons que (4.13) admet une solution unique dans $H^1(\Omega)$ si $f \in L^2(\Omega)$, $\alpha, \beta \in \mathbb{R}$, et $a \in L^\infty(\Omega)$ tel que $a(x) \geq a_0 > 0$ p.p. dans Ω (voir le Chapitre 3). La formulation variationnelle (4.2) de l'approximation interne devient ici : trouver $u_h \in V_h$ tel que

$$\int_0^1 (u_h'(x)v_h'(x) + a(x)u_h(x)v_h(x))\,dx = \int_0^1 f(x)v_h(x)\,dx - \alpha v_h(0) + \beta v_h(1),$$

pour tout $v_h \in V_h$. En décomposant u_h sur la base des $(\phi_j)_{0\leq j\leq n+1}$, la formulation variationnelle dans V_h revient à résoudre dans $\mathbb{R}^{n+2}$ le système linéaire

$$\mathcal{K}_h U_h = b_h,$$

avec $U_h = (u_h(x_j))_{0\leq j\leq n+1}$, et une nouvelle matrice de rigidité

$$\mathcal{K}_h = \left(\int_0^1 \left(\phi_j'(x)\phi_i'(x) + a(x)\phi_j(x)\phi_i(x)\right)\,dx \right)_{0\leq i,j\leq n+1},$$

et

$$\begin{aligned} (b_h)_i &= \textstyle\int_0^1 f(x)\phi_i(x)\,dx \quad \text{si } 1 \leq i \leq n, \\ (b_h)_0 &= \textstyle\int_0^1 f(x)\phi_0(x)\,dx - \alpha, \\ (b_h)_{n+1} &= \textstyle\int_0^1 f(x)\phi_{n+1}(x)\,dx + \beta. \end{aligned}$$

Lorsque $a(x)$ n'est pas une fonction constante, il est aussi nécessaire en pratique d'utiliser des formules de quadrature pour évaluer les coefficients de la matrice $\mathcal{K}_h$ (comme nous l'avons fait dans l'exemple précédent pour le second membre b_h).

Exercice 4.2.1 Appliquer la méthode des éléments finis $\mathbb{P}_1$ au problème

$$\begin{cases} -u'' = f \text{ dans }]0,1[\\ u(0) = \alpha, u(1) = \beta, \end{cases}$$

Vérifier que les conditions aux limites de Dirichlet non-homogènes apparaissent dans le second membre du système linéaire qui en découle.

Exercice 4.2.2 On reprend le problème de Neumann (4.13) en supposant que la fonction $a(x) = 0$ dans Ω. Montrer que la matrice du système linéaire issu de la méthode des éléments finis $\mathbb{P}_1$ est singulière. Montrer qu'on peut néanmoins résoudre le système linéaire si les données vérifient la condition de compatibilité

$$\int_0^1 f(x)\,dx = \alpha - \beta.$$

Comparer ce résultat avec le Théorème 3.2.18.

Exercice 4.2.3 Appliquer la méthode des différences finies au problème de Dirichlet (4.7). Vérifier qu'avec un schéma centré d'ordre deux, on obtient un système linéaire à résoudre avec la même matrice $\mathcal{K}_h$ (à un coefficient multiplicatif près) mais avec un second membre b_h différent. Même question pour le problème de Neumann (4.13).

4.2.2 Convergence et estimation d'erreur

Pour démontrer la convergence de la méthode des éléments finis $\mathbb{P}_1$ en une dimension d'espace nous suivons la démarche esquissée dans la Sous-section 4.1.2. Nous définissons tout d'abord un **opérateur d'interpolation** r_h (comme dans le Lemme 4.1.3).

Définition 4.2.4 *On appelle opérateur d'interpolation* $\mathbb{P}_1$ *l'application linéaire* r_h *de* $H^1(0,1)$ *dans* V_h *définie, pour tout* $v \in H^1(0,1)$*, par*

$$(r_h v)(x) = \sum_{j=0}^{n+1} v(x_j)\phi_j(x).$$

Cette définition a bien un sens car, en vertu du Lemme 2.3.3, les fonctions de $H^1(0,1)$ sont continues et leurs valeurs ponctuelles sont donc bien définies. L'interpolée $r_h v$ d'une fonction v est simplement la fonction affine par morceaux qui coïncide avec v sur les sommets du maillage x_j (voir la Figure 4.2). Remarquons qu'en une dimension d'espace l'interpolée est définie pour toute fonction de $H^1(0,1)$, et non pas seulement pour les fonctions régulières de $H^1(0,1)$ (ce qui sera le cas en dimension supérieure).

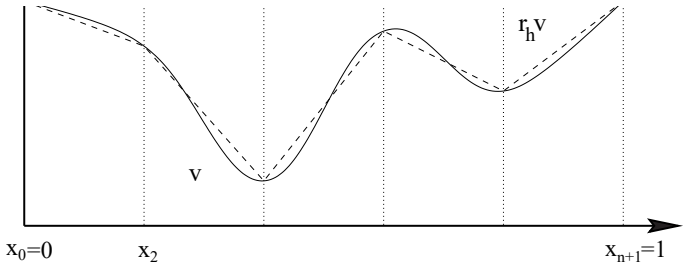


FIGURE 4.2 – Interpolation $\mathbb{P}_1$ d'une fonction de $H^1(0,1)$.

La convergence de la méthode des éléments finis $\mathbb{P}_1$ repose sur le lemme suivant.

Lemme 4.2.5 (d'interpolation) *Soit r_h l'opérateur d'interpolation $\mathbb{P}_1$. Pour tout $v \in H^1(0,1)$, il vérifie*

$$\lim_{h\to 0} \|v - r_h v\|_{H^1(0,1)} = 0.$$

De plus, si $v \in H^2(0,1)$, alors il existe une constante C indépendante de h telle que

$$\|v - r_h v\|_{H^1(0,1)} \leq Ch\|v''\|_{L^2(0,1)}.$$

Nous repoussons momentanément la démonstration de ce lemme pour énoncer tout de suite le résultat principal de cette sous-section qui établit la convergence de la méthode des éléments finis $\mathbb{P}_1$ pour le problème de Dirichlet.

Théorème 4.2.6 *Soit $u \in H_0^1(0,1)$ et $u_h \in V_{0h}$ les solutions de (4.7) et (4.11), respectivement. Alors, la méthode des éléments finis $\mathbb{P}_1$ converge, c'est-à-dire que*

$$\lim_{h\to 0} \|u - u_h\|_{H^1(0,1)} = 0. \tag{4.14}$$

De plus, si $u \in H^2(0,1)$ (ce qui est vrai si $f \in L^2(0,1)$), alors il existe une constante C indépendante de h telle que

$$\|u - u_h\|_{H^1(0,1)} \leq Ch\|u''\|_{L^2(0,1)} = Ch\|f\|_{L^2(0,1)}. \tag{4.15}$$

Remarque 4.2.7 On peut faire une analogie entre la convergence d'une méthode d'éléments finis et la convergence d'une méthode de différences finies. Rappelons que, d'après le Théorème de Lax, la convergence d'un schéma aux différences finies découle de sa stabilité et de sa consistance. Indiquons quels sont les équivalents (formels) de ces ingrédients dans le contexte des éléments finis. Le rôle de la consistance pour les éléments finis est joué par la propriété d'interpolation du Lemme 4.2.5, tandis que le rôle de la stabilité est tenu par la propriété de coercivité de la forme bilinéaire qui assure la résolution (stable) de toute approximation interne. •

Démonstration. Le Lemme 4.2.5 permet d'appliquer le résultat de convergence du Lemme 4.1.3 qui entraîne immédiatement (4.14). Pour obtenir (4.15), on majore l'estimation du Lemme 4.1.2 de Céa

$$\|u - u_h\|_{H^1(0,1)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(0,1)} \leq C\|u - r_h u\|_{H^1(0,1)},$$

ce qui permet de conclure grâce au Lemme 4.2.5. □

Nous donnons maintenant la démonstration du Lemme 4.2.5 sous la forme de deux autres lemmes techniques.

Lemme 4.2.10 *Il existe une constante C indépendante de h telle que, pour tout $v \in H^2(0,1)$,*

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch^2 \|v''\|_{L^2(0,1)}, \tag{4.16}$$

et

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq Ch \|v''\|_{L^2(0,1)}. \tag{4.17}$$

Démonstration. Soit $v \in C^\infty([0,1])$. Par définition, l'interpolée $r_h v$ est une fonction affine et, pour tout $x \in]x_j, x_{j+1}[$, on a

$$\begin{aligned} v(x) - r_h v(x) &= v(x) - \left(v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j}(x - x_j) \right) \\ &= \int_{x_j}^{x} v'(t)\, dt - \frac{x - x_j}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v'(t)\, dt \\ &= (x - x_j) v'(x_j + \theta_x) - (x - x_j) v'(x_j + \theta_j) \\ &= (x - x_j) \int_{x_j + \theta_j}^{x_j + \theta_x} v''(t)\, dt, \end{aligned} \tag{4.18}$$

par application de la formule des accroissements finis avec $0 \leq \theta_x \leq x - x_j$ et $0 \leq \theta_j \leq h$. On en déduit en utilisant l'inégalité de Cauchy-Schwarz

$$|v(x) - r_h v(x)|^2 \leq h^2 \left(\int_{x_j}^{x_{j+1}} |v''(t)|\, dt \right)^2 \leq h^3 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt. \tag{4.19}$$

En intégrant (4.19) par rapport à x sur l'intervalle $[x_j, x_{j+1}]$, on obtient

$$\int_{x_j}^{x_{j+1}} |v(x) - r_h v(x)|^2\, dx \leq h^4 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt,$$

ce qui, par sommation en j, donne exactement (4.16). Par densité ce résultat est encore vrai pour tout $v \in H^2(0,1)$. La démonstration de (4.17) est tout à fait similaire : pour $v \in C^\infty([0,1])$ et $x \in]x_j, x_{j+1}[$ on écrit

$$\begin{aligned} v'(x) - (r_h v)'(x) \;=\; v'(x) - \frac{v(x_{j+1}) - v(x_j)}{h} \;&=\; \frac{1}{h} \int_{x_j}^{x_{j+1}} \Big(v'(x) - v'(t) \Big) dt \\ &=\; \frac{1}{h} \int_{x_j}^{x_{j+1}} \int_t^x v''(y)\, dy. \end{aligned}$$

Élevant au carré cette inégalité, appliquant Cauchy-Schwarz deux fois et sommant en j on obtient (4.17), qui est aussi valide pour tout $v \in H^2(0,1)$ par densité. $\square$

Lemme 4.2.11 *Il existe une constante C indépendante de h telle que, pour tout $v \in H^1(0,1)$,*

$$\|r_h v\|_{H^1(0,1)} \leq C\|v\|_{H^1(0,1)}, \tag{4.20}$$

et

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch\|v'\|_{L^2(0,1)}. \tag{4.21}$$

De plus, pour tout $v \in H^1(0,1)$, on a

$$\lim_{h\to 0} \|v' - (r_h v)'\|_{L^2(0,1)} = 0. \tag{4.22}$$

Démonstration. Les preuves de (4.20) et (4.21) sont dans le même esprit que celles du lemme précédent. Soit $v \in H^1(0,1)$. Tout d'abord on a

$$\|r_h v\|_{L^2(0,1)} \leq \max_{x\in[0,1]} |r_h v(x)| \leq \max_{x\in[0,1]} |v(x)| \leq C\|v\|_{H^1(0,1)},$$

en vertu du Lemme 2.3.3. D'autre part, comme $r_h v$ est affine, et grâce à la propriété (2.8) du Lemme 2.3.3 qui affirme que v est bien la primitive de v', on a

$$\begin{aligned}\int_{x_j}^{x_{j+1}} |(r_h v)'(x)|^2\, dx &= \frac{(v(x_{j+1}) - v(x_j))^2}{h} \\ &= \frac{1}{h}\left(\int_{x_j}^{x_{j+1}} v'(x)\, dx\right)^2 \leq \int_{x_j}^{x_{j+1}} |v'(x)|^2\, dx,\end{aligned}$$

par Cauchy-Schwarz, ce qui, par sommation en j, conduit à (4.20). Pour obtenir (4.21) on reprend la deuxième égalité de (4.18) d'où l'on déduit

$$|v(x) - r_h v(x)| \leq 2 \int_{x_j}^{x_{j+1}} |v'(t)|\, dt.$$

En élevant au carré, en utilisant Cauchy-Schwarz, en intégrant par rapport à x, puis en sommant en j, on obtient bien (4.21).

Passons à la démonstration de (4.22). Soit $\epsilon > 0$. Comme $C^\infty([0,1])$ est dense dans $H^1(0,1)$, pour tout $v \in H^1(0,1)$ il existe $\phi \in C^\infty([0,1])$ tel que

$$\|v' - \phi'\|_{L^2(0,1)} \leq \epsilon.$$

Or r_h est une application linéaire qui vérifie (4.20), donc on en déduit

$$\|(r_h v)' - (r_h \phi)'\|_{L^2(0,1)} \leq C\|v' - \phi'\|_{L^2(0,1)} \leq C\epsilon.$$

Le choix de ϕ et de ϵ étant fixé, on déduit de (4.17) appliqué à ϕ que, pour h suffisamment petit,

$$\|\phi' - (r_h \phi)'\|_{L^2(0,1)} \leq \epsilon.$$

Par conséquent, en sommant ces trois dernières inégalités on obtient

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq \|v' - \phi'\|_{L^2} + \|\phi' - (r_h \phi)'\|_{L^2} + \|(r_h v)' - (r_h \phi)'\|_{L^2} \leq C\epsilon,$$

ce qui implique (4.22). □

4.2.3 Éléments finis $\mathbb{P}_2$

La méthode des éléments finis $\mathbb{P}_2$ repose sur l'espace discret

$$V_h = \left\{ v \in C([0,1]) \text{ tel que } v\,\big|_{[x_j,x_{j+1}]} \in \mathbb{P}_2 \text{ pour tout } 0 \le j \le n \right\}, \tag{4.23}$$

et sur son sous-espace

$$V_{0h} = \{ v \in V_h \text{ tel que } v(0) = v(1) = 0 \}. \tag{4.24}$$

La méthode des éléments finis $\mathbb{P}_2$ est la méthode d'approximation variationnelle interne de la Sous-section 4.1.2 appliquée à ces espaces V_h ou V_{0h}. Ceux-ci sont composés de fonctions continues, paraboliques par morceaux qu'on peut représenter à l'aide de fonctions de base très simples.

Introduisons tout d'abord les points milieux des segments $[x_j, x_{j+1}]$ définis par $x_{j+1/2} = x_j + h/2$ pour $0 \le j \le n$. On définit aussi deux fonctions "mères"

$$\phi(x) = \begin{cases} (1+x)(1+2x) & \text{si } -1 \le x \le 0, \\ (1-x)(1-2x) & \text{si } 0 \le x \le 1, \\ 0 & \text{si } |x| > 1, \end{cases}$$

et

$$\psi(x) = \begin{cases} 1 - 4x^2 & \text{si } |x| \le 1/2, \\ 0 & \text{si } |x| > 1/2. \end{cases}$$

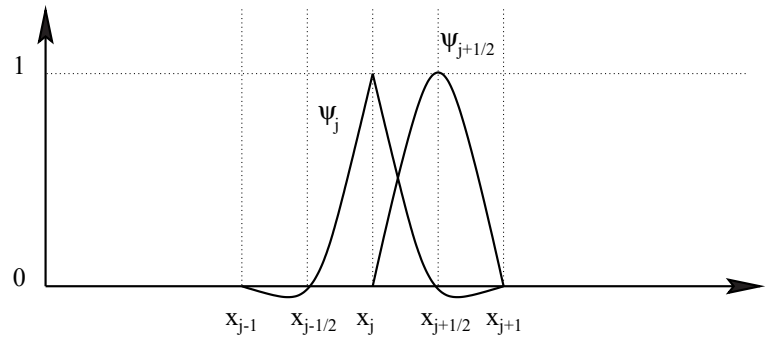


FIGURE 4.3 – Les fonctions de base des éléments finis $\mathbb{P}_2$.

