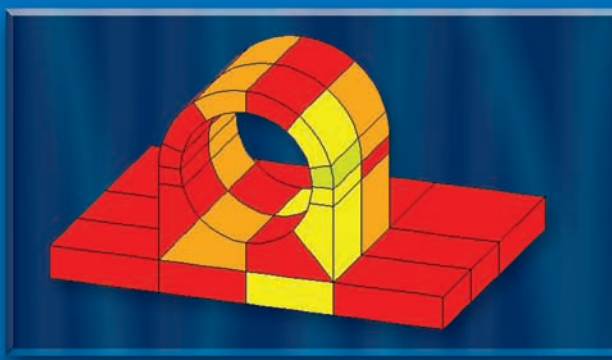
L'USINE NOUVELLE

Alexandre Ern

Aide-mémoire
Éléments finis



DUNOD

Alexandre Ern

Aide-mémoire
Éléments finis

L'USINE NOUVELLE

DUNOD

ISBN 2 10 007303 6

TABLE DES MATIÈRES

AVANT-PROPOS

Les origines de la méthode des éléments finis remontent aux années 1950 lorsque des ingénieurs l'utilisèrent afin de simuler des problèmes de mécanique des milieux continus déformables. Depuis, le champ d'applications s'est considérablement étendu et les fondements théoriques de la méthode se sont amplement consolidés. Il existe de nos jours un nombre important de logiciels commerciaux et académiques qui utilisent la méthode des éléments finis comme un outil de simulation robuste pour des problèmes de mécanique des milieux continus, de mécanique des fluides, de thermique, d'électromagnétisme ou de finance, pour ne citer que quelques exemples.

L'essor de la méthode des éléments finis repose sur deux ingrédients fondamentaux. D'une part, les propriétés interpolantes des éléments finis : ceux-ci permettent d'approcher des fonctions définies sur un domaine en maillant ce domaine puis en choisissant sur chaque maille des combinaisons linéaires de fonctions de forme (par exemple polynômiales). D'autre part, la méthode de Galerkin, qui fournit un cadre d'approximation général pour une large classe de problèmes où l'inconnue est une fonction qui doit satisfaire une ou plusieurs équations aux dérivées partielles et des conditions aux limites.

Cet aide-mémoire s'adresse en premier lieu aux ingénieurs en bureaux d'études qui utilisent ou développent des modèles numériques basés sur la méthode des éléments finis. Son objectif est de rappeler (sans démonstration) les principaux résultats théoriques fondant la méthode, d'en analyser des applications à divers problèmes modèles des sciences de l'ingénieur et, enfin, d'en étudier la mise en œuvre numérique et les bases de sa programmation. Cet aide-mémoire constitue également un outil de travail pour les élèves-ingénieurs et étudiants de niveau master. En particulier, certains chapitres correspondent à des cours

dispensés par l'auteur en première et deuxième années d'Écoles d'ingénieurs. Une annexe qui résume les bases mathématiques de la méthode des éléments finis permet au lecteur de faire le point sur son bagage mathématique afin de tirer le meilleur profit de la lecture de cet ouvrage.

Cet aide-mémoire peut également servir d'introduction au livre de Ern et Guermond, *Theory and Practice of Finte elements*, Applied Mathematical Series, volume 159, Springer, New York (2004), qui s'adresse aux étudiants de troisième cycle et aux chercheurs. Le lecteur désireux d'approfondir l'étude de la méthode des éléments finis est invité à consulter cette référence. Il y trouvera en particulier la preuve des résultats qui sont ici énoncés sans démonstration. Par ailleurs, cet aide-mémoire propose une bibliographie comprenant quatre-vingts références à la littérature spécialisée dont une quarantaine d'ouvrages de référence dans le domaine.

J'adresse mes plus vifs remerciements à Erik Burman, Linda El Alaoui, Jean-Frédéric Gerbeau, Tony Lelièvre et Pierre Tardif d'Hamonville pour avoir relu cet ouvrage et m'avoir fait part de leurs suggestions.

1 • PRÉLUDE : ÉLÉMENTS FINIS EN DIMENSION UN

Ce chapitre introductif a pour but d'éclairer les fondements théoriques de la méthode des éléments finis et les grandes étapes intervenant dans sa mise en œuvre numérique à travers l'étude d'un exemple relativement simple : un problème aux limites d'ordre deux en dimension un. Cette étude permettra d'introduire d'une part quelques mots clés essentiels pour la compréhension de la méthode et d'autre part quelques éléments finis classiques en une dimension d'espace.

1.1 Le problème modèle

On considère un intervalle $\Omega =]a, b[$. Étant donné deux fonctions $\alpha : \Omega \to \mathbb{R}$ et $f : \Omega \to \mathbb{R}$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-(\alpha u')' = f \qquad \text{dans } \Omega, \tag{1.1}$$

$$u(a) = u(b) = 0. \tag{1.2}$$

Le problème modèle (1.1)–(1.2) admet plusieurs interprétations physiques.

- **Équilibre mécanique d'une corde tendue.** On considère une corde horizontale tendue entre ses deux extrémités situées aux points a et b. On applique à cette corde une densité linéique d'efforts verticaux. Ces efforts sont décrits par la fonction f : pour $x \in \Omega$, $f(x)\delta x$ représente l'intensité des efforts appliqués sur le segment $(x, x + \delta x)$ de la corde. La fonction u représente le déplacement vertical de la corde à l'équilibre ; voir la figure 1.1. Enfin, la fonction α décrit les propriétés mécaniques de la corde. Si la corde est homogène, la fonction α est constante.

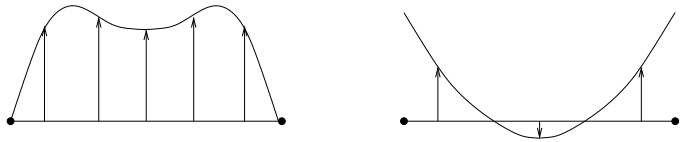


Figure 1.1 – Équilibre mécanique d'une corde tendue : déplacement de la corde à l'équilibre (à gauche) et densité linéique d'efforts appliqués (à droite). Lorsque la fonction α est constante et égale à 1, la fonction de droite est égale à l'opposé de la dérivée seconde de la fonction de gauche.

- **Équilibre thermique d'une barre chauffée.** La barre occupe le domaine Ω, la fonction inconnue u représente la distribution de température dans la barre, la fonction f la puissance linéique fournie et la fonction α la conductivité thermique de la barre. Si la barre est homogène, la fonction α est constante.

Les équations (1.1)–(1.2) interviennent également dans des modèles de diffusion et dans des modèles d'électrostatique.

Dans le problème (1.1)–(1.2), l'inconnue u est une fonction de Ω dans $\mathbb{R}$. La méthode des éléments finis permet de construire une approximation de cette fonction, c'est-à-dire une fonction de Ω dans $\mathbb{R}$ que l'on note u_h et telle que la différence $u - u_h$ en une certaine norme puisse être rendue suffisamment petite. Toutefois, avant d'étudier l'approximation du problème (1.1)–(1.2) par la méthode des éléments finis, il convient de préciser le cadre mathématique dans lequel on se place. L'objectif est de s'assurer que le problème (1.1)–(1.2) est *bien posé*, c'est-à-dire qu'il admet *une et une seule solution*. Pour cela, on reformule ce problème sous la forme suivante, appelée *forme faible*,

$$\left\{ \begin{array}{l} \text{Chercher } u \in V \text{ tel que} \\ \int_\Omega \alpha u' v' = \int_\Omega f v, \quad \forall v \in V, \end{array} \right. \tag{1.3}$$

où V est un espace fonctionnel (un espace vectoriel dont les éléments sont des fonctions) qui sera précisé par la suite. On suppose que les éléments de V s'annulent en a et en b. Formellement, l'équivalence entre (1.1)–(1.2) et (1.3) repose sur une intégration par parties. En effet, si u est solution de (1.1)–(1.2), alors en multipliant (1.1) par une *fonction test* v arbitraire dans V, en intégrant

par parties et en utilisant le fait que v s'annule en a et en b, il vient

$$\int_\Omega fv = \int_\Omega -(\alpha u')'v = \int_\Omega \alpha u'v' - [\alpha u'v]_a^b = \int_\Omega \alpha u'v'. \qquad (1.4)$$

Réciproquement, si u est solution de (1.3), on obtient en intégrant par parties le membre de gauche de (1.3),

$$\forall v \in V, \qquad \int_\Omega [f + (\alpha u')']v = 0. \qquad (1.5)$$

Puisque v est arbitraire dans V, on en déduit (1.1). De plus, par construction, $u \in V$ implique que u s'annule en a et en b, si bien que l'équation (1.2), qu'on appelle *condition aux limites*, est également satisfaite.

Afin d'établir le caractère bien posé de (1.3), il est nécessaire de préciser l'espace fonctionnel V. Un point important concerne le sens à donner aux dérivées. En effet, si le coefficient α est discontinu (ce qui est le cas dans les exemples ci-dessus lorsque la corde ou la barre est hétérogène), on ne peut pas donner un sens classique à la dérivée de $\alpha u'$ même si la fonction u est régulière. Pour remédier cette difficulté, on introduit la notion de distribution sur Ω et celle de dérivée au sens des distributions. Les distributions sur Ω constituent une généralisation naturelle de la notion de fonction : toute fonction intégrable (au sens de Lebesgue) sur Ω est une distribution sur Ω, mais il existe des distributions sur Ω qui ne peuvent pas être représentées par des fonctions (par exemple, la masse de Dirac). De plus, toute distribution sur Ω est *dérivable au sens des distributions*. Cette notion fournit une extension naturelle de la notion de dérivation au sens classique puisque pour toute fonction continûment différentiable sur Ω, sa dérivée usuelle et sa dérivée au sens des distributions coïncident. De plus, pour une fonction continue sur Ω et différentiable par morceaux, sa dérivée au sens des distributions s'évalue simplement en dérivant au sens usuel la fonction là où elle est dérivable. Ainsi, la dérivée au sens des distributions de la fonction $1-|x|$ sur $\Omega =]-1, 1[$ est la fonction valant 1 sur $]-1, 0[$ et -1 sur $]0, 1[$; voir la figure 1.2. Pour des rappels sur les bases mathématiques de la méthode des éléments finis, on renvoie à l'annexe A.

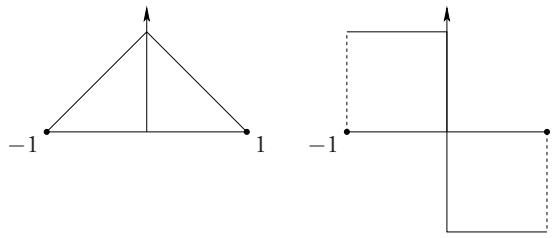


Figure 1.2 – Fonction $1-|x|$ (à gauche) et sa dérivée au sens des distributions (à droite).

On introduit les espaces fonctionnels suivants :

$$H^1(\Omega) = \{v \in L^2(\Omega)\ ;\ v' \in L^2(\Omega)\}, \tag{1.6}$$

$$H^1_0(\Omega) = \{v \in H^1(\Omega)\ ;\ v(a) = v(b) = 0\}, \tag{1.7}$$

les dérivées étant entendues au sens des distributions[1]. On équipe les espaces $H^1(\Omega)$ et $H^1_0(\Omega)$ de la norme

$$\|v\|_{1,\Omega} = \left(\|v\|^2_{0,\Omega} + \|v'\|^2_{0,\Omega}\right)^{\frac{1}{2}}, \tag{1.8}$$

où $\|\cdot\|_{0,\Omega}$ désigne la norme canonique de $L^2(\Omega)$: pour $v \in L^2(\Omega)$, on a $\|v\|_{0,\Omega} = (\int_\Omega v^2)^{\frac{1}{2}}$. Un résultat classique d'analyse fonctionnelle montre qu'équipés de cette norme, les espaces $H^1(\Omega)$ et $H^1_0(\Omega)$ sont des espaces de Hilbert. Par ailleurs, on pose

$$|v|_{1,\Omega} = \|v'\|_{0,\Omega}. \tag{1.9}$$

On notera que $|\cdot|_{1,\Omega}$ est une *semi-norme* (et non une norme) sur $H^1(\Omega)$ car $|v|_{1,\Omega} = 0$ n'implique pas $v = 0$ (la fonction v peut être constante sur Ω).

1. Les éléments de $H^1(\Omega)$ sont des fonctions définies presque partout sur Ω, c'est-à-dire que ces fonctions sont définies partout sur Ω sauf sur un ensemble de mesure nulle. Il n'est donc pas évident *a priori* que l'on puisse parler de la valeur de ces fonctions en a ou en b. En fait, un résultat classique d'analyse fonctionnelle montre que si $v \in H^1(\Omega)$, les valeurs prises par v en a et en b ont bien un sens ; voir la section A.4.

Par la suite, on considère également les espaces fonctionnels suivants, qu'on appelle *espaces de Sobolev* : pour un entier $s \geqslant 1$,

$$H^s(\Omega) = \{v \in L^2(\Omega)\ ;\ \forall k \in \{1, \ldots, s\},\ v^{(k)} \in L^2(\Omega)\}, \tag{1.10}$$

où $v^{(k)}$ désigne la dérivée d'ordre k de v (au sens des distributions). Équipé de la norme

$$\|v\|_{s,\Omega} = \left(\|v\|^2_{0,\Omega} + \|v'\|^2_{0,\Omega} + \ldots + \|v^{(s)}\|^2_{0,\Omega}\right)^{\frac{1}{2}} = \left(\sum_{k=0}^{s} \|v^{(k)}\|^2_{0,\Omega}\right)^{\frac{1}{2}}, \tag{1.11}$$

$H^s(\Omega)$ est un espace de Hilbert. Par ailleurs, on introduit la semi-norme

$$|v|_{s,\Omega} = \|v^{(s)}\|_{0,\Omega}. \tag{1.12}$$

Il s'agit d'une norme et non d'une semi-norme car $|v|_{s,\Omega} = 0$ si la fonction v est un polynôme de degré inférieur ou égal à $(s-1)$.
On formule le problème (1.3) sous la forme suivante :

$$\begin{cases} \text{Chercher } u \in H^1_0(\Omega) \text{ tel que} \\ \int_\Omega \alpha u' v' = \int_\Omega f v, \quad \forall v \in H^1_0(\Omega). \end{cases} \tag{1.13}$$

On suppose que $f \in L^2(\Omega)$ et que la fonction $\alpha : \Omega \to \mathbb{R}$ est d'une part minorée sur Ω par un réel α_0 strictement positif et d'autre part majorée sur Ω par un réel α_1. On a le résultat suivant.

Proposition 1.1. *Avec les hypothèses ci-dessus, le problème* (1.13) *est bien posé.*

Le caractère bien posé du problème (1.13) résulte du lemme de Lax–Milgram ; voir la section 2.1.1. En particulier, on utilise le fait que

$$\forall v \in H^1_0(\Omega), \qquad \int_\Omega \alpha (v')^2 \geqslant \alpha_0 |v|^2_{1,\Omega} \geqslant c_\Omega \alpha_0 \|v\|^2_{1,\Omega}, \tag{1.14}$$

où c_Ω est une constante strictement positive ne dépendant que de la mesure de l'intervalle Ω. La première minoration résulte de l'hypothèse sur la fonction α. La deuxième minoration est une conséquence de *l'inégalité de Poincaré* ; voir le lemme 5.1 et la section A.4.

1.2 Principes de la méthode des éléments finis

La méthode des éléments finis repose sur deux principes : d'une part, la formulation d'un problème approché par la méthode de Galerkin (ou une variante de celle-ci) ; d'autre part, la construction d'un espace d'approximation (de dimension finie) à l'aide d'un maillage, de fonctions polynômiales par morceaux et de degrés de liberté sur chaque maille.

1.2.1 Le problème approché

On cherche une solution approchée du problème (1.13) en y remplaçant l'espace de dimension infinie $H_0^1(\Omega)$ par un sous-espace de dimension finie. En notant $V_h \subset H_0^1(\Omega)$ ce sous-espace de dimension finie, qu'on appelle *espace d'approximation*, le problème approché consiste à

$$\left\{ \begin{array}{l} \text{Chercher } u_h \in V_h \text{ tel que} \\ \displaystyle\int_\Omega \alpha u_h' v_h' = \int_\Omega f v_h, \quad \forall v_h \in V_h. \end{array} \right. \tag{1.15}$$

La méthode d'approximation introduite ci-dessus porte le nom de *méthode de Galerkin*. Elle est présentée ici sous sa forme la plus simple. Plusieurs variantes sont étudiées dans le chapitre 2.

Le problème approché (1.15) n'est rien d'autre qu'un système linéaire. En effet, soit $\{\varphi_1, \ldots, \varphi_N\}$ une base de V_h où N désigne la dimension de V_h. On décompose la solution approchée u_h dans cette base selon

$$u_h = \sum_{i=1}^{N} U_i \varphi_i, \tag{1.16}$$

et on introduit le vecteur U de $\mathbb{R}^N$ formé par les composantes de u_h dans cette base, $U = (U_i)_{1 \leqslant i \leqslant N}$. Soit $\mathcal{A} \in \mathbb{R}^{N,N}$ la *matrice de rigidité* dont les composantes sont

$$\mathcal{A}_{ij} = \int_\Omega \alpha \varphi_i' \varphi_j', \qquad i,j \in \{1, \ldots, N\}, \tag{1.17}$$

et soit $F \in \mathbb{R}^N$ le vecteur de composantes

$$F_i = \int_\Omega f\varphi_i, \qquad i \in \{1, \ldots, N\}. \tag{1.18}$$

Un calcul élémentaire montre que u_h est solution de (1.15) si et seulement si

$$AU = F. \tag{1.19}$$

La méthode de Galerkin permet donc de remplacer un problème posé en dimension infinie par un système linéaire.

Grâce à l'inégalité (1.14), on montre que la matrice de rigidité A est *définie positive*. Le système linéaire (1.19) admet donc une et une seule solution et il en va de même du problème approché (1.15).

La prochaine question qui se pose est de savoir si la solution approchée u_h est une bonne approximation de la solution exacte u. Pour répondre à cette question, on dispose de *l'estimation d'erreur* suivante. Il s'agit d'un cas particulier du lemme de Céa qui sera énoncé au chapitre 2 sous une forme un peu plus abstraite.

Proposition 1.2. *Il existe une constante c, indépendante du choix de l'espace d'approximation V_h, telle que*

$$\|u - u_h\|_{1,\Omega} \leqslant c \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega}. \tag{1.20}$$

La quantité $\inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega}$ s'interprète comme la distance de la solution exacte u à l'espace d'approximation V_h (pour la distance induite par la norme $\|\cdot\|_{1,\Omega}$). L'estimation d'erreur (1.20) montre que la solution approchée n'est pas « trop loin » de la plus proche fonction à u dans V_h.

L'estimation (1.20) résulte de la *relation d'orthogonalité de Galerkin* que l'on retrouvera au chapitre 2.

Lemme 1.3. *Pour tout $v_h \in V_h$, on a*

$$\int_\Omega \alpha(u - u_h)' v_h' = 0. \tag{1.21}$$

La preuve du lemme 1.3 est immédiate. Pour tout $v_h \in V_h$, il vient

$$\int_\Omega \alpha u_h' v_h' = \int_\Omega f v_h = \int_\Omega \alpha u' v_h', \tag{1.22}$$

la dernière égalité résultant du fait que $V_h \subset H_0^1(\Omega)$. En utilisant l'inégalité (1.14), on déduit que pour tout $v_h \in V_h$,

$$c_\Omega \alpha_0 \|u - u_h\|_{1,\Omega}^2 \leqslant \int_\Omega \alpha (u - u_h)'(u - u_h)' \leqslant \int_\Omega \alpha (u - u_h)'(u - v_h)' \leqslant \alpha_1 \|u - u_h\|_{1,\Omega} \|u - v_h\|_{1,\Omega}, \tag{1.23}$$

d'où l'estimation (1.20) avec $c = \frac{1}{c_\Omega} \frac{\alpha_1}{\alpha_0}$ puisque la fonction v_h est arbitraire dans V_h.

1.2.2 Construction de l'espace d'approximation

La première étape dans la construction de l'espace d'approximation V_h consiste à *mailler* l'intervalle Ω. En une dimension d'espace, un maillage de $\Omega =]a, b[$ est une collection indexée d'intervalles, $\{I_i = [x_{1,i}, x_{2,i}]\}_{1 \leqslant i \leqslant N_{\rm ma}}$, tous de mesure non-nulle, et formant une partition de Ω. En d'autres termes, on a

$$[a, b] = \bigcup_{i=1}^{N_{\rm ma}} [x_{1,i}, x_{2,i}] \quad \text{et} \quad]x_{1,i}, x_{2,i}[\cap]x_{1,j}, x_{2,j}[= \varnothing \quad \text{pour } i \neq j. \tag{1.24}$$

Les intervalles I_i sont appelés les *mailles* (ou les *éléments* ou les *cellules* du maillage) et l'entier $N_{\rm ma}$ désigne le nombre total de mailles. La façon la plus simple de construire un maillage est de choisir $(N_{\rm ma} + 1)$ points distincts de $\overline{\Omega}$ tels que

$$a = x_1 < x_2 < \ldots < x_{N_{\rm ma}} < x_{N_{\rm ma}+1} = b, \tag{1.25}$$

et de poser $x_{1,i} = x_i$ et $x_{2,i} = x_{i+1}$ pour tout $i \in \{1, \ldots, N_{\rm ma}\}$. Les points de l'ensemble $\{x_1, \ldots, x_{N_{\rm ma}+1}\}$ sont appelés les *sommets* du maillage. On désigne par $N_{\rm so}$ le nombre de sommets du maillage. En une dimension d'espace, on a donc

$$N_{\rm so} = N_{\rm ma} + 1. \tag{1.26}$$

Le maillage est *a priori* de pas variable. On pose pour tout $i \in \{1, \ldots, N_{\rm ma}\}$,

$$h_i = x_{i+1} - x_i \quad \text{et} \quad h = \max_{1 \leqslant i \leqslant N_{\rm ma}} h_i. \tag{1.27}$$

On dit que le maillage est *uniforme* lorsque $h_i = h$ pour tout $i \in \{1, \ldots, N_{\mathrm{ma}}\}$. Par la suite, le maillage est désigné sous la forme $\mathcal{T}_h = \{I_i\}_{1 \leqslant i \leqslant N_{\mathrm{ma}}}$, l'indice h indiquant la finesse globale du maillage.

La deuxième étape dans la construction de l'espace d'approximation consiste à choisir des *fonctions de forme* sur chaque maille. En d'autres termes, les fonctions de V_h sont telles que leur restriction à chaque maille $I_i \in \mathcal{T}_h$ est dans tel ou tel espace polynômial.