Si le maillage est uniforme, pour $0 \leq j \leq n+1$ on définit les fonctions de base (voir la Figure 4.3)

$$\psi_j(x) = \phi\left(\frac{x - x_j}{h}\right),\ 0 \leq j \leq n+1,\ \text{et}\ \psi_{j+1/2}(x) = \psi\left(\frac{x - x_{j+1/2}}{h}\right),\ 0 \leq j \leq n.$$

Lemme 4.2.12 *L'espace V_h, défini par (4.23), est un sous-espace de $H^1(0,1)$ de dimension $2n+3$, et toute fonction $v_h \in V_h$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{0\leq j\leq n+1}$ et aux milieux $(x_{j+1/2})_{0\leq j\leq n}$*

$$\overbrace{\qquad}^{n+1} \qquad \overbrace{\qquad}^{n}$$

$$v_h(x) = \sum_{j=0}^{n} v_h(x_j)\psi_j(x) + \sum_{j=0}^{n} v_h(x_{j+1/2})\psi_{j+1/2}(x) \quad \forall x \in [0,1].$$

De même, V_{0h}, défini par (4.24), est un sous-espace de $H_0^1(0,1)$ de dimension $2n+1$, et toute fonction $v_h \in V_{0h}$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{1\le j\le n}$ et aux milieux $(x_{j+1/2})_{0\le j\le n}$

$$v_h(x) = \sum_{j=1}^{n} v_h(x_j)\psi_j(x) + \sum_{j=0}^{n} v_h(x_{j+1/2})\psi_{j+1/2}(x) \quad \forall x \in [0,1].$$

Remarque 4.2.13 Ici encore, V_h est un espace d'éléments finis de Lagrange (cf. la Remarque 4.2.2). Comme les fonctions sont localement $\mathbb{P}_2$, on dit que l'espace V_h, défini par (4.23), est l'espace des éléments finis de Lagrange d'ordre 2. •

Démonstration. Clairement V_h et V_{0h} sont bien des sous-espaces de $H^1(0,1)$. Leur dimension et les bases proposées se trouvent facilement en remarquant que $\psi_j(x_i) = \delta_{ij}$, $\psi_{j+1/2}(x_{i+1/2}) = \delta_{ij}$, $\psi_j(x_{i+1/2}) = 0$, $\psi_{j+1/2}(x_i) = 0$ (voir la Figure 4.3). □

Décrivons la **résolution pratique** du problème de Dirichlet (4.7) par la méthode des éléments finis $\mathbb{P}_2$. La formulation variationnelle (4.2) de l'approximation interne revient à résoudre dans $\mathbb{R}^{2n+1}$ un système linéaire $\mathcal{K}_h U_h = b_h$. On note $(x_{k/2})_{1\le k\le 2n+1}$ les points du maillage et $(\psi_{k/2})_{1\le k\le 2n+1}$ les fonctions de base correspondantes dans V_{0h}. Dans cette base $U_h \in \mathbb{R}^{2n+1}$ est le vecteur des coordonnées de

la solution approchée u_h et la matrice de rigidité est ici pentadiagonale

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 16/3 & -8/3 & 0 & & & & & & \\ -8/3 & 14/3 & -8/3 & 1/3 & & & 0 & & \\ 0 & -8/3 & 16/3 & -8/3 & 0 & & & & \\ & 1/3 & -8/3 & 14/3 & -8/3 & 1/3 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 0 & -8/3 & 16/3 & -8/3 & 0 & \\ & 0 & & & 1/3 & -8/3 & 14/3 & -8/3 & \\ & & & & & 0 & -8/3 & 16/3 & \end{pmatrix}.$$

Remarquons que cette matrice est plus "pleine" que celle obtenue par la méthode des éléments finis $\mathbb{P}_1$, et donc que la résolution du système linéaire coûtera plus cher en temps de calcul. Pour évaluer le second membre b_h on a recours aux même formules de quadrature que celles présentées dans la méthode $\mathbb{P}_1$.

Théorème 4.2.14 *Soit $u \in H_0^1(0,1)$ la solution exacte de (4.7) et $u_h \in V_{0h}$ la solution approchée par la méthode des éléments finis $\mathbb{P}_2$. La méthode des éléments finis $\mathbb{P}_2$ converge, c'est-à-dire que*

$$\lim_{h \to 0} \|u - u_h\|_{H^1(0,1)} = 0.$$

De plus, si $u \in H^3(0,1)$ (ce qui est vrai si $f \in H^1(0,1)$), alors il existe une constante

C indépendante de h telle que

$$\|u - u_h\|_{H^1(0,1)} \leq Ch^2 \|u'''\|_{L^2(0,1)}.$$

Le Théorème 4.2.14 montre l'avantage principal des éléments finis $\mathbb{P}_2$: si la solution est régulière, alors la convergence de la méthode est **quadratique** (la vitesse de convergence est proportionnelle à h^2) alors que la convergence pour les éléments finis $\mathbb{P}_1$ est seulement linéaire (proportionnelle à h). Bien sûr cet avantage a un prix : il y a deux fois plus d'inconnues (exactement $2n + 1$ au lieu de n pour les éléments finis $\mathbb{P}_1$) donc la matrice est deux fois plus grande, et en plus la matrice a cinq diagonales non nulles au lieu de trois dans le cas $\mathbb{P}_1$. Remarquons que si la solution n'est pas régulière ($u \in H^3(0,1)$) il n'y a aucun avantage théorique (mais aussi pratique) à utiliser des éléments finis $\mathbb{P}_2$ plutôt que $\mathbb{P}_1$.

4.2.4 Propriétés qualitatives

Nous savons que la solution d'un problème de Dirichlet vérifie le principe du maximum (voir le Théorème 3.2.22). Il est important de savoir si cette propriété est conservée par l'approximation variationnelle interne.

Proposition 4.2.15 (principe du maximum discret) *On suppose que $f \geq 0$ presque partout dans $]0,1[$. Alors, la solution u_h de l'approximation variationnelle (4.11) par la méthode des éléments finis $\mathbb{P}_1$ vérifie $u_h \geq 0$ dans $[0,1]$.*

Démonstration. Soit u_h la solution de (4.11). En vertu du Lemme 4.2.1, on a

$$u_h(x) = \sum_{j=1}^{n} u_h(x_j)\phi_j(x),$$

où les fonctions ϕ_j sont les fonctions de base des éléments finis $\mathbb{P}_1$ dans V_{0h} et $U_h = (u_h(x_j))_{1\leq j\leq n}$ est solution du système linéaire

$$\mathcal{K}_h U_h = b_h. \tag{4.27}$$

Les fonctions ϕ_j sont des fonctions "chapeaux" (voir la Figure 4.1) qui sont positives : il suffit donc de montrer que toutes les composantes du vecteur $U_h = (U_h^j)_{1\leq j\leq n}$ sont positives pour prouver que la fonction u_h est positive sur $[0,1]$. Rappelons que, en posant $U_h^0 = U_h^{n+1} = 0$, le système linéaire (4.27) est équivalent à

$$-U_h^{j-1} + 2U_h^j - U_h^{j+1} = hb_h^j \text{ pour tout } 1 \leq j \leq n. \tag{4.28}$$

Soit $U_h^{j_0} = \min_j U_h^j$ la plus petite composante de U_h : s'il y a plusieurs plus petites composantes, on choisit celle de plus petit indice j_0. Si $j_0 = 0$, alors $U_h^j \geq U_h^0 = 0$ pour tout j, ce qui est le résultat recherché. Si $j_0 \geq 1$, alors $U_h^{j_0} < U_h^0 = 0$, et comme $U_h^{n+1} = 0$ on en déduit que $j_0 \leq n$. Comme $b_h^j = \int_0^1 f\psi_j\,dx \geq 0$ par hypothèse sur f, on peut alors déduire de la relation (4.28) pour j_0 que

$$\left(U_h^{j_0} - U_h^{j_0-1}\right) + \left(U_h^{j_0} - U_h^{j_0+1}\right) \geq 0,$$

ce qui est une contradiction avec le caractère minimal (strict) de $U_h^{j_0}$. Par conséquent la méthode des éléments finis $\mathbb{P}_1$ vérifie le principe du maximum discret. $\square$

4.3 Éléments finis en dimension $N \geq 2$

Nous nous plaçons maintenant en dimension d'espace $N \geq 2$ (en pratique $N = 2, 3$). Pour simplifier l'exposé, certains résultats ne seront démontrés qu'en dimension $N = 2$, mais ils s'étendent à la dimension $N = 3$ (au prix, parfois, de complications techniques et pratiques importantes).

Nous considérons le problème modèle de Dirichlet

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{4.36}$$

dont nous savons qu'il admet une solution unique dans $H_0^1(\Omega)$, si $f \in L^2(\Omega)$ (voir le Chapitre 3).

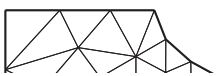
Dans tout ce qui suit nous supposerons que le domaine Ω est **polyédrique** (polygonal si $N = 2$), c'est-à-dire que $\overline{\Omega}$ est une réunion finie de polyèdres de $\mathbb{R}^N$. Rappelons qu'un polyèdre est une intersection finie de demi-espaces de $\mathbb{R}^N$ et que les parties de son bord qui appartiennent à un seul hyperplan sont appelées ses faces. La raison de cette hypothèse est qu'il n'est possible de mailler exactement que de tels ouverts. Nous dirons plus loin ce qui se passe pour des domaines généraux à bords "courbes".

4.3.1 Éléments finis triangulaires

Tout commence par la définition d'un maillage du domaine Ω par des triangles en dimension $N = 2$ et des tétraèdres en dimension $N = 3$. On regroupe les triangles et les tétraèdres dans la famille plus générale des N-simplexes. On appelle N-simplexe K de $\mathbb{R}^N$ l'enveloppe convexe de $(N + 1)$ points $(a_j)_{1\leq j\leq N+1}$ de $\mathbb{R}^N$, appelés sommets de K. Bien sûr un 2-simplexe est simplement un triangle et un 3-simplexe un tétraèdre (voir la Figure 4.9). On dit que le N-simplexe K est non dégénéré si les points $(a_j)_{1\leq j\leq N+1}$ n'appartiennent pas à un même hyperplan de $\mathbb{R}^N$ (le triangle ou le tétraèdre est non "plat"). Si on note $(a_{i,j})_{1\leq i\leq N}$ les coordonnées du vecteur a_j, la condition de non dégénérescence de K est que la matrice

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,N+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,N+1} \\ \vdots & \vdots & & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,N+1} \\ 1 & 1 & \dots & 1 \end{pmatrix} \tag{4.37}$$

soit inversible (ce que l'on supposera toujours par la suite). Un N-simplexe a autant de faces que de sommets, qui sont elles-mêmes des $(N - 1)$-simplexes.



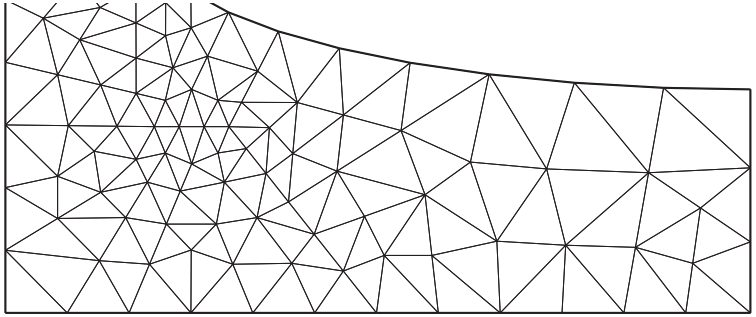


FIGURE 4.7 – Exemple de maillage triangulaire en dimension $N = 2$.

Définition 4.3.1 *Soit Ω un ouvert connexe polyédrique de $\mathbb{R}^N$. Un* **maillage** *triangulaire ou une triangulation de $\overline{\Omega}$ est un ensemble $\mathcal{T}_h$ de N-simplexes (non dégénérés) $(K_i)_{1\leq i\leq n}$ qui vérifient*

1. *$K_i \subset \overline{\Omega}$ et $\overline{\Omega} = \cup_{i=1}^n K_i$,*
2. *l'intersection $K_i \cap K_j$ de deux N-simplexes distincts est un m-simplexe, avec $0 \leq m \leq N-1$, dont tous les sommets sont aussi des sommets de K_i et K_j. (En dimension $N = 2$, l'intersection de deux triangles est soit vide, soit réduite à un sommet commun, soit une arête commune* **entière** *; en dimension $N = 3$, l'intersection de deux tétraèdres est soit vide, soit un sommet commun, soit une arête commune entière, soit une face commune entière.)*

Les **sommets** *ou* **noeuds** *du maillage $\mathcal{T}_h$ sont les sommets des N-simplexes K_i qui le composent. Par convention, le paramètre h désigne le maximum des diamètres des N-simplexes K_i.*

Il est clair que la Définition 4.3.1 ne peut s'appliquer qu'à un ouvert polyédrique et pas à un ouvert quelconque. La Définition 4.3.1 contient un certain nombre de restrictions sur le maillage : dans ce cas on parle souvent de **maillage conforme**. Un exemple de maillage conforme est donné à la Figure 4.7, tandis que la Figure 4.8 présente des situations interdites par la Définition 4.3.1.

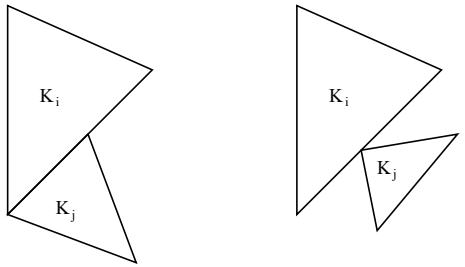


FIGURE 4.8 – Exemples de situations interdites pour un maillage triangulaire.

Remarque 4.3.2 Nous ne disons rien ici des algorithmes qui permettent de construire un maillage triangulaire. Contentons nous de dire que, s'il est relativement facile de mailler des domaines plans (il existe de nombreux logiciels libres qui per-

facile de mailler des domaines plans (il existe de nombreux logiciels libres qui permettent de le faire), il est encore assez compliqué de mailler des domaines tridimensionnels. Nous renvoyons à l'ouvrage [17] le lecteur intéressé par ce sujet. •

Dans un N-simplexe K il est commode d'utiliser des coordonnées barycentriques au lieu des coordonnées cartésiennes usuelles. Rappelons que, si K est un N-simplexe non dégénéré de sommets $(a_j)_{1\leq j\leq N+1}$, les **coordonnées barycentriques** $(\lambda_j)_{1\leq j\leq N+1}$ de $x\in\mathbb{R}^N$ sont définies par

$$\sum_{j=1}^{N+1}\lambda_j=1,\quad \sum_{j=1}^{N+1}a_{i,j}\lambda_j=x_i\quad \text{pour } 1\leq i\leq N, \tag{4.38}$$

qui admet bien une unique solution car la matrice A, définie par (4.37), est inversible. Remarquons que les λ_j sont des fonctions affines de x. On vérifie alors que

$$K=\left\{x\in\mathbb{R}^N \text{ tel que } \lambda_j(x)\geq 0 \ \text{ pour } 1\leq j\leq N+1\right\},$$

et que les $(N+1)$ faces de K sont les intersections de K et des hyperplans $\lambda_j(x)=0$, $1\leq j\leq N+1$. On peut alors définir un ensemble de points de K qui vont jouer un rôle particulier pour la suite : pour tout entier $k\geq 1$ on appelle **treillis d'ordre** k l'ensemble

$$\Sigma_k=\left\{x\in K \text{ tel que } \lambda_j(x)\in\{0,\frac{1}{k},...,\frac{k-1}{k},1\} \text{ pour } 1\leq j\leq N\right\}. \tag{4.39}$$

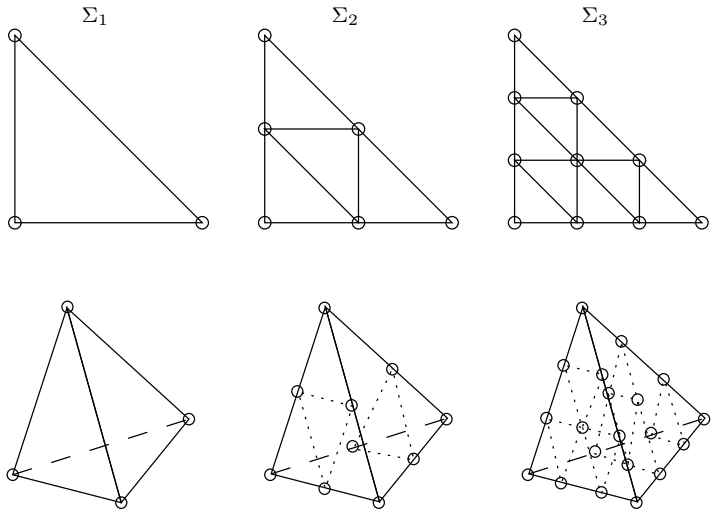


FIGURE 4.9 – Treillis d'ordre 1, 2, et 3 pour un triangle (en haut) et un tétraèdre (en bas). Les ronds représentent les points du treillis.

Pour $k = 1$ il s'agit de l'ensemble des sommets de K, et pour $k = 2$ des sommets et des points milieux des arêtes reliant deux sommets (voir la Figure 4.9). Dans le cas général, Σ_k est un ensemble fini de points $(\sigma_j)_{1\leq j\leq n_k}$.

Nous définissons maintenant l'ensemble $\mathbb{P}_k$ des polynômes à coefficients réels de $\mathbb{R}^N$ dans $\mathbb{R}$ de degré inférieur ou égal à k, c'est-à-dire que tout $p \in \mathbb{P}_k$ s'écrit sous

la forme

$$p(x) = \sum_{\substack{i_1,\dots,i_N \geq 0 \\ i_1+\dots+i_N \leq k}} \alpha_{i_1,\dots,i_N} x_1^{i_1} \cdots x_N^{i_N} \text{ avec } x = (x_1, ..., x_N).$$

L'intérêt de la notion de treillis Σ_k d'un N-simplexe K est qu'il permet de caractériser tous les polynômes de $\mathbb{P}_k$ (on dit que Σ_k est **unisolvant** pour $\mathbb{P}_k$).

Lemme 4.3.3 *Soit K un N-simplexe. Pour un entier $k \geq 1$, soit Σ_k le treillis d'ordre k, défini par (4.39), dont les points sont notés $(\sigma_j)_{1\leq j\leq n_k}$. Alors, tout polynôme de $\mathbb{P}_k$ est déterminé de manière unique par ses valeurs aux points $(\sigma_j)_{1\leq j\leq n_k}$. Autrement dit, il existe une base $(\psi_j)_{1\leq j\leq n_k}$ de $\mathbb{P}_k$ telle que*

$$\psi_j(\sigma_i) = \delta_{ij} \quad 1 \leq i, j \leq n_k.$$

Démonstration. Le cardinal de Σ_k et la dimension de $\mathbb{P}_k$ coïncident

$$\text{card}(\Sigma_k) = \dim(\mathbb{P}_k) = \frac{(N+k)!}{N!\,k!}$$

(le vérifier en guise d'exercice, au moins pour $k = 1, 2$). Comme l'application qui, à tout polynôme de $\mathbb{P}_k$, fait correspondre ses valeurs sur le treillis Σ_k est linéaire, il suffit de montrer qu'elle est injective pour montrer qu'elle est bijective. Soit donc un polynôme $p \in \mathbb{P}_k$ qui s'annule sur Σ_k. Montrons, par récurrence sur la dimension N, que p est identiquement nul sur $\mathbb{R}^N$. Pour $N = 1$, il est clair qu'un polynôme de degré k qui s'annule en $(k + 1)$ points distincts est nul. Supposons le résultat

vrai à l'ordre $N-1$. Comme x dépend linéairement des coordonnées barycentriques $(\lambda_j(x))_{1\leq j\leq N+1}$, on peut définir un polynôme $q(\lambda) = p(x)$ de degré au plus k en la variable $\lambda \in \mathbb{R}^{N+1}$. Si l'on fixe une coordonnée λ_j dans l'ensemble $\{0, 1/k, ..., (k-1)/k, 1\}$ et que l'on pose $\lambda = (\lambda', \lambda_j)$, on obtient un polynôme $q_j(\lambda') = q(\lambda)$ qui dépend de $N-1$ variables indépendantes (car on a la relation $\sum_{j=1}^{N+1} \lambda_j = 1$) et qui est nul sur la section du treillis Σ_k correspondant à la valeur fixée de λ_j. Comme cette section est aussi le treillis d'ordre k d'un $(N-1)$-simplexe dans l'hyperplan λ_j fixé, on peut appliquer l'hypothèse de récurrence et en déduire que $q_j = 0$. Autrement dit, le facteur $\lambda_j(\lambda_j - 1/k)\cdots(\lambda_j-(k-1)/k)(\lambda_j-1)$ divise q, ce qui est une contradiction avec le fait que le degré de $q(\lambda)$ est inférieur ou égal à k, sauf si $q = 0$, ce qui est le résultat désiré. □

Lemme 4.3.4 *Soit K et K' deux N-simplexes ayant une face commune $\Gamma = \partial K \cap \partial K'$. Soit un entier $k \geq 1$. Alors, leurs treillis d'ordre k, Σ_k et Σ'_k coïncident sur cette face Γ. De plus, étant donné p_K et $p_{K'}$ deux polynômes de $\mathbb{P}_k$, la fonction v définie par*

$$v(x) = \begin{cases} p_K(x) & \text{si } x \in K \\ p_{K'}(x) & \text{si } x \in K' \end{cases}$$

est continue sur $K \cup K'$, si et seulement si p_K et $p_{K'}$ ont des valeurs qui coïncident aux points du treillis sur la face commune Γ.

Démonstration. Il est clair que la restriction à une face de K de son treillis d'ordre

Σ_k est aussi un treillis d'ordre k dans l'hyperplan contenant cette face, qui ne dépend que des sommets de cette face. Par conséquent, les treillis Σ_k et Σ'_k coïncident sur leur face commune Γ. Si les polynômes p_K et $p_{K'}$ coïncident aux points de $\Sigma_k \cap \Gamma$, alors par application du Lemme 4.3.3 ils sont égaux sur Γ, ce qui prouve la continuité de v. □

En pratique, on utilise surtout des polynômes de degré 1 ou 2. Dans ce cas on a les caractérisations suivantes de $\mathbb{P}_1$ et $\mathbb{P}_2$ dans un N-simplexe K.

Exercice 4.3.2 Soit K un N-simplexe de sommets $(a_j)_{1\leq j\leq N+1}$. Montrer que tout polynôme $p \in \mathbb{P}_1$ se met sous la forme

$$p(x) = \sum_{j=1}^{N+1} p(a_j)\lambda_j(x),$$

où les $(\lambda_j(x))_{1\leq j\leq N+1}$ sont les coordonnées barycentriques de $x \in \mathbb{R}^N$.

Exercice 4.3.3 Soit K un N-simplexe de sommets $(a_j)_{1\leq j\leq N+1}$. On définit les points milieux $(a_{jj'})_{1\leq j<j'\leq N+1}$ des arêtes de K par leur coordonnées barycentriques

$$\lambda_j(a_{jj'}) = \lambda_{j'}(a_{jj'}) = \frac{1}{2}, \quad \lambda_l(a_{jj'}) = 0 \text{ pour } l \neq j, j'.$$

Vérifier que Σ_2 est précisément constitué des sommets et des points milieux des arêtes et que tout polynôme $p \in \mathbb{P}_2$ se met sous la forme

$$p(x) = \sum_{j=1}^{N+1} p(a_j)\lambda_j(x)\left(2\lambda_j(x) - 1\right) + \sum_{1\le j<j'\le N+1} 4p(a_{jj'})\lambda_j(x)\lambda_{j'}(x),$$

où les $(\lambda_j(x))_{1\le j\le N+1}$ sont les coordonnées barycentriques de $x \in \mathbb{R}^N$.

Nous avons maintenant tous les outils pour définir la méthode des éléments finis $\mathbb{P}_k$.

Définition 4.3.5 *Étant donné un maillage $\mathcal{T}_h$ d'un ouvert connexe polyédrique Ω, la méthode des éléments finis $\mathbb{P}_k$, ou* **éléments finis triangulaires de Lagrange d'ordre** *k, associée à ce maillage, est définie par l'espace discret*

$$V_h = \left\{ v \in C(\overline{\Omega}) \text{ tel que } v\,|_{K_i} \in \mathbb{P}_k \text{ pour tout } K_i \in \mathcal{T}_h \right\}. \tag{4.40}$$

On appelle **noeuds des degrés de liberté** *l'ensemble des points $(\hat{a}_i)_{1\le i\le n_{dl}}$ des treillis d'ordre k de chacun des N-simplexes $K_i \in \mathcal{T}_h$. On ne compte qu'une seule fois les points qui coïncident et n_{dl} est le nombre de degrés de liberté de la méthode des éléments finis $\mathbb{P}_k$. On appelle* **degrés de liberté** *d'une fonction $v \in V_h$ l'ensemble des valeurs de v en ces noeuds $(\hat{a}_i)_{1\le i\le n_{dl}}$. On définit aussi le sous-espace V_{0h} par*

$$V_{0h} = \{ v \in V_h \text{ tel que } v = 0 \text{ sur } \partial\Omega \}. \tag{4.41}$$

Lorsque $k = 1$ les noeuds des degrés de liberté coïncident avec les sommets du maillage. Lorsque $k = 2$ ces noeuds sont constitués d'une part des sommets du maillage et d'autre part des points milieux des arêtes reliant deux sommets.

Remarque 4.3.6 L'appellation "éléments finis de Lagrange" correspond aux éléments finis dont les degrés de liberté sont des valeurs ponctuelles des fonctions de l'espace V_h. On peut définir d'autres types d'éléments finis, par exemple les éléments finis de Hermite pour lesquels les degrés de liberté sont les valeurs ponctuelles de la fonction et de ses dérivées. •

Proposition 4.3.7 *L'espace V_h, défini par (4.40), est un sous-espace de $H^1(\Omega)$ dont la dimension est finie, égale au nombre de degrés de liberté. De plus, il existe une base de V_h $(\phi_i)_{1\leq i\leq n_{dl}}$ définie par*

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

telle que

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i)\phi_i(x).$$

Démonstration. Les éléments de V_h, étant réguliers sur chaque maille K_i et continus sur $\overline{\Omega}$, appartiennent à $H^1(\Omega)$. Grâce au Lemme 4.3.4 les éléments de V_h sont exactement obtenus en assemblant sur chaque $K_i \in \mathcal{T}_h$ des polynômes de $\mathbb{P}_k$ qui

coïncident sur les degrés de liberté des faces (ce qui prouve au passage que V_h n'est pas réduit aux seules fonctions constantes). Enfin, en assemblant les bases $(\psi_j)_{1\leq j\leq n_k}$ de $\mathbb{P}_k$ sur chaque maille K_i (fournies par le Lemme 4.3.3) on obtient la base annoncée $(\phi_i)_{1\leq i\leq n_{dl}}$ de V_h. □

Remarque 4.3.8 On obtient un résultat semblable pour le sous-espace V_{0h}, défini par (4.41), qui est un sous-espace de $H_0^1(\Omega)$ de dimension finie égale au nombre de degrés de liberté intérieurs (on ne compte pas les noeuds sur le bord $\partial\Omega$). •

Décrivons la **résolution pratique** du problème de Dirichlet (4.36) par la méthode des éléments finis $\mathbb{P}_k$. La formulation variationnelle (4.2) de l'approximation interne devient ici :

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_\Omega \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx \;\forall\, v_h \in V_{0h}. \tag{4.42}$$

On décompose u_h sur la base des $(\phi_j)_{1\leq j\leq n_{dl}}$ et on prend $v_h = \phi_i$ ce qui donne

$$\sum_{j=1}^{n_{dl}} u_h(\hat{a}_j) \int_\Omega \nabla\phi_j \cdot \nabla\phi_i \, dx = \int_\Omega f\phi_i \, dx.$$

En notant $U_h = (u_h(\hat{a}_j))_{1\leq j\leq n_{dl}}$, $b_h = \left(\int_\Omega f\phi_i \, dx\right)_{1\leq i\leq n_{dl}}$, et en introduisant la **matrice de rigidité**

$$K_h = \left(\int \nabla\phi_j \cdot \nabla\phi_i \, dx \right)$$

$$\mathcal{K}_h = \left(\int_\Omega \nabla\phi_j \cdot \nabla\phi_i \, dx \right)_{1 \le i,j \le n_{dl}},$$

la formulation variationnelle dans V_{0h} revient à résoudre dans $\mathbb{R}^{n_{dl}}$ le système linéaire

$$\mathcal{K}_h U_h = b_h.$$

Comme les fonctions de base ϕ_j ont un "petit" support autour du noeud $\hat{a}_i$ (voir la Figure 4.10), l'intersection des supports de ϕ_j et ϕ_i est souvent vide et la plupart des coefficients de $\mathcal{K}_h$ sont nuls. On dit que la matrice $\mathcal{K}_h$ est **creuse**.

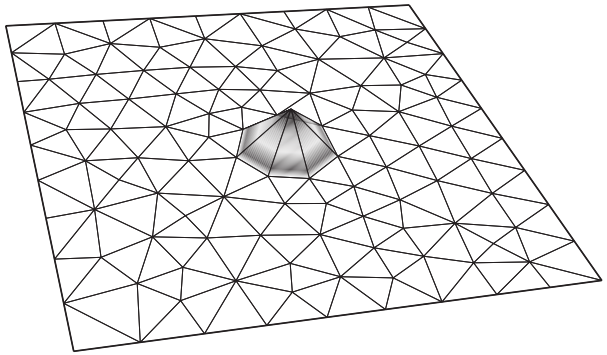


FIGURE 4.10 – Fonction de base $\mathbb{P}_1$ en dimension $N = 2$.

Pour calculer les coefficients de $\mathcal{K}_h$, on peut utiliser la formule d'intégration exacte suivante. On note $(\lambda_i(x))_{1 \le i \le N+1}$ les coordonnées barycentriques du point

courant x d'un N-simplexe K. Pour tout $\alpha_1, ..., \alpha_{N+1} \in \mathbb{N}$, on a

$$\int_K \lambda_1(x)^{\alpha_1} \cdots \lambda_{N+1}(x)^{\alpha_{N+1}}\, dx = \text{Volume}(K) \frac{\alpha_1! \cdots \alpha_{N+1}!\, N!}{(\alpha_1 + ... + \alpha_{N+1} + N)!}. \tag{4.43}$$

Pour calculer le second membre b_h (et même éventuellement la matrice $\mathcal{K}_h$), on utilise des **formules de quadrature** (ou formules d'intégration numérique) qui donnent une approximation des intégrales sur chaque N-simplexe $K_i \in \mathcal{T}_h$. Par exemple, si K est un N-simplexe de sommets $(a_i)_{1 \le i \le N+1}$, les formules suivantes généralisent les formules en dimension 1, dites du "point milieu" et des "trapèzes" :

$$\int_K \psi(x)\, dx \approx \text{Volume}(K) \psi(a_0), \tag{4.44}$$

avec $a_0 = (N+1)^{-1} \sum_{i=1}^{N+1} a_i$, le barycentre de K, et

$$\int_K \psi(x)\, dx \approx \frac{\text{Volume}(K)}{N+1} \sum_{i=1}^{N+1} \psi(a_i). \tag{4.45}$$

Comme le montre l'Exercice 4.3.6, ces formules sont exactes pour des fonctions affines et sont donc approchées à l'ordre 2 en h pour des fonctions régulières.

La construction de la matrice $\mathcal{K}_h$ est appelée **assemblage de la matrice**. La mise en oeuvre informatique de cette étape du calcul peut être assez compliquée,

mais son coût en terme de temps de calcul est faible. Ce n'est pas le cas de la résolution du système linéaire $\mathcal{K}_h U_h = b_h$ qui est l'étape la **plus coûteuse** de la méthode en temps de calcul (et en place mémoire). En particulier, les calculs tridimensionnels sont encore très chers de nos jours dès que l'on utilise des maillages fins. L'Exercice 4.3.11 permet de s'en rendre compte. Heureusement, la matrice de rigidité $\mathcal{K}_h$ est **creuse** (c'est-à-dire que la plupart de ses éléments sont nuls), ce qui permet de minimiser les calculs (pour plus de détails voir les algorithmes de résolution de systèmes linéaires dans la Section 4.4 de l'annexe). Rappelons que la matrice $\mathcal{K}_h$ est nécessairement inversible par application du Lemme 4.1.1 et qu'elle est symétrique.

Exercice 4.3.5 Démontrer la formule (4.43) en dimension $N = 2$.

Exercice 4.3.6 Montrer que les formules (4.44) et (4.45) sont exactes pour $\psi \in \mathbb{P}_1$.

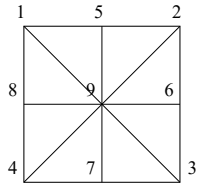


FIGURE 4.11 – Exemple de maillage et de numérotation des noeuds.

Exercice 4.3.9 On considère le carré $\Omega =]-1,+1[^2$ maillé suivant la Figure 4.11. Calculer la matrice de rigidité $\mathcal{K}_h$ des éléments finis $\mathbb{P}_1$ appliqués au Laplacien avec condition aux limites de Neumann (on utilisera les symétries du maillage).

Exercice 4.3.10 Appliquer la méthode des éléments finis $\mathbb{P}_1$ au problème de Dirichlet (4.36) dans le carré $\Omega =]0,1[^2$ avec le maillage triangulaire uniforme de la Figure 4.12. Montrer que la matrice de rigidité $\mathcal{K}_h$ est la même matrice que celle que l'on obtiendrait par application de la méthode des différences finies (à un facteur multiplicatif h^2 près), mais que le second membre b_h est différent.

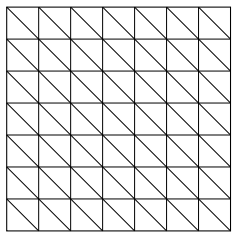


FIGURE 4.12 – Maillage triangulaire uniforme d'un carré.

Exercice 4.3.11 On reprend les notations de l'Exercice 4.3.10. On note n le nombre de points du maillage sur un coté du carré (supposé être le même pour chaque coté). On numérote "ligne par ligne" les noeuds du maillage (ou les degrés de liberté). Montrer

On numérote ligne par ligne les noeuds du maillage (ou les degrés de liberté). Montrer que la matrice de rigidité $\mathcal{K}_h$ des éléments finis $\mathbb{P}_1$ est de taille de l'ordre de n^2 et de largeur de bande de l'ordre de $2n$ (pour n grand).

Montrer que la même méthode et le même type de maillage pour le cube $\Omega =]0,1[^3$ conduisent à une matrice de taille de l'ordre de n^3 et de largeur de bande de l'ordre de $2n^2$ (où n est le nombre de noeuds le long d'une arête du cube Ω).

L'exercice suivant montre que la méthode des éléments finis $\mathbb{P}_1$ vérifie le principe du maximum.

Exercice 4.3.12 On dit qu'une matrice carrée réelle $B = (b_{ij})_{1\leq i,j\leq n}$ est une M-matrice si, pour tout i,

$$b_{ii} > 0, \quad \sum_{k=1}^{n} b_{ik} > 0, \quad b_{ij} \leq 0 \quad \forall\, j \neq i.$$

Montrer que toute M-matrice est inversible et que tous les coefficients de son inverse sont positifs ou nuls.

Exercice 4.3.13 On se place en dimension $N = 2$. Soit u_h la solution approchée du problème de Dirichlet (4.36) obtenue par la méthode des éléments finis $\mathbb{P}_1$. On suppose que tous les angles des triangles $K_i \in \mathcal{T}_h$ sont inférieurs ou égaux à $\pi/2$. Montrer que $u_h(x) \geq 0$ dans Ω si $f(x) \geq 0$ dans Ω. Indication : on montrera que, pour tout $\epsilon > 0$, $\mathcal{K}_h + \epsilon\,\mathrm{Id}$ est une M-matrice, où $\mathcal{K}_h$ est la matrice de rigidité.

4.3.2 Convergence et estimation d'erreur

Nous démontrons la convergence des méthodes d'éléments finis $\mathbb{P}_k$ pour le problème de Dirichlet (4.36). Insistons sur le fait qu'il s'agit seulement d'un problème modèle, et que ces méthodes convergent pour d'autres problèmes, comme celui de Neumann. Nous allons avoir besoin d'hypothèses géométriques sur la qualité du maillage. Pour tout N-simplexe K on introduit deux paramètres géométriques : le **diamètre** $\mathrm{diam}(K)$ et la **rondeur** $\rho(K)$, définie comme le diamètre de la plus grande boule contenue dans K,

$$\mathrm{diam}(K) = \max_{x,y\in K} \|x-y\|, \quad \rho(K) = \max_{B_r\subset K}(2r).$$

Bien sûr, on a toujours $\mathrm{diam}(K)/\rho(K) > 1$. Ce rapport est d'autant plus grand que K est "aplati" : il mesure en quelque sorte la tendance à la dégénérescence de K. En pratique, comme en théorie, il faut éviter d'utiliser des N-simplexes K trop aplatis.