Définition 1.4. *Soit un entier $k \geqslant 1$. En une dimension d'espace, on désigne par $\mathbb{P}_k$ l'espace vectoriel des polynômes à coefficients réels de degré inférieur ou égal à k.*

On pose

$$W_h = \{w_h \in L^2(\Omega) \, ; \, \forall i \in \{1, \ldots, N_{\mathrm{ma}}\}, \; w_h|_{I_i} \in \mathbb{P}_k\}. \tag{1.28}$$

Il est clair que W_h est un espace de dimension finie, sa dimension étant égale à $(k + 1) \times N_{\mathrm{ma}}$. Toutefois, W_h ne peut pas être utilisé tel quel dans le problème approché (1.15) car il n'est pas inclus dans $H_0^1(\Omega)$. En effet, une fonction $w_h \in W_h$ peut être discontinue aux interfaces entre les mailles et un résultat classique d'analyse fonctionnelle montre que dans ces conditions, $w_h \notin H^1(\Omega)$. De plus, une fonction $w_h \in W_h$ n'est pas nécessairement nulle en a et en b. On pose donc

$$V_h = W_h \cap H_0^1(\Omega). \tag{1.29}$$

Les sections 1.3 et 1.4 présentent des exemples concrets d'espaces d'approximation V_h.

1.3 Élément fini de Lagrange $\mathbb{P}_1$

On considère les espaces vectoriels suivants :

$$P^1_{\mathrm{c},h} = \{v_h \in \mathcal{C}^0(\overline{\Omega}) \, ; \, \forall i \in \{1, \ldots, N_{\mathrm{ma}}\}, \; v_h|_{I_i} \in \mathbb{P}_1\}, \tag{1.30}$$

$$P^1_{\mathrm{c},h,0} = \{v_h \in P^1_{\mathrm{c},h} \, ; \, v_h(a) = v_h(b) = 0\}, \tag{1.31}$$

dont les éléments sont des fonctions *continues* et affines par morceaux. Les fonctions de $P^1_{c,h}$ sont dérivables (au sens classique) sur chaque maille ; elles sont de plus continues aux interfaces entre les mailles. Un résultat d'analyse fonctionnelle conduit alors au résultat suivant.

Proposition 1.5. *$P^1_{c,h} \subset H^1(\Omega)$ et $P^1_{c,h,0} \subset H^1_0(\Omega)$.*

On introduit la famille de fonctions $\{\varphi_1, \ldots, \varphi_{N_{so}}\}$ que l'on définit localement sur chaque maille de la manière suivante : pour tout $i \in \{2, \ldots, N_{so} - 1\}$,

$$\varphi_i(x) = \begin{cases} \frac{1}{h_{i-1}}(x - x_{i-1}) & \text{si } x \in I_{i-1}, \\ \frac{1}{h_i}(x_{i+1} - x) & \text{si } x \in I_i, \\ 0 & \text{sinon,} \end{cases} \tag{1.32}$$

et (on rappelle que $N_{so} - 1 = N_{ma}$)

$$\begin{aligned} \varphi_1(x) &= \begin{cases} \frac{1}{h_1}(x_2 - x) & \text{si } x \in I_1, \\ 0 & \text{sinon,} \end{cases} \\ \varphi_{N_{so}}(x) &= \begin{cases} \frac{1}{h_{N_{so}-1}}(x - x_{N_{so}-1}) & \text{si } x \in I_{N_{so}-1}, \\ 0 & \text{sinon.} \end{cases} \end{aligned} \tag{1.33}$$

Il est clair que $\varphi_i \in P^1_{c,h}$ pour tout $i \in \{1, \ldots, N_{so}\}$ et que $\varphi_i \in P^1_{c,h,0}$ pour tout $i \in \{2, \ldots, N_{so} - 1\}$.

Pour tout $i \in \{1, \ldots, N_{so}\}$, la fonction φ_i vaut 1 au sommet x_i et 0 aux autres sommets du maillage. On a donc

$$\varphi_i(x_j) = \delta_{ij}, \qquad i, j \in \{1, \ldots, N_{so}\}, \tag{1.34}$$

où δ_{ij} désigne le symbole de Kronecker tel que $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$. Les fonctions φ_i sont appelées *fonctions chapeau* en référence à la forme de leur graphe ; voir la figure 1.3. La dérivée au sens des distributions de la

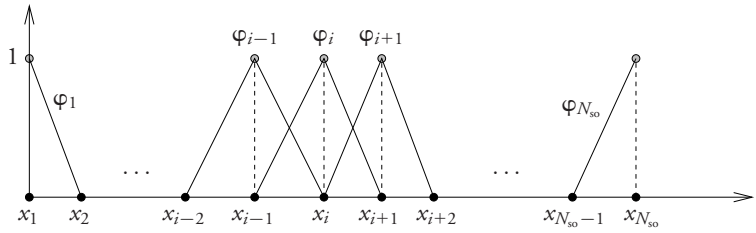


Figure 1.3 – Fonctions de forme dans l'espace d'approximation $P^1_{c,h}$: fonctions chapeau.

fonction φ_i s'exprime sous la forme

$$\varphi_i'(x) = \begin{cases} \frac{1}{h_{i-1}} & \text{si } x \in I_{i-1}, \\ -\frac{1}{h_i} & \text{si } x \in I_i, \\ 0 & \text{sinon.} \end{cases} \tag{1.35}$$

Il s'agit donc d'une fonction constante par morceaux.

La famille $\{\varphi_1, \ldots, \varphi_{N_{so}}\}$ est une base de $P^1_{c,h}$. En effet, cette famille est clairement libre puisque si la fonction

$$w = \sum_{j=1}^{N_{so}} \alpha_j \varphi_j \tag{1.36}$$

est identiquement nulle sur Ω, il est clair que pour tout $i \in \{1, \ldots, N_{so}\}$, on a

$$w(x_i) = \sum_{j=1}^{N_{so}} \alpha_j \varphi_j(x_i) = \sum_{j=1}^{N_{so}} \alpha_j \delta_{ij} = \alpha_i = 0. \tag{1.37}$$

Cette famille est également génératrice de $P^1_{c,h}$. En effet, soit $v_h \in P^1_{c,h}$.
On pose

$$w_h = \sum_{j=1}^{N_{so}} v_h(x_j) \varphi_j. \tag{1.38}$$

Sur chaque intervalle $I_i \in \mathcal{T}_h$, $i \in \{1, \ldots, N_{\mathrm{ma}}\}$, les fonctions v_h et w_h sont affines et coïncident en deux points (les extrémités de la maille I_i). Par conséquent, ces fonctions sont égales. On en déduit que $v_h = w_h$ sur Ω, ce qui montre que toute fonction de $P^1_{\mathrm{c},h}$ peut s'écrire comme une combinaison linéaire des fonctions $\{\varphi_1, \ldots, \varphi_{N_{\mathrm{so}}}\}$.

Pour tout $i \in \{1, \ldots, N_{\mathrm{so}}\}$, on définit la forme linéaire

$$\gamma_i : \mathcal{C}^0(\overline{\Omega}) \ni v \longmapsto v(x_i) \in \mathbb{R}. \tag{1.39}$$

Il est clair que pour tout $i, j \in \{1, \ldots, N_{\mathrm{so}}\}$,

$$\gamma_i(\varphi_j) = \delta_{ij}. \tag{1.40}$$

Proposition 1.6.

(i) *La famille* $\{\varphi_1, \ldots, \varphi_{N_{\mathrm{so}}}\}$ *est une base de* $P^1_{\mathrm{c},h}$ *et la famille* $\{\gamma_1, \ldots, \gamma_{N_{\mathrm{so}}}\}$ *est une base de* $\mathcal{L}(P^1_{\mathrm{c},h}; \mathbb{R})$.

(ii) *La famille* $\{\varphi_2, \ldots, \varphi_{N_{\mathrm{so}}-1}\}$ *est une base de* $P^1_{\mathrm{c},h,0}$ *et la famille* $\{\gamma_2, \ldots, \gamma_{N_{\mathrm{so}}-1}\}$ *est une base de* $\mathcal{L}(P^1_{\mathrm{c},h,0}; \mathbb{R})$.

Corollaire 1.7.

$$\dim P^1_{\mathrm{c},h} = N_{\mathrm{so}} = N_{\mathrm{ma}} + 1 \quad \textit{et} \quad \dim P^1_{\mathrm{c},h,0} = N_{\mathrm{so}} - 2 = N_{\mathrm{ma}} - 1.$$

Définition 1.8.

(i) *Les formes linéaires* $\{\gamma_1, \ldots, \gamma_{N_{\mathrm{so}}}\}$ *sont appelées les* degrés de liberté *dans* $P^1_{\mathrm{c},h}$ *et les fonctions* $\{\varphi_1, \ldots, \varphi_{N_{\mathrm{so}}}\}$ *sont appelées les* fonctions de forme *dans* $P^1_{\mathrm{c},h}$.

(ii) *Les formes linéaires* $\{\gamma_2, \ldots, \gamma_{N_{\mathrm{so}}-1}\}$ *sont appelées les* degrés de liberté *dans* $P^1_{\mathrm{c},h,0}$ *et les fonctions* $\{\varphi_2, \ldots, \varphi_{N_{\mathrm{so}}-1}\}$ *sont appelées les* fonctions de forme *dans* $P^1_{\mathrm{c},h,0}$.

On introduit *l'opérateur d'interpolation* suivant :

$$\mathcal{I}^1_{\mathrm{c},h} : \mathcal{C}^0(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{N_{\mathrm{so}}} \gamma_i(v)\varphi_i \in P^1_{\mathrm{c},h}. \tag{1.41}$$

Pour une fonction $v \in \mathcal{C}^0(\overline{\Omega})$, $\mathcal{I}^1_{\mathrm{c},h} v$ est l'unique fonction continue et affine par morceaux qui prend les mêmes valeurs que v aux N_{so} sommets du

maillage ; voir la figure 1.4. La fonction $\mathcal{I}^1_{c,h}v$ est appelée *l'interpolé de Lagrange* de v de degré 1.

En une dimension d'espace, les fonctions de $H^1(\Omega)$ sont continues. Par conséquent, $\mathcal{I}^1_{c,h}$ peut également être vu comme un opérateur de $H^1(\Omega)$ dans $H^1(\Omega)$. On montre que cet opérateur est continu et que sa norme $\|\mathcal{I}^1_{c,h}\|_{\mathcal{L}(H^1(\Omega);H^1(\Omega))}$ est uniformément bornée en h. En d'autres termes, il existe une constante c, indépendante de h, telle que pour tout $v \in H^1(\Omega)$,

$$\|\mathcal{I}^1_{c,h}v\|_{1,\Omega} \leqslant c\,\|v\|_{1,\Omega}. \tag{1.42}$$

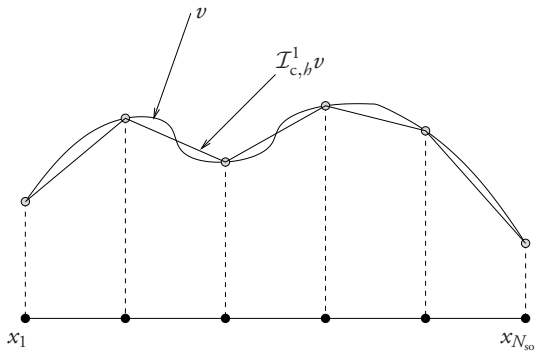


Figure 1.4 – Interpolé de Lagrange de degré 1.

Par ailleurs, on souhaite connaître la précision de l'opérateur d'interpolation $\mathcal{I}^1_{c,h}$, c'est-à-dire que pour toute fonction v suffisamment régulière, on souhaite estimer l'erreur d'interpolation $v - \mathcal{I}^1_{c,h}v$ dans une certaine norme. On a le résultat suivant.

Proposition 1.9. *Pour tout h et pour tout $v \in H^2(\Omega)$, on a*

$$\|v - \mathcal{I}^1_{c,h}v\|_{0,\Omega} \leqslant h^2|v|_{2,\Omega} \qquad \textit{et} \qquad |v - \mathcal{I}^1_{c,h}v|_{1,\Omega} \leqslant h|v|_{2,\Omega}. \tag{1.43}$$

On dit que l'erreur d'interpolation en norme L^2 est d'ordre 2 en h et qu'elle est d'ordre 1 en h en semi-norme H^1 (et donc également en norme H^1). La preuve de la proposition 1.9 est à la fois relativement simple et instructive.

(i) On considère un intervalle $I_i \in \mathcal{T}_h$. Soit $w \in H^1(I_i)$ une fonction qui s'annule en (au moins) un point ξ dans I_i. Alors, pour tout $x \in I_i$, on a

$$|w(x)| = |w(x) - w(\xi)| \leqslant \int_\xi^x |w'(s)|\,\mathrm{d}s \leqslant \left(\int_\xi^x \mathrm{d}s\right)^{\frac{1}{2}} \left(\int_\xi^x |w'(s)|^2\,\mathrm{d}s\right)^{\frac{1}{2}} \leqslant h_i^{\frac{1}{2}} |w|_{1,I_i},$$

grâce à l'inégalité de Cauchy–Schwarz. On en déduit $\|w\|_{0,I_i} \leqslant h_i |w|_{1,I_i}$.

(ii) Soit $v \in H^2(\Omega)$ et soit $i \in \{1, \ldots, N_{\mathrm{ma}}\}$. On pose $\theta_i = (v - \mathcal{I}^1_{\mathrm{c},h} v)|_{I_i}$ et $w_i = \theta_i'$. Il est clair que $w_i \in H^1(I_i)$ et d'après le théorème des accroissements finis, w_i s'annule en (au moins) un point ξ dans I_i. On déduit de l'étape (i) ci-dessus que $\|w_i\|_{0,I_i} \leqslant h_i |w_i|_{1,I_i}$. Par conséquent,

$$|v - \mathcal{I}^1_{\mathrm{c},h} v|_{1,I_i} = \|w_i\|_{0,I_i} \leqslant h_i |w_i|_{1,I_i} = h_i |v|_{2,I_i},$$

puisque la fonction $(\mathcal{I}^1_{\mathrm{c},h} v)''$ est identiquement nulle sur I_i. En sommant les estimations ci-dessus sur toutes les mailles, on obtient la deuxième majoration dans (1.43).

(iii) Afin de prouver la première majoration dans (1.43), on observe que l'estimation de l'étape (i) ci-dessus peut être appliquée à $(v - \mathcal{I}^1_{\mathrm{c},h} v)|_{I_i}$ avec $\xi = x_i$, ce qui donne

$$\|v - \mathcal{I}^1_{\mathrm{c},h} v\|_{0,I_i} \leqslant h_i |v - \mathcal{I}^1_{\mathrm{c},h} v|_{1,I_i} \leqslant h_i^2 |v|_{2,I_i}.$$

On conclut en sommant sur les mailles.

Cette preuve illustre très clairement le fait que les propriétés interpolantes de l'opérateur $\mathcal{I}^1_{\mathrm{c},h}$ sont purement *locales*. On établit d'abord une estimation de l'erreur d'interpolation sur chaque maille, puis on déduit les estimations globales (1.43) en sommant les contributions des différentes mailles. Cette observation motive l'approche adoptée dans les chapitres 3 et 4 où :

(i) on définit un élément fini et l'opérateur d'interpolation associé en adoptant un point de vue local sur une maille ;

(ii) puis, on construit un opérateur d'interpolation global en maillant le domaine Ω.

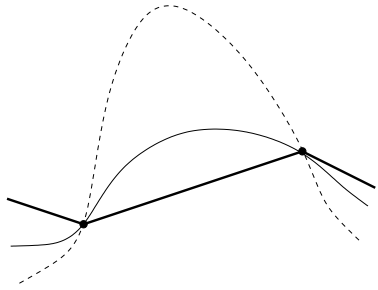


Figure 1.5 – Une fonction à interpoler dont la dérivée seconde n'est pas grande (trait fin), une fonction à interpoler dont le graphe présente une forte courbure (trait pointillé) ; ces deux fonctions ont ici le même interpolé de Lagrange de degré 1 (trait gras).

Remarque 1.10

Le fait que la dérivée seconde de v intervienne dans les estimations (1.43) est relativement naturel dans la mesure où plus cette dérivée seconde est grande, plus le graphe de la fonction v est courbe, d'où une plus grande déviation par rapport à un interpolé affine par morceaux ; voir la figure 1.5 pour une illustration graphique. Par ailleurs, si la fonction à interpoler n'est pas suffisamment régulière pour être dans $H^2(\Omega)$, on dispose des majorations suivantes :

$$\forall h,\ \|v - \mathcal{I}^1_{c,h} v\|_{0,\Omega} \leqslant h|v|_{1,\Omega} \qquad \text{et} \qquad \lim_{h \to 0} |v - \mathcal{I}^1_{c,h} v|_{1,\Omega} = 0,$$

qui montrent que l'erreur d'interpolation en norme H^1 tend vers zéro et que l'erreur d'interpolation en norme L^2 converge à l'ordre 1 en h.

1.4 Élément fini de Lagrange $\mathbb{P}_k$

Soit un entier $k \geqslant 1$. On considère les espaces vectoriels suivants :

$$P^k_{c,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega})\ ;\ \forall i \in \{1, \ldots, N_{\text{ma}}\},\ v_h|_{I_i} \in \mathbb{P}_k\}, \tag{1.44}$$

$$P^k_{c,h,0} = \{v_h \in P^k_{c,h}\ ;\ v_h(a) = v_h(b) = 0\}, \tag{1.45}$$

dont les éléments sont des fonctions *continues* et polynômiales de degré k par morceaux. On a le résultat suivant.

Proposition 1.11. *$P^k_{c,h} \subset H^1(\Omega)$ et $P^k_{c,h,0} \subset H^1_0(\Omega)$.*

Afin d'exhiber les fonctions de forme dans $P^k_{c,h}$ et $P^k_{c,h,0}$, on introduit les polynômes d'interpolation de Lagrange.

Définition 1.12 (Polynômes d'interpolation de Lagrange). *Soit un entier $k \geqslant 1$. On considère une famille $\mathcal{F} = \{s_0, \ldots, s_k\}$ constituée de $(k+1)$ réels distincts. Les polynômes d'interpolation de Lagrange $\{\mathcal{L}^{\mathcal{F}}_0, \ldots, \mathcal{L}^{\mathcal{F}}_k\}$ associés à la famille $\mathcal{F}$ sont définis comme suit :*

$$\mathcal{L}^{\mathcal{F}}_m(t) = \frac{\prod_{l \neq m}(t - s_l)}{\prod_{l \neq m}(s_m - s_l)}, \qquad m \in \{0, \ldots, k\}. \tag{1.46}$$

Par construction, on a

$$\mathcal{L}^{\mathcal{F}}_m(s_n) = \delta_{mn}, \qquad m, n \in \{0, \ldots, k\}. \tag{1.47}$$

Par la suite, les polynômes d'interpolation de Lagrange associés à la famille de réels $\{\frac{m}{k}\}_{0 \leqslant m \leqslant k}$ équirépartis sur l'intervalle $[0, 1]$ sont notés $\{\mathcal{L}^k_0, \ldots, \mathcal{L}^k_k\}$. Le tableau 1.1 contient une représentation graphique ainsi que l'expression analytique de ces polynômes pour $k \in \{1, 2, 3\}$.

Soit $i \in \{1, \ldots, N_{\text{ma}}\}$. On considère la famille de $(k+1)$ réels équirépartis sur la maille $I_i \in \mathcal{T}_h$ telle que

$$\mathcal{F}_i = \{x_i + \tfrac{m}{k} h_i\}_{0 \leqslant m \leqslant k}. \tag{1.48}$$

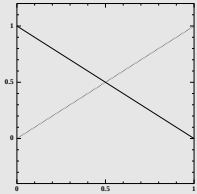
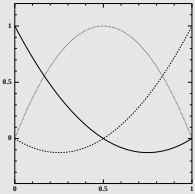
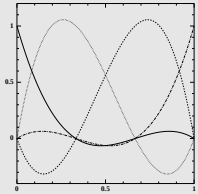
Un simple changement de variables montre que les polynômes d'interpolation de Lagrange associés à la famille $\mathcal{F}_i$ s'expriment sous la forme

$$\mathcal{L}^{\mathcal{F}_i}_m(t) = \mathcal{L}^k_m\left(\tfrac{t - x_i}{h_i}\right), \qquad m \in \{0, \ldots, k\}. \tag{1.49}$$

On regroupe les N_{ma} familles de réels $\mathcal{F}_i$ en une seule grande famille. En comptant une seule fois les réels qui se trouvent aux extrémités des intervalles (voir la figure 1.6 pour un exemple avec $k = 3$), on obtient une famille de $(kN_{\text{ma}} + 1)$ réels distincts que l'on note

$$\{a_1, \ldots, a_{kN_{\text{ma}}+1}\}. \tag{1.50}$$

Tableau 1.1 – Polynômes d'interpolation de Lagrange $\{\mathcal{L}_0^k, \ldots, \mathcal{L}_k^k\}$ pour $k \in \{1, 2, 3\}$: représentation graphique et expression analytique.

$k = 1$	$k = 2$	$k = 3$
		
$\mathcal{L}_0^1(t) = 1 - t$ $\mathcal{L}_1^1(t) = t$	$\mathcal{L}_0^2(t) = (2t-1)(t-1)$ $\mathcal{L}_1^2(t) = 4t(1-t)$ $\mathcal{L}_2^2(t) = t(2t-1)$	$\mathcal{L}_0^3(t) = -\frac{1}{2}(3t-1)(3t-2)(t-1)$ $\mathcal{L}_1^3(t) = \frac{9}{2}t(3t-2)(t-1)$ $\mathcal{L}_2^3(t) = -\frac{9}{2}t(3t-1)(t-1)$ $\mathcal{L}_3^3(t) = \frac{1}{2}t(3t-1)(3t-2)$

Les réels a_j sont appelés les *nœuds* du maillage[1]. On note N_{no} le nombre de nœuds du maillage. On a donc

$$N_{\mathrm{no}} = kN_{\mathrm{ma}} + 1. \tag{1.51}$$

Pour un nœud a_j avec $j \in \{1, \ldots, N_{\mathrm{no}}\}$, on effectue la division euclidienne de $(j-1)$ par k sous la forme

$$j - 1 = k(\mathtt{i}(j) - 1) + \mathtt{m}(j), \tag{1.52}$$

avec $\mathtt{i}(j) \in \{1, \ldots, N_{\mathrm{so}}\}$ et $\mathtt{m}(j) \in \{0, \ldots, k-1\}$. Lorsque $\mathtt{m}(j) \neq 0$, le nœud a_j se trouve à l'intérieur de l'intervalle $I_{\mathtt{i}(j)}$. Lorsque $\mathtt{m}(j) = 0$, le nœud a_j coïncide avec le sommet $x_{\mathtt{i}(j)}$ du maillage.

On introduit la famille de fonctions $\{\varphi_1, \ldots, \varphi_{N_{\mathrm{no}}}\}$ définies de la manière suivante : pour tout $j \in \{1, \ldots, N_{\mathrm{no}}\}$,

1. Pour $k = 1$, la notion de nœud coïncide avec celle de sommet.

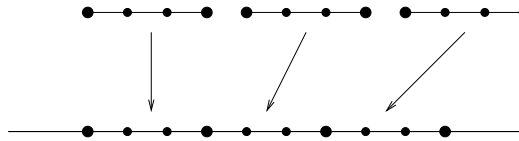


Figure 1.6 – Assemblage des nœuds du maillage pour l'espace d'approximation $P^3_{c,h}$.

(i) si $\mathtt{m}(j) \neq 0$, la fonction φ_j est définie localement sur chaque maille par

$$\varphi_j(x) = \begin{cases} \mathcal{L}_{\mathtt{m}(j)}^{\mathcal{F}_{\mathtt{i}(j)}}(x) & \text{si } x \in I_{\mathtt{i}(j)}, \\ 0 & \text{sinon;} \end{cases} \tag{1.53}$$

(ii) si $\mathtt{m}(j) = 0$, la fonction φ_j est définie localement sur chaque maille par

$$\varphi_j(x) = \begin{cases} \mathcal{L}_{k}^{\mathcal{F}_{\mathtt{i}(j)-1}}(x) & \text{si } x \in I_{\mathtt{i}(j)-1}, \\ \mathcal{L}_{0}^{\mathcal{F}_{\mathtt{i}(j)}}(x) & \text{si } x \in I_{\mathtt{i}(j)}, \\ 0 & \text{sinon.} \end{cases} \tag{1.54}$$

Si $\mathtt{i}(j) = 1$ ou $\mathtt{i}(j) = N_{\mathrm{so}}$, seulement un des deux intervalles intervient dans la définition ci-dessus.

Il est clair que $\varphi_j \in P^k_{c,h}$ pour $j \in \{1, \ldots, N_{\mathrm{no}}\}$ et que $\varphi_j \in P^k_{c,h,0}$ pour $j \in \{2, \ldots, N_{\mathrm{no}} - 1\}$. De plus, par construction, on a

$$\varphi_j(a_{j'}) = \delta_{jj'}, \qquad j, j' \in \{1, \ldots, N_{\mathrm{no}}\}. \tag{1.55}$$

Enfin, on notera que la dérivée au sens des distributions de φ_j est une fonction polynômiale de degré $(k-1)$ par morceaux. Cette fonction est discontinue aux sommets du maillage.

La figure 1.7 présente une illustration graphique des fonctions φ_j pour $k = 2$. Pour $i \in \{1, \ldots, N_{\mathrm{ma}}\}$, on note $x_{i+\frac{1}{2}}$ le point milieu de l'intervalle I_i. On observera la différence de support entre les fonctions associées aux sommets du maillage (le support est constitué de deux mailles) et celles associées aux milieux des mailles (le support est réduit à la maille correspondante).

La famille $\{\varphi_1, \ldots, \varphi_{N_{\mathrm{no}}}\}$ est une base de $P^k_{c,h}$. En effet, cette famille est clairement libre puisque si la fonction $w = \sum_{j=1}^{N_{\mathrm{no}}} \alpha_j \varphi_j$ est identiquement nulle sur

Ω, il est clair que pour tout $i \in \{1, \ldots, N_{\rm no}\}$, on a $w(a_i) = \alpha_i = 0$. Cette famille est également génératrice de $P^k_{{\rm c},h}$. En effet, soit $v_h \in P^k_{{\rm c},h}$. On pose $w_h = \sum_{j=1}^{N_{\rm no}} v_h(a_j)\varphi_j$. Sur chaque intervalle $I_i \in \mathcal{T}_h$, $i \in \{1, \ldots, N_{\rm ma}\}$, les fonctions v_h et w_h sont des polynômes de degré k qui coïncident en $(k+1)$ points (les nœuds $\{a_{k(i-1)+m+1}\}_{0 \leqslant m \leqslant k}$ situés dans I_i). Par conséquent, ces fonctions sont égales. On en déduit que $v_h = w_h$ sur Ω, ce qui montre que toute fonction de $P^k_{{\rm c},h}$ peut s'écrire comme une combinaison linéaire des fonctions $\{\varphi_1, \ldots, \varphi_{N_{\rm no}}\}$.

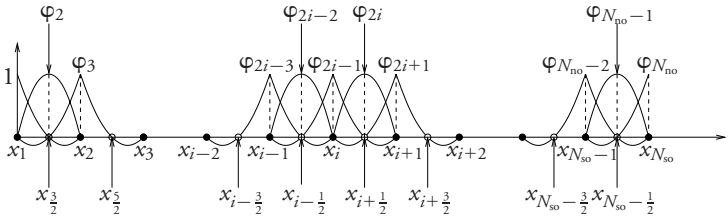


Figure 1.7 – Fonctions de forme dans l'espace d'approximation $P^2_{{\rm c},h}$.

Pour tout $j \in \{1, \ldots, N_{\rm no}\}$, on introduit la forme linéaire

$$\gamma_j : \mathcal{C}^0(\overline{\Omega}) \ni v \longmapsto v(a_j) \in \mathbb{R}. \tag{1.56}$$

Il est clair que $\gamma_j(\varphi_{j'}) = \delta_{jj'}$ pour tout $j, j' \in \{1, \ldots, N_{\rm no}\}$.

Proposition 1.13.

(i) *La famille* $\{\varphi_1, \ldots, \varphi_{N_{\rm no}}\}$ *est une base de* $P^k_{{\rm c},h}$
et la famille $\{\gamma_1, \ldots, \gamma_{N_{\rm no}}\}$ *est une base de* $\mathcal{L}(P^k_{{\rm c},h}; \mathbb{R})$.

(ii) *La famille* $\{\varphi_2, \ldots, \varphi_{N_{\rm no}-1}\}$ *est une base de* $P^k_{{\rm c},h,0}$
et la famille $\{\gamma_2, \ldots, \gamma_{N_{\rm no}-1}\}$ *est une base de* $\mathcal{L}(P^k_{{\rm c},h,0}; \mathbb{R})$.

Corollaire 1.14.

$$\dim P^k_{{\rm c},h} = N_{\rm no} = kN_{\rm ma} + 1 \quad \textit{et} \quad \dim P^k_{{\rm c},h,0} = N_{\rm no} - 2 = kN_{\rm ma} - 1.$$

Définition 1.15.

(i) *Les formes linéaires* $\{\gamma_1, \ldots, \gamma_{N_{\rm no}}\}$ *sont appelées les* degrés de liberté *dans* $P^k_{{\rm c},h}$ *et les fonctions* $\{\varphi_1, \ldots, \varphi_{N_{\rm no}}\}$ *sont appelées les* fonctions de forme *dans* $P^k_{{\rm c},h}$.

(ii) *Les formes linéaires* $\{\gamma_2, \ldots, \gamma_{N_{\rm no}-1}\}$ *sont appelées les* degrés de liberté *dans* $P^k_{{\rm c},h,0}$ *et les fonctions* $\{\varphi_2, \ldots, \varphi_{N_{\rm no}-1}\}$ *sont appelées les* fonctions de forme *dans* $P^k_{{\rm c},h,0}$.

On introduit *l'opérateur d'interpolation* suivant :

$$\mathcal{I}^k_{{\rm c},h} : \mathcal{C}^0(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{N_{\rm no}} \gamma_i(v)\varphi_i \in P^k_{{\rm c},h}. \tag{1.57}$$

Pour une fonction $v \in \mathcal{C}^0(\overline{\Omega})$, $\mathcal{I}^k_{{\rm c},h}v$ est l'unique fonction continue et polynômiale de degré k par morceaux qui prend les mêmes valeurs que v aux $N_{\rm no}$ nœuds du maillage. La fonction $\mathcal{I}^k_{{\rm c},h}v$ est appelée *l'interpolé de Lagrange* de v de degré k.

L'opérateur d'interpolation $\mathcal{I}^k_{{\rm c},h}$ peut également être vu comme un opérateur de $H^1(\Omega)$ dans $H^1(\Omega)$. On peut montrer que cet opérateur est continu et qu'on a la propriété de stabilité suivante : il existe une constante c, indépendante de h (mais dépendant de k), telle que pour tout $v \in H^1(\Omega)$,

$$\|\mathcal{I}^k_{{\rm c},h}v\|_{1,\Omega} \leqslant c\,\|v\|_{1,\Omega}. \tag{1.58}$$

Par ailleurs, le résultat suivant permet d'estimer la précision de l'opérateur d'interpolation $\mathcal{I}^k_{{\rm c},h}$.

Proposition 1.16. *Il existe une constante c (dépendant de k) telle que pour tout h et pour tout $v \in H^{k+1}(\Omega)$,*

$$\|v - \mathcal{I}^k_{{\rm c},h}v\|_{0,\Omega} + h|v - \mathcal{I}^k_{{\rm c},h}v|_{1,\Omega} \leqslant c\,h^{k+1}|v|_{k+1,\Omega}, \tag{1.59}$$

et

$$\sum_{m=2}^{k+1} h^m \left(\sum_{i=0}^{N} |v - \mathcal{I}^k_{{\rm c},h}v|^2_{m,I_i}\right)^{\frac{1}{2}} \leqslant c\,h^{k+1}|v|_{k+1,\Omega}. \tag{1.60}$$

L'estimation (1.59) montre que l'erreur d'interpolation est d'ordre $(k+1)$ en norme $\|\cdot\|_{0,\Omega}$ et qu'elle est d'ordre k en semi-norme $|\cdot|_{1,\Omega}$; elle est donc également d'ordre k en norme $\|\cdot\|_{1,\Omega}$. La preuve de la proposition 1.16 est analogue à celle de la proposition 1.9. Le point important est qu'à nouveau, les propriétés interpolantes de l'opérateur $\mathcal{I}^k_{c,h}$ sont purement *locales*.

Une comparaison entre les estimations de la proposition 1.16 et celles de la proposition 1.9 montre que, à maillage fixé, l'erreur d'interpolation est plus petite si on utilise des polynômes de degré élevé *pourvu que la fonction à interpoler soit suffisamment régulière*. En particulier, si $v \in H^s(\Omega)$ et $v \notin H^{s+1}(\Omega)$ pour un entier $s \leqslant k$, on montre que l'estimation (1.59) devient

$$\|v - \mathcal{I}^k_{c,h} v\|_{0,\Omega} + h|v - \mathcal{I}^k_{c,h} v|_{1,\Omega} \leqslant c\, h^s |v|_{s,\Omega}. \tag{1.61}$$

On obtient donc la même estimation que pour une interpolation par des polynômes de degré $(s-1)$.

1.5 Analyse de convergence

L'objectif de cette section est l'analyse de convergence de la solution u_h du problème approché (1.15) vers la solution u du problème exact (1.3) lorsque l'espace d'approximation V_h dans (1.15) est pris égal à $P^1_{c,h,0}$ ou plus généralement à $P^k_{c,h,0}$ pour un entier $k \geqslant 1$.

On cherche d'abord à estimer l'erreur $u - u_h$ dans la norme H^1. Pour cela, on utilise l'estimation (1.20), ce qui conduit à

$$\begin{aligned} \|u - u_h\|_{1,\Omega} &\leqslant c \inf_{v_h \in P^k_{c,h,0}} \|u - v_h\|_{1,\Omega} \\ &\leqslant c\, \|u - \mathcal{I}^k_{c,h} u\|_{1,\Omega} \\ &\leqslant c\, h^k |u|_{k+1,\Omega}, \end{aligned} \tag{1.62}$$

pourvu que la solution exacte soit suffisamment régulière, à savoir $u \in H^{k+1}(\Omega)$. On notera que $\mathcal{I}^k_{c,h} u \in P^k_{c,h,0}$ puisque $u \in H^1_0(\Omega)$; en d'autres termes, $\mathcal{I}^k_{c,h} u$ est bien nul en a et en b. On a ainsi montré le résultat suivant.

Proposition 1.17. *Soit un entier $k \geqslant 1$. On suppose que la solution de* (1.3) *est dans $H^{k+1}(\Omega)$. On désigne par u_h la solution du problème approché* (1.15) *avec l'espace d'approximation $V_h = P^k_{c,h,0}$. Alors, il existe une constante c, indépendante de h, telle que*

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^k |u|_{k+1,\Omega}. \tag{1.63}$$

On dit que l'estimation d'erreur (1.63) est *optimale* car elle est du même ordre en h que l'erreur d'interpolation en norme H^1 ; voir la proposition 1.16.

Si la solution exacte n'est pas suffisamment régulière, par exemple si $u \in H^s(\Omega)$ mais $u \notin H^{s+1}(\Omega)$ pour un entier $s \leqslant k$, on montre la majoration d'erreur suivante :

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^{s-1} |u|_{s,\Omega}. \tag{1.64}$$

En d'autres termes, l'approximation basée sur l'élément fini de Lagrange $\mathbb{P}_k$ est sous-optimale car on obtient le même ordre de convergence que si on avait choisi des polynômes de degré $(s-1)$ par morceaux. Ce résultat permet d'établir un lien direct entre l'ordre de convergence de la méthode des éléments finis, la régularité de la solution exacte (quantifiée par le plus grand entier s tel que $u \in H^s(\Omega)$ mais $u \notin H^{s+1}(\Omega)$) et le degré polynômial des fonctions de V_h sur chaque maille. En notant δ cet ordre de convergence, on a

$$\delta = \min(k, s-1). \tag{1.65}$$

On s'intéresse maintenant à une estimation de l'erreur $u - u_h$ en norme L^2. Pour cela, on utilise la technique suivante, dite de *dualité*. On introduit un problème adjoint qui consiste à

$$\begin{cases} \text{Chercher } z \in H^1_0(\Omega) \text{ tel que} \\ \int_\Omega \alpha v' z' = \int_\Omega (u - u_h) v, \quad \forall v \in H^1_0(\Omega). \end{cases} \tag{1.66}$$

Ce problème est clairement bien posé. De plus, en supposant que $\alpha \in \mathcal{C}^1(\overline{\Omega})$, on déduit de la relation $-(\alpha z')' = u - u_h$ que $z \in H^2(\Omega) \cap H^1_0(\Omega)$ et que

$$|z|_{2,\Omega} \leqslant c\, \|u - u_h\|_{0,\Omega}, \tag{1.67}$$

pour une constante c indépendante de h. En prenant $v = u - u_h$ dans (1.66), il vient

$$\begin{aligned} \|u - u_h\|_{0,\Omega}^2 &= \int_\Omega \alpha (u - u_h)' z' \\ &= \int_\Omega \alpha (u - u_h)'(z - z_h)', \qquad \forall z_h \in P^k_{c,h,0}, \end{aligned} \tag{1.68}$$

d'après la relation d'orthogonalité de Galerkin (1.21). On en déduit

$$\|u - u_h\|_{0,\Omega}^2 \leqslant \alpha_1 \|z - z_h\|_{1,\Omega} \|u - u_h\|_{1,\Omega}, \qquad \forall z_h \in P^k_{c,h,0}. \tag{1.69}$$

Puisque $z \in H^2(\Omega) \cap H_0^1(\Omega)$, en prenant $z_h = \mathcal{I}^1_{c,h} z \in P^k_{c,h,0}$ dans la majoration ci-dessus et en utilisant les estimations (1.43) et (1.67), on obtient[1]

$$\begin{aligned} \|u - u_h\|_{0,\Omega}^2 &\leqslant c\, h |z|_{2,\Omega} \|u - u_h\|_{1,\Omega} \\ &\leqslant c\, h \|u - u_h\|_{0,\Omega} \|u - u_h\|_{1,\Omega}. \end{aligned} \tag{1.70}$$

D'où finalement,

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h \|u - u_h\|_{1,\Omega}, \tag{1.71}$$

ce qui, grâce à l'estimation (1.63), conduit au résultat suivant.

Proposition 1.18. *Avec les hypothèses de la proposition* 1.17 *et en supposant que* $\alpha \in \mathcal{C}^1(\overline{\Omega})$*, il existe une constante c, indépendante de h, telle que*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^{k+1} |u|_{k+1,\Omega}. \tag{1.72}$$

L'estimation d'erreur (1.72) est *optimale* car elle est du même ordre en h que l'erreur d'interpolation en norme L^2 ; voir la proposition 1.16. Enfin, si la solution exacte n'est pas suffisamment régulière, par exemple si $u \in H^s(\Omega)$ mais $u \notin H^{s+1}(\Omega)$ pour un entier $s \leqslant k$, on montre la majoration d'erreur suivante :

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^s |u|_{s,\Omega}. \tag{1.73}$$

1. Dans cet aide-mémoire, on adopte la convention de notation suivante : c désigne une constante générique, indépendante de h, mais dont la valeur numérique peut changer à chaque occurrence.

1.6 Résolution numérique

Dans la section 1.2.1, on a vu que la méthode des éléments finis consiste à approcher la solution du problème modèle (1.3) par la solution du problème approché (1.15) en introduisant un espace d'approximation V_h de dimension finie. De plus, en choisissant une base $\{\varphi_1, \ldots, \varphi_N\}$ de V_h, où $N = \dim V_h$, les composantes de la solution approchée dans cette base s'obtiennent par la résolution du système linéaire (1.19). L'objet de cette section est d'examiner brièvement l'évaluation de la matrice de rigidité $\mathcal{A}$ intervenant dans (1.19) et la résolution du système linéaire. On rappelle que les coefficients de la matrice de rigidité s'expriment sous la forme

$$\mathcal{A}_{ij} = \int_\Omega \alpha \varphi_i' \varphi_j', \qquad i,j \in \{1, \ldots, N\}. \tag{1.74}$$

Pour simplifier, on suppose que l'approximation est basée sur l'élément fini de Lagrange $\mathbb{P}_1$. On a donc $V_h = P^1_{c,h,0}$ et $N = N_{ma} - 1 = N_{so} - 2$; voir la section 1.3. Soit $j \in \{2, \ldots, N_{so} - 1\}$. Le support de la fonction de forme φ_j associée au j-ième sommet du maillage est réduit aux deux mailles partageant ce sommet. Par conséquent,

$$\mathcal{A}_{ij} = 0 \qquad \text{si} \qquad |i - j| > 1. \tag{1.75}$$

En d'autres termes, la matrice $\mathcal{A}$ est *tridiagonale* :

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & & & & & \\ \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & & & & \\ & & & & \ddots & \ddots & \ddots \\ & & & & \mathcal{A}_{N-1,N-2} & \mathcal{A}_{N-1,N-1} & \mathcal{A}_{N-1,N} \\ & & & & & \mathcal{A}_{N,N-1} & \mathcal{A}_{N,N} \end{pmatrix}. \tag{1.76}$$

Afin d'évaluer les coefficients non-nuls de cette matrice, on pose, pour tout $i \in \{1, \ldots, N_{ma}\}$,

$$\overline{\alpha}_i = \frac{1}{h_i} \int_{I_i} \alpha. \tag{1.77}$$

Un calcul direct montre que

$$\mathcal{A}_{i,i+1} = \mathcal{A}_{i+1,i} = -\frac{1}{h_{i+1}}\overline{\alpha}_{i+1}, \qquad i \in \{1, \ldots, N-1\}, \tag{1.78}$$

$$\mathcal{A}_{i,i} = -\mathcal{A}_{i,i-1} - \mathcal{A}_{i,i+1}, \qquad i \in \{2, \ldots, N-1\}, \tag{1.79}$$

ainsi que

$$\mathcal{A}_{11} = \frac{1}{h_1}\overline{\alpha}_1 + \frac{1}{h_2}\overline{\alpha}_2 \qquad \text{et} \qquad \mathcal{A}_{NN} = \frac{1}{h_N}\overline{\alpha}_N + \frac{1}{h_{N+1}}\overline{\alpha}_{N+1}. \tag{1.80}$$

Dans le cas particulier où la fonction α est constante sur Ω et vaut α_0 et où le maillage est uniforme de pas h, on obtient

$$\mathcal{A} = \frac{\alpha_0}{h}\,\text{tridiag}(-1, 2, -1). \tag{1.81}$$

Lorsque la fonction α n'est pas constante, on ne dispose pas nécessairement d'une expression explicite permettant d'évaluer sa valeur moyenne sur les mailles. Dans ces conditions, les coefficients de la matrice de rigidité sont évalués de façon approchée par une formule de *quadrature*. Le chapitre 9 présente diverses formules de quadrature en une, deux et trois dimensions d'espace. L'utilisation de quadratures est également nécessaire afin d'évaluer le membre de droite du système linéaire (1.19).

Une fois calculés les coefficients de la matrice de rigidité, il s'agit d'évaluer la solution U du système linéaire (1.19). Lorsque la matrice $\mathcal{A}$ est tridiagonale, on dispose d'un algorithme de résolution particulièrement efficace, connu sous le nom *d'algorithme de Crout* ; voir l'algorithme 1.1. On procède en deux étapes.

(i) la matrice $\mathcal{A}$ est décomposée sous la forme d'un produit d'une matrice bidiagonale inférieure et d'une matrice bidiagonale supérieure (dont les coefficients diagonaux valent 1) :

$$\mathcal{A} = \begin{pmatrix} d_1 & & & \\ l_2 & d_2 & & \\ & \ddots & \ddots & \\ & & l_N & d_N \end{pmatrix} \begin{pmatrix} 1 & u_1 & & \\ & \ddots & \ddots & \\ & & 1 & u_{N-1} \\ & & & 1 \end{pmatrix}. \tag{1.82}$$

On désigne par $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$, respectivement, la matrice bidiagonale inférieure et supérieure dans le membre de droite de (1.82).

(ii) La solution du système $\mathcal{A}U = F$ s'évalue en deux étapes : on forme d'abord le vecteur de travail $Y \in \mathbb{R}^N$ tel que $\mathcal{T}^{\inf} Y = F$; pour cela, on résout le système bidiagonal inférieur en balayant les lignes dans l'ordre croissant. Puis, on inverse le système bidiagonal supérieur $\mathcal{T}^{\sup} U = Y$ en balayant les lignes dans l'ordre décroissant. À l'arrivée, on obtient

$$\mathcal{A}U = \mathcal{T}^{\inf}\mathcal{T}^{\sup} U = \mathcal{T}^{\inf} Y = F, \tag{1.83}$$

si bien que U est effectivement la solution du système linéaire $\mathcal{A}U = F$.