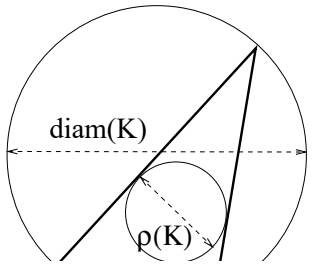


FIGURE 4.17 – Diamètre $\mathrm{diam}(K)$ et rondeur $\rho(K)$ d'un triangle K.

Définition 4.3.11 *Soit $(\mathcal{T}_h)_{h>0}$ une suite de maillages de Ω. On dit qu'il s'agit d'une suite de* **maillages réguliers** *si*

1. *la suite $h = \max_{K_i \in \mathcal{T}_h} \mathrm{diam}(K_i)$ tend vers 0,*
2. *il existe une constante C telle que, pour tout $h > 0$ et tout $K \in \mathcal{T}_h$,*

$$\frac{\mathrm{diam}(K)}{\rho(K)} \leq C. \tag{4.46}$$

Remarque 4.3.12 En dimension $N = 2$ la condition (4.46) est équivalente à la condition suivante sur les angles du triangle K : il existe un angle minimum $\theta_0 > 0$ qui minore (uniformément en h) tous les angles de tout $K \in \mathcal{T}_h$. Insistons sur le fait que la condition (4.46) est tout aussi importante en pratique que pour l'analyse de convergence qui va suivre. •

Nous pouvons maintenant énoncer le résultat principal de cette sous-section qui affirme la convergence de la méthode des éléments finis $\mathbb{P}_k$ et qui donne une estimation de la vitesse de convergence si la solution est régulière.

Théorème 4.3.13 *Soit $(\mathcal{T}_h)_{h>0}$ une suite de maillages réguliers de Ω. Soit $u \in H_0^1(\Omega)$, la solution du problème de Dirichlet (4.36), et $u_h \in V_{0h}$, celle de son approximation interne (4.42) par la méthode des éléments finis $\mathbb{P}_k$. Alors la méthode des éléments finis $\mathbb{P}_k$ converge, c'est-à-dire que*

$$\lim_{h\to 0} \|u - u_h\|_{H^1(\Omega)} = 0. \tag{4.47}$$

De plus, si $u \in H^{k+1}(\Omega)$ et si $k+1 > N/2$, alors on a l'estimation d'erreur

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^k \|u\|_{H^{k+1}(\Omega)}, \tag{4.48}$$

où C est une constante indépendante de h et de u.

Remarque 4.3.14 Le Théorème 4.3.13 s'applique en fait à toute méthode d'éléments finis de type Lagrange (par exemple, les éléments finis rectangulaires de la Sous-section 4.3.3). En effet, le seul argument utilisé est la construction d'un opérateur d'interpolation basé sur la caractérisation des fonctions de V_h par leurs valeurs aux noeuds des degrés de liberté, ce qui est toujours possible pour des éléments finis de type Lagrange (voir la Remarque 4.3.6). Remarquons que, pour les cas physiquement pertinents $N = 2$ ou $N = 3$, la condition $k + 1 > N/2$ est toujours satisfaite dès que $k \geq 1$. •

La démonstration du Théorème 4.3.13 repose sur la définition suivante d'un

opérateur d'interpolation r_h et sur le résultat d'interpolation de la Proposition 4.3.16. Rappelons que nous avons noté $(\hat{a}_i)_{1\leq i\leq n_{dl}}$ la famille des noeuds des degrés de liberté et $(\phi_i)_{1\leq i\leq n_{dl}}$ la base de V_{0h} de la méthode des éléments finis $\mathbb{P}_k$ (voir la Proposition 4.3.7). Pour toute fonction continue v, on définit son interpolée

$$r_h v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i)\phi_i(x). \tag{4.49}$$

La différence principale avec l'étude faite en dimension $N = 1$ est que, les fonctions de $H^1(\Omega)$ n'étant pas continues lorsque $N \geq 2$, l'opérateur d'interpolation r_h n'est pas défini sur $H^1(\Omega)$ (les valeurs ponctuelles d'une fonction de $H^1(\Omega)$ n'ont a priori pas de sens). Néanmoins, et c'est la raison de l'hypothèse $k+1 > N/2$, r_h **est bien défini sur** $H^{k+1}(\Omega)$ car les fonctions de $H^{k+1}(\Omega)$ sont continues ($H^{k+1}(\Omega) \subset C(\overline{\Omega})$ d'après le Théorème 2.3.25).

Proposition 4.3.16 *Soit $(\mathcal{T}_h)_{h>0}$ une suite de maillages réguliers de Ω. On suppose que $k+1 > N/2$. Alors, pour tout $v \in H^{k+1}(\Omega)$ l'interpolée $r_h v$ est bien définie, et il existe une constante C, indépendante de h et de v, telle que*

$$\|v - r_h v\|_{H^1(\Omega)} \leq C h^k \|v\|_{H^{k+1}(\Omega)}. \tag{4.50}$$

En admettant la Proposition 4.3.16 nous pouvons conclure quant à la convergence de la méthode des éléments finis $\mathbb{P}_k$.

Démonstration du Théorème 4.3.13. On applique le cadre abstrait de la Sous-section 4.1.2. Pour démontrer (4.47) on utilise le Lemme 4.1.3 avec $\mathcal{V} = C_c^\infty(\Omega)$ qui est bien dense dans $H_0^1(\Omega)$. Comme $C_c^\infty(\Omega) \subset H^{k+1}(\Omega)$, l'estimation (4.50) de la Proposition 4.3.16 permet de vérifier l'hypothèse (4.5) du Lemme 4.1.3 (pour des fonctions régulières on n'a pas besoin de la condition $k+1 > N/2$ dans la Proposition 4.3.16).

Pour obtenir l'estimation d'erreur (4.48) on utilise le Lemme de Céa 4.1.2 qui nous dit que

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_{0h}} \|u - v_h\|_{H^1(\Omega)} \leq C\|u - r_h u\|_{H^1(\Omega)},$$

si $r_h u$ appartient bien à $H^1(\Omega)$. Par application de la Proposition 4.3.16 à u on obtient (4.48). □

4.3.3 Éléments finis rectangulaires

Si le domaine Ω est de type rectangulaire (c'est-à-dire que Ω est un ouvert polyédrique dont les faces sont perpendiculaires aux axes), on peut le mailler par des rectangles (voir la Figure 4.20) et utiliser une méthode d'éléments finis adaptée. Nous allons définir des éléments finis de type Lagrange (c'est-à-dire dont les degrés de liberté sont des valeurs ponctuelles de fonctions), dits éléments finis $\mathbb{Q}_k$. Commençons par définir un N-rectangle K de $\mathbb{R}^N$ comme le pavé (non-dégénéré)

$\prod_{i=1}^{N}]l_i, L_i[$ avec $-\infty < l_i < L_i < +\infty$. On note $(a_j)_{1\leq j\leq 2^N}$ les sommets de K.

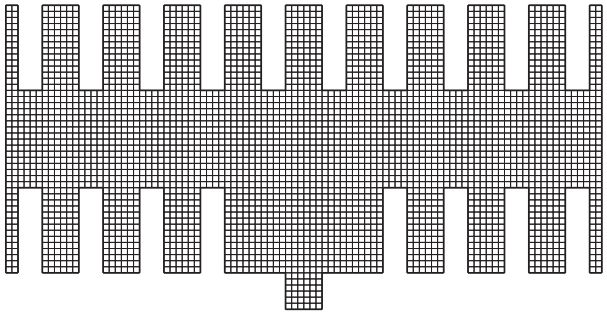


FIGURE 4.20 – Exemple de maillage rectangulaire en dimension $N = 2$.

Définition 4.3.21 *Soit Ω un ouvert connexe polyédrique de $\mathbb{R}^N$. Un* **maillage rectangulaire** *de $\overline{\Omega}$ est un ensemble $\mathcal{T}_h$ de N-rectangles (non dégénérés) $(K_i)_{1\leq i\leq n}$ qui vérifient*

1. *$K_i \subset \overline{\Omega}$ et $\overline{\Omega} = \cup_{i=1}^{n} K_i$,*
2. *l'intersection $K_i \cap K_j$ de deux N-rectangles distincts est un m-rectangle, avec $0 \leq m \leq N-1$, dont tous les sommets sont aussi des sommets de K_i et K_j. (En dimension $N = 2$, l'intersection de deux rectangles est soit vide, soit un sommet commun, soit une face commune entière.)*

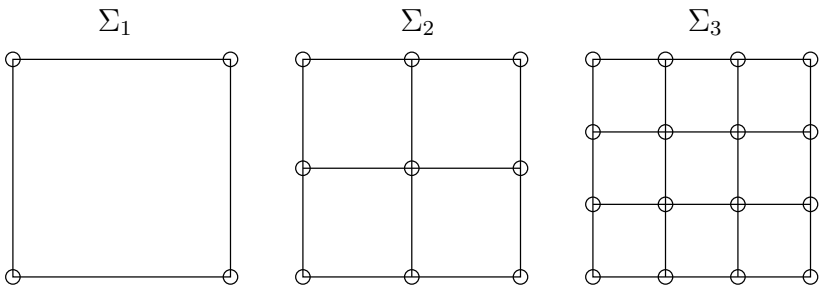


FIGURE 4.21 – Treillis d'ordre 1, 2, et 3 pour un rectangle (les ronds représentent les points du treillis).

Les **sommets** *ou* **noeuds** *du maillage $\mathcal{T}_h$ sont les sommets des N-rectangles K_i qui le composent. Par convention, le paramètre h désigne le maximum des diamètres des N-rectangles K_i.*

Nous définissons l'ensemble $\mathbb{Q}_k$ des polynômes à coefficients réels de $\mathbb{R}^N$ dans $\mathbb{R}$ de degré inférieur ou égal à k **par rapport à chaque variable**, c'est-à-dire que tout $p \in \mathbb{Q}_k$ s'écrit sous la forme

$$p(x) = \sum_{0 \leq i_1 \leq k, \ldots, 0 \leq i_N \leq k} \alpha_{i_1,\ldots,i_N} x_1^{i_1} \cdots x_N^{i_N} \text{ avec } x = (x_1, ..., x_N).$$

Remarquons que le degré total de p peut être supérieur à k, ce qui différencie l'espace

Remarquons que le degré total de p peut être supérieur à k, ce qui différencie l'espace $\mathbb{Q}_k$ de $\mathbb{P}_k$.

Pour tout entier $k \geq 1$ on définit le **treillis d'ordre** k du N-rectangle K comme l'ensemble

$$\Sigma_k = \left\{ x \in K \text{ tel que } \frac{x_j - l_j}{L_j - l_j} \in \{0, \frac{1}{k}, ..., \frac{k-1}{k}, 1\} \text{ pour } 1 \leq j \leq N \right\}. \qquad (4.51)$$

Pour $k = 1$ il s'agit de l'ensemble des sommets de K, et pour $k = 2$ et $N = 2$ des sommets, des points milieux des arêtes reliant deux sommets, et du barycentre (voir la Figure 4.21).

Le treillis Σ_k d'un N-rectangle K est **unisolvant** pour $\mathbb{Q}_k$, c'est-à-dire qu'il permet de caractériser tous les polynômes de $\mathbb{Q}_k$.

Lemme 4.3.22 *Soit K un N-rectangle. Soit un entier $k \geq 1$. Alors, tout polynôme de $\mathbb{Q}_k$ est déterminé de manière unique par ses valeurs aux points du treillis d'ordre k, Σ_k, défini par (4.51).*

Démonstration. On vérifie que le cardinal de Σ_k et la dimension de $\mathbb{Q}_k$ coïncident

$$\text{card}(\Sigma_k) = \dim(\mathbb{Q}_k) = (k+1)^N.$$

Comme l'application qui, à tout polynôme de $\mathbb{Q}_k$, fait correspondre ses valeurs sur le treillis Σ_k est linéaire, il suffit d'exhiber une base de $\mathbb{Q}_k$ dont les éléments valent

1 en un point du treillis et 0 ailleurs pour démontrer le résultat. Soit un point x^μ de Σ_k défini par

$$\frac{x_j^\mu - l_j}{L_j - l_j} = \frac{\mu_j}{k} \quad \text{avec } 0 \leq \mu_j \leq k, \ \forall\, j \in \{1, ..., N\}.$$

On définit le polynôme $p \in \mathbb{Q}_k$ par

$$p(x) = \prod_{j=1}^{N} \left(\prod_{\substack{i=0 \\ i \neq \mu_j}}^{k} \frac{k(x_j - l_j) - i(L_j - l_j)}{(\mu_j - i)(L_j - l_j)} \right) \qquad \text{avec } x = (x_1, ..., x_N).$$

On vérifie facilement que $p(x^\mu) = 1$ tandis que p s'annule sur tous les autres points de Σ_k, ce qui est le résultat désiré. □

Comme dans le cas triangulaire nous avons la condition suivante de continuité à travers une face (nous laissons la démonstration, tout à fait similaire à celle du Lemme 4.3.4, au lecteur).

Lemme 4.3.23 *Soit K et K' deux N-rectangles ayant une face commune $\Gamma = \partial K \cap \partial K'$. Soit un entier $k \geq 1$. Alors, leur treillis d'ordre k Σ_k et Σ'_k coïncident sur cette face Γ. De plus, étant donné p_K et $p_{K'}$ deux polynômes de $\mathbb{Q}_k$, la fonction v définie par*

$$v(x) = \begin{cases} p_K(x) & si\ x \in K \\ p_{K'}(x) & si\ x \in K' \end{cases}$$

est continue sur $K \cup K'$*, si et seulement si* p_K *et* $p_{K'}$ *ont des valeurs qui coïncident aux points du treillis sur la face commune* Γ*.*

En pratique, on utilise surtout les espaces $\mathbb{Q}_1$ et $\mathbb{Q}_2$. La Figure 4.22 montre une fonction de base $\mathbb{Q}_1$ en dimension $N = 2$ (on peut y vérifier que les fonctions de $\mathbb{Q}_1$ ne sont pas affines par morceaux comme celles de $\mathbb{P}_1$).

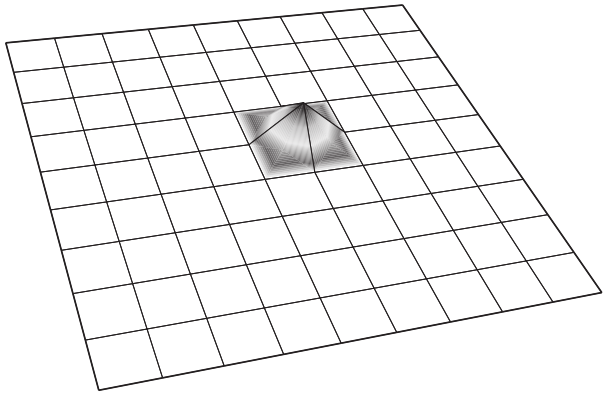


FIGURE 4.22 – Fonction de base $\mathbb{Q}_1$ en dimension $N = 2$.

Définition 4.3.25 *Étant donné un maillage rectangulaire $\mathcal{T}_h$ d'un ouvert Ω, la méthode des éléments finis $\mathbb{Q}_k$ est définie par l'espace discret*

$$V_h = \left\{ v \in C(\overline{\Omega}) \text{ tel que } v\,|_{K_i} \in \mathbb{Q}_k \text{ pour tout } K_i \in \mathcal{T}_h \right\}. \tag{4.52}$$

On appelle noeuds des **degrés de liberté** *l'ensemble des points $(\hat{a}_i)_{1\leq i\leq n_{dl}}$ des treillis d'ordre k de chacun des N-rectangles $K_i \in \mathcal{T}_h$.*

Comme dans le cas triangulaire, la Définition 4.3.25 a un sens grâce à la proposition suivante (dont nous laissons la démonstration au lecteur en guise d'exercice).

Proposition 4.3.26 *L'espace V_h, défini par (4.52), est un sous-espace de $H^1(\Omega)$ dont la dimension est le nombre de degrés de liberté n_{dl}. De plus, il existe une base de V_h $(\phi_i)_{1\leq i\leq n_{dl}}$ définie par*

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

telle que

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i)\phi_i(x).$$

Comme les éléments finis $\mathbb{Q}_k$ sont des éléments finis de type Lagrange, on peut démontrer les mêmes résultats de convergence que pour la méthode des éléments finis $\mathbb{P}_k$.

$\|\cdot\|_k$.

4.4 Résolution des systèmes linéaires

Cette section est consacrée à la résolution des systèmes linéaires. Pour plus de détails nous renvoyons aux ouvrages [2] et [9]. On appelle système linéaire le problème qui consiste à trouver la ou les solutions $x \in \mathbb{R}^n$ (si elle existe) de l'équation algébrique suivante

$$Ax = b, \tag{4.53}$$

où A appartient à l'ensemble $\mathcal{M}_n(\mathbb{R})$ des matrices réelles carrées d'ordre n, et $b \in \mathbb{R}^n$ est un vecteur appelé second membre. Nous allons voir deux types de méthodes de résolution de systèmes linéaires : celles dites directes, c'est-à-dire qui permettent de calculer la solution exacte en un nombre fini d'opérations, et celles dites **itératives**, c'est-à-dire qui calculent une suite de solutions approchées qui converge vers la solution exacte.

4.4.1 Rappels sur les normes matricielles

Nous commençons par rappeler la notion de **norme subordonnée** pour les matrices. Même si l'on considère des matrices réelles, il est nécessaire, pour des raisons techniques qui seront exposées à la Remarque 4.4.4, de les traiter comme des matrices complexes.