Algorithme 1.1 Algorithme de Crout pour résoudre le système tridiagonal $\mathcal{A}U = F$

Input : $F \in \mathbb{R}^N$ et $\mathcal{A} \in \mathbb{R}^{N,N}$
======Décomposition de la matrice $\mathcal{A}$ selon (1.82)
for $i \in \{2, \ldots, N\}$ **do**
  $l_i = \mathcal{A}_{i,i-1}$
end for
$d_1 = \mathcal{A}_{11}$
for $i \in \{2, \ldots, N\}$ **do**
  $u_{i-1} = \frac{\mathcal{A}_{i-1,i}}{d_{i-1}}$
  $d_i = \mathcal{A}_{ii} - l_i u_{i-1}$
end for
======Résolution du système linéaire $\mathcal{T}^{\inf} Y = F$
$Y_1 = \frac{F_1}{d_1}$
for $i \in \{2, \ldots, N\}$ **do**
  $Y_i = \frac{1}{d_i}(F_i - l_i Y_{i-1})$
end for
======Résolution du système linéaire $\mathcal{T}^{\sup} U = Y$
$U_N = Y_N$
for $i \in \{N-1, \ldots, 1\}$ **do**
  $U_i = Y_i - u_i U_{i+1}$
end for

Le cadre idéal des matrices tridiagonales est toutefois très limité. On verra dans le chapitre 10 que la structure de la matrice de rigidité est nettement plus complexe lorsque le problème modèle est posé en deux ou trois dimensions d'espace. La résolution du système linéaire (1.19) se fait, en général, en utilisant une méthode itérative. De telles méthodes sont décrites dans le chapitre 11.

Remarque 1.19

Même en une dimension d'espace, la matrice de rigidité n'est plus tridiagonale, mais bloc-tridiagonale, lorsqu'on emploie un élément fini de Lagrange $\mathbb{P}_k$ avec $k \geqslant 2$.

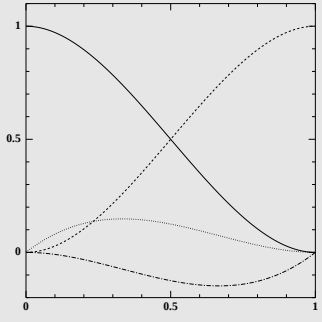
1.7 Complément : élément fini de Hermite

On considère l'espace vectoriel

$$P_h^{\mathrm{Her}} = \{v_h \in \mathcal{C}^1(\overline{\Omega})\,;\, \forall i \in \{1,\ldots,N_{\mathrm{ma}}\},\ v_h|_{I_i} \in \mathbb{P}_3\}, \tag{1.84}$$

dont les éléments sont des fonctions de classe $\mathcal{C}^1$ et polynômiales de degré 3 par morceaux. On vérifie que $P_h^{\mathrm{Her}} \subset H^2(\Omega)$.

Tableau 1.2 – Polynômes de Hermite sur $[0, 1]$: représentation graphique et expression analytique.



$$\theta_1(t) = (2t+1)(t-1)^2$$
$$\theta_2(t) = t(t-1)^2$$
$$\theta_3(t) = (3-2t)t^2$$
$$\theta_4(t) = (t-1)t^2$$

Afin d'exhiber les fonctions de forme dans P_h^{Her}, on introduit les polynômes de Hermite $\{\theta_1, \theta_2, \theta_3, \theta_4\}$ sur l'intervalle de référence $[0, 1]$; voir le tableau 1.2 pour leur représentation graphique et leur expression analytique. En introduisant les formes linéaires $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$ sur $\mathbb{P}_3$ telles que

$$\sigma_1(p) = p(0), \quad \sigma_2(p) = p'(0), \quad \sigma_3(p) = p(1), \quad \sigma_4(p) = p'(1), \tag{1.85}$$

on observe que

$$\sigma_m(\theta_n) = \delta_{mn}, \qquad m, n \in \{1, 2, 3, 4\}. \tag{1.86}$$

On introduit les fonctions $\{\varphi_{1,0}, \ldots, \varphi_{N_{\mathrm{so}},0}, \varphi_{1,1}, \ldots, \varphi_{N_{\mathrm{so}},1}\}$ telles que

$$\begin{aligned}
\varphi_{i,0}(x) &= \begin{cases} \theta_3\left(\frac{x - x_{i-1}}{h_{i-1}}\right) & \text{si } x \in I_{i-1}, \\ \theta_1\left(\frac{x - x_i}{h_i}\right) & \text{si } x \in I_i, \\ 0 & \text{sinon}, \end{cases} \\
\varphi_{i,1}(x) &= \begin{cases} h_{i-1}\theta_4\left(\frac{x - x_{i-1}}{h_{i-1}}\right) & \text{si } x \in I_{i-1}, \\ h_i\theta_2\left(\frac{x - x_i}{h_i}\right) & \text{si } x \in I_i, \\ 0 & \text{sinon}, \end{cases}
\end{aligned} \tag{1.87}$$

avec des modifications élémentaires si $i = 1$ ou si $i = N_{\mathrm{so}}$. Pour tout $i \in \{1, \ldots, N_{\mathrm{so}}\}$, on considère les formes linéaires suivantes :

$$\gamma_{i,0} : \mathcal{C}^1(\overline{\Omega}) \ni v \longmapsto v(x_i) \in \mathbb{R}, \tag{1.88}$$

$$\gamma_{i,1} : \mathcal{C}^1(\overline{\Omega}) \ni v \longmapsto v'(x_i) \in \mathbb{R}. \tag{1.89}$$

On constate que

(i) pour tout $i, j \in \{1, \ldots, N_{\mathrm{so}}\}$ et pour tout $m, n \in \{0, 1\}$,

$$\gamma_{i,m}(\varphi_{j,n}) = \delta_{ij}\delta_{mn};$$

(ii) la famille $\{\varphi_{1,0}, \ldots, \varphi_{N_{\mathrm{so}},0}, \varphi_{1,1}, \ldots, \varphi_{N_{\mathrm{so}},1}\}$ est une base de P_h^{Her} ;

(iii) la famille $\{\gamma_{1,0}, \ldots, \gamma_{N_{\mathrm{so}},0}, \gamma_{1,1}, \ldots, \gamma_{N_{\mathrm{so}},1}\}$ est une base de $\mathcal{L}(P_h^{\mathrm{Her}}; \mathbb{R})$;

(iv) $\dim P_h^{\mathrm{Her}} = 2N_{\mathrm{so}}$.

Les formes linéaires $\{\gamma_{1,0}, \ldots, \gamma_{N_{so},0}, \gamma_{1,1}, \ldots, \gamma_{N_{so},1}\}$ sont appelées les *degrés de liberté* dans P_h^{Her} et les fonctions $\{\varphi_{1,0}, \ldots, \varphi_{N_{so},0}, \varphi_{1,1}, \ldots, \varphi_{N_{so},1}\}$ sont appelées les *fonctions de forme* dans P_h^{Her}.

On introduit *l'opérateur d'interpolation* suivant :

$$\mathcal{I}_h^{\mathrm{Her}} : \mathcal{C}^1(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{N_{so}} \gamma_{i,0}(v)\varphi_{i,0} + \sum_{i=1}^{N_{so}} \gamma_{i,1}(v)\varphi_{i,1} \in P_h^{\mathrm{Her}}. \quad (1.90)$$

La fonction $\mathcal{I}_h^{\mathrm{Her}} v$ est appelée *l'interpolé de Hermite* de v. Comme les fonctions de $H^2(\Omega)$ sont de classe $\mathcal{C}^1$, $\mathcal{I}_h^{\mathrm{Her}}$ peut être vu comme un opérateur de $H^2(\Omega)$ dans $H^2(\Omega)$. Cet opérateur est uniformément continu en h. De plus, on a le résultat suivant.

Proposition 1.20. *Il existe une constante c, indépendante de h, telle que pour tout* $v \in H^4(\Omega)$,

$$\|v - \mathcal{I}_h^{\mathrm{Her}} v\|_{0,\Omega} + h|v - \mathcal{I}_h^{\mathrm{Her}} v|_{1,\Omega} + h^2|v - \mathcal{I}_h^{\mathrm{Her}} v|_{2,\Omega} \leqslant c\, h^4 |v|_{4,\Omega}. \quad (1.91)$$

L'opérateur d'interpolation de Hermite est donc particulièrement précis, pourvu que la fonction à interpoler soit suffisamment régulière. L'élément fini de Hermite peut être utilisé pour approcher le problème modèle (1.3). Il peut être également considéré dans l'approximation du problème suivant :

$$\begin{cases} \text{Chercher } u \in H_0^2(\Omega) \text{ tel que} \\ \int_\Omega \alpha u'' v'' = \int_\Omega f v, \quad \forall v \in H_0^2(\Omega), \end{cases} \quad (1.92)$$

où $H_0^2(\Omega) = \{v \in H^2(\Omega)\ ;\ v(a) = v'(a) = v(b) = v'(b) = 0\}$. Ce problème intervient par exemple dans la modélisation des poutres en flexion et encastrées à leurs deux extrémités. On observera que (modulo des modifications triviales au bord) $P_h^{\mathrm{Her}} \subset H_0^2(\Omega)$ mais que $P_{\mathrm{c},h,0}^k \not\subset H_0^2(\Omega)$ car les dérivées des fonctions de $P_{\mathrm{c},h,0}^k$ sont discontinues aux interfaces entre les mailles.

2 • LA MÉTHODE DE GALERKIN

La méthode de Galerkin permet d'approcher la solution de problèmes modèles dont la formulation abstraite est la suivante :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u, w) = f(w), \quad \forall w \in W, \end{cases} \tag{2.1}$$

où V et W sont des espaces fonctionnels (des espaces vectoriels dont les éléments sont des fonctions), a est une forme bilinéaire définie sur $V \times W$ et f est une forme linéaire définie sur W. On dit que V est *l'espace solution* et que W est *l'espace test*. Les éléments de W sont appelés des *fonctions tests*.

Les espaces fonctionnels V et W sont équipés de normes, notées $\|\cdot\|_V$ et $\|\cdot\|_W$ respectivement, qui leur confèrent une structure d'espace de Banach (V et W sont des espaces vectoriels normés où toute suite de Cauchy est convergente). Dans de nombreuses applications, les normes $\|\cdot\|_V$ et $\|\cdot\|_W$ sont induites par des produits scalaires, notés $(\cdot,\cdot)_V$ et $(\cdot,\cdot)_W$ respectivement, si bien que V et W sont en fait des espaces de Hilbert. Pour simplifier, on conserve cette hypothèse par la suite.

On suppose que la forme bilinéaire a est continue sur $V \times W$, ce qu'on note $a \in \mathcal{L}(V \times W; \mathbb{R})$. On rappelle que cette hypothèse consiste à supposer qu'il existe une constante c_1 telle que pour tout $(v, w) \in V \times W$,

$$a(v, w) \leqslant c_1 \|v\|_V \|w\|_W. \tag{2.2}$$

De même, on suppose que la forme linéaire f est continue sur W, ce qu'on note $f \in \mathcal{L}(W; \mathbb{R}) := W'$, c'est-à-dire qu'il existe une constante c_2 telle que pour tout $w \in W$,

$$f(w) \leqslant c_2 \|w\|_W. \tag{2.3}$$

On introduit les normes de a et f (dans $\mathcal{L}(V \times W; \mathbb{R})$ et W' respectivement) définies par

$$\|a\|_{V,W} = \sup_{(v,w)\in V\times W} \frac{a(v,w)}{\|v\|_V \|w\|_W}, \qquad \|f\|_{W'} = \sup_{w\in W} \frac{f(w)}{\|w\|_W}, \tag{2.4}$$

étant entendu que les arguments des suprema sont pris non-nuls. On renvoie à la section A.1 pour des compléments.

2.1 Le problème modèle est-il bien posé ?

L'objet de cette section est de rappeler brièvement les deux principaux résultats qui permettent d'étudier le caractère bien posé du problème (2.1). La notion de problème bien posé est entendue au sens de la définition suivante.

Définition 2.1 (Hadamard). *On dit que le problème* (2.1) *est* bien posé *s'il admet une et une seule solution.*

Lorsque le problème (2.1) est bien posé, son unique solution u satisfait l'estimation *a priori* suivante : il existe une constante c tel que pour tout $f \in W'$,

$$\|u\|_V \leqslant c\,\|f\|_{W'}. \tag{2.5}$$

Cette estimation découle des propriétés générales des opérateurs bijectifs dans les espaces de Banach ; voir la section A.1.4.

2.1.1 Le lemme de Lax–Milgram

On considère d'abord le cas particulier où l'espace solution et l'espace test dans (2.1) sont identiques : $V = W$. Le problème modèle consiste donc à

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u,w) = f(w), \quad \forall w \in V. \end{cases} \tag{2.6}$$

Définition 2.2 (Coercivité). *Soit V un espace de Hilbert. On dit qu'une forme bilinéaire $a \in \mathcal{L}(V \times V; \mathbb{R})$ est* V-coercive*, ou* coercive sur V*, si*

$$\exists \alpha > 0, \quad \forall v \in V, \quad a(v,v) \geqslant \alpha \|v\|_V^2. \tag{2.7}$$

Lemme 2.3 (Lax–Milgram). *Soit V un espace de Hilbert, $a \in \mathcal{L}(V \times V; \mathbb{R})$ et $f \in V'$. On suppose que la forme bilinéaire a est V-coercive. Alors, le problème* (2.6) *est bien posé.*

Lorsque la forme bilinéaire a n'est pas coercive sur V, peut-on en déduire que le problème (2.6) n'est pas bien posé ? La réponse est négative : le lemme de Lax–Milgram ne fournit que des conditions *suffisantes* pour analyser le caractère bien posé de (2.6).[1]

2.1.2 Le théorème de Banach–Nečas–Babuška (BNB)

Le théorème BNB est le résultat fondamental pour analyser le caractère bien posé des problèmes (2.1) et (2.6). Contrairement au lemme de Lax–Milgram qui ne fournit que des conditions suffisantes, le théorème BNB fournit des conditions *nécessaires et suffisantes* pour que le problème modèle soit bien posé.

Théorème 2.4 (Banach–Nečas–Babuška). *Soit V et W deux espaces de Hilbert,*[2] *$a \in \mathcal{L}(V \times W; \mathbb{R})$ et $f \in W'$. Alors, le problème* (2.1) *est bien posé si et seulement si*

$$\exists \alpha > 0, \qquad \inf_{v \in V} \sup_{w \in W} \frac{a(v,w)}{\|v\|_V \|w\|_W} \geqslant \alpha, \tag{BNB1}$$

$$\forall w \in W, \quad (\forall v \in V,\ a(v,w) = 0) \implies (w = 0). \tag{BNB2}$$

La terminologie adoptée pour ce théorème a été introduite par Ern et Guermond [**38**]. Elle fait référence au fait que le théorème BNB est une reformulation de deux résultats fondamentaux dus à Banach : le théorème de l'image fermée et le théorème de l'application ouverte ; voir la section A.1.4. Le théorème BNB a été énoncé dans sa forme ci-dessous par Nečas en 1962 [**58**]. Son importance pour l'analyse des méthodes d'éléments finis a été soulignée par Babuška en 1972 [**9**].

1. Si la forme bilinéaire a est symétrique ($a(v,w) = a(w,v)$ pour tout $(v,w) \in V \times V$) et positive ($a(v,v) \geqslant 0$ pour tout $v \in V$), la V-coercivité est une condition nécessaire et suffisante pour que le problème (2.6) soit bien posé.

2. Voir la section A.1.4 pour le cadre général du théorème BNB qui est celui des espaces de Banach.

La condition inf-sup (BNB1) se reformule de la façon suivante : il existe $\alpha > 0$ tel que pour tout $v \in V$,

$$\alpha \|v\|_V \leqslant \sup_{w \in W} \frac{a(v,w)}{\|w\|_W}. \tag{2.8}$$

Pour prouver la condition inf-sup, on peut procéder comme suit : on considère une fonction $v \in V$ et on construit une fonction $w_v \in W$ telle que $a(v, w_v) \geqslant \alpha_1 \|v\|_V^2$ et $\|w_v\|_W \leqslant \alpha_2 \|v\|_V$. Ceci permet de montrer que la condition (BNB1) est satisfaite avec $\alpha = \frac{\alpha_1}{\alpha_2}$.

2.2 Principe de la méthode de Galerkin

On considère le problème modèle (2.1) et on suppose qu'il est bien posé. La méthode de Galerkin permet d'approcher la solution u de ce problème. L'idée consiste à remplacer dans (2.1) les espaces fonctionnels V et W par des espaces de *dimension finie*, notés V_h et W_h, ce qui conduit à

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in W_h. \end{cases} \tag{2.9}$$

On dit que (2.9) est le *problème approché* ou le *problème discret* et que u_h est la *solution approchée*. On notera que sous sa forme la plus générale, le problème approché (2.9) fait intervenir une forme bilinéaire $a_h \in \mathcal{L}(V_h \times W_h; \mathbb{R})$ qui est une approximation de la forme bilinéaire a et une forme linéaire $f_h \in W_h'$ qui est une approximation de la forme linéaire f. L'espace V_h, qu'on appellera *espace d'approximation*, et l'espace W_h, qu'on appellera *espace test discret*, sont construits à l'aide de la méthode des éléments finis selon les techniques présentées dans le chapitre 1 pour les problèmes en dimension 1 et dans les chapitres 3 et 4 pour les problèmes en dimension supérieure. L'indice h fait référence à la finesse des maillages employés pour construire ces espaces. Les éléments de W_h sont appelés des *fonctions tests discrètes*.

Un choix particulier dans (2.9) consiste à utiliser le même espace V_h comme espace d'approximation et comme espace test discret, ce qui conduit au problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{2.10}$$

Dans ce cas, on parle de *méthode de Galerkin standard*, alors que si les espaces discrets V_h et W_h sont différents, on parle de *méthode de Galerkin non-standard* (dans la littérature, on rencontre également la terminologie « méthode de Petrov–Galerkin »).

Définition 2.5 (Conformité). *L'approximation* (2.9) *est dite* conforme *si* $V_h \subset V$ *et* $W_h \subset W$ *; elle est dite* non-conforme *si* $V_h \not\subset V$ *ou* $W_h \not\subset W$. *On dit que l'espace* V_h *est* V-conforme *lorsque* $V_h \subset V$ *et que l'espace* W_h *est* W-conforme *lorsque* $W_h \subset W$.

Définition 2.6 (Consistance). *Soit* u *la solution unique de* (2.1). *On suppose que la forme bilinéaire* a_h *peut être étendue à* $(V + V_h) \times W_h$. *L'approximation* (2.9) *est dite* consistante *si*

$$\forall w_h \in W_h, \qquad a_h(u, w_h) = f_h(w_h). \tag{2.11}$$

Si tel n'est pas le cas, l'approximation est dite non-consistante.

En d'autres termes, l'approximation est consistante si la solution exacte satisfait les équations discrètes. La non-consistance de la méthode d'approximation peut, par exemple, provenir de l'utilisation de quadratures pour évaluer les intégrales dans la forme bilinéaire a et la forme linéaire f.

Le problème approché (2.9) est un système linéaire. En effet, on pose

$$N = \dim V_h \qquad \text{et} \qquad M = \dim W_h. \tag{2.12}$$

Soit $\{\varphi_1, \ldots, \varphi_N\}$ une base de V_h et soit $\{\psi_1, \ldots, \psi_M\}$ une base de W_h. On décompose la solution approchée u_h dans la base de V_h selon

$$u_h = \sum_{i=1}^{N} U_i \varphi_i, \tag{2.13}$$

et on introduit le vecteur U de $\mathbb{R}^N$ formé par les composantes de u_h dans cette base, $U = (U_i)_{1 \leqslant i \leqslant N}$. Soit $\mathcal{A} \in \mathbb{R}^{M,N}$ la *matrice de rigidité* dont les composantes sont

$$\mathcal{A}_{ij} = a_h(\varphi_j, \psi_i), \qquad i \in \{1, \ldots, M\},\ j \in \{1, \ldots, N\}, \tag{2.14}$$

et soit $F \in \mathbb{R}^M$ le vecteur de composantes

$$F_i = f_h(\psi_i), \qquad i \in \{1, \ldots, M\}.$$

Il est clair que u_h est solution de (2.9) si et seulement si

$$\mathcal{A}U = F. \tag{2.15}$$

2.3 Le problème approché est-il bien posé ?

L'objet de cette section est d'analyser le caractère bien posé du problème approché (2.9). On retiendra les résultats suivants.

- Pour une approximation consistante et conforme d'un problème dont la forme bilinéaire est *coercive*, le problème approché est automatiquement bien posé. De plus, la matrice de rigidité est *définie positive*.
- Lorsque le caractère bien posé du problème modèle repose sur les conditions inf-sup (BNB1) et (BNB2), celles-ci ne sont pas transférées automatiquement au cadre discret. Pour montrer que le problème discret est bien posé, il faut (et il suffit de) prouver une *condition inf-sup discrète* et vérifier que l'espace d'approximation et l'espace test discret ont la même dimension.

2.3.1 Approximation consistante et conforme d'un problème coercif

Soit V un espace de Hilbert, soit $a \in \mathcal{L}(V \times V; \mathbb{R})$ une forme bilinéaire et V-coercive et soit $f \in V'$. Dans ce cadre, le problème modèle (2.6) admet une unique solution u. Pour approcher cette solution, on considère le problème discret suivant :

$$\left\{ \begin{array}{l} \text{Chercher } u_h \in V_h \text{ tel que} \\ a(u_h, w_h) = f(w_h), \quad \forall w_h \in V_h, \end{array} \right. \tag{2.16}$$

et on suppose que $V_h \subset V$. On notera que le problème discret (2.16) fait intervenir la même forme bilinéaire a et la même forme linéaire f que le problème modèle (2.6).

Proposition 2.7. *Avec les hypothèses ci-dessus, la matrice de rigidité $\mathcal{A}$ est définie positive ; par conséquent, le problème discret* (2.16) *est bien posé.*

Le caractère défini positif de la matrice $\mathcal{A}$ résulte du fait que pour tout $X = (X_i)_{1 \leqslant i \leqslant N} \in \mathbb{R}^N$ où $N = \dim V_h$, on a

$$\sum_{1 \leqslant i,j \leqslant N} \mathcal{A}_{ij} X_i X_j = a(\xi, \xi) \geqslant \alpha \|\xi\|_V^2, \tag{2.17}$$

avec $\xi = \sum_{i=1}^{N} X_i \varphi_i \in V_h$, $\{\varphi_1, \ldots, \varphi_N\}$ étant une base de V_h. Par suite, $\sum_{1 \leqslant i,j \leqslant N} \mathcal{A}_{ij} X_i X_j = 0$ implique $\xi = 0$ et donc $X = 0$.

Remarque 2.8.
Si la forme bilinéaire a est symétrique, la matrice de rigidité l'est également.

2.3.2 Cas général

On considère maintenant le cas général, c'est-à-dire que l'on considère le problème modèle (2.1), que l'on suppose bien posé, et on souhaite utiliser le problème discret (2.9) pour obtenir une solution approchée u_h. On notera $\|\cdot\|_{V_h}$ et $\|\cdot\|_{W_h}$ les normes dont sont équipés les espaces discrets V_h et W_h, respectivement. L'approximation pouvant être non-conforme, il n'est pas possible d'équiper *a priori* les espaces discrets V_h et W_h des normes induites par V et W, respectivement.

Clairement, en vertu du théorème BNB, le caractère bien posé de (2.9) est équivalent aux deux conditions suivantes :

$$\exists \alpha_h > 0, \qquad \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{a_h(v_h, w_h)}{\|v_h\|_{V_h} \|w_h\|_{W_h}} \geqslant \alpha_h, \tag{BNB1$_h$}$$

$$\forall w_h \in W_h, \quad (\forall v_h \in V_h,\ a_h(v_h, w_h) = 0) \implies (w_h = 0). \tag{BNB2$_h$}$$

La condition (BNB1$_h$) est une *condition inf-sup discrète*. Même si l'approximation est conforme et consistante, rien ne garantit *a priori* que la condition inf-sup (BNB1) implique la condition inf-sup discrète (BNB1$_h$). La même difficulté se pose entre les conditions (BNB2) et (BNB2$_h$).

On constate que l'interprétation des conditions (BNB1$_h$) et (BNB2$_h$) en termes matriciels est la suivante :

(i) (BNB1$_h$) équivaut au fait que la matrice $\mathcal{A}$ est injective ;

(ii) (BNB2$_h$) équivaut au fait que la matrice $\mathcal{A}$ est de rang maximal.

Par conséquent, les conditions (BNB1$_h$) et (BNB2$_h$) sont équivalentes à (BNB1$_h$) et $\dim V_h = \dim W_h$. En résumé, on a le résultat suivant.

Théorème 2.9. *Le problème approché* (2.9) *est bien posé si et seulement si la condition inf-sup discrète* (BNB1$_h$) *est satisfaite et si* $\dim V_h = \dim W_h$.

Remarque 2.10.

La constante α_h intervenant dans (BNB1$_h$) est la plus petite valeur propre de $\mathcal{A}^T\mathcal{A}$.

2.4 Analyse d'erreur

On considère le problème modèle (2.1) et son approximation (2.9) par la méthode de Galerkin. On suppose que ces deux problèmes sont bien posés, c'est-à-dire que :

(i) la forme bilinéaire a est dans $\mathcal{L}(V \times W; \mathbb{R})$ et elle satisfait les conditions inf-sup (BNB1) et (BNB2) ;

(ii) la forme bilinéaire a_h est dans $\mathcal{L}(V_h \times W_h; \mathbb{R})$, elle satisfait la condition inf-sup discrète (BNB1$_h$) et $\dim V_h = \dim W_h$.

On note u et u_h la solution unique de (2.1) et (2.9), respectivement. L'objectif de cette section est d'estimer l'erreur $u - u_h$. Cette quantité est appelée *l'erreur d'approximation*. En particulier, on souhaite préciser sous quelles hypothèses l'erreur d'approximation tend vers zéro lorsque h tend vers zéro (on rappelle que le paramètre h fait référence à la finesse du maillage qui est utilisé pour construire les espaces V_h et W_h). On s'intéresse donc à des familles d'espaces $\{V_h\}_{h>0}$ et $\{W_h\}_{h>0}$ obtenues en raffinant le maillage.

2.4.1 Approximation consistante et conforme

On suppose dans cette section que l'approximation est consistante et conforme. On a donc $V_h \subset V$ et $W_h \subset W$ et la relation (2.11) est satisfaite. On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in W_h. \end{cases} \tag{2.18}$$

L'hypothèse $V_h \subset V$ implique en particulier que l'erreur $u - u_h$ est dans V. On peut donc utiliser la norme $\|\cdot\|_V$ pour la mesurer. Une conséquence immédiate de (2.11) est la suivante.

Lemme 2.11 (Orthogonalité de Galerkin). *Avec les hypothèses ci-dessus, on a la relation, dite* d'orthogonalité de Galerkin,

$$\forall w_h \in W_h, \qquad a_h(u - u_h, w_h) = 0. \tag{2.19}$$

On suppose en outre que $a_h = a$ et $f_h = f$. Le résultat suivant est connu sous le nom de lemme de Céa.

Lemme 2.12 (Céa). *Avec les hypothèses ci-dessus, on a*

$$||u - u_h||_V \leqslant \left(1 + \frac{\|a\|_{V,W}}{\alpha_h}\right) \inf_{v_h \in V_h} ||u - v_h||_V. \tag{2.20}$$

On suppose en outre que $V = W$, $V_h = W_h$ et que la forme bilinéaire a est V-coercive. Dans ces conditions, on montre que l'estimation d'erreur devient

$$||u - u_h||_V \leqslant \frac{\|a\|_{V,V}}{\alpha} \inf_{v_h \in V_h} ||u - v_h||_V, \tag{2.21}$$

où α est la constante de coercivité de a. Si la forme bilinéaire a est de plus symétrique, cette estimation peut encore être améliorée en

$$\|u - u_h\|_V \leqslant \left(\frac{\|a\|_{V,V}}{\alpha}\right)^{\frac{1}{2}} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{2.22}$$

Afin d'établir la convergence de u_h vers u, on doit contrôler la quantité $\inf_{v_h \in V_h} \|u - v_h\|_V$. Il s'agit donc d'estimer la distance de u à V_h pour la norme $\|\cdot\|_V$. Pour cela, on introduit la notion suivante.

Définition 2.13. *On dit que la famille d'espaces* $\{V_h\}_{h>0}$ *est* asymptotiquement dense *dans* V *si*

$$\forall v \in V, \qquad \lim_{h \to 0} \left(\inf_{v_h \in V_h} \|v - v_h\|_V \right) = 0. \tag{2.23}$$

Théorème 2.14. *On suppose que :*

(i) *la condition* $(\mathrm{BNB1_h})$ *est satisfaite uniformément en* h *;*

(ii) *l'approximation est consistante et conforme ;*

(iii) *la famille* $\{V_h\}_{h>0}$ *est asymptotiquement dense.*

Alors, on a

$$\lim_{h \to 0} \|u - u_h\|_V = 0. \tag{2.24}$$

2.4.2 Le cadre général pour l'analyse de convergence

L'approximation pouvant être non-conforme, l'erreur $u - u_h$ n'appartient pas nécessairement à l'espace V mais à l'espace étendu

$$V(h) = V + V_h. \tag{2.25}$$

On équipe cet espace d'une *norme étendue* $\|\cdot\|_{V(h)}$ et afin d'effectuer l'analyse d'erreur, on fait les hypothèses suivantes :

$$\forall v_h \in V_h, \qquad \|v_h\|_{V(h)} = \|v_h\|_{V_h}, \tag{2.26}$$

$$\forall v \in V, \qquad \|v\|_{V(h)} \leqslant c\,\|v\|_V, \tag{2.27}$$

pour une constante c indépendante de v et de h. Ces deux hypothèses signifient que la norme étendue est une extension de la norme de V_h et que V s'injecte continûment dans $V(h)$ pour la norme étendue. La notion de densité asymptotique se reformule à l'aide de la norme étendue $\|\cdot\|_{V(h)}$ de la manière suivante.

Définition 2.15 (Densité asymptotique). *On dit que la famille d'espaces* $\{V_h\}_{h>0}$ *est* asymptotiquement dense *dans* V *si*

$$\forall v \in V, \qquad \lim_{h\longrightarrow 0}\left(\inf_{v_h\in V_h}\|v - v_h\|_{V(h)}\right) = 0. \tag{2.28}$$

Afin d'estimer la quantité $\|u - u_h\|_{V(h)}$, on introduit une nouvelle notion : la consistance asymptotique.

Définition 2.16 (Consistance asymptotique). *On suppose que la forme bilinéaire* a_h *est uniformément continue en* h *sur* $V_h \times W_h$, *c'est-à-dire que* $\|a_h\|_{V_h,W_h}$ *est majoré uniformément en* h. *L'approximation* (2.9) *est dite* asymptotiquement consistante *s'il existe un opérateur* $\Pi_h : V \longrightarrow V_h$ *tel que* (i) *pour tout* $v \in V$, $\|\Pi_h v - v\|_{V(h)} \leqslant c\,\inf_{v_h\in V_h}\|v - v_h\|_{V(h)}$ *où* c *est une constante indépendante de* v *et de* h, *et* (ii)

$$\lim_{h\longrightarrow 0}\left(\sup_{w_h\in W_h}\frac{|f_h(w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_{W_h}}\right) = 0. \tag{2.29}$$

Dans ces conditions, on définit l'erreur de consistance $R_h(u)$ *par*

$$R_h(u) = \sup_{w_h \in W_h} \frac{|f_h(w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_{W_h}}. \tag{2.30}$$

Remarque 2.17

Lorsque la famille $\{V_h\}_{h>0}$ est *asymptotiquement dense*, la notion de consistance asymptotique est indépendante du choix de l'opérateur Π_h. En effet, soit $\Pi'_h : V \to V_h$ un deuxième opérateur tel que pour tout $v \in V$, $\|\Pi'_h v - v\|_{V(h)} \leqslant c \inf_{v_h \in V_h} \|v - v_h\|_{V(h)}$. Soit $R'_h(u)$ l'erreur de consistance mesurée avec l'opérateur Π'_h. Alors, pour tout $w_h \in W_h$,

$$|f_h(w_h) - a_h(\Pi'_h u, w_h)| \leqslant |f_h(w_h) - a_h(\Pi_h u, w_h)| + |a_h(\Pi_h u - \Pi'_h u, w_h)|,$$

ce qui implique, grâce à l'uniforme continuité de a_h et à l'inégalité triangulaire, que

$$R'_h(u) \leqslant R_h(u) + c\,\|a_h\|_{V_h, W_h} \left(\inf_{v_h \in V_h} \|u - v_h\|_{V(h)} \right).$$

En utilisant la densité asymptotique de $\{V_h\}_{h>0}$, cette inégalité implique que l'approximation (2.9) est asymptotiquement consistante en utilisant l'opérateur Π'_h. En d'autres termes, l'erreur de consistance est indépendante, à un facteur près contrôlé par la propriété de densité asymptotique de la famille $\{V_h\}_{h>0}$, de l'opérateur Π_h utilisé pour l'évaluer.

Théorème 2.18. *On suppose que :*

(i) *la condition* (BNB1_h) *est satisfaite uniformément en h ;*

(ii) *la forme bilinéaire a_h est uniformément continue en h sur $V_h \times W_h$;*

(iii) *la famille $\{V_h\}_{h>0}$ est asymptotiquement dense ;*

(iv) *l'approximation est asymptotiquement consistante.*

Alors, en notant $R_h(u)$ l'erreur de consistance, on a

$$\|u - u_h\|_{V(h)} \leqslant \tfrac{1}{\alpha_h} R_h(u) + c \left(\inf_{v_h \in V_h} \|u - v_h\|_{V(h)} \right), \tag{2.31}$$

si bien que

$$\lim_{h \to 0} \|u - u_h\|_{V(h)} = 0. \tag{2.32}$$

Ce théorème montre quelles sont les quatre propriétés à satisfaire pour garantir la convergence de l'approximation dans la méthode de Galerkin : stabilité de a_h uniforme en h, continuité de a_h uniforme en h, densité asymptotique et consistance asymptotique. Un principle général de l'analyse numérique, connu sous le nom de *principe de Lax*, est que stabilité et consistance impliquent convergence. Le fait que ce principe ne mentionne pas explicitement la continuité et la densité asymptotique ne veut pas dire que ces deux propriétés doivent être considérées comme aquises dans tous les cas. On pourra par exemple consulter [**38**, p. 97] pour un contre-exemple à la densité asymptotique dans le contexte des équations de Maxwell.

2.4.3 Cas particuliers : lemmes de Strang

Les deux estimations d'erreur présentées dans cette section sont connues sous le nom de premier et deuxième lemme de Strang.

On suppose d'abord que l'approximation est conforme, c'est-à-dire que $V_h \subset V$ et $W_h \subset W$. On a donc $V(h) = V$, ce qui permet de mesurer l'erreur dans la norme $\|\cdot\|_V$.

Lemme 2.19 (Strang 1). *On suppose que $V_h \subset V$ et $W_h \subset W$. Alors, on a*

$$\|u - u_h\|_V \leqslant \frac{1}{\alpha_h} \sup_{w_h \in W_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_{W_h}} \tag{2.33}$$

$$+ \inf_{v_h \in V_h} \left[\left(1 + \frac{\|a\|_{V,W}}{\alpha_h} \right) \|u - v_h\|_V + \frac{1}{\alpha_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_{W_h}} \right].$$

La preuve du lemme 2.19 est relativement simple. Soit $v_h \in V_h$. On déduit de la condition (BNB1$_h$) que

$$\alpha_h \|u_h - v_h\|_V \leqslant \sup_{w_h \in W_h} \frac{a_h(u_h - v_h, w_h)}{\|w_h\|_{W_h}}. \tag{2.34}$$

Un calcul direct montre que

$$a_h(u_h - v_h, w_h) = a(u - v_h, w_h) + a(v_h, w_h) - a_h(v_h, w_h) + f_h(w_h) - f(w_h). \tag{2.35}$$

Par conséquent,

$$\alpha_h\|u_h - v_h\|_V \leqslant \|a\|_{V,W}\|u - v_h\|_V + \sup_{w_h\in W_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_{W_h}} + \sup_{w_h\in W_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_{W_h}}. \tag{2.36}$$

On conclut en utilisant l'inégalité triangulaire

$$\|u - u_h\|_V \leqslant \|u - v_h\|_V + \|u_h - v_h\|_V, \tag{2.37}$$

puis en prenant l'infimum sur $v_h \in V_h$.

On ne suppose plus maintenant que l'approximation est conforme ; par contre, on suppose que la forme bilinéaire a_h peut être étendue à $V(h) \times W_h$; par suite, $a_h(v, w_h)$ est bien défini pour $v \in V(h)$ et $w_h \in W_h$.

Lemme 2.20 (Strang 2). *On suppose que a_h est continue sur $V(h) \times W_h$. Alors, on a*

$$\|u - u_h\|_{V(h)} \leqslant \left(1 + \frac{\|a_h\|_{V(h),W_h}}{\alpha_h}\right) \inf_{v_h\in V_h} \|u - v_h\|_{V(h)} + \frac{1}{\alpha_h} \sup_{w_h\in W_h} \frac{|f_h(w_h) - a_h(u, w_h)|}{\|w_h\|_{W_h}}. \tag{2.38}$$

De plus, si l'approximation est consistante, alors

$$\|u - u_h\|_{V(h)} \leqslant \left(1 + \frac{\|a_h\|_{V(h),W_h}}{\alpha_h}\right) \inf_{v_h\in V_h} \|u - v_h\|_{V(h)}. \tag{2.39}$$

La preuve du lemme 2.20 est relativement simple. Soit $v_h \in V_h$ et soit $w_h \in W_h$. On a

$$\begin{aligned} a_h(u_h - v_h, w_h) &= a_h(u_h - u, w_h) + a_h(u - v_h, w_h) \\ &= f_h(w_h) - a_h(u, w_h) + a_h(u - v_h, w_h). \end{aligned} \tag{2.40}$$

Grâce à la condition (BNB1$_h$) on obtient

$$\alpha_h\|u_h - v_h\|_{V(h)} \leqslant \sup_{w_h\in W_h} \frac{|f_h(w_h) - a_h(u, w_h)|}{\|w_h\|_{W_h}} + \|a_h\|_{V(h),W_h}\|u - v_h\|_{V(h)}. \tag{2.41}$$

On conclut en utilisant une inégalité triangulaire puis en prenant l'infimum sur $v_h \in V_h$.

2.4.4 Le lemme de Aubin–Nitsche

On suppose que $V = W$ et qu'il existe deux espaces de Hilbert Z et L tels que

$$Z \subset V \subset L, \tag{2.42}$$

avec injections continues. On considère une forme bilinéaire l sur $L \times L$ que l'on suppose continue, symétrique et positive. On désigne par $|\cdot|_L = \sqrt{l(\cdot,\cdot)}$ la semi-norme induite par l. L'objectif de cette section est d'estimer l'erreur dans la semi-norme $|\cdot|_L$. Pour simplifier, on se restreint au cadre d'une approximation consistante et conforme par la méthode de Galerkin standard ($V_h = W_h$ et $V_h \subset V$) ; voir, par exemple, Braess [**18**, p. 108] pour un cas plus général. On suppose que :

(i) il existe une constante de stabilité c_S telle que pour tout $g \in L$, la solution $\varsigma(g)$ du problème adjoint

$$\begin{cases} \text{Chercher } \varsigma(g) \in V \text{ tel que} \\ a(v, \varsigma(g)) = l(g, v), \quad \forall v \in V, \end{cases} \tag{2.43}$$

satisfait l'estimation *a priori*

$$\|\varsigma(g)\|_Z \leqslant c_S |g|_L; \tag{2.44}$$

(ii) il existe une constante c_i telle que

$$\forall h,\ \forall v \in Z, \qquad \inf_{v_h \in V_h} \|v - v_h\|_V \leqslant c_i h \|v\|_Z. \tag{2.45}$$

Le résultat ci-dessous est connu sous le nom de lemme de Aubin–Nitsche ; voir, par exemple, Aubin [**8**].

Lemme 2.21 (Aubin–Nitsche). *Avec les hypothèses ci-dessus, on a*

$$\forall h, \qquad |u - u_h|_L \leqslant c\, h \|u - u_h\|_V, \tag{2.46}$$

avec $c = c_i c_S \|a\|_{V,V}$.

Définition 2.22. *Lorsque la propriété* (2.44) *est satisfaite, le problème* (2.43) *est dit* régularisant.

On utilisera la notion de problème régularisant et le lemme de Aubin–Nitsche dans le chapitre 5 pour le Laplacien et les modèles de mécanique des milieux continus ($|\cdot|_L$ correspondra à la norme $L^2(\Omega)$ ou $[L^2(\Omega)]^3$) et dans le chapitre 6 pour le problème de Stokes ($|\cdot|_L$ correspondra à la semi-norme $|\cdot|_{1,\Omega}$ de la vitesse).

3 • ÉLÉMENTS FINIS DE LAGRANGE

L'objet de ce chapitre est d'étudier les éléments finis les plus couramment rencontrés dans la pratique, à savoir les éléments finis de Lagrange. On effectue cette étude en adoptant un point de vue local. Cette approche permet d'appréhender les éléments finis de Lagrange (et plus généralement tout élément fini ; voir le chapitre 4) comme la brique élémentaire permettant d'interpoler des fonctions définies sur un domaine Ω : on maille ce domaine (par des triangles, des quadrangles ou d'autres types de mailles) puis on génère des éléments finis sur chaque maille à partir d'un élément fini de référence dont les propriétés sont purement locales.

Ce chapitre est organisé comme suit. On donne d'abord une définition générale d'un élément fini de Lagrange et on introduit les notions de degrés de liberté, de fonctions de forme et d'opérateur d'interpolation local. On présente ensuite les exemples classiques d'éléments finis de Lagrange. Enfin, on étudie comment mailler un domaine et comment utiliser ce maillage afin de construire d'une part un opérateur d'interpolation global et d'autre part des espaces vectoriels de dimension finie qui ont vocation à être utilisés comme espaces d'approximation dans le cadre de la méthode de Galerkin afin d'approcher la solution de problèmes modèles.

3.1 Notion locale d'élément fini de Lagrange

Soit K une partie de $\mathbb{R}^d$; pour simplifier, on suppose que K est un intervalle en dimension 1, un polygone en dimension 2 ou un polyèdre en dimension

3. Soit P un espace vectoriel de fonctions (en général polynômiales) définies sur K et à valeurs dans $\mathbb{R}$. Soit $\{a_1, \ldots, a_{n_f}\}$ un ensemble de points dans K où n_f est un entier strictement positif. Pour $i \in \{1, \ldots, n_f\}$, on introduit la forme linéaire

$$\sigma_i : P \ni p \longmapsto p(a_i) \in \mathbb{R}. \tag{3.1}$$

On pose $\Sigma = \{\sigma_1, \ldots, \sigma_{n_f}\}$.

Définition 3.1 (Élément fini de Lagrange). *Si l'application linéaire*

$$P \ni p \longmapsto \big(\sigma_1(p), \ldots, \sigma_{n_f}(p)\big)^T \in \mathbb{R}^{n_f}, \tag{3.2}$$

est bijective, on dit que le triplet $\{K, P, \Sigma\}$ est un élément fini de Lagrange. Les points $\{a_1, \ldots, a_{n_f}\}$ sont appelés les nœuds *de l'élément fini et les formes linéaires $\{\sigma_1, \ldots, \sigma_{n_f}\}$ sont appelées les* degrés de liberté *de l'élément fini.*

La bijectivité de l'application linéaire définie en (3.2) signifie que pour tout $(\alpha_1, \ldots, \alpha_{n_f})^T \in \mathbb{R}^{n_f}$, il existe un et un seul polynôme $p \in P$ tel que $p(a_i) = \alpha_i$ pour tout $i \in \{1, \ldots, n_f\}$. Ceci équivaut au fait que

$$\begin{cases} \dim P = \operatorname{card} \Sigma = n_f, \\ \forall p \in P, \quad \big(p(a_i) = 0,\ i \in \{1, \ldots, n_f\}\big) \implies (p = 0), \end{cases}$$

ou encore au fait qu'il existe une base de P, notée $\{\theta_1, \ldots, \theta_{n_f}\}$, telle que

$$\sigma_i(\theta_j) = \theta_j(a_i) = \delta_{ij}, \qquad i, j \in \{1, \ldots, n_f\}. \tag{3.3}$$

On rappelle que δ_{ij} désigne le symbole de Kronecker tel que $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$. Les fonctions $\{\theta_1, \ldots, \theta_{n_f}\}$ sont appelées les *fonctions de forme* de l'élément fini. Pour $i \in \{1, \ldots, n_f\}$, la fonction de forme θ_i vaut 1 au nœud a_i et 0 aux autres nœuds.

Définition 3.2 (Opérateur d'interpolation local). *L'opérateur d'interpolation local $\mathcal{I}_K^{\mathrm{Lag}}$ est défini comme suit :*

$$\mathcal{I}_K^{\mathrm{Lag}} : \mathcal{C}^0(K) \ni v \longmapsto \sum_{i=1}^{n_f} v(a_i)\theta_i \in P. \tag{3.4}$$

On dit que $\mathcal{I}_K^{\mathrm{Lag}} v$ est *l'interpolé de Lagrange* de v sur K. L'interpolé de Lagrange est tel que sa valeur aux nœuds $\{a_1, \ldots, a_{n_f}\}$ coïncide avec celle de la fonction à interpoler v.