Définition 4.4.1 *Soit $\|\cdot\|$ une norme vectorielle sur $\mathbb{C}^n$. On lui associe une norme matricielle, dite subordonnée à cette norme vectorielle, définie par*

$$\|A\| = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Par abus de langage on note de la même façon les normes vectorielle et matricielle subordonnée. On vérifie aisément qu'une norme subordonnée ainsi définie est bien une norme matricielle et qu'elle vérifie le résultat suivant.

Lemme 4.4.2 *Soit $\|\cdot\|$ une norme matricielle subordonnée sur $\mathcal{M}_n(\mathbb{C})$.*

1. *Pour toute matrice A, la norme $\|A\|$ est aussi définie par*

$$\|A\| = \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\| = \sup_{x\in\mathbb{C}^n, \|x\|\leq 1} \|Ax\|.$$

2. *Il existe $x_A \in \mathbb{C}^n, x_A \neq 0$ tel que $\|A\| = \dfrac{\|Ax_A\|}{\|x_A\|}$.*
3. *La matrice identité vérifie $\|\mathrm{Id}\| = 1$.*
4. *Soient A et B deux matrices. On a $\|AB\| \leq \|A\|\,\|B\|$.*

On note $\|A\|_p$ la norme matricielle subordonnée à la norme vectorielle sur $\mathbb{C}^n$ définie pour $p \geq 1$ par $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, et pour $p = +\infty$ par $\|x\|_\infty = \max_{1\leq i\leq} |x_i|$. On peut calculer explicitement certaines de ces normes subordonnées

$\max_{1\leq i\leq n}|x_i|$. On peut calculer explicitement certaines de ces normes subordonnées. (Dans tout ce qui suit on note A^* la matrice adjointe de A.)

Exercice 4.4.1 Montrer que

1. $\|A\|_2 = \|A^*\|_2 =$ maximum des valeurs singulières de A,
2. $\|A\|_1 = \max_{1\leq j\leq n}\left(\sum_{i=1}^n |a_{ij}|\right)$,
3. $\|A\|_\infty = \max_{1\leq i\leq n}\left(\sum_{j=1}^n |a_{ij}|\right)$.

Remarque 4.4.4 Une matrice réelle peut être considérée soit comme une matrice de $\mathcal{M}_n(\mathbb{R})$, soit comme une matrice de $\mathcal{M}_n(\mathbb{C})$ car $\mathbb{R}\subset\mathbb{C}$. Si $\|\cdot\|_{\mathbb{C}}$ est une norme vectorielle dans $\mathbb{C}^n$, on peut définir sa restriction $\|\cdot\|_{\mathbb{R}}$ à $\mathbb{R}^n$ qui est aussi une norme vectorielle dans $\mathbb{R}^n$. Pour une matrice réelle $A\in\mathcal{M}_n(\mathbb{R})$, on peut donc définir deux normes matricielles subordonnées $\|A\|_{\mathbb{C}}$ et $\|A\|_{\mathbb{R}}$ par

$$\|A\|_{\mathbb{C}} = \sup_{x\in\mathbb{C}^n, x\neq 0}\frac{\|Ax\|_{\mathbb{C}}}{\|x\|_{\mathbb{C}}} \text{ et } \|A\|_{\mathbb{R}} = \sup_{x\in\mathbb{R}^n, x\neq 0}\frac{\|Ax\|_{\mathbb{R}}}{\|x\|_{\mathbb{R}}}.$$

A priori ces deux définitions peuvent être distinctes. Grâce aux formules explicites de l'Exercice 4.4.1, on sait qu'elles coïncident si $\|x\|_{\mathbb{C}}$ est une des normes $\|x\|_1$, $\|x\|_2$, ou $\|x\|_\infty$. Cependant, pour d'autres normes vectorielles on peut avoir $\|A\|_{\mathbb{C}} > \|A\|_{\mathbb{R}}$. Par ailleurs, dans la preuve de la Proposition 4.4.7 on a besoin de la définition sur $\mathbb{C}$ de la norme subordonnée même si la matrice est réelle. C'est pourquoi on utilise $\mathbb{C}$ dans la Définition 4.4.1 de la norme subordonnée. •

Définition 4.4.5 *Soit A une matrice dans $\mathcal{M}_n(\mathbb{C})$. On appelle rayon spectral de A, et on note $\rho(A)$, le maximum des modules des valeurs propres de A.*

Proposition 4.4.7 *Soit $\|\cdot\|$ une norme subordonnée sur $\mathcal{M}_n(\mathbb{C})$. On a*

$$\rho(A) \leq \|A\|.$$

Réciproquement, pour toute matrice A et pour tout réel $\epsilon > 0$, il existe une norme subordonnée $\|\cdot\|$ (qui dépend de A et ϵ) telle que

$$\|A\| \leq \rho(A) + \epsilon. \tag{4.54}$$

Lemme 4.4.8 *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. Les quatre conditions suivantes sont équivalentes*

1. $\lim_{i\to+\infty} A^i = 0$,
2. $\lim_{i\to+\infty} A^i x = 0$ *pour tout vecteur* $x \in \mathbb{C}^n$,
3. $\rho(A) < 1$,
4. *il existe au moins une norme matricielle subordonnée telle que* $\|A\| < 1$.

4.4.2 Conditionnement et stabilité

Avant de décrire les algorithmes de résolution de systèmes linéaires, il nous faut évoquer les problèmes de précision et de stabilité dus aux erreurs d'arrondi. En effet, dans un ordinateur il n'y a pas de calculs exacts, et la précision est limitée à cause du

nombre de bits utilisés pour représenter les nombres réels : d'habitude 32 ou 64 bits (ce qui fait à peu près 8 ou 16 chiffres significatifs). Il faut donc faire très attention aux inévitables erreurs d'arrondi et à leur propagation au cours d'un calcul. Les méthodes numériques de résolution de systèmes linéaires qui n'amplifient pas ces erreurs sont dites stables. En pratique, on utilisera donc des algorithmes qui sont à la fois **efficaces et stables**. Cette amplification des erreurs dépend de la matrice considérée. Pour quantifier ce phénomène, on introduit la notion de conditionnement d'une matrice.

Définition 4.4.9 *Soit une norme matricielle subordonnée $\|A\|$. On appelle conditionnement d'une matrice $A \in \mathcal{M}_n(\mathbb{C})$, relatif à cette norme, la valeur définie par*

$$\mathrm{cond}(A) = \|A\|.\|A^{-1}\|.$$

Cette notion de conditionnement va permettre de mesurer l'amplification des erreurs des données (second membre ou matrice) au résultat.

Proposition 4.4.10 *Soit A une matrice inversible. Soit $b \neq 0$ un vecteur non nul.*

1. *Soit x et $x + \delta x$ les solutions respectives des systèmes*

$$Ax = b \ , \ et \ A(x + \delta x) = b + \delta b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}. \tag{4.55}$$

2. Soit x et $x + \delta x$ les solutions respectives des systèmes

$$Ax = b \ , \ et \ (A + \delta A)(x + \delta x) = b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}. \tag{4.56}$$

De plus, ces inégalités sont optimales.

Remarque 4.4.11 On dira qu'une matrice est bien conditionnée si son conditionnement est proche de 1 (sa valeur minimale) et qu'elle est mal conditionnée si son conditionnement est grand. A cause des résultats de la Proposition 4.4.10, en pratique il faudra faire attention aux erreurs d'arrondi si on résout un système linéaire pour une matrice mal conditionnée. •

Démonstration. Pour montrer le premier résultat, on remarque que $A\delta x = \delta b$, et donc $\|\delta x\| \leq \|A^{-1}\|.\|\delta b\|$. Or, on a aussi $\|b\| \leq \|A\|\|x\|$, ce qui donne (4.55). Cette inégalité est optimale au sens suivant : pour toute matrice A, il existe δb et x (qui dépendent de A) tels que (4.55) est en fait une égalité. En effet, d'après une propriété des normes matricielles subordonnées (voir le Lemme 4.4.2) il existe x tel que $\|b\| = \|A\|\|x\|$ et il existe δb tel que $\|\delta x\| = \|A^{-1}\|\|\delta b\|$.

Pour obtenir (4.56) on remarque que $A\delta x + \delta A(x+\delta x) = 0$, et donc $\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x+\delta x\|$, ce qui implique (4.56). Pour en démontrer l'optimalité, on va montrer que pour toute matrice A il existe une perturbation δA et un second membre b pour lesquels il y a égalité. Grâce au Lemme 4.4.2 il existe $y \neq 0$ tel que $\|A^{-1}y\| = \|A^{-1}\|\|y\|$. Soit ϵ un scalaire non nul. On pose $\delta A = \epsilon Id$ et $b = (A+\delta A)y$. On vérifie alors que $y = y + \delta x$ et $\delta x = -\epsilon A^{-1}y$, et comme $\|\delta A\| = |\epsilon|$ on obtient l'égalité dans (4.56). □

Les conditionnements les plus utilisés en pratique sont ceux associés aux normes $\|A\|_p$ avec $p = 1, 2, +\infty$.

Exercice 4.4.2 Soit une matrice $A \in \mathcal{M}_n(\mathbb{C})$. Vérifier que

1. $\mathrm{cond}(A) = \mathrm{cond}(A^{-1}) \geq 1,\ \mathrm{cond}(\alpha A) = \mathrm{cond}(A)\ \forall \alpha \neq 0$,
2. pour une matrice quelconque, $\mathrm{cond}_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$, où $\mu_1(A), \mu_n(A)$ sont respectivement la plus petite et la plus grande valeur singulière de A,
3. pour une matrice normale, $\mathrm{cond}_2(A) = \frac{|\lambda_n(A)|}{|\lambda_1(A)|}$, où $|\lambda_1(A)|, |\lambda_n(A)|$ sont respectivement la plus petite et la plus grande valeur propre en module de A,
4. pour toute matrice unitaire U, $\mathrm{cond}_2(U) = 1$,
5. pour toute matrice unitaire U, $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(A)$.

4.4.3 Méthodes directes

Méthode d'élimination de Gauss

L'idée principale de cette méthode est de se ramener à la résolution d'un système linéaire dont la matrice est triangulaire. En effet, la résolution d'un système linéaire, $Tx = b$, où la matrice T est triangulaire et inversible, est très facile par simple substitution récursive. On appelle ce procédé **remontée** dans le cas d'une matrice triangulaire supérieure et **descente** dans le cas d'une matrice triangulaire inférieure. Remarquons que l'on résout ainsi le système $Tx = b$ sans inverser la matrice T. De la même manière, la méthode d'élimination de Gauss va résoudre le système $Ax = b$ sans calculer l'inverse de la matrice A.

La méthode d'élimination de Gauss se décompose en trois étapes :

(i) élimination : calcul d'une matrice M inversible telle que $MA = T$ soit triangulaire supérieure,

(ii) mise à jour du second membre : calcul simultané de Mb,

(iii) substitution : résolution du système triangulaire $Tx = Mb$ par simple remontée.

L'existence d'une telle matrice M est garantie par le résultat suivant dont on va donner une démonstration constructive qui n'est rien d'autre que la méthode d'élimination de Gauss.

Proposition 4.4.13 *Soit A une matrice carrée (inversible ou non). Il existe au*

moins une matrice inversible M telle que la matrice $T = MA$ soit triangulaire supérieure.

Démonstration. Le principe est de construire une suite de matrices A^k, $1 \leq k \leq n$, dont les $(k-1)$ premières colonnes sont remplies de zéros sous la diagonale. Par modifications successives, on passe de $A^1 = A$ à $A^n = T$ qui est triangulaire supérieure. On note $\left(a_{ij}^k\right)_{1 \leq i,j \leq n}$ les éléments de la matrice A^k, et on appelle pivot de A^k l'élément a_{kk}^k. Pour passer de la matrice A^k à la matrice A^{k+1}, on s'assure tout d'abord que le pivot a_{kk}^k n'est pas nul. S'il l'est, on permute la k-ème ligne avec une autre ligne pour amener en position de pivot un élément non nul. Puis on procède à l'élimination de tous les éléments de la k-ème colonne en dessous de la k-ème ligne en faisant des combinaisons linéaires de la ligne courante avec la k-ème ligne. □

Méthode de la factorisation LU

La méthode LU consiste à factoriser la matrice A en un produit de deux matrices triangulaires $A = LU$, où L est triangulaire inférieure (L pour "lower" en anglais) et U est triangulaire supérieure (U pour "upper" en anglais). Il s'agit en fait du même algorithme que celui de l'élimination de Gauss dans le cas particulier où **on ne pivote jamais**. Une fois établie la factorisation LU de A, la résolution du système linéaire $Ax = b$ est équivalente à la simple résolution de deux systèmes triangulaires $Ly = b$ puis $Ux = y$.

Proposition 4.4.15 *Soit une matrice $A = (a_{ij})_{1\leq i,j\leq n}$ d'ordre n telle que toutes les sous-matrices diagonales d'ordre k, définies par*

$$\Delta^k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$

soient inversibles. Il existe un unique couple de matrices (L, U), avec U triangulaire supérieure, et L triangulaire inférieure ayant une diagonale de 1, tel que

$$A = LU.$$

Calcul pratique de la factorisation LU. On peut calculer la factorisation LU (si elle existe) d'une matrice A par identification de A au produit LU. En posant $A = (a_{ij})_{1\leq i,j\leq n}$, et

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}, \; U = \begin{pmatrix} u_{1,1} & \dots & \dots & u_{1,n} \\ 0 & u_{2,2} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{n,n} \end{pmatrix},$$

comme L est triangulaire inférieure et U triangulaire supérieure, pour $1 \leq i, j \leq n$ il vient

$$\sum_{n} \qquad \sum_{min(i,j)}$$

$$a_{i,j} = \sum_{k=1} l_{i,k} u_{k,j} = \sum_{k=1}^{(i,j)} l_{i,k} u_{k,j}.$$

En identifiant par ordre croissant les colonnes de A on en déduit les colonnes de L et de U. Ainsi, après avoir calculé les $(j-1)$ premières colonnes de L et de U en fonction des $(j-1)$ premières colonnes de A, on lit la j-ème colonne de A

$$a_{i,j} = \sum_{k=1}^{i} l_{i,k} u_{k,j} \quad \Rightarrow \quad u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} u_{k,j} \qquad \text{pour } 1 \leq i \leq j,$$

$$a_{i,j} = \sum_{k=1}^{j} l_{i,k} u_{k,j} \quad \Rightarrow \quad l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} u_{k,j}}{u_{jj}} \quad \text{pour } j+1 \leq i \leq n.$$

On calcule donc les j premières composantes de la j-ème colonne de U et les $n-j$ dernières composantes de la j-ème colonne de L en fonction de leurs $(j-1)$ premières colonnes. On divise par le pivot u_{jj} qui doit donc être non nul !

Compte d'opérations. Pour mesurer l'efficacité de l'algorithme de la décomposition LU on compte le nombre d'opérations nécessaires à son accomplissement (qui sera proportionnel à son temps d'exécution sur un ordinateur). On ne calcule pas exactement ce nombre d'opérations, et on se contente du premier terme de son développement asymptotique lorsque la dimension n est grande. De plus, pour simplifier on ne compte que les multiplications et divisions (et pas les additions dont le nombre est en général du même ordre de grandeur).

– Élimination ou décomposition LU : le nombre d'opérations N_{op} est

$$N_{op} = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} (1 + \sum_{k=j+1}^{n} 1),$$

qui, au premier ordre, donne $N_{op} \approx n^3/3$.

– Substitution (ou remontée-descente sur les deux systèmes triangulaires) : le nombre d'opérations N_{op} est donné par la formule

$$N_{op} = 2 \sum_{j=1}^{n} j,$$

qui, au premier ordre, donne $N_{op} \approx n^2$.

Au total la résolution d'un système linéaire $Ax = b$ par la méthode de la factorisation LU demande $N_{op} \approx n^3/3$ opérations car n^2 est négligeable devant n^3 quand n est grand.

Méthode de Cholesky

C'est une méthode qui ne s'applique qu'aux matrices symétriques réelles, définies positives. Elle consiste à factoriser une matrice A sous la forme $A = BB^*$ où B est une matrice triangulaire inférieure (et B^* son adjointe ou transposée).

Proposition 4.4.19 *Soit A une matrice symétrique réelle, définie positive. Il existe une unique matrice réelle B triangulaire inférieure, telle que tous ses éléments diagonaux soient positifs, et qui vérifie*

$$A = BB^*.$$

Calcul pratique de la factorisation de Cholesky. En pratique, on calcule le facteur de Cholesky B par identification dans l'égalité $A = BB^*$. Soit $A = (a_{ij})_{1\leq i,j\leq n}$, $B = (b_{ij})_{1\leq i,j\leq n}$ avec $b_{ij} = 0$ si $i < j$. Pour $1 \leq i, j \leq n$, il vient

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{min(i,j)} b_{ik}b_{jk}.$$

En identifiant par ordre croissant les colonnes de A (ou ses lignes, ce qui revient au même puisque A est symétrique) on en déduit les colonnes de B. Ainsi, après avoir calculé les $(j-1)$ premières colonnes de B en fonction des $(j-1)$ premières colonnes de A, on lit la j-ème colonne de A en dessous de la diagonale

$$\begin{aligned} a_{jj} = \sum_{k=1}^{j} (b_{jk})^2 \quad &\Rightarrow \quad b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2} \\ a_{i,j} = \sum_{k=1}^{j} b_{jk}b_{i,k} \quad &\Rightarrow \quad b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{jk}b_{i,k}}{b_{jj}} \text{ pour } j+1 \leq i \leq n. \end{aligned}$$

On calcule donc la j-ème colonne de B en fonction de ses $(j-1)$ premières colonnes. A cause du théorème précédent, on est sûr que, si A est symétrique définie positive, les termes sous les racines carrées sont strictement positifs. Au contraire, si A n'est pas définie positive, on trouvera que $a_{jj} - \sum_{k=1}^{j-1}(b_{jk})^2 \leq 0$ pour un certain rang j, ce qui empêche de terminer l'algorithme.