L'opérateur d'interpolation $\mathcal{I}_K^{\mathrm{Lag}}$ est une *projection* de $\mathcal{C}^0(K)$ dans P. En effet, pour tout $p \in P$, en décomposant p dans la base des fonctions de forme selon $p = \sum_{j=1}^{n_{\mathrm{f}}} x_j \theta_j$, on obtient

$$\mathcal{I}_K^{\mathrm{Lag}} p = \sum_{i=1}^{n_{\mathrm{f}}} p(a_i)\theta_i = \sum_{i,j=1}^{n_{\mathrm{f}}} x_j \theta_j(a_i)\theta_i = \sum_{i,j=1}^{n_{\mathrm{f}}} x_j \delta_{ij}\theta_i = p. \tag{3.5}$$

Par suite, pour tout $v \in \mathcal{C}^0(K)$, il vient $\mathcal{I}_K^{\mathrm{Lag}}(\mathcal{I}_K^{\mathrm{Lag}} v) = \mathcal{I}_K^{\mathrm{Lag}} v$. Cette propriété se généralise à tout type d'élément fini ; voir la section 4.2.

3.2 Exemples classiques d'éléments finis de Lagrange

L'objet de cette section est de dresser le catalogue des principaux éléments finis de Lagrange utilisés en pratique. On considère :

(i) les éléments finis de Lagrange unidimensionnels ;

(ii) les éléments finis de Lagrange simplectiques (K est un triangle, un tétraèdre ou plus généralement un simplexe) : on parle d'éléments finis de Lagrange $\mathbb{P}_k$;

(iii) les éléments finis de Lagrange à structure tensorielle (K est un carré, un cube ou plus généralement un hypercube) : on parle d'éléments finis de Lagrange $\mathbb{Q}_k$;

(iv) les éléments finis de Lagrange prismatiques (K est un prisme).

On désigne par x le point courant de $\mathbb{R}^d$ et on note $(x_1, \ldots, x_d)$ ses coordonnées cartésiennes.

3.2.1 Éléments finis de Lagrange unidimensionnels

Définition 3.3. *Soit un entier $k \geqslant 1$. Soit $K = [c, d]$ un intervalle de mesure non-nulle. En une dimension d'espace, l'élément fini de Lagrange $\mathbb{P}_k$ est défini comme le triplet $\{K, P, \Sigma\}$ tel que $P = \mathbb{P}_k$ et $\Sigma = \{\sigma_0, \ldots, \sigma_k\}$ où les formes linéaires σ_m, $m \in \{0, \ldots, k\}$, associent à $p \in \mathbb{P}_k$ sa valeur au nœud $a_m = c + \frac{m}{k}(d - c)$.*

Les nœuds $\{a_0, \ldots, a_k\}$ associés à l'élément fini de Lagrange $\mathbb{P}_k$ sont illustrés sur la figure 3.1 pour $k \in \{1, 2, 3\}$. Lorsque $K = [0, 1]$, les fonctions de forme de l'élément fini de Lagrange $\mathbb{P}_k$ sont les polynômes d'interpolation de Lagrange $\{\mathcal{L}_0^k, \ldots, \mathcal{L}_k^k\}$ introduits dans la section 1.4. Le tableau 1.1 page 17 contient une représentation graphique de ces polynômes ainsi que leur expression analytique. Lorsque $K = [c, d]$, les fonctions de forme sont les polynômes $\theta_m(x) = \mathcal{L}_m^k(\frac{x-c}{d-c})$, $m \in \{0, \ldots, k\}$.

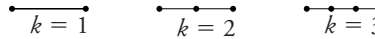


Figure 3.1 – Nœuds $\{a_0, \ldots, a_k\}$ de l'élément fini de Lagrange $\mathbb{P}_k$, $k \in \{1, 2, 3\}$, en dimension 1.

3.2.2 Éléments finis de Lagrange simplectiques

Soit $\{s_0, \ldots, s_d\}$ une famille de points dans $\mathbb{R}^d$ avec $d \geqslant 2$. On suppose que les vecteurs $\{s_1 - s_0, \ldots, s_d - s_0\}$ sont linéairement indépendants.

Définition 3.4. *L'enveloppe convexe des points $\{s_0, \ldots, s_d\}$ est appelée un* simplexe *de $\mathbb{R}^d$ et les points $\{s_0, \ldots, s_d\}$ sont appelés les* sommets *du simplexe. En particulier, le* simplexe unité *de $\mathbb{R}^d$ est l'ensemble de points*

$$\left\{ x \in \mathbb{R}^d;\ x_i \geqslant 0,\ i \in \{1, \ldots, d\},\ \text{et}\ \sum_{i=1}^{d} x_i \leqslant 1 \right\}, \tag{3.6}$$

dont les $(d + 1)$ sommets ont pour coordonnées cartésiennes $(0, \ldots, 0)$ et pour tout $i \in \{1, \ldots, d\}$, $(0, \ldots, 0, 1, 0, \ldots, 0)$, le 1 étant en i-ième position.

De manière équivalente, on peut définir un simplexe de $\mathbb{R}^d$ comme l'image par une transformation affine bijective du simplexe unité. En dimension 2, un simplexe est appelé un *triangle* et en dimension 3, un simplexe est appelé un *tétraèdre*.

Pour $i \in \{0, \ldots, d\}$, on définit F_i comme la *face* de K opposée à un sommet s_i et on définit n_i comme la normale extérieure à K sur F_i ; voir la figure 3.2. En dimension 2, une face est également appelée *arête*.

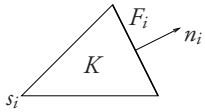


Figure 3.2 – Un triangle K, un sommet s_i, la face opposée F_i et la normale extérieure n_i.

Définition 3.5. *Dans un simplexe K de $\mathbb{R}^d$, on définit les* coordonnées barycentriques $(\lambda_0, \ldots, \lambda_d)$ *telles que pour tout $i \in \{0, \ldots, d\}$,*

$$\lambda_i : \mathbb{R}^d \ni x \longmapsto 1 - \frac{(x - s_i) \cdot n_i}{(s_j - s_i) \cdot n_i} \in \mathbb{R}, \tag{3.7}$$

où s_j est un des sommets de K situés sur F_i, $(x - s_i)$ le vecteur reliant s_i à x et $(s_j - s_i)$ le vecteur reliant s_i à s_j.

La définition (3.7) de λ_i est clairement indépendante du choix particulier du sommet s_j sur la face F_i. La coordonnée barycentrique λ_i est une fonction affine qui vaut 1 en s_i et s'annule sur la face F_i. De plus, ses courbes de niveau sont des hyperplans (des droites si on est en dimension 2) qui sont parallèles à la face F_i. Le barycentre de K a toutes ses coordonnées barycentriques égales à $\frac{1}{d+1}$. Si K est le simplexe unité, on a

en dimension 2, $\quad \lambda_0 = 1 - x_1 - x_2, \lambda_1 = x_1$ et $\lambda_2 = x_2$;
en dimension 3, $\quad \lambda_0 = 1 - x_1 - x_2 - x_3, \lambda_1 = x_1, \lambda_2 = x_2$ et $\lambda_3 = x_3$.

Les coordonnées barycentriques satisfont les propriétés suivantes :

(i) pour tout $x \in K$, $0 \leqslant \lambda_i(x) \leqslant 1$;

(ii) pour tout $x \in \mathbb{R}^d$,

$$\sum_{i=0}^{d} \lambda_i(x) = 1 \qquad \text{et} \qquad x = \sum_{i=0}^{d} \lambda_i(x) s_i. \tag{3.8}$$

On considère l'espace vectoriel des polynômes en les variables $(x_1, \ldots, x_d)$, à coefficients réels et de degré global inférieur ou égal à k. On pose

$$\mathbb{P}_k = \left\{ p(x) = \sum_{\substack{0 \leqslant i_1, \ldots, i_d \leqslant k \\ i_1 + \ldots + i_d \leqslant k}} \alpha_{i_1 \ldots i_d} x_1^{i_1} \ldots x_d^{i_d}; \quad \alpha_{i_1 \ldots i_d} \in \mathbb{R} \right\}. \tag{3.9}$$

$\mathbb{P}_k$ est un espace vectoriel dont la dimension s'exprime en fonction des coefficients binomiaux de Pascal sous la forme

$$\dim \mathbb{P}_k = C_{k+d}^k = \begin{cases} k+1 & \text{si } d = 1, \\ \frac{1}{2}(k+1)(k+2) & \text{si } d = 2, \\ \frac{1}{6}(k+1)(k+2)(k+3) & \text{si } d = 3. \end{cases} \tag{3.10}$$

Proposition 3.6. *Soit K un simplexe de $\mathbb{R}^d$. Soit un entier $k \geqslant 1$. On considère l'ensemble de nœuds $\{a_i\}_{1 \leqslant i \leqslant n_{\mathrm{f}}}$ dont les coordonnées barycentriques sont*

$$\left(\tfrac{i_0}{k}, \ldots, \tfrac{i_d}{k}\right), \quad 0 \leqslant i_0, \ldots, i_d \leqslant k, \;\; i_0 + \ldots + i_d = k, \tag{3.11}$$

et on note Σ les degrés de liberté associés à ces nœuds. Alors, $n_{\mathrm{f}} = \dim \mathbb{P}_k$ et le triplet $\{K, \mathbb{P}_k, \Sigma\}$ est un élément fini de Lagrange, appelé élément fini de Lagrange $\mathbb{P}_k$.

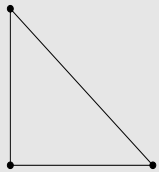
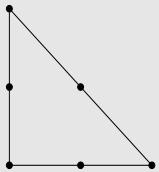
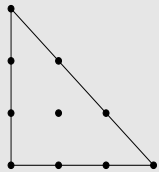
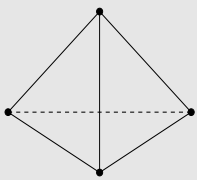
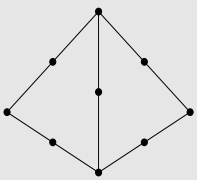
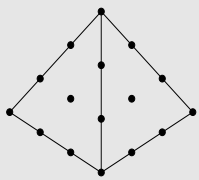
Le tableau 3.1 présente les nœuds et les fonctions de forme pour les éléments finis de Lagrange $\mathbb{P}_1$, $\mathbb{P}_2$ et $\mathbb{P}_3$ en dimension 2 et 3. On notera que pour $k = 1$, les $(d+1)$ fonctions de forme sont les coordonnées barycentriques. En dimension 2 et pour $k = 2$, les fonctions de forme sont

$$\{\lambda_0(2\lambda_0 - 1), \lambda_1(2\lambda_1 - 1), \lambda_2(2\lambda_2 - 1), 4\lambda_0\lambda_1, 4\lambda_0\lambda_2, 4\lambda_1\lambda_2\}.$$

L'espace polynômial $\mathbb{P}_k$ sert, de manière plus générale, à définir le *degré* d'un élément fini de Lagrange quelconque.

Définition 3.7. *Soit $\{K, P, \Sigma\}$ un élément fini de Lagrange. Le plus grand entier k tel que $\mathbb{P}_k \subset P$ est appelé le* degré *de l'élément fini.*

Tableau 3.1 – Degrés de liberté et fonctions de forme pour les éléments finis de Lagrange $\mathbb{P}_1$, $\mathbb{P}_2$ et $\mathbb{P}_3$ en dimension 2 et 3 ; en dimension 3, seuls les degrés de liberté visibles sont représentés.

$\mathbb{P}_1$	$\mathbb{P}_2$	$\mathbb{P}_3$
		
		
$\lambda_i \quad [0 \leqslant i \leqslant d]$	$\lambda_i(2\lambda_i - 1) \quad [0 \leqslant i \leqslant d]$ $4\lambda_i\lambda_j \quad [0 \leqslant i < j \leqslant d]$	$\frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2) \quad [0 \leqslant i \leqslant d]$ $\frac{9}{2}\lambda_i(3\lambda_i - 1)\lambda_j \quad [0 \leqslant i,j \leqslant d,\ i \neq j]$ $27\lambda_i\lambda_j\lambda_k \quad [0 \leqslant i < j < k \leqslant d]$

3.2.3 Éléments finis de Lagrange à structure tensorielle

Définition 3.8. *On considère un ensemble de d intervalles* $\{[c_i, d_i]\}_{1 \leqslant i \leqslant d}$*, tous de mesure non-nulle. L'ensemble* $K = \prod_{i=1}^{d} [c_i, d_i]$ *est appelé un* hypercube *(ou, également, un* pavé*).*

Définition 3.9. *Soit K un hypercube. Pour* $x \in K$*, il existe un unique vecteur de composantes* $(t_1, \ldots, t_d) \in [0,1]^d$ *tel que pour tout* $i \in \{1, \ldots, d\}$*,* $x_i = c_i + t_i(d_i - c_i)$*. Les réels* $(t_1, \ldots, t_d)$ *sont appelées les* coordonnées locales *de x dans K.*

On considère l'espace vectoriel des polynômes en les variables $(x_1, \ldots, x_d)$, à coefficients réels et de degré inférieur ou égal à k en chaque variable. Cet espace est noté

$$\mathbb{Q}_k = \left\{ q(x) = \sum_{0 \leqslant i_1, \ldots, i_d \leqslant k} \alpha_{i_1 \ldots i_d} x_1^{i_1} \ldots x_d^{i_d}; \quad \alpha_{i_1 \ldots i_d} \in \mathbb{R} \right\}. \tag{3.12}$$

$\mathbb{Q}_k$ est un espace vectoriel de dimension

$$\dim \mathbb{Q}_k = (k+1)^d. \tag{3.13}$$

On notera les inclusions $\mathbb{P}_k \subset \mathbb{Q}_k \subset \mathbb{P}_{kd}$ et le fait qu'en dimension un, $\mathbb{P}_k = \mathbb{Q}_k$.

Proposition 3.10. *Soit K un hypercube de $\mathbb{R}^d$. Soit un entier $k \geqslant 1$. On considère l'ensemble de nœuds $\{a_i\}_{1 \leqslant i \leqslant n_{\mathrm{f}}}$ dont les coordonnées locales dans K sont*

$$\left(\tfrac{i_1}{k}, \ldots, \tfrac{i_d}{k}\right), \quad 0 \leqslant i_1, \ldots, i_d \leqslant k, \tag{3.14}$$

et on note Σ les degrés de liberté associés à ces nœuds. Alors, $n_{\mathrm{f}} = \dim \mathbb{Q}_k$ et le triplet $\{K, \mathbb{Q}_k, \Sigma\}$ est un élément fini de Lagrange, appelé élément fini de Lagrange $\mathbb{Q}_k$. *Cet élément fini est de degré k.*

Le tableau 3.2 présente les nœuds et les fonctions de forme pour les éléments finis de Lagrange $\mathbb{Q}_1$, $\mathbb{Q}_2$ et $\mathbb{Q}_3$ en dimension 2 et 3. On rappelle que $\{\mathcal{L}_0^k, \ldots, \mathcal{L}_k^k\}$ sont les polynômes d'interpolation de Lagrange associés aux nœuds $\{a_0, \ldots, a_k\}$ de l'intervalle $[0, 1]$ tels que $a_m = \frac{m}{k}$ pour $m \in \{0, \ldots, k\}$. On notera que les fonctions de forme pour l'élément fini de Lagrange $\mathbb{Q}_k$ s'obtiennent simplement par produit tensoriel des polynômes d'interpolation de Lagrange évalués en les coordonnées locales $(t_1, \ldots, t_d)$ de x dans K.

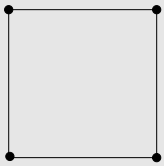
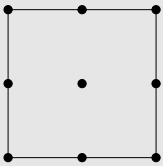
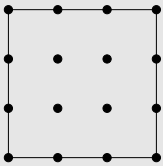
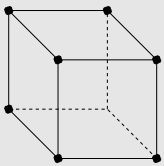
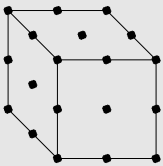
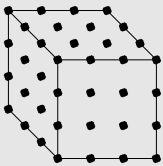
En dimension 2, les fonctions de forme sont pour $k = 1$,

$$\{\mathcal{L}_0^1(t_1)\mathcal{L}_0^1(t_2), \mathcal{L}_1^1(t_1)\mathcal{L}_0^1(t_2), \mathcal{L}_0^1(t_1)\mathcal{L}_1^1(t_2), \mathcal{L}_1^1(t_1)\mathcal{L}_1^1(t_2)\},$$

et pour $k = 2$,

$$\{\mathcal{L}_0^2(t_1)\mathcal{L}_0^2(t_2), \mathcal{L}_0^2(t_1)\mathcal{L}_1^2(t_2), \mathcal{L}_0^2(t_1)\mathcal{L}_2^2(t_2), \mathcal{L}_1^2(t_1)\mathcal{L}_0^2(t_2), \ldots, \mathcal{L}_2^2(t_1)\mathcal{L}_2^2(t_2)\}.$$

Tableau 3.2 – Degrés de liberté et fonctions de forme pour les éléments finis de Lagrange $\mathbb{Q}_1$, $\mathbb{Q}_2$ et $\mathbb{Q}_3$ en dimension 2 et 3 ; en dimension 3, seuls les degrés de liberté visibles sont représentés. $(t_1, \ldots, t_d)$ sont les coordonnées locales du point courant de l'hypercube.

$\mathbb{Q}_1$	$\mathbb{Q}_2$	$\mathbb{Q}_3$
		
		
$\mathcal{L}^1_{i_1}(t_1)\ldots\mathcal{L}^1_{i_d}(t_d)$	$\mathcal{L}^2_{i_1}(t_1)\ldots\mathcal{L}^2_{i_d}(t_d)$	$\mathcal{L}^3_{i_1}(t_1)\ldots\mathcal{L}^3_{i_d}(t_d)$
$[0 \leqslant i_1, \ldots, i_d \leqslant 1]$	$[0 \leqslant i_1, \ldots, i_d \leqslant 2]$	$[0 \leqslant i_1, \ldots, i_d \leqslant 3]$

3.2.4 Éléments finis de Lagrange prismatiques

Dans cette section, on se place en dimension $d \geqslant 3$. Pour $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, on pose $x' = (x_1, \ldots, x_{d-1})$. Soit K' un simplexe de $\mathbb{R}^{d-1}$ et soit $[a, b]$ un intervalle de mesure non-nulle.

Définition 3.11. *L'ensemble* $K = \{x \in \mathbb{R}^d\,;\, x' \in K' \quad \text{et} \quad x_d \in [a, b]\}$ *est appelé un* prisme.

On introduit les coordonnées barycentriques $(\lambda_0, \ldots, \lambda_{d-1})$ de x' dans K' et on considère $t \in [0, 1]$ tel que $x_d = a + t(b - a)$.

Définition 3.12. *On appelle* coordonnées prismatiques *de $x \in K$ le vecteur de composantes* $(\lambda_0, \ldots, \lambda_{d-1}; t)$.

Soit $\mathbb{P}_k[x']$ (respectivement, $\mathbb{P}_k[x_d]$) l'espace des polynômes à coefficients réels en la variable x' (respectivement, x_d) de degré global inférieur ou égal à k. On pose

$$\mathbb{PR}_k = \{p(x) = p_1(x')\, p_2(x_d)\ ;\ p_1 \in \mathbb{P}_k[x'],\ p_2 \in \mathbb{P}_k[x_d]\}. \tag{3.15}$$

On constate que $\mathbb{P}_k \subset \mathbb{PR}_k$ et que $\dim \mathbb{PR}_k = \frac{1}{2}(k+1)^2(k+2)$ en dimension 3.

Proposition 3.13. *Soit K un prisme de $\mathbb{R}^d$. Soit un entier $k \geqslant 1$. On considère l'ensemble de nœuds $\{a_i\}_{1 \leqslant i \leqslant n_{\mathrm{f}}}$ de coordonnées prismatiques*

$$\left(\tfrac{i_0}{k}, \ldots, \tfrac{i_{d-1}}{k}; \tfrac{i_d}{k}\right), \quad 0 \leqslant i_0, \ldots, i_{d-1}, i_d \leqslant k, \quad i_0 + \ldots + i_{d-1} = k, \tag{3.16}$$

et on note Σ les degrés de liberté associés à ces nœuds. Alors, $n_{\mathrm{f}} = \dim \mathbb{PR}_k$ et le triplet $\{K, \mathbb{PR}_k, \Sigma\}$ est un élément fini de Lagrange, appelé élément fini de Lagrange prismatique. *Cet élément fini est de degré k.*

Le tableau 3.3 présente les nœuds et les fonctions de forme pour les éléments finis de Lagrange prismatiques de degré $k \in \{1, 2, 3\}$ en dimension 3. En notant $(\lambda_0, \lambda_1, \lambda_2; t)$ les coordonnées prismatiques du point courant dans le prisme, les fonctions de forme s'expriment comme un produit tensoriel des fonctions de forme associées au triangle K' (et qui font intervenir les coordonnées barycentriques $(\lambda_0, \lambda_1, \lambda_2)$) et des polynômes d'interpolation de Lagrange unidimensionnels en la variable t.

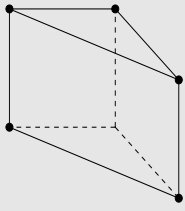
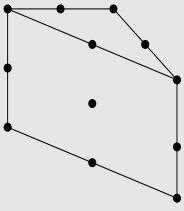
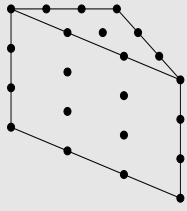
Par exemple, pour $k = 1$, les fonctions de forme sont les suivantes :

$$\{\lambda_0 \mathcal{L}_0^1(t), \lambda_1 \mathcal{L}_0^1(t), \lambda_2 \mathcal{L}_0^1(t), \lambda_0 \mathcal{L}_1^1(t), \lambda_1 \mathcal{L}_1^1(t), \lambda_2 \mathcal{L}_1^1(t)\}.$$

3.3 Notions élémentaires sur les maillages

Intuitivement, un maillage d'un domaine Ω est une partition de Ω en *mailles*. Pour simplifier, on suppose que ces mailles sont des intervalles en dimension 1, des triangles ou des quadrangles en dimension 2 et des tétraèdres, des prismes ou des pavés en dimension 3. Les mailles sont également appelées les *cellules* du maillage.

Tableau 3.3 – Degrés de liberté et les fonctions de forme pour les éléments finis de Lagrange prismatiques de degré $k \in \{1, 2, 3\}$; seuls les degrés de liberté visibles sont représentés. $(\lambda_0, \lambda_1, \lambda_2; t)$ sont les coordonnées prismatiques du point courant dans le prisme.

$\mathbb{PR}_1$	$\mathbb{PR}_2$	$\mathbb{PR}_3$
		
$\lambda_i \mathcal{L}_0^1(t)$ $[0 \leqslant i \leqslant 2]$	$\lambda_i(2\lambda_i - 1)\mathcal{L}_0^2(t)$ $[0 \leqslant i \leqslant 2]$	$\frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2)\mathcal{L}_m^2(t)$ $[0 \leqslant i \leqslant 2]$
$\lambda_i \mathcal{L}_1^1(t)$ $[0 \leqslant i \leqslant 2]$	$\lambda_i(2\lambda_i - 1)\mathcal{L}_1^2(t)$ $[0 \leqslant i \leqslant 2]$	$[0 \leqslant m \leqslant 2]$
	$\lambda_i(2\lambda_i - 1)\mathcal{L}_2^2(t)$ $[0 \leqslant i \leqslant 2]$	$\frac{9}{2}\lambda_i(3\lambda_i - 1)\lambda_j \mathcal{L}_m^2(t)$ $[0 \leqslant i, j \leqslant 2,\ i \neq j]$
	$4\lambda_i\lambda_j \mathcal{L}_0^2(t)$ $[0 \leqslant i < j \leqslant 2]$	$[0 \leqslant m \leqslant 2]$
	$4\lambda_i\lambda_j \mathcal{L}_1^2(t)$ $[0 \leqslant i < j \leqslant 2]$	$27\lambda_0\lambda_1\lambda_2 \mathcal{L}_m^2(t)$ $[0 \leqslant m \leqslant 2]$
	$4\lambda_i\lambda_j \mathcal{L}_2^2(t)$ $[0 \leqslant i < j \leqslant 2]$	

La famille de mailles constituant le maillage sera notée $\{K_m\}_{1 \leqslant m \leqslant N_{\text{ma}}}$, où N_{ma} est le nombre de mailles. Par hypothèse, les mailles sont des fermés et leurs intérieurs sont deux à deux disjoints (il n'y a pas de recouvrement entre les mailles). Par la suite, on pose

$$h_{K_m} = \operatorname{diam}(K_m) = \max_{x_1, x_2 \in K_m} \|x_1 - x_2\|_{\mathbb{R}^d}, \qquad m \in \{1, \ldots, N_{\text{ma}}\}, \quad (3.17)$$

où $\|\cdot\|_{\mathbb{R}^d}$ désigne la norme euclidienne sur $\mathbb{R}^d$. On pose également

$$h = \max_{1 \leqslant m \leqslant N_{\text{ma}}} h_{K_m}, \quad (3.18)$$

et

$$\mathcal{T}_h = \{K_m\}_{1 \leqslant m \leqslant N_{\text{ma}}}. \quad (3.19)$$

Dans les applications, on est souvent amené à considérer une suite de maillages de plus en plus fins, ce qu'on notera conventionnellement $\{\mathcal{T}_h\}_{h>0}$. On parle de *famille de maillages*.

Définition 3.14 (Famille de maillages quasi-uniformes). *On dit que la famille* $\{\mathcal{T}_h\}_{h>0}$ *est* quasi-uniforme *s'il existe une constante* c *telle que*

$$\forall h,\ \forall K \in \mathcal{T}_h, \qquad h_K \geqslant c\,h. \tag{3.20}$$

Lorsque le domaine Ω est un polygone ou un polyèdre, le maillage peut être construit de façon à recouvrir exactement Ω ; on a donc

$$\overline{\Omega} = \bigcup_{m=1}^{N_{\mathrm{ma}}} K_m, \tag{3.21}$$

où $\overline{\Omega}$ désigne l'adhérence du domaine Ω. Par contre, si le domaine Ω est à frontière courbe, le recouvrement n'est pas exact en général ; dans ces conditions, on note Ω_h l'intérieur de $\bigcup_{m=1}^{N_{\mathrm{ma}}} K_m$, si bien que

$$\overline{\Omega}_h = \bigcup_{m=1}^{N_{\mathrm{ma}}} K_m. \tag{3.22}$$

Qu'est-ce qu'au juste un *domaine* ? En dimension 1, un domaine est un intervalle ouvert et borné. En dimension $d \geqslant 2$, les polygones dans $\mathbb{R}^2$ et les polyèdres dans $\mathbb{R}^3$ sont des domaines. Plus généralement, un domaine de $\mathbb{R}^d$ est un ouvert borné et connexe dont la frontière $\partial\Omega$ satisfait certaines propriétés de régularité, à savoir qu'il existe :

(i) deux réels $\alpha > 0$ et $\beta > 0$;

(ii) une famille finie, de cardinal R, de systèmes de coordonnées locales $x^r = (x^{r\prime}, x_d^r)$ pour $r \in \{1, \ldots, R\}$, avec $x^{r\prime} \in \mathbb{R}^{d-1}$ et $x_d^r \in \mathbb{R}$;

(iii) une famille finie de R cartes locales ϕ^r qui sont *lipschitziennes*[1] sur leur domaine de définition $\{x^{r\prime} \in \mathbb{R}^{d-1}\ ;\ |x^{r\prime}| < \alpha\}$;

1. On dit qu'une fonction $f : D \to \mathbb{R}^n$ est lipschitzienne sur son domaine de définition $D \subset \mathbb{R}^m$ s'il existe un réel L tel que pour tout $(x, y) \in D \times D$, $\|f(x) - f(y)\|_{\mathbb{R}^m} \leqslant L\|x - y\|_{\mathbb{R}^n}$.

tels que

$$\partial\Omega = \bigcup_{r=1}^{R} \{(x^{r\prime}, x_d^r)\ ;\ x_d^r = \phi^r(x^{r\prime})\ ;\ |x^{r\prime}| < \alpha\},$$

et pour tout $r \in \{1, \ldots, R\}$,

$$\{(x^{r\prime}, x_d^r)\ ;\ \phi^r(x^{r\prime}) < x_d^r < \phi^r(x^{r\prime}) + \beta\ ;\ |x^{r\prime}| < \alpha\} \subset \Omega,$$
$$\{(x^{r\prime}, x_d^r)\ ;\ \phi^r(x^{r\prime}) - \beta < x_d^r < \phi^r(x^{r\prime})\ ;\ |x^{r\prime}| < \alpha\} \subset \mathbb{R}^d \backslash \overline{\Omega},$$

où $|x^{r\prime}| \leqslant \alpha$ signifie que $|x_i^{r\prime}| \leqslant \alpha$ pour tout $i \in \{1, \ldots, d-1\}$. On notera qu'un domaine est nécessairement situé d'un seul côté de sa frontière. Lorsque les propriétés ci-dessus sont satisfaites, on dit que la frontière de Ω est *lipschitzienne*.

Pour un entier $m \geqslant 1$, on dit qu'un domaine Ω est de *classe* $\mathcal{C}^m$ si toutes les cartes locales ϕ^r sont de classe $\mathcal{C}^m$. Pour un domaine de classe $\mathcal{C}^m$, sa normale extérieure n est définie en tout point de sa frontière.

3.3.1 Génération du maillage

Un maillage est généré à partir d'une *maille de référence*, qu'on note $\widehat{K}$, et d'une famille de *transformations géométriques* envoyant $\widehat{K}$ dans les cellules du maillage. Par la suite, on fait l'hypothèse que ces transformations sont des $\mathcal{C}^1$-difféomorphismes et pour une maille $K \in \mathcal{T}_h$ (on omet l'indice m pour alléger les notations), on note

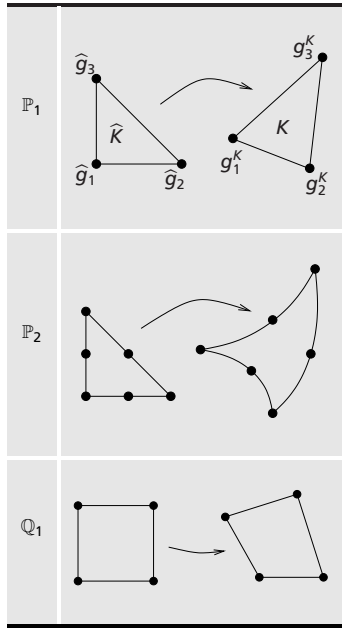
$$T_K : \widehat{K} \to K,$$

la transformation géométrique correspondante.

Comment spécifier la transformation géométrique T_K ? Une façon simple de procéder consiste à utiliser un élément fini de Lagrange. Par la suite, cet élément fini sera noté $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$. On pose $n_{\text{géo}} = \text{card}(\widehat{\Sigma}_{\text{géo}})$, on désigne par $\{\widehat{g}_1, \ldots, \widehat{g}_{n_{\text{géo}}}\}$ les nœuds de $\widehat{K}$ et par $\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{n_{\text{géo}}}\}$ les fonctions de forme correspondantes.

Définition 3.15. *On dit que* $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$ *est* l'élément fini géométrique, $\{\widehat{g}_1, \ldots, \widehat{g}_{n_{\text{géo}}}\}$ *sont les* nœuds géométriques *et* $\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{n_{\text{géo}}}\}$ *sont les* fonctions de forme géométriques.

Tableau 3.4 – Exemples de transformations géométriques générées à partir d'un élément fini de Lagrange.



Pour chaque maille $K \in \mathcal{T}_h$, on dispose d'un $n_{\text{géo}}$-uplet $\{g_1^K, \ldots, g_{n_{\text{géo}}}^K\}$. La transformation géométrique envoyant $\widehat{K}$ dans K est alors définie comme suit :

$$T_K : \widehat{K} \ni \widehat{x} \longmapsto \sum_{i=1}^{n_{\text{géo}}} g_i^K \widehat{\psi}_i(\widehat{x}) \in K, \tag{3.23}$$

si bien que $T_K(\widehat{g}_i) = g_i^K$ pour tout $i \in \{1, \ldots, n_{\text{géo}}\}$. On observera que $T_K \in [\widehat{P}_{\text{géo}}]^d$. Les points $\{g_1^K, \ldots, g_{n_{\text{géo}}}^K\}$ sont appelés les *nœuds géométriques* de K.

Le tableau 3.4 présente des exemples de transformations géométriques générées à partir d'un élément fini de Lagrange en dimension 2. L'élément fini

de Lagrange $\mathbb{P}_1$ permet de transformer le triangle unité de $\mathbb{R}^2$ en un triangle non-dégénéré ; l'élément fini de Lagrange $\mathbb{P}_2$ permet de transformer le triangle unité de $\mathbb{R}^2$ en un triangle courbe ; l'élément fini de Lagrange $\mathbb{Q}_1$ permet de transformer le carré unité de $\mathbb{R}^2$ en un quadrangle non-dégénéré. Par la suite et pour des raisons de simplicité, on supposera que toutes les mailles sont générées à partir du même élément fini géométrique. Lorsque $\widehat{K}$ est un triangle, on dit que le maillage est une *triangulation*.

Remarque 3.16

La transformation géométrique T_K définie en (3.23) étant un $\mathcal{C}^1$-difféomorphisme, la numérotation des nœuds $\{g_1^K, \ldots, g_{n_{\text{géo}}}^K\}$ doit être compatible avec celle adoptée dans l'élément fini géométrique. La figure 3.3 présente un exemple et un contre-exemple en dimension 2 pour l'élément fini de Lagrange $\mathbb{P}_2$.

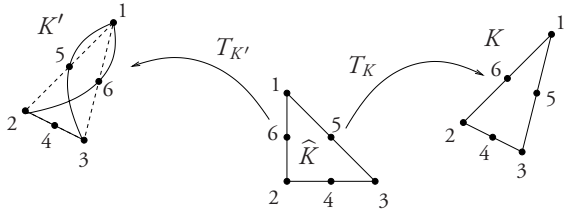


Figure 3.3 – À droite, la numérotation des nœuds de K est compatible avec celle de $\widehat{K}$; à gauche elle ne l'est pas.

Définition 3.17 (Maillage affine). *Lorsque toutes les transformations* $\{T_K\}_{K\in\mathcal{T}_h}$ *sont affines, c'est-à-dire lorsque pour tout* $K \in \mathcal{T}_h$*, il existe un vecteur* $b_K \in \mathbb{R}^d$ *et une matrice* $J_K \in \mathbb{R}^{d,d}$ *tels que*

$$T_K : \widehat{K} \ni \widehat{x} \longmapsto b_K + J_K\widehat{x} \in K, \tag{3.24}$$

on dit que le maillage est affine.

Des exemples de maillages affines sont les suivants :

(i) lorsque l'élément fini géométrique est l'élément fini de Lagrange $\mathbb{P}_1$, toutes les mailles sont des triangles en dimension 2 et des tétraèdres en dimension 3 ;

(ii) lorsque l'élément fini géométrique est l'élément fini de Lagrange $\mathbb{Q}_1$ et qu'on se restreint à des transformations affines, toutes les mailles sont des parallélogrammes (et non des quadrangles) en dimension 2 et des parallélépipèdes (et non des héxaèdres) en dimension 3.

3.3.2 Faces, arêtes et sauts

La maille de référence $\widehat{K}$ étant un polygone en dimension 2 ou un polyèdre en dimension 3, on convient d'appeler *sommets*, *arêtes* et *faces* d'une maille $K = T_K(\widehat{K})$, l'image par T_K des sommets, arêtes et faces de $\widehat{K}$. Lorsque la dimension d'espace n'est pas explicitée, on convient de désigner indifféremment par « faces » les arêtes de K en dimension 2 et les faces de K en dimension 3.

Par la suite, on désigne par $\mathcal{F}_h^{\mathrm{i}}$ l'ensemble des faces intérieures (ou interfaces) dans le maillage et par $\mathcal{F}_h^{\partial}$ l'ensemble des faces situées sur la frontière de Ω_h (on rappelle que Ω_h est le domaine recouvert par le maillage). On pose $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\partial}$. En dimension $d \geqslant 3$, on désigne par $\mathcal{E}_h^{\mathrm{i}}$ et $\mathcal{E}_h^{\partial}$ l'ensemble des arêtes intérieures et des arêtes situées sur la frontière, respectivement, et on pose $\mathcal{E}_h = \mathcal{E}_h^{\mathrm{i}} \cup \mathcal{E}_h^{\partial}$.

Pour $F \in \mathcal{F}_h^{\mathrm{i}}$, il existe deux mailles K_1 et K_2 dans $\mathcal{T}_h$ telles que $F = K_1 \cap K_2$. On désigne par n_1 et n_2 la normale extérieure à K_1 et K_2, respectivement (noter que $n_1 + n_2 = 0$ par définition). Soit v une fonction à valeurs scalaires et définie localement sur chaque maille. On suppose que v est suffisamment régulière pour admettre une limite des deux côtés de F (ces limites n'étant pas forcément les mêmes). Dans ces conditions, on pose $v_1 = v|_{K_1}$ et $v_2 = v|_{K_2}$ et on définit le *saut* de v à-travers F par

$$[\![v]\!]_F = v_1 n_1 + v_2 n_2. \tag{3.25}$$

On notera que $[\![v]\!]_F$ est une quantité vectorielle[1]. Par la suite, s'il n'y a pas d'ambiguïté, on pourra omettre l'indice F. Lorsque la fonction v est à valeurs dans $\mathbb{R}^d$, on notera

$$[\![v{\cdot}n]\!]_F = v_1{\cdot}n_1 + v_2{\cdot}n_2, \tag{3.26}$$

1. L'avantage de ce formalisme par rapport à celui où le saut est défini par la différence $(v_2 - v_1)$ est qu'il n'est pas nécessaire de choisir *a priori* une orientation à-travers F pour calculer le saut.

le saut de la composante normale de v à-travers F. En dimension 3, on définit le saut de la composante tangentielle de v à-travers F par

$$[\![v\times n]\!]_F = v_1\times n_1 + v_2\times n_2. \tag{3.27}$$

3.3.3 Maillages conformes et relations d'Euler

Définition 3.18 (Maillages conformes). *Un maillage* conforme *est un maillage où l'intersection de deux mailles distinctes est soit vide, soit un sommet, soit une arête, soit une face.*

La figure 3.4 présente un exemple et un contre-exemple de maillage conforme. Les maillages conformes présentent deux intérêts : d'une part, ils sont utiles dans la construction d'espaces d'approximation H^1-conformes (voir la section 3.5) et d'autre part, sur de tels maillages, on dispose des relations d'Euler.

Figure 3.4 – À gauche : deux mailles triangulaires dans un maillage conforme ; à droite : trois mailles triangulaires dans un maillage non-conforme.

Lemme 3.19 (Relations d'Euler). *Soit $\mathcal{T}_h = \{K_m\}_{1\leqslant m\leqslant N_{\rm ma}}$ un maillage conforme. On note Ω_h l'intérieur de $\bigcup_{m=1}^{N_{\rm ma}} K_m$.*

(i) *En dimension 2, soit I le degré de multiple connexité*[1] *de Ω_h, $N_{\rm ar}$ le nombre d'arêtes du maillage, $N_{\rm so}$ le nombre de sommets, $N_{\rm ar}^{\partial}$ le nombre d'arêtes situées sur la frontière de Ω_h et $N_{\rm so}^{\partial}$ le nombre de sommets situés sur la frontière. Alors, on a*

$$N_{\rm ma} - N_{\rm ar} + N_{\rm so} = 1 - I, \tag{3.28}$$

$$N_{\rm so}^{\partial} - N_{\rm ar}^{\partial} = 0. \tag{3.29}$$

1. I est le nombre de « trous » dans Ω_h.

De plus, si les mailles sont des polygones à ν sommets, on a

$$2N_{\mathrm{ar}} - N_{\mathrm{ar}}^{\partial} = \nu N_{\mathrm{ma}}. \tag{3.30}$$

En particulier, $2N_{\mathrm{ar}} - N_{\mathrm{ar}}^{\partial} = 3N_{\mathrm{ma}}$ pour un maillage par des triangles et $2N_{\mathrm{ar}} - N_{\mathrm{ar}}^{\partial} = 4N_{\mathrm{ma}}$ pour un maillage par des quadrangles.

(ii) *En dimension 3, soit I le degré de multiple connexité de Ω_h, J le nombre de composantes connexes de la frontière de Ω_h, N_{fa} le nombre de faces du maillage, N_{ar} le nombre d'arêtes, N_{so} le nombre de sommets, $N_{\mathrm{fa}}^{\partial}$ le nombre de faces situées sur la frontière de Ω_h, $N_{\mathrm{ar}}^{\partial}$ le nombre d'arêtes situées sur la frontières et $N_{\mathrm{so}}^{\partial}$ le nombre de sommets situés sur la frontière. Alors, on a*

$$N_{\mathrm{ma}} - N_{\mathrm{fa}} + N_{\mathrm{ar}} - N_{\mathrm{so}} = -1 + I - J, \tag{3.31}$$

$$N_{\mathrm{fa}}^{\partial} - N_{\mathrm{ar}}^{\partial} + N_{\mathrm{so}}^{\partial} = 2(J - I). \tag{3.32}$$

De plus, si les mailles sont des polyèdres à ν faces, on a

$$2N_{\mathrm{fa}} - N_{\mathrm{fa}}^{\partial} = \nu N_{\mathrm{ma}}. \tag{3.33}$$

En particulier, $2N_{\mathrm{fa}} - N_{\mathrm{fa}}^{\partial} = 4N_{\mathrm{ma}}$ pour un maillage par des tétraèdres et $2N_{\mathrm{fa}} - N_{\mathrm{fa}}^{\partial} = 6N_{\mathrm{ma}}$ pour un maillage par des hexaèdres.

3.4 Génération d'éléments finis de Lagrange

On considère un maillage $\mathcal{T}_h$ construit à partir d'un élément fini géométrique $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$. Un maille $K \in \mathcal{T}_h$ est donc l'image de $\widehat{K}$ par la transformation géométrique T_K définie en (3.23).

Étant donné un élément fini de Lagrange $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$, qu'on appelle *élément fini de référence*, l'objectif de cette section est de générer une famille d'éléments finis de Lagrange

$$\{K, P_K, \Sigma_K\}_{K \in \mathcal{T}_h}. \tag{3.34}$$

On désigne par $\{\widehat{a}_1, \ldots, \widehat{a}_{n_{\mathrm{f}}}\}$ les nœuds de l'élément fini de référence et par $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_{\mathrm{f}}}\}$ les fonctions de forme correspondantes.

Proposition 3.20. *Soit $K \in \mathcal{T}_h$. On pose $P_K = \{\widehat{p} \circ T_K^{-1} \, ; \widehat{p} \in \widehat{P}\}$. Pour tout $i \in \{1, \ldots, n_{\mathrm{f}}\}$, on pose $a_{K,i} = T_K(\widehat{a}_i)$. On désigne par Σ_K les degrés de liberté associés aux nœuds $\{a_{K,1}, \ldots, a_{K,n_{\mathrm{f}}}\}$ dans K. Alors, le triplet $\{K, P_K, \Sigma_K\}$ est un élément fini de Lagrange.*

Les *fonctions de forme* de l'élément fini $\{K, P_K, \Sigma_K\}$ sont définies par

$$\theta_{K,i} = \widehat{\theta}_i \circ T_K^{-1}, \qquad i \in \{1, \ldots, n_{\mathrm{f}}\}, \tag{3.35}$$

et *l'opérateur d'interpolation local* par

$$\mathcal{I}_K^{\mathrm{Lag}} : \mathcal{C}^0(K) \ni v \longmapsto \sum_{i=1}^{n_{\mathrm{f}}} v(a_{K,i})\, \theta_{K,i} \in P_K. \tag{3.36}$$

Une propriété importante de $\mathcal{I}_K^{\mathrm{Lag}}$ est que pour tout $v \in \mathcal{C}^0(K)$,

$$\mathcal{I}_{\widehat{K}}^{\mathrm{Lag}}(v \circ T_K) = \mathcal{I}_K^{\mathrm{Lag}}(v) \circ T_K, \tag{3.37}$$

où $\mathcal{I}_{\widehat{K}}^{\mathrm{Lag}}$ désigne l'opérateur d'interpolation local associé à l'élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$. Une deuxième propriété importante de $\mathcal{I}_K^{\mathrm{Lag}}$ est la suivante :

Théorème 3.21. *On suppose que :*

(i) *la transformation T_K est affine ;*

(ii) $\mathbb{P}_k \subset \widehat{P}$ *et* $k + 1 > \frac{d}{2}$.

On note h_K le diamètre de K et ρ_K le diamètre de la plus grande boule inscrite dans K ; voir la figure 3.5. *On pose*

$$\varpi_K = \frac{h_K}{\rho_K}. \tag{3.38}$$

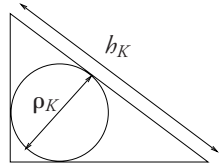


Figure 3.5 – Paramètres géométriques h_K et ρ_K associés à un triangle.