Compte d'opérations. Pour mesurer l'efficacité de la méthode de Cholesky on compte le nombre d'opérations (uniquement les multiplications) nécessaires à son accomplissement. Le nombre de racines carrées est n qui est négligeable dans ce compte d'opérations.

- Factorisation de Cholesky : le nombre d'opérations N_{op} est

$$N_{op} = \sum_{j=1}^{n} \left((j-1) + \sum_{i=j+1}^{n} j \right),$$

  qui, au premier ordre, donne $N_{op} \approx n^3/6$.
- Substitution : il faut effectuer une remontée et une descente sur les systèmes triangulaire associés à B et B^*. Le nombre d'opérations est au premier ordre $N_{op} \approx n^2$.

La méthode de Cholesky est donc approximativement **deux fois plus rapide** que celle de Gauss pour une matrice symétrique définie positive.

Matrices bandes et matrices creuses

Lorsqu'une matrice a beaucoup de coefficients nuls, on dit qu'elle est **creuse**. Si les éléments non nuls sont répartis à proximité de la diagonale, on dit que la matrice a une structure **bande**. Pour ces deux types de matrices (qui apparaissent naturellement dans la méthode des éléments finis comme dans la plupart des autres méthodes), on peut améliorer le compte d'opérations et la taille de stockage nécessaire pour résoudre un système linéaire. Ce gain est très important en pratique.

Définition 4.4.20 *Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est dite matrice bande, de demie largeur de bande (hors diagonale) $p \in \mathbb{N}$ si ses éléments vérifient $a_{i,j} = 0$ pour $|i-j| > p$. La largeur de la bande est alors $2p+1$.*

L'intérêt des matrices bandes vient de la propriété suivante.

Exercice 4.4.4 Montrer que les factorisations LU et de Cholesky conservent la structure bande des matrices.

Remarque 4.4.21 Si les factorisations LU et de Cholesky préservent la structure bande des matrices, il n'en est pas de même de leur structure creuse. En général, si A est creuse (même à l'intérieur d'une bande), les facteurs L et U, ou B et B^* sont "pleins" (le contraire de creux) à l'intérieur de la même bande. •

L'exercice suivant permet de quantifier le gain qu'il y a à utiliser des matrices bandes.

Exercice 4.4.5 Montrer que, pour une matrice bande d'ordre n et de demie largeur de bande p, le compte d'opérations de la factorisation LU est $\mathcal{O}(np^2/3)$ et celui de la factorisation de Cholesky est $\mathcal{O}(np^2/6)$.

4.4.4 Méthodes itératives

Les méthodes itératives sont particulièrement intéressantes pour les très grandes matrices ou les matrices creuses. En effet, dans ce cas les méthodes directes peuvent avoir un coût de calcul et de stockage en mémoire prohibitif (se rappeler que la factorisation LU ou de Cholesky demande de l'ordre de n^3 opérations). Commençons par une classe très simple de méthodes itératives.

Définition 4.4.24 *Soit A une matrice inversible. On introduit une décomposition régulière de A (en anglais "splitting"), c'est-à-dire un couple de matrices (M, N) avec M inversible (et facile à inverser dans la pratique) tel que $A = M - N$. La méthode itérative basée sur le splitting (M, N) est définie par*

$$\left\{ \begin{array}{l} x_0 \text{ donné dans } \mathbb{R}^n, \\ Mx_{k+1} = Nx_k + b \quad \forall k \geq 1. \end{array} \right. \tag{4.60}$$

Si la suite de solutions approchées x_k converge vers une limite x quand k tend vers l'infini, alors, par passage à la limite dans la relation de récurrence (4.60), on obtient

$$(M - N)x = Ax = b$$

$$(M - N)x = Ax = 0.$$

Par conséquent, si la suite de solutions approchées converge, sa limite est forcément la solution du système linéaire.

D'un point de vue pratique, il faut savoir quand on peut arrêter les itérations, c'est-à-dire à quel moment x_k est suffisamment proche de la solution inconnue x. Comme on ne connaît pas x, on ne peut pas décider d'arrêter le calcul dès que $\|x - x_k\| \leq \epsilon$ où ϵ est la précision désirée. Par contre on connaît Ax (qui vaut b), et un critère d'arrêt fréquemment utilisé est $\|b - Ax_k\| \leq \epsilon$. Cependant, si la norme de A^{-1} est grande ce critère peut être trompeur car

$$\|x - x_k\| \leq \|A^{-1}\| \, \|b - Ax_k\| \leq \epsilon \|A^{-1}\|$$

qui peut ne pas être petit.

Définition 4.4.25 *On dit qu'une méthode itérative est convergente si,* **quel que soit le choix du vecteur initial** $x_0 \in \mathbb{R}^n$*, la suite de solutions approchées* x_k *converge vers la solution exacte* x*.*

On commence par donner une condition nécessaire et suffisante de convergence d'une méthode itérative à l'aide du rayon spectral de la matrice d'itération (voir la Définition 4.4.5 pour la notion de rayon spectral).

Lemme 4.4.26 *La méthode itérative définie par (4.60) converge si et seulement si le rayon spectral de la matrice d'itération* $M^{-1}N$ *vérifie* $\rho(M^{-1}N) < 1$*.*

Démonstration. On définit l'erreur $e_k = x_k - x$. On a

$$e_k = (M^{-1}Nx_{k-1} + M^{-1}b) - (M^{-1}Nx + M^{-1}b) = M^{-1}Ne_{k-1} = (M^{-1}N)^k e_0.$$

Par application du Lemme 4.4.8, on en déduit que e_k tend vers 0, quel que soit e_0, si et seulement si $\rho(M^{-1}N) < 1$. □

Définition 4.4.29 (méthode de Jacobi) *Soit $A = (a_{ij})_{1\leq i,j\leq n}$. On note $D = \mathrm{diag}(a_{ii})$ la diagonale de A. On appelle méthode de Jacobi la méthode itérative associée à la décomposition*

$$M = D, \quad N = D - A.$$

Définition 4.4.30 (méthode de Gauss-Seidel) *Soit $A = (a_{ij})_{1\leq i,j\leq n}$. On décompose A sous la forme $A = D - E - F$ où $D = \mathrm{diag}(a_{ii})$ est la diagonale, $-E$ est la partie triangulaire inférieure (strictement), et $-F$ est la partie triangulaire supérieure (strictement) de A. On appelle méthode de Gauss-Seidel la méthode itérative associée à la décomposition*

$$M = D - E, \quad N = F.$$

Définition 4.4.31 (méthode de relaxation (SOR)) *Soit $\omega \in \mathbb{R}^+$. On appelle méthode de relaxation (SOR en anglais pour "Successive Over Relaxation"), pour le paramètre ω, la méthode itérative associée à la décomposition*

$$M = \frac{D}{\omega} - E, \quad N = \frac{1-\omega}{\omega} D + F.$$

$$M = \frac{}{\omega} - E, \quad N = \frac{}{\omega} D + F$$

Définition 4.4.32 (méthode du gradient) *Soit un paramètre réel $\alpha \neq 0$. On appelle méthode du gradient la méthode itérative associée à la décomposition*

$$M = \frac{1}{\alpha}\,\mathrm{Id} \quad \textit{et} \quad N = \left(\frac{1}{\alpha}\,\mathrm{Id} - A\right).$$

La méthode du gradient semble encore plus primitive que les méthodes précédentes, mais elle a une interprétation en tant que méthode de minimisation de la fonction $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ qui lui donne une plus grande applicabilité.

4.4.5 Méthode du gradient conjugué

La méthode du gradient conjugué est la méthode itérative de choix pour résoudre des systèmes linéaires dont la matrice est symétrique réelle définie positive. Il s'agit d'une amélioration spectaculaire de la méthode du gradient (surtout si elle combinée avec un préconditionnement, voir la Définition 4.4.43).

Proposition 4.4.39 *Soit A une matrice symétrique définie positive, et $x_0 \in \mathbb{R}^n$. Soit (x_k, r_k, p_k) trois suites définies par les relations de récurrence*

$$p_0 = r_0 = b - Ax_0, \ \textit{et pour } 0 \leq k \ \begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases} \tag{4.63}$$

avec

$$\alpha_k = \frac{\|r_k\|^2}{Ap_k \cdot p_k} \text{ et } \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Alors, la suite (x_k) de la méthode du gradient conjugué converge en moins de n itérations vers la solution du système linéaire $Ax = b$.

On peut montrer que r_k est la suite des résidus, c'est-à-dire $r_k = b - Ax_k$. Par conséquent, dès que $r_k = 0$, l'algorithme a convergé, c'est-à-dire que x_k est la solution du système $Ax = b$. On sait que la convergence est atteinte en moins de n itérations. Cependant dans la pratique, les calculs sur ordinateurs sont toujours sujets à des erreurs d'arrondi, et on ne trouve pas exactement $r_n = 0$. C'est pourquoi, on introduit un "petit" paramètre ϵ (typiquement 10^{-4} ou 10^{-8} selon la précision désirée), et on décide que l'algorithme a convergé dès que

$$\frac{\|r_k\|}{\|r_0\|} \leq \epsilon.$$

Par ailleurs, pour des systèmes de grande taille (pour lesquels n est "grand", de l'ordre de 10^4 à 10^6), la méthode du gradient conjugué est utilisée comme une méthode itérative, c'est-à-dire qu'elle converge, au sens du critère ci-dessus, en un nombre d'itérations bien inférieur à n.

Remarque 4.4.41

1. En général, si on n'a pas d'indications sur la solution, on choisit d'initialiser

1. En général, si on n'a pas d'indications sur la solution, on choisit d'initialiser la méthode du gradient conjugué par $x_0 = 0$. Si on résout une suite de problèmes peu différents les uns des autres, on peut initialiser x_0 par la solution précédente.
2. A chaque itération on n'a besoin de faire qu'un seul produit matrice-vecteur, à savoir Ap_k, car r_k est calculé par la formule de récurrence et non par la relation $r_k = b - Ax_k$.
3. Pour mettre en oeuvre la méthode du gradient conjugué, il n'est pas nécessaire de stocker la matrice A dans un tableau si on sait calculer le produit matrice vecteur Ay pour tout vecteur y.
4. La méthode du gradient conjugué est très efficace et très utilisée. Elle a beaucoup de variantes ou de généralisations, notamment au cas des matrices non symétriques définies positives.

•

La vitesse de convergence de la méthode du gradient conjugué dépend du conditionnement de la matrice A, l'idée du préconditionnement est de pré-multiplier le système linéaire $Ax = b$ par une matrice C^{-1} telle que le conditionnement de $(C^{-1}A)$ soit plus petit que celui de A. En pratique on choisit une matrice C "proche" de A mais plus facile à inverser.

Définition 4.4.43 *Soit à résoudre le système linéaire $Ax = b$. On appelle préconditionnement de A, une matrice C (facile à inverser) telle que* $\mathrm{cond}_2(C^{-1}A)$ *soit*

plus petit que $\mathrm{cond}_2(A)$. *On appelle système préconditionné le système équivalent* $C^{-1}Ax = C^{-1}b$.

La technique du préconditionnement est très efficace et essentielle en pratique pour converger rapidement. Nous indiquons trois choix possibles de C du plus simple au plus compliqué. Le préconditionnement le plus simple est le "préconditionnement diagonal" : il consiste à prendre $C = \mathrm{diag}(A)$. Il est malheureusement peu efficace, et on lui préfère souvent le "préconditionnement SSOR" (pour Symmetric SOR). En notant $D = \mathrm{diag}(A)$ la diagonale d'une matrice symétrique A et $-E$ sa partie strictement inférieure telle que $A = D - E - E^*$, pour $\omega \in]0, 2[$, on pose

$$C_\omega = \frac{\omega}{2-\omega}\left(\frac{D}{\omega} - E\right) D^{-1} \left(\frac{D}{\omega} - E^*\right).$$

On vérifie que, si A est définie positive, alors C l'est aussi. Le système $Cz = r$ est facile à résoudre car C est déjà sous une forme factorisée en produit de matrices triangulaires. Le nom de ce préconditionnement vient du fait qu'inverser C revient à effectuer deux itérations successives de la méthode itérative de relaxation (SOR), avec deux matrices d'itérations symétriques l'une de l'autre.

Un dernier exemple est le "préconditionnement de Cholesky incomplet". La matrice C est cherchée sous la forme BB^* où B est le facteur "incomplet" de la factorisation de Cholesky de A (voir la Proposition 4.4.19). Cette matrice triangulaire inférieure B est obtenue en appliquant l'algorithme de factorisation de Cholesky à A en forçant l'égalité $b_{ij} = 0$ si $a_{ij} = 0$. Cette modification de l'algorithme assure

A en forçant l'égalité $b_{ij} = 0$ si $a_{ij} = 0$. Cette modification de l'algorithme assure, d'une part que le facteur B sera aussi creux que la matrice A, et d'autre part que le calcul de ce facteur incomplet sera beaucoup moins cher (en temps de calcul) que le calcul du facteur exact si A est creuse (ce qui est le cas pour des matrices de discrétisation par éléments finis). Le préconditionnement de Cholesky incomplet est souvent le préconditionnement le plus efficace en pratique.

Chapitre 5

PROBLÈMES AUX VALEURS PROPRES

5.1 Motivation et exemples

5.1.1 Introduction

Ce chapitre est consacré à la théorie spectrale des équations aux dérivées partielles, c'est-à-dire à l'étude des valeurs propres et des fonctions propres de ces équations. La motivation de cette étude est double. D'une part, cela va nous permettre d'étudier des solutions particulières, dites oscillantes en temps (ou vibrantes), des

d'étudier des solutions particulières, dites stationnaires en temps (ou vibrantes), des problèmes d'évolution associés à ces équations. D'autre part, cela permet de déduire une méthode de résolution générale de ces mêmes problèmes d'évolution qui dépasse l'objet de ce cours.

Donnons tout de suite un exemple de **problème aux valeurs propres** pour le Laplacien avec condition aux limites de Dirichlet. Si Ω est un ouvert borné de $\mathbb{R}^N$ on cherche les couples $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$, avec $u \neq 0$, solutions de

$$\begin{cases} -\Delta u = \lambda u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \tag{5.1}$$

Le réel λ est appelé **valeur propre**, et la fonction $u(x)$ **mode propre ou fonction propre**. L'ensemble des valeurs propres est appelé le spectre de (5.1). On peut faire l'analogie entre (5.1) et le problème plus simple de détermination des valeurs et vecteurs propres d'une matrice A d'ordre n,

$$Au = \lambda u \quad \text{avec} \quad (\lambda, u) \in \mathbb{R} \times \mathbb{R}^n, \tag{5.2}$$

en affirmant que l'opérateur $-\Delta$ est une "généralisation" en dimension infinie d'une matrice A en dimension finie. La résolution de (5.1) sera utile pour résoudre les problèmes d'évolution, de type parabolique ou hyperbolique, associés au Laplacien, c'est-à-dire l'équation de la chaleur (5.5) ou l'équation des ondes (5.7). Néanmoins, les solutions de (5.1) ont aussi une interprétation physique qui leur est propre, par exemple comme modes propres de vibration.

Le plan de ce chapitre est le suivant. Après avoir motivé plus amplement le problème aux valeurs propres (5.1), nous développons dans la Section 5.3 une **théorie spectrale abstraite** dans les espaces de Hilbert. Le but de cette section est de généraliser en dimension infinie le résultat bien connu en dimension finie qui affirme que toute matrice symétrique réelle est diagonalisable dans une base orthonormée. Cette section relève en partie d'un cours de mathématiques "pures", aussi nous insistons à nouveau sur le fait que c'est l'esprit des résultats plus que la lettre des démonstrations qui importe ici. Nous appliquons cette théorie spectrale aux équations aux dérivées partielles elliptiques dans la Section 5.3. En particulier, nous démontrons que le problème spectral (5.1) **admet une infinité dénombrable de solutions**. Enfin, la Section 5.4 est consacrée aux questions **d'approximation numérique** des valeurs propres et fonctions propres d'une équation aux dérivées partielles. En particulier, nous introduisons la notion de **matrice de masse** $\mathcal{M}$ qui vient compléter celle de matrice de rigidité $\mathcal{K}$, et nous montrons que des valeurs propres approchées de (5.1) se calculent comme les valeurs propres du système $\mathcal{K}u = \lambda\mathcal{M}u$, ce qui confirme l'analogie entre (5.1) et sa version discrète (5.2).

5.1.2 Résolution des problèmes instationnaires

Avant de nous lancer dans les développements abstraits de la prochaine section, montrons en quoi la résolution d'un problème aux valeurs propres permet de résoudre aussi un problème d'évolution. Pour cela nous allons faire une analogie avec la résolution de systèmes différentiels en dimension finie. Dans tout ce qui suit A

désigne une matrice symétrique réelle, définie positive, d'ordre n. On note λ_k ses valeurs propres et r_k ses vecteurs propres, $1 \leq k \leq n$, tels que $Ar_k = \lambda_k r_k$.