Alors, il existe une constante c, indépendante de K, telle que pour tout $v \in H^{k+1}(K)$ et pour tout $m \in \{0, \ldots, k+1\}$,

$$|v - \mathcal{I}_K^{\mathrm{Lag}} v|_{m,K} \leqslant c\, h_K^{k+1-m} \varpi_K^m |v|_{k+1,K}. \tag{3.39}$$

L'estimation (3.39) montre que la précision de l'opérateur d'interpolation $\mathcal{I}_K^{\mathrm{Lag}}$ est d'autant meilleure que le paramètre ϖ_K est petit (ϖ_K est, par définition, supérieur à 1). Par exemple, si K est un triangle, le paramètre ϖ_K est inversement proportionnel à $\sin \varphi_K$ où φ_K est le plus petit angle de K ; il empire donc avec l'aplatissement du triangle K. Par ailleurs, l'hypothèse $k + 1 > \frac{d}{2}$ est une hypothèse technique qui garantit que $H^{k+1}(K) \subset \mathcal{C}^0(K)$; on observera que cette hypothèse n'est pas restrictive en dimension 2 et 3 dès que $k \geqslant 1$. Enfin, lorsque la fonction à interpoler n'est pas suffisamment régulière, par exemple si $v \in H^s(K)$ et $v \notin H^{s+1}(K)$ pour un entier $s \leqslant k$, l'estimation (3.39) devient

$$|v - \mathcal{I}_K^{\mathrm{Lag}} v|_{m,K} \leqslant c\, h_K^{s-m} \varpi_K^m |v|_{s,K}, \tag{3.40}$$

pour tout $m \in \{0, \ldots, s\}$.

3.5 Espaces H^1-conformes

Définition 3.22. *On dit qu'un espace vectoriel V_h de fonctions définies sur un domaine Ω_h est H^1-conforme si $V_h \subset H^1(\Omega_h)$.*

3.5.1 Construction

Soit $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ un élément fini de référence. Soit $\mathcal{T}_h$ un maillage de Ω. On désigne par $\{K, P_K, \Sigma_K\}_{K \in \mathcal{T}_h}$ la famille d'éléments finis de Lagrange générés selon la proposition 3.20. On pose

$$W_h = \{v_h \in L^2(\Omega_h)\ ;\ \forall K \in \mathcal{T}_h,\ v_h|_K \in P_K\}. \tag{3.41}$$

Il est important d'observer qu'à ce stade, l'espace W_h n'est pas H^1-conforme car les fonctions de W_h peuvent être discontinues aux interfaces entre les mailles. On pose

$$V_h = W_h \cap \mathcal{C}^0(\overline{\Omega}_h) = \{v_h \in W_h\ ;\ \forall F \in \mathcal{F}_h^{\mathrm{i}},\ [\![v_h]\!]_F = 0\}. \tag{3.42}$$

On a le résultat suivant.

Proposition 3.23. $V_h \subset H^1(\Omega_h)$.

Afin de construire simplement une base nodale de l'espace V_h, on suppose que le maillage $\mathcal{T}_h$ est affine et conforme (voir les définitions 3.17 et 3.18) et on fait les hypothèses suivantes sur l'élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$.

(i) Toutes les faces de $\widehat{K}$ ont le même nombre de nœuds et la disposition de ces nœuds est la même sur chaque face[1].

(ii) La restriction d'une fonction de $\widehat{P}$ à une face de $\widehat{K}$ est déterminée de manière univoque par les valeurs que cette fonction prend aux nœuds de $\widehat{K}$ situés sur cette face.

Tous les éléments finis de Lagrange présentés dans la section 3.2 satisfont ces deux hypothèses.

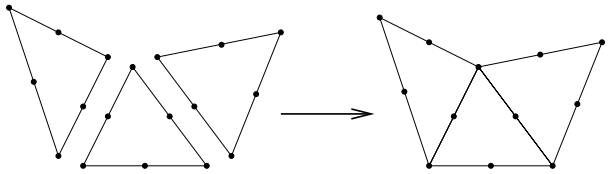


Figure 3.6 – Réunion des nœuds de Lagrange associés aux différentes mailles ; exemple pour l'élément fini de Lagrange $\mathbb{P}_2$ en dimension 2.

On note $\{a_1, \ldots, a_N\} = \bigcup_{K \in \mathcal{T}_h} \{a_{K,1}, \ldots, a_{K,n_\mathrm{f}}\}$ la réunion des nœuds de tous les éléments du maillage, les nœuds situés aux interfaces n'étant comptabilisés qu'une fois ; voir la figure 3.6. Soit a_i un nœud du maillage.

(i) Si le nœud a_i appartient à l'intérieur d'une maille K, on définit φ_i comme la fonction qui coïncide avec la fonction de forme associée au nœud a_i dans K et qui vaut zéro sur les autres mailles.

1. Cette hypothèse signifie qu'en considérant deux faces $\widehat{F}_1$ et $\widehat{F}_2$ de $\widehat{K}$, pour toute transformation affine envoyant $\widehat{F}_1$ sur $\widehat{F}_2$, l'image de l'ensemble des nœuds de $\widehat{F}_1$ est l'ensemble des nœuds de $\widehat{F}_2$.

(ii) Si le nœud a_i est partagé par plusieurs mailles, on définit φ_i comme la fonction qui coïncide avec la fonction de forme associée au nœud a_i sur chacune de ces mailles et qui vaut zéro sur les autres mailles.

Proposition 3.24. *Pour tout $i \in \{1, \dots, N\}$, $\varphi_i \in V_h$. De plus, la famille $\{\varphi_1, \dots, \varphi_N\}$ forme une base de V_h.*

On dit que les fonctions $\{\varphi_1, \dots, \varphi_N\}$ forment la *base nodale* de V_h. Ces fonctions sont également appelées les *fonctions de forme* dans V_h. Pour tout $i \in \{1, \dots, N\}$, on définit la forme linéaire

$$\gamma_i : V_h \in v_h \longmapsto v_h(a_i) \in \mathbb{R}. \tag{3.43}$$

Il est clair que la famille $\{\gamma_1, \dots, \gamma_N\}$ forme une base de $\mathcal{L}(V_h; \mathbb{R})$. Les formes linéaires $\{\gamma_1, \dots, \gamma_N\}$ sont appelées les *degrés de liberté* dans V_h.

3.5.2 Exemples fondamentaux

Deux exemples fondamentaux d'espaces H^1-conformes sont les suivants[1] :

$$P^k_{c,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega}_h)\ ;\ \forall K \in \mathcal{T}_h,\ v_h \circ T_K \in \mathbb{P}_k\}, \tag{3.44}$$

$$Q^k_{c,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega}_h)\ ;\ \forall K \in \mathcal{T}_h,\ v_h \circ T_K \in \mathbb{Q}_k\}. \tag{3.45}$$

Le tableau 3.5 indique la dimension de ces espaces pour $k \in \{1, 2, 3\}$. La figure 3.7 illustre les fonctions de forme dans les espaces $P^1_{c,h}$ et $P^2_{c,h}$ en dimension 2. Les fonctions de forme dans $P^1_{c,h}$ sont également appelées *fonctions chapeau* en référence à la forme de leur graphe. Par la suite, on considérera également les espaces suivants :

$$P^k_{c,h,0} = \{v_h \in P^k_{c,h}\ ;\ v_h|_{\partial\Omega_h} = 0\}, \tag{3.46}$$

$$Q^k_{c,h,0} = \{v_h \in Q^k_{c,h}\ ;\ v_h|_{\partial\Omega_h} = 0\}. \tag{3.47}$$

1. Les transformations T_K étant affines, on a :

$$P^k_{c,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega}_h)\ ;\ \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{P}_k\}.$$

Tableau 3.5 – Dimension des espaces $P^k_{c,h}$ et $Q^k_{c,h}$ en dimension 2 et 3 pour $k \in \{1, 2, 3\}$. Les notations sont définies dans le lemme 3.19.

espace	d	$k = 1$	$k = 2$	$k = 3$
$P^k_{c,h}$	2	N_{so}	$N_{so} + N_{ar}$	$N_{so} + 2N_{ar} + N_{ma}$
$Q^k_{c,h}$	2	N_{so}	$N_{so} + N_{ar} + N_{ma}$	$N_{so} + 2N_{ar} + 4N_{ma}$
$P^k_{c,h}$	3	N_{so}	$N_{so} + N_{ar}$	$N_{so} + 2N_{ar} + N_{fa}$
$Q^k_{c,h}$	3	N_{so}	$N_{so} + N_{ar} + N_{fa} + N_{ma}$	$N_{so} + 2N_{ar} + 4N_{fa} + 8N_{ma}$

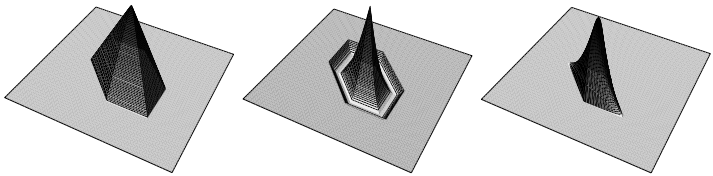


Figure 3.7 – Fonctions de forme dans les espaces $P^1_{c,h}$ et $P^2_{c,h}$ en dimension 2 ; à gauche : fonction chapeau dans l'espace $P^1_{c,h}$; au milieu : fonction de forme dans $P^2_{c,h}$ associée à un sommet du maillage ; à droite : fonction de forme dans $P^2_{c,h}$ associée à un milieu d'arête du maillage (son support est réduit aux deux mailles partageant l'arête en question).

3.5.3 Projections orthogonales

Soit un entier $k \geqslant 1$. On considère les projections orthogonales

$$\Pi^{0,k}_{c,h} : L^2(\Omega) \longrightarrow P^k_{c,h}, \tag{3.48}$$

$$\Pi^{1,k}_{c,h} : H^1(\Omega) \longrightarrow P^k_{c,h}, \tag{3.49}$$

associées aux produits scalaires :

$$(v, w)_{0,\Omega} = \int_\Omega vw \quad \text{et} \quad (v, w)_{1,\Omega} = \int_\Omega vw + \int_\Omega \nabla v \cdot \nabla w, \text{ respectivement.}$$

En d'autres termes, on a

$$\forall v \in L^2(\Omega), \quad \forall v_h \in P^k_{c,h}, \qquad (\Pi^{0,k}_{c,h}(v), v_h)_{0,\Omega} = (v, v_h)_{0,\Omega}, \tag{3.50}$$

$$\forall v \in H^1(\Omega), \quad \forall v_h \in P^k_{c,h}, \qquad (\Pi^{1,k}_{c,h}(v), v_h)_{1,\Omega} = (v, v_h)_{1,\Omega}. \tag{3.51}$$

L'opérateur $\Pi^{1,k}_{c,h}$ est appelé *projecteur elliptique*. On peut également considérer la projection orthogonale de $H^1_0(\Omega)$ dans $P^k_{c,h,0}$ avec le produit scalaire $\int_\Omega \nabla v \cdot \nabla w$. Ce projecteur satisfait les mêmes propriétés que celles énoncées ci-dessous pour l'opérateur $\Pi^{1,k}_{c,h}$.

Les projecteurs $\Pi^{0,k}_{c,h}$ et $\Pi^{1,k}_{c,h}$ satisfont les propriétés de stabilité suivantes :

$$\forall v \in L^2(\Omega), \quad \|\Pi^{0,k}_{c,h} v\|_{0,\Omega} \leqslant \|v\|_{0,\Omega}, \tag{3.52}$$

$$\forall v \in H^1(\Omega), \quad \|\Pi^{1,k}_{c,h} v\|_{1,\Omega} \leqslant \|v\|_{1,\Omega}. \tag{3.53}$$

De plus, si la famille $\{\mathcal{T}_h\}_{h>0}$ est quasi-uniforme, il existe une constante c, indépendante de h, telle que

$$\forall v \in H^1(\Omega), \qquad \|\Pi^{0,k}_{c,h} v\|_{1,\Omega} \leqslant c\, \|v\|_{1,\Omega}. \tag{3.54}$$

De plus, les projecteurs $\Pi^{0,k}_{c,h}$ et $\Pi^{1,k}_{c,h}$ sont tels que :

(i) il existe une constante c telle que pour tout h, pour tout $l \in \{1, \ldots, k\}$ et pour tout $v \in H^{l+1}(\Omega)$,

$$\|v - \Pi^{0,k}_{c,h}(v)\|_{0,\Omega} \leqslant c\, h^{l+1} |v|_{l+1,\Omega}, \tag{3.55}$$

$$\|v - \Pi^{1,k}_{c,h}(v)\|_{1,\Omega} \leqslant c\, h^{l} |v|_{l+1,\Omega}; \tag{3.56}$$

(ii) si Ω est convexe, il existe une constante c telle que pour tout h, pour tout $l \in \{1, \ldots, k\}$ et pour tout $v \in H^{l+1}(\Omega)$,

$$\|v - \Pi^{1,k}_{c,h}(v)\|_{0,\Omega} \leqslant c\, h^{l+1} |v|_{l+1,\Omega}; \tag{3.57}$$

(iii) si la famille $\{\mathcal{T}_h\}_{h>0}$ est quasi-uniforme, il existe une constante c telle que pour tout h, pour tout $l \in \{1, \ldots, k\}$ et pour tout $v \in H^{l+1}(\Omega)$,

$$\|v - \Pi^{0,k}_{c,h}(v)\|_{1,\Omega} \leqslant c\, h^{l} |v|_{l+1,\Omega}. \tag{3.58}$$

Des résultats analogues existent pour des opérateurs de projection orthogonale sur les espaces $Q^k_{c,h}$ définis en (3.45).

3.6 Interpolé de Lagrange sur un maillage

Soit $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ un élément fini de référence. Soit $\mathcal{T}_h$ un maillage affine et conforme de Ω. On désigne par $\{K, P_K, \Sigma_K\}_{K\in\mathcal{T}_h}$ la famille d'éléments finis de Lagrange générés selon la proposition 3.20. Soit V_h l'espace H^1-conforme construit dans la section 3.5. On désigne par $\{\varphi_1, \ldots, \varphi_N\}$ la base nodale de V_h et par $\{a_1, \ldots, a_N\}$ les nœuds associés. L'opérateur d'interpolation de Lagrange est défini comme suit :

$$\mathcal{I}_h^{\mathrm{Lag}} : \mathcal{C}^0(\overline{\Omega}_h) \ni v \longmapsto \sum_{i=1}^{N} v(a_i)\varphi_i \in V_h. \tag{3.59}$$

On observera que pour tout $K \in \mathcal{T}_h$ et pour tout $v \in \mathcal{C}^0(\overline{\Omega}_h)$,

$$(\mathcal{I}_h^{\mathrm{Lag}} v)|_K = \mathcal{I}_K^{\mathrm{Lag}}(v|_K). \tag{3.60}$$

En d'autres termes, la restriction de l'interpolé de Lagrange à une cellule du maillage coïncide avec l'interpolé de Lagrange local appliqué à la restriction de la fonction à interpoler.

Définition 3.25 (Régularité d'une famille de maillages). *On dit que la famille de maillages* $\{\mathcal{T}_h\}_{h>0}$ *est* régulière *s'il existe une constante* ϖ_0 *telle que*

$$\forall h,\ \forall K \in \mathcal{T}_h, \qquad \varpi_K \leqslant \varpi_0, \tag{3.61}$$

où ϖ_K *est défini en* (3.38).

Théorème 3.26. *On suppose que :*

(i) $\{\mathcal{T}_h\}_{h>0}$ *est une famille régulière de maillages affines et conformes ;*

(ii) $\mathbb{P}_k \subset \widehat{P}$ *et* $k + 1 > \frac{d}{2}$.

Alors, il existe une constante c telle que pour tout h et pour tout $v \in H^{k+1}(\Omega_h)$,

$$\|v - \mathcal{I}_h^{\mathrm{Lag}} v\|_{0,\Omega_h} + h|v - \mathcal{I}_h^{\mathrm{Lag}} v|_{1,\Omega_h} \leqslant c\, h^{k+1} |v|_{k+1,\Omega_h}. \tag{3.62}$$

De plus, pour tout $m \in \{2, \ldots, k+1\}$, *on a*

$$\left(\sum_{K\in\mathcal{T}_h} |v - \mathcal{I}_h^{\mathrm{Lag}} v|_{m,K}^2 \right)^{\frac{1}{2}} \leqslant c\, h^{k+1-m} |v|_{k+1,\Omega_h}. \tag{3.63}$$

Ce théorème est une conséquence directe du théorème d'interpolation local 3.21. On notera que l'hypothèse portant sur la régularité de la famille $\{\mathcal{T}_h\}_{h>0}$ permet de contrôler uniformément les rapports ϖ_K qui apparaissent dans (3.39). Enfin, lorsque la fonction à interpoler n'est pas suffisamment régulière, par exemple si $v \in H^s(\Omega_h)$ et $v \notin H^{s+1}(\Omega_h)$ pour un entier $s \in \{1, \ldots, k\}$, l'estimation (3.62) devient

$$\|v - \mathcal{I}_h^{\mathrm{Lag}} v\|_{0,\Omega_h} + h|v - \mathcal{I}_h^{\mathrm{Lag}} v|_{1,\Omega_h} \leqslant c\, h^s |v|_{s,\Omega_h}. \tag{3.64}$$

3.7 Interpolation isoparamétrique

Lorsque le problème modèle est posé sur un domaine à frontière courbe et qu'on souhaite utiliser un élément fini de référence de degré élevé, on n'obtiendra des propriétés d'interpolation optimales que si la frontière du domaine est décrite avec une précision suffisante. On doit donc utiliser un élément fini géométrique de degré élevé. Ces considérations motivent la définition suivante.

Définition 3.27 (Interpolation iso- et subparamétrique). *Soit* $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ *l'élément fini de référence et soit* $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$ *l'élément fini géométrique. Lorsque ces deux éléments finis sont les mêmes, on parle d'interpolation* isoparamétrique. *On parle d'interpolation* subparamétrique *lorsque* $\widehat{P}_{\text{géo}} \subset \widehat{P}$ *et* $\widehat{P}_{\text{géo}} \neq \widehat{P}$.

On suppose par exemple que l'élément fini de référence est l'élément fini de Lagrange $\mathbb{P}_2$ en dimension deux. Si l'élément fini géométrique est l'élément fini de Lagrange $\mathbb{P}_1$ (le maillage est alors composé de triangles), l'interpolation est subparamétrique. Si l'élément fini géométrique est l'élément fini de Lagrange $\mathbb{P}_2$ (le maillage peut alors contenir des triangles courbes ; voir ci-dessous pour un exemple de construction), l'interpolation est isoparamétrique.

3.7.1 Un exemple de maillage par des triangles courbes

Cette section présente une technique simple permettant de générer à partir d'un maillage affine de triangles un maillage contenant des triangles courbes qui décrivent mieux la frontière du domaine lorsque celle-ci est elle-même courbe. Le principe consiste à ne modifier que les triangles ayant une arête de

bord, c'est-à-dire une arête dont les deux sommets sont sur la frontière, et à conserver tels quels les autres triangles.

Pour fixer les idées, on présente l'algorithme dans le cas où l'élément fini géométrique est l'élément fini de Lagrange $\mathbb{P}_2$. Les arêtes des triangles courbes sont donc des coniques.

(i) Soit $\widetilde{K}$ un triangle dont une des arêtes a ses deux sommets sur la frontière $\partial\Omega$. On note $\{b_1, \ldots, b_{n_{\text{géo}}}\}$ les nœuds géométriques de $\widetilde{K}$. Dans l'exemple considéré, $n_{\text{géo}} = 6$ et les nœuds géométriques du triangle $\widetilde{K}$ sont ses trois sommets et les milieux de ses trois arêtes.

(ii) Pour chaque b_i avec $i \in \{1, \ldots, n_{\text{géo}}\}$, on construit un nouveau nœud g_i de la manière suivante :
- Si b_i est situé sur une arête ayant ses deux sommets sur la frontière, on construit le point g_i comme l'intersection avec $\partial\Omega$ de la normale à l'arête en question passant par b_i.
- Sinon, on pose simplement $g_i = b_i$.

(iii) On remplace $\widetilde{K}$ par le triangle courbe $K = T_K(\widehat{K})$ où

$$\forall \widehat{x} \in \widehat{K}, \qquad T_K(\widehat{x}) = \sum_{i=1}^{n_{\text{géo}}} g_i \widehat{\psi}_i(\widehat{x}), \tag{3.65}$$

$\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{n_{\text{géo}}}\}$ étant les fonctions de forme de l'élément fini de Lagrange $\mathbb{P}_2$.

L'algorithme ci-dessus est illustré sur la figure 3.8.

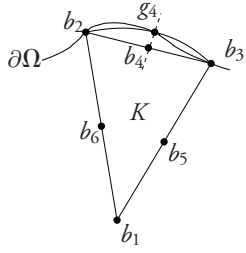


Figure 3.8 – Construction d'un triangle courbe près de la frontière.

3.7.2 Un résultat d'interpolation isoparamétrique

En s'inspirant de la construction présentée dans la section précédente, on considère ici des maillages non-affines obtenus par transformation de maillages affines. La question est alors de savoir comment étendre la notion de régularité (voir la définition 3.25) à de tels maillages.

Soit $\{\widetilde{\mathcal{T}}_h\}_{h>0}$ une famille de maillages *affines*. On pose $\widetilde{\Omega}_h = \bigcup_{\widetilde{K}\in\widetilde{\mathcal{T}}_h} \widetilde{K}$. Soit un entier $k_{\text{géo}} \geqslant 2$. On considère une transformation $F_h : \widetilde{\Omega}_h \to \Omega_h = F_h(\widetilde{\Omega}_h)$ telle que pour tout $\widetilde{K} \in \widetilde{\mathcal{T}}_h$, $F_h|_{\widetilde{K}} \in [\mathbb{P}_{k_{\text{géo}}}]^d$. Pour tout $h > 0$, cette transformation permet de construire un nouveau maillage $\mathcal{T}_h = \{F_h(\widetilde{K})\}_{\widetilde{K}\in\widetilde{\mathcal{T}}_h}$.

Définition 3.28. *On dit que la famille de maillages $\{\mathcal{T}_h\}_{h>0}$ est* régulière *si la famille $\{\widetilde{\mathcal{T}}_h\}_{h>0}$ est régulière selon la définition* 3.25 *et si les transformations $\{F_h\}_{h>0}$ satisfont les propriétés suivantes :*

(i) *la restriction de F_h à une maille $\widetilde{K}$ de $\widetilde{\mathcal{T}}_h$ est l'identité si $\partial\widetilde{K} \cap \partial\widetilde{\Omega}_h = \varnothing$;*

(ii) $\sup_{x\in\widetilde{\partial\Omega}_h} \operatorname{dist}(x, \partial\Omega_h) \leqslant c\, h^{k_{\text{géo}}+1}$ *où la constante c est indépendante de h ;*

(iii) *la matrice jacobienne de F_h et son inverse sont bornées sur Ω_h uniformément en h, ainsi que toutes leurs dérivées jusqu'à l'ordre $k_{\text{géo}}$.*

Soit $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ un élément fini de Lagrange tel que $\mathbb{P}_k \subset \widehat{P}$. On note $\widetilde{\mathcal{I}}_h^{\text{Lag}}$ l'opérateur d'interpolation construit en utilisant $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ comme