On commence par un système différentiel du premier ordre

$$\begin{cases} \dfrac{\partial u}{\partial t} + Au = 0 \quad \text{pour } t \geq 0 \\ u(t=0) = u_0, \end{cases} \tag{5.3}$$

où $u(t)$ est une fonction de classe C^1 de $\mathbb{R}^+$ dans $\mathbb{R}^n$, et $u_0 \in \mathbb{R}^n$. Il est bien connu que (5.3) admet une solution unique obtenue en diagonalisant la matrice A. Plus précisément, la donnée initiale se décompose sous la forme $u_0 = \sum_{k=1}^n u_k^0 r_k$, ce qui donne

$$u(t) = \sum_{k=1}^n u_k^0 e^{-\lambda_k t} r_k.$$

Un deuxième exemple est le système différentiel du deuxième ordre

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + Au = 0 \quad \text{pour } t \geq 0 \\ u(t=0) = u_0, \\ \dfrac{\partial u}{\partial t}(t=0) = u_1, \end{cases} \tag{5.4}$$

où $u(t)$ est une fonction de classe C^2 de $\mathbb{R}^+$ dans $\mathbb{R}^n$, et $u_0, u_1 \in \mathbb{R}^N$. En décomposant les données initiales sous la forme $u_0 = \sum_{k=1}^n u_k^0 r_k$ et $u_1 = \sum_{k=1}^n u_k^1 r_k$, (5.4) admet

comme solution unique

$$u(t) = \sum_{k=1}^{n} \left(u_k^0 \cos\left(\sqrt{\lambda_k} t\right) + \frac{u_k^1}{\sqrt{\lambda_k}} \sin\left(\sqrt{\lambda_k} t\right) \right) r_k.$$

Il est clair sur ces deux exemples que la connaissance du spectre de la matrice A permet de résoudre les problèmes d'évolution (5.3) et (5.4). Aussi évident soient-ils, ces exemples sont tout à fait représentatifs de la démarche que nous allons suivre dans la suite. Nous allons remplacer la matrice A par l'opérateur $-\Delta$, l'espace $\mathbb{R}^n$ par l'espace de Hilbert $L^2(\Omega)$, et nous allons "diagonaliser" le Laplacien pour résoudre l'équation de la chaleur ou l'équation des ondes.

Afin de se convaincre que (5.1) est bien la "bonne" formulation du problème aux valeurs propres pour le Laplacien, on peut passer par un argument de "séparation des variables" dans l'équation de la chaleur ou l'équation des ondes que nous décrivons formellement. En l'absence de terme source, et en "oubliant" (provisoirement) la condition initiale et les conditions aux limites, nous cherchons une solution $\mathbf{u}$ de ces équations qui s'écrive sous la forme

$$\mathbf{u}(x,t) = \phi(t)u(x),$$

c'est-à-dire que l'on sépare les variables de temps et d'espace. Si $\mathbf{u}$ est solution de l'équation de la chaleur

$$\frac{\partial \mathbf{u}}{\partial t} - \Delta \mathbf{u} = 0, \tag{5.5}$$

on trouve (au moins formellement) que

$$\frac{\phi'(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

où $\lambda \in \mathbb{R}$ est une constante indépendante de t et de x. On en déduit que $\phi(t) = e^{-\lambda t}$ et que u doit être solution du problème aux valeurs propres

$$-\Delta u = \lambda u \tag{5.6}$$

muni de conditions aux limites adéquates.

De la même manière, si $\mathbf{u}$ est solution de l'équation des ondes

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} - \Delta \mathbf{u} = 0, \tag{5.7}$$

on trouve que

$$\frac{\phi''(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

où $\lambda \in \mathbb{R}$ est une constante. Cette fois-ci on en déduit que, si $\lambda > 0$ (ce qui sera effectivement le cas), alors $\phi(t) = a\cos(\sqrt{\lambda}t) + b\sin(\sqrt{\lambda}t)$ et que u doit encore être solution de (5.6). Remarquons que, si le comportement en espace de la solution $\mathbf{u}$ est le même pour l'équation de la chaleur et pour l'équation des ondes, il n'en est pas de même pour son comportement en temps : elle oscille en temps pour les ondes alors qu'elle décroît exponentiellement en temps (car $\lambda > 0$) pour la chaleur.

Exercice 5.1.1 Soit $\Omega = \mathbb{R}^N$. Montrer que $u(x) = \exp(ik \cdot x)$ est une solution de (5.6) si $|k|^2 = \lambda$. Une telle solution est appelée onde plane.

Exercice 5.1.2 Soit un potentiel régulier $V(x)$. Montrer que, si $\mathbf{u}(x,t) = e^{-i\omega t}u(x)$ est solution de

$$i\frac{\partial \mathbf{u}}{\partial t} + \Delta \mathbf{u} - V\mathbf{u} = 0 \ \text{ dans } \mathbb{R}^N \times \mathbb{R}^+_*, \tag{5.8}$$

alors $u(x)$ est solution de

$$-\Delta u + Vu = \omega u \ \text{ dans } \mathbb{R}^N. \tag{5.9}$$

On retrouve le même type de problème spectral que (5.6), à l'addition d'un terme d'ordre zéro près. Pour l'équation de Schrödinger la valeur propre ω s'interprète comme une énergie. La plus petite valeur possible de cette énergie correspond à l'énergie de l'état fondamental du système décrit par (5.8). Les autres valeurs, plus grandes, donnent les énergies des états excités. Sous des conditions "raisonnables" sur le potentiel V, ces niveaux d'énergie sont discrets en nombre infini dénombrable (ce qui est cohérent avec la vision physique des *quanta*).

Exercice 5.1.3 Soit $V(x) = Ax \cdot x$ avec A matrice symétrique réelle définie positive. Montrer que $u(x) = \exp(-A^{1/2}x \cdot x/2)$ est une solution de (5.9) si $\omega = \operatorname{tr}(A^{1/2})$. Une telle solution est appelée état fondamental.

5.3 Valeurs propres d'un problème elliptique

5.3.1 Problème variationnel

Nous revenons au cadre variationnel introduit au Chapitre 1. L'intérêt de ce cadre assez général est qu'il s'appliquera à de nombreux modèles différents. Dans un espace de Hilbert V nous considérons une forme bilinéaire $a(\cdot,\cdot)$, **symétrique**, continue et coercive, c'est-à-dire que $a(w,v) = a(v,w)$, et il existe $M > 0$ et $\nu > 0$ tels que

$$|a(w,v)| \leq M\|w\|_V \, \|v\|_V \text{ pour tout } w, v \in V$$

et

$$a(v,v) \geq \nu\|v\|_V^2 \text{ pour tout } v \in V.$$

Remarquons que l'hypothèse de symétrie de la forme bilinéaire $a(\cdot,\cdot)$ est nouvelle puisqu'elle n'est pas nécessaire dans le cadre du théorème de Lax-Milgram. De plus, nous introduisons un nouvel ingrédient, à savoir un autre espace de Hilbert H sur lequel nous faisons l'hypothèse fondamentale suivante

$$\begin{cases} V \subset H \text{ avec injection compacte} \\ V \text{ est dense dans } H. \end{cases} \tag{5.11}$$

L'expression "injection compacte" veut dire que de toute suite bornée de V on peut extraire une sous-suite convergente dans H. Les espaces H et V ne partagent pas

le même produit scalaire, et nous les noterons $\langle \cdot, \cdot \rangle_H$ et $\langle \cdot, \cdot \rangle_V$ pour éviter toute confusion.

Nous considérons le problème variationnel de valeurs propres suivant (ou problème spectral) : trouver $\lambda \in \mathbb{R}$ et $u \in V \setminus \{0\}$ tels que

$$a(u, v) = \lambda \langle u, v \rangle_H \quad \forall\, v \in V. \tag{5.12}$$

On dira que λ est une valeur propre du problème variationnel (5.12) (ou de la forme bilinéaire a) et que u est le vecteur propre associé.

Les solutions de (5.12) sont données par le résultat suivant que nous admettrons.

Théorème 5.3.2 *Soit V et H deux espaces de Hilbert réels de dimension infinie. On suppose que $V \subset H$ avec injection compacte et que V est dense dans H. Soit $a(\cdot, \cdot)$ une forme bilinéaire symétrique continue et coercive sur V. Alors les valeurs propres de (5.12) forment une suite croissante $(\lambda_k)_{k \geq 1}$ de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de H $(u_k)_{k \geq 1}$ de vecteurs propres associés, c'est-à-dire que*

$$u_k \in V, \quad \text{et } a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall\, v \in V.$$

De plus, $(u_k / \sqrt{\lambda_k})_{k \geq 1}$ est une base hilbertienne de V pour le produit scalaire $a(\cdot, \cdot)$.

Remarque 5.3.3 Dans le Théorème 5.3.2 on peut remplacer l'hypothèse de coercivité de la forme bilinéaire $a(\cdot, \cdot)$ par l'hypothèse plus faible qu'il existe deux

civité de la forme bilinéaire $a(\cdot,\cdot)$ par l'hypothèse plus faible qu'il existe deux constantes positives $\eta > 0$ et $\nu > 0$ telles que

$$a(v,v) + \eta\|v\|_H^2 \geq \nu\|v\|_V^2 \text{ pour tout } v \in V.$$

Dans ce cas les valeurs propres $(\lambda_k)_{k\geq 1}$ ne sont pas forcément positives, mais vérifient seulement $\lambda_k + \eta > 0$. •

Nous donnons au passage une caractérisation très utile des valeurs propres du problème variationnel (5.12), appelée **principe du min-max ou de Courant-Fisher**. Pour cela on introduit le quotient de Rayleigh défini, pour chaque fonction $v \in V \setminus \{0\}$, par

$$R(v) = \frac{a(v,v)}{\|v\|_H^2}.$$

Proposition 5.3.4 (Courant-Fisher) *Soit V et H deux espaces de Hilbert réels de dimension infinie. On suppose que $V \subset H$ avec injection compacte et que V est dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique continue et coercive sur V. Pour $k \geq 0$ on note $\mathcal{E}_k$ l'ensemble des sous-espaces vectoriels de dimension k de V. On note $(\lambda_k)_{k\geq 1}$ la suite* **croissante** *des valeurs propres du problème variationnel (5.12). Alors, pour tout $k \geq 1$, la k-ème valeur propre est donnée par*

$$\lambda_k = \min_{W\in\mathcal{E}_k} \left(\max_{v\in W\setminus\{0\}} R(v) \right) = \max_{W\in\mathcal{E}_{k-1}} \left(\min_{v\in W^\perp\setminus\{0\}} R(v) \right). \tag{5.15}$$

En particulier, la première valeur propre vérifie

$$\lambda_1 = \min_{v \in V \setminus \{0\}} R(v), \tag{5.16}$$

et tout point de minimum dans (5.16) est un vecteur propre associé à λ_1.

Démonstration. Soit $(u_k)_{k\geq 1}$ la base hilbertienne de H formée des vecteurs propres de (5.12). D'après le Théorème 5.3.2, $(u_k/\sqrt{\lambda_k})_{k\geq 1}$ est une base hilbertienne de V. On peut donc caractériser les espaces H et V à partir de leur décomposition spectrale

$$H = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k,\ \|v\|_H^2 = \sum_{k=1}^{+\infty} \alpha_k^2 < +\infty \right\},$$
$$V = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k,\ \|v\|_V^2 = \sum_{k=1}^{+\infty} \lambda_k \alpha_k^2 < +\infty \right\}.$$

On remarque au passage que, comme les valeurs propres λ_k sont minorées par $\lambda_1 > 0$, cette caractérisation fait bien apparaître V comme un sous-espace de H. On peut alors réécrire le quotient de Rayleigh

$$R(v) = \frac{\sum_{k=1}^{+\infty} \lambda_k \alpha_k^2}{\sum_{k=1}^{+\infty} \alpha_k^2},$$

ce qui démontre immédiatement le résultat pour la première valeur propre. Introduisons le sous-espace $W_k \in \mathcal{E}_k$ engendré par $(u_1, u_2, ..., u_k)$. On a

$$R(v) = \frac{\sum_{j=1}^{k} \lambda_j \alpha_j^2}{\sum_{j=1}^{k} \alpha_j^2} \quad \forall\, v \in W_k \quad \text{et} \quad R(v) = \frac{\sum_{j=k}^{+\infty} \lambda_j \alpha_j^2}{\sum_{j=k}^{+\infty} \alpha_j^2} \quad \forall\, v \in W_{k-1}^{\perp},$$

d'où l'on déduit

$$\lambda_k = \max_{v \in W_k \setminus \{0\}} R(v) = \min_{v \in W_{k-1}^{\perp} \setminus \{0\}} R(v).$$

Soit W un sous-espace quelconque dans $\mathcal{E}_k$. Comme W est de dimension k et W_{k-1} de dimension $k-1$, l'intersection $W \cap W_{k-1}^{\perp}$ n'est pas réduite à $\{0\}$. Par conséquent,

$$\max_{v \in W \setminus \{0\}} R(v) \geq \max_{v \in W \cap W_{k-1}^{\perp} \setminus \{0\}} R(v) \geq \min_{v \in W \cap W_{k-1}^{\perp} \setminus \{0\}} R(v) \geq \min_{v \in W_{k-1}^{\perp} \setminus \{0\}} R(v) = \lambda_k,$$

ce qui prouve la première égalité dans (5.15). De même, si W est un sous-espace dans $\mathcal{E}_{k-1}$, alors $W^{\perp} \cap W_k$ n'est pas réduit à $\{0\}$, et

$$\min_{v \in W^{\perp} \setminus \{0\}} R(v) \leq \min_{v \in W^{\perp} \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W^{\perp} \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W_k \setminus \{0\}} R(v) = \lambda_k,$$

ce qui prouve la deuxième égalité dans (5.15). Soit maintenant u un point de minimum dans (5.16). Pour $v \in V$, on introduit la fonction $f(t) = R(u + tv)$ d'une variable réelle $t \in \mathbb{R}$ qui admet un minimum en $t = 0$. Par conséquent sa dérivée

s'annule en $t = 0$. En tenant compte de ce que $f(0) = \lambda_1$, un simple calcul montre que

$$f'(0) = 2\frac{a(u,v) - \lambda_1 \langle u, v\rangle_H}{\|u\|_H^2}.$$

Comme v est quelconque dans V, la condition $f'(0) = 0$ n'est rien d'autre que la formulation variationnelle (5.12), c'est-à-dire que u est un vecteur propre associé à la valeur propre λ_1. □

5.3.2 Valeurs propres du Laplacien

On peut immédiatement appliquer le Théorème 5.3.2 à la formulation variationnelle du Laplacien avec conditions aux limites de Dirichlet, ce qui nous donne le résultat suivant.

Théorème 5.3.5 *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^1$ de $\mathbb{R}^N$. Il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de $L^2(\Omega)$ $(u_k)_{k\geq 1}$, telle que chaque u_k appartient à $H_0^1(\Omega)$ et vérifie*

$$\begin{cases} -\Delta u_k = \lambda_k u_k & p.p.\ dans\ \Omega \\ u_k = 0 & p.p.\ sur\ \partial\Omega. \end{cases} \tag{5.17}$$

Démonstration. Pour le Laplacien avec conditions aux limites de Dirichlet, on

choisit $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, et la forme bilinéaire symétrique est définie par

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx,$$

et le produit scalaire sur $L^2(\Omega)$ est bien sûr

$$\langle u, v \rangle_H = \int_\Omega uv \, dx.$$

On vérifie aisément les hypothèses du Théorème 5.3.2. Grâce au Théorème 2.3.21 de Rellich, V est bien compactement inclus dans H. Comme $C_c^\infty(\Omega)$ est dense à la fois dans H et dans V, V est bien dense dans H. Enfin, on a vu au Chapitre 3 que la forme bilinéaire a est bien continue et coercive sur V. Par conséquent, il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de $L^2(\Omega)$ $(u_k)_{k\geq 1}$, tels que $u_k \in H_0^1(\Omega)$ et

$$\int_\Omega \nabla u_k \cdot \nabla v \, dx = \lambda_k \int_\Omega u_k v \, dx \quad \forall\, v \in H_0^1(\Omega).$$

Par une simple intégration par parties (du même type que celle pratiquée dans la démonstration du Théorème 3.2.2) on obtient (5.17). Remarquons que nous n'utilisons la régularité de Ω que pour pouvoir appliquer le Théorème de trace 2.3.13 et donner un sens "presque partout" à la condition aux limites de Dirichlet. $\square$

Remarque 5.3.6 L'hypothèse sur le caractère borné de l'ouvert Ω est absolument fondamentale dans le Théorème 5.3.5. Si elle n'est pas satisfaite, le Théorème 2.3.21 de Rellich (sur l'injection compacte de $H^1(\Omega)$ dans $L^2(\Omega)$) est en général faux, et on peut montrer que le Théorème 5.3.5 n'a pas lieu. En fait, il se peut qu'il existe une infinité (non dénombrable) de valeurs propres "généralisées" au sens où les fonctions propres n'appartiennent pas à $L^2(\Omega)$. A la lumière de l'Exercice 5.1.1 on méditera le cas du Laplacien dans $\Omega = \mathbb{R}^N$. •

Exercice 5.3.2 En dimension $N = 1$, on considère $\Omega =]0, 1[$. Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Dirichlet (5.17). A l'aide de la décomposition spectrale de ce problème, montrer que la série

$$\sum_{k=1}^{+\infty} a_k \sin(k\pi x)$$

converge dans $L^2(0,1)$ si et seulement si $\sum_{k=1}^{+\infty} a_k^2 < +\infty$, et dans $H^1(0,1)$ si et seulement si $\sum_{k=1}^{+\infty} k^2 a_k^2 < +\infty$.

Exercice 5.3.3 On considère un parallélépipède $\Omega =]0, L_1[\times]0, L_2[\times \cdots \times]0, L_N[$, où les $(L_i > 0)_{1 \leq i \leq N}$ sont des constantes positives. Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Dirichlet (5.17).

Exercice 5.3.4 On considère à nouveau un ouvert Ω parallélépipédique comme dans

Exercice 5.3.4 On considère à nouveau un ouvert Ω parallélépipédique comme dans l'Exercice 5.3.3. Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Neumann sur tout le bord $\partial\Omega$.

On peut montrer que les fonctions propres du Laplacien, avec conditions aux limites de Dirichlet ou de Neumann, sont régulières.

Proposition 5.3.9 *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^\infty$. Alors les fonctions propres solutions de (5.17) appartiennent à $C^\infty(\overline{\Omega})$.*

Démonstration. Soit u_k la k-ème fonction propre solution dans $H_0^1(\Omega)$ de (5.17). On peut considérer que u_k est solution du problème aux limites suivant

$$\begin{cases} -\Delta u_k = f_k & \text{dans } \Omega \\ u_k = 0 & \text{sur } \partial\Omega, \end{cases}$$

avec $f_k = \lambda_k u_k$. Comme f_k appartient à $H^1(\Omega)$, par application du Théorème 3.2.26 de régularité on en déduit que la solution u_k appartient à $H^3(\Omega)$. Du coup, le second membre f_k est plus régulier ce qui permet d'augmenter encore la régularité de u_k. Par une récurrence facile on montre ainsi que u_k appartient à $H^m(\Omega)$ pour tout $m \geq 1$. En vertu du Théorème 2.3.25 sur la continuité des fonctions de $H^m(\Omega)$ (voir aussi la Remarque 2.3.26), on en déduit que u_k appartient donc $C^\infty(\overline{\Omega})$. □

Nous admettons un résultat qualitatif important à propos de la première valeur propre.

Théorème 5.3.10 (de Krein-Rutman) *On reprend les notations et les hypothèses du Théorème 5.3.5. On suppose que l'ouvert Ω est connexe. Alors la première valeur propre λ_1 est simple (i.e. le sous-espace propre correspondant est de dimension 1) et le premier vecteur propre peut être choisi positif presque partout dans Ω.*

Remarque 5.3.11 Le Théorème 5.3.10 de Krein-Rutman est spécifique au cas des équations "scalaires" (c'est-à-dire que l'inconnue u est à valeurs dans $\mathbb{R}$). Ce résultat est faux en général si l'inconnue u est à valeurs vectorielles (voir plus loin l'exemple du système de l'élasticité). La raison de cette différence entre le cas scalaire et vectoriel est que ce théorème s'appuie sur le principe du maximum (voir le Théorème 3.2.22) qui n'est valable que dans le cas scalaire. •

Exercice 5.3.6 Soit Ω un ouvert borné régulier et connexe. Montrer que la première valeur propre du Laplacien dans Ω avec condition aux limites de Neumann est nulle et qu'elle est simple.

Remarque 5.3.12 L'ensemble des résultats de cette sous-section se généralise sans difficulté à d'autres conditions aux limites et à des opérateurs elliptiques généraux du second ordre (voir la Sous-section 3.2.3). •

5.3.3 Autres modèles

L'extension des résultats de la sous-section précédente à des équations aux dé-

rivées partielles elliptiques plus compliquées que le Laplacien ne pose pas de problèmes conceptuels nouveaux. Nous décrivons brièvement cette généralisation pour un exemple significatif : le système de l'élasticité linéarisée.

Les équations (3.56) de l'élasticité linéarisée décrivent en fait le régime stationnaire des équations dynamiques suivantes (très semblables à l'équation des ondes)

$$\begin{cases} \rho\frac{\partial^2 u}{\partial t^2} - \operatorname{div}\left(2\mu e(u) + \lambda \operatorname{tr}(e(u))\operatorname{Id}\right) = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+, \end{cases} \tag{5.18}$$

où $\rho > 0$ est la densité volumique du matériau et $e(u) = \left(\nabla u + (\nabla u)^t\right)/2$. Rappelons que les coefficients de Lamé du matériau vérifient $\mu > 0$ et $2\mu + N\lambda > 0$. En l'absence de forces extérieures f (et en ne tenant pas compte d'éventuelles conditions initiales) on peut aussi chercher des solutions oscillantes en temps de (5.18) comme nous l'avons décrit pour l'équation des ondes dans Sous-section 5.1.2. Cela conduit à chercher des solutions (ℓ, u) du problème aux valeurs propres suivant

$$\begin{cases} -\operatorname{div}\left(2\mu e(u) + \lambda \operatorname{tr}(e(u))\operatorname{Id}\right) = \ell u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{5.19}$$

où $\ell = \omega^2$ est le carré de la fréquence de vibration (nous avons changé la notation de la valeur propre pour éviter une confusion avec le coefficient de Lamé λ). En mécanique, la fonction propre u est aussi appelée **mode propre de vibration**.

En suivant la méthode appliquée ci-dessus au Laplacien on peut démontrer le résultat suivant (nous laissons les détails au lecteur en guise d'exercice).

Proposition 5.3.13 *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^1$ de $\mathbb{R}^N$. Il existe une suite croissante $(\ell_k)_{k\geq 1}$ de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de $L^2(\Omega)^N$ $(u_k)_{k\geq 1}$, telle que chaque u_k appartient à $H_0^1(\Omega)^N$ et vérifie*

$$\begin{cases} -\mathrm{div}\,(2\mu e(u_k) + \lambda\,\mathrm{tr}(e(u_k))\,\mathrm{Id}) = \ell_k u_k & \text{p.p. dans } \Omega \\ u_k = 0 & \text{p.p. sur } \partial\Omega. \end{cases}$$

Rang du mode propre	1	2	3	4
Valeur propre	102.54	102.54	1885.2	2961.2

TABLE 5.1 – Valeurs propres correspondant aux modes propres de la Figure 5.1.

Le résultat de régularité sur les fonctions propres u_k de la Proposition 5.3.9 s'étend aussi facilement au cas de l'élasticité et du problème (5.19). Par contre le Théorème 5.3.10 sur la simplicité de la première valeur propre et la positivité de la première fonction propre est faux en général (comme est faux le principe du maximum). Comme exemple nous calculons par la méthode des éléments finis Q_1 les 4 premiers modes propres d'une "tour" dont la base est fixée (condition aux limites de Dirichlet) et dont les autres parois sont libres (condition aux limites de Neumann). Les 2 premiers modes, dits de battement, correspondent à la même valeur propre (ils sont indépendants mais symétriques par rotation de 90° suivant l'axe z) (voir la Figure 5.1 et le Tableau 5.3.3). La première valeur propre est donc "double".

Exercice 5.3.8 On considère le problème aux valeurs propres pour l'équation de Schrödinger avec un potentiel quadratique $V(x) = Ax \cdot x$ où A est une matrice symétrique définie positive (modèle de l'oscillateur harmonique)

$$-\Delta u + Vu = \lambda u \quad \text{dans } \mathbb{R}^N. \tag{5.21}$$

On définit les espaces $H = L^2(\mathbb{R}^N)$ et

$$V = \left\{ v \in H^1(\mathbb{R}^N) \text{ tel que } |x|v(x) \in L^2(\mathbb{R}^N) \right\}.$$

Montrer que V est un espace de Hilbert pour le produit scalaire

$$\langle u, v \rangle_V = \int_{\mathbb{R}^N} \nabla u(x) \cdot \nabla v(x)\, dx + \int_{\mathbb{R}^N} |x|^2 u(x) v(x)\, dx,$$

et que l'injection de V dans H est compacte. En déduire qu'il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de réels positifs qui tend vers l'infini et une base hilbertienne de $L^2(\mathbb{R}^N)$ $(u_k)_{k\geq 1}$ qui sont les valeurs propres et les fonctions propres de (5.21). Calculer explicitement ses valeurs et fonctions propres (on cherchera u_k sous la forme $p_k(x)\exp(-Ax \cdot x/2)$ où p_k est un polynôme de degré $k-1$). Interpréter physiquement les résultats.

Exercice 5.3.9 Soit Ω un ouvert borné régulier de $\mathbb{R}^N$. On considère le problème de vibrations pour l'équation des plaques avec condition aux limites d'encastrement

$$\begin{cases} \Delta(\Delta u) = \lambda u & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = u = 0 & \text{sur } \partial\Omega. \end{cases}$$

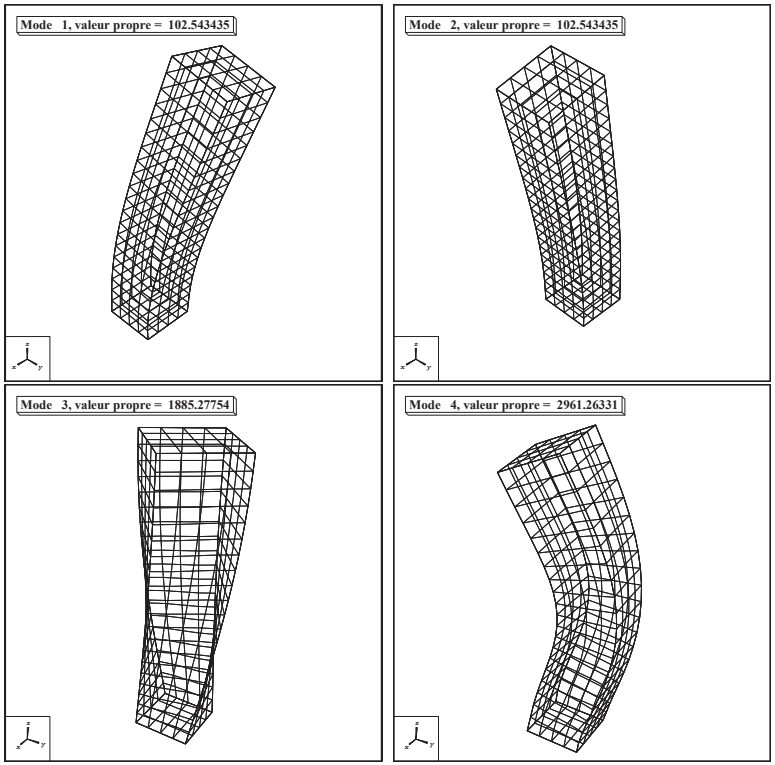


FIGURE 5.1 – Les 4 premiers modes propres d'une "tour" en élasticité.

Montrer qu'il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de valeurs propres positives qui tend vers l'infini et une base hilbertienne dans $L^2(\Omega)$ de fonctions propres $(u_k)_{k\geq 1}$ qui appartiennent à $H_0^2(\Omega)$.

5.4 Méthodes numériques

5.4.1 Discrétisation par éléments finis

On va considérer une approximation interne de la formulation variationnelle introduite à la Sous-section 5.3.1. Étant donné un sous-espace V_h de l'espace de Hilbert V, de dimension finie, on cherche les solutions $(\lambda_h, u_h) \in \mathbb{R} \times V_h$ de

$$a(u_h, v_h) = \lambda_h \langle u_h, v_h \rangle_H \quad \forall\, v_h \in V_h. \tag{5.22}$$

Typiquement, V_h est un espace d'éléments finis comme ceux introduits par les Définitions 4.3.5 et 4.3.25, et H est l'espace $L^2(\Omega)$. La résolution de l'approximation interne (5.22) est facile comme le montre le lemme suivant.

Lemme 5.4.1 *On se place sous les hypothèses du Théorème 5.3.2. Alors les valeurs propres de (5.22) forment une suite croissante finie*

$$0 < \lambda_1 \leq \cdots \leq \lambda_{n_{dl}} \quad \textit{avec } n_{dl} = \dim V_h,$$

et il existe une base de V_h, orthonormale dans H, $(u_{k,h})_{1\leq k\leq n_{dl}}$ de vecteurs propres associés, c'est-à-dire que

$$u_{k,h} \in V_h, \quad et\ a(u_{k,h}, v_h) = \lambda_k \langle u_{k,h}, v_h \rangle_H \quad \forall\, v_h \in V_h.$$

Démonstration. Ce lemme peut être considéré comme une variante évidente du Théorème 5.3.2 (à la différence près qu'en dimension finie il existe un nombre fini de valeurs propres). Néanmoins nous en donnons une démonstration différente, purement algébrique, qui correspond plus à la démarche suivie en pratique. Soit $(\phi_i)_{1\leq i\leq n_{dl}}$ une base de V_h (par exemple, les fonctions de base d'une méthode d'éléments finis, voir la Proposition 4.3.7). On cherche u_h solution de (5.22) sous la forme

$$u_h(x) = \sum_{i=1}^{n_{dl}} U_i^h \phi_i(x).$$

Introduisant la **matrice de masse** $\mathcal{M}_h$ définie par

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_H \quad 1 \leq i, j \leq n_{dl},$$

et la **matrice de rigidité** $\mathcal{K}_h$ définie par

$$(\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl},$$

le problème (5.22) est équivalent à trouver $(\lambda_h, U_h) \in \mathbb{R} \times \mathbb{R}^{n_{dl}}$ solution de

$$\mathcal{K}_h U_h = \lambda_h \mathcal{M}_h U_h. \tag{5.23}$$

Les appellations "matrices de masse et de rigidité" proviennent des applications en mécanique des solides. Remarquons que, dans le cas où V_h est un espace d'éléments finis, la matrice de rigidité $\mathcal{K}_h$ est exactement la même matrice que celle rencontrée au Chapitre 4 dans l'application de la méthode des éléments finis aux problèmes elliptiques. On vérifie immédiatement que les matrices $\mathcal{M}_h$ et $\mathcal{K}_h$ sont symétriques et définies positives. Le système (5.23) est un problème matriciel aux valeurs propres "généralisé". Le théorème de réduction simultanée (voir par exemple le théorème 2.3.6 dans [2]) affirme qu'il existe une matrice inversible P_h telle que

$$\mathcal{M}_h = P_h P_h^*, \text{ et } \mathcal{K}_h = P_h \operatorname{diag}(\lambda_k) P_h^*.$$

Par conséquent, les solutions de (5.23) sont les valeurs propres (λ_k) et les vecteurs propres $(U_{k,h})_{1\leq k\leq n_{dl}}$ qui sont les vecteurs colonnes de l'inverse de P_h^*. Ces vecteurs colonnes forment donc une base, orthogonale pour $\mathcal{K}_h$ et orthonormale pour $\mathcal{M}_h$ (nous indiquerons brièvement à la Remarque 5.4.3 comment calculer cette base). Finalement, les vecteurs $U_{k,h}$ sont simplement les vecteurs des coordonnées dans la base $(\phi_i)_{1\leq i\leq n_{dl}}$ des fonctions $u_{k,h}$ qui forment une base orthonormale de V_h pour le produit scalaire de H. □

Remarque 5.4.2 Dans le Lemme 5.4.1 on a repris les hypothèses du Théorème 5.3.2 : en particulier, la forme bilinéaire $a(u, v)$ est supposée **symétrique**. On voit bien l'importance de cette hypothèse dans la démonstration. En effet, si elle n'était pas symétrique, on ne saurait pas si le système (5.23) est diagonalisable, c'est-à-dire s'il existe des solutions du problème aux valeurs propres (5.22). •

L'application du Lemme 5.4.1 à l'approximation variationnelle par éléments finis du problème de Dirichlet (5.17) est immédiate. On prend $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, et l'espace discret V_{0h} de la Définition 4.3.5 (rappelons que V_{0h} contient la condition aux limites de Dirichlet).

Exercice 5.4.1 On considère le problème aux valeurs propres en dimension $N = 1$

$$\begin{cases} -u_k'' = \lambda_k u_k \quad \text{pour } 0 < x < 1 \\ u_k(0) = u_k(1) = 0. \end{cases}$$

On se propose de calculer la matrice de masse pour la méthode des éléments finis P_1. On reprend les notations de la Section 4.2. Montrer que la matrice de masse $\mathcal{M}_h$ est donnée par

$$\mathcal{M}_h = h \begin{pmatrix} 2/3 & 1/6 & & & 0 \\ 1/6 & 2/3 & 1/6 & & \\ & \ddots & \ddots & \ddots & \\ & & 1/6 & 2/3 & 1/6 \\ 0 & & & 1/6 & 2/3 \end{pmatrix},$$

et que ses valeurs propres sont

$$\lambda_k(\mathcal{M}_h) = \frac{h}{3}\left(2 + \cos(k\pi h)\right) \text{ pour } 1 \leq k \leq n.$$

Montrer que, si on utilise la formule de quadrature (4.45), alors on trouve que $\mathcal{M}_h = h\,\mathrm{Id}$. Dans ce dernier cas, calculer les valeurs propres du problème spectral discret.

Remarque 5.4.3 Pour calculer les valeurs et vecteurs propres du problème spectral matriciel (5.23) il faut, en général, commencer par calculer la factorisation de Cholesky de la matrice de masse $\mathcal{M}_h = \mathcal{L}_h \mathcal{L}_h^*$, pour se ramener au cas classique

$$\tilde{\mathcal{K}}_h \tilde{U}_h = \lambda_h \tilde{U}_h \quad \text{avec } \tilde{\mathcal{K}}_h = \mathcal{L}_h^{-1} \mathcal{K}_h (\mathcal{L}_h^*)^{-1} \text{ et } \tilde{U}_h = \mathcal{L}_h^* U_h,$$

pour lequel on dispose d'algorithmes de calcul de valeurs et vecteurs propres. Nous renvoyons à la Section 13.2 pour plus de détails sur ces algorithmes : disons seulement que c'est l'étape la plus coûteuse en temps de calcul.

On peut éviter de construire la matrice $\tilde{\mathcal{K}}_h$ et faire l'économie de la factorisation de Cholesky de $\mathcal{M}_h$ si on utilise une formule de quadrature pour évaluer les coefficients de la matrice $\mathcal{M}_h$ qui la rende **diagonale**. Ce procédé d'intégration numérique est appelé **condensation de masse** (ou "mass lumping" en anglais) et est fréquemment utilisé. Par exemple, si on utilise la formule de quadrature (4.45) (qui utilise uniquement les valeurs aux noeuds d'une fonction pour calculer une intégrale), on voit facilement que la matrice de masse $\mathcal{M}_h$ ainsi obtenue est diagonale (voir l'Exercice 5.4.1). •

Une analyse numérique précise de la méthode des éléments finis montre que **seules les premières valeurs propres** discrètes $\lambda_{k,h}$ (les plus petites) sont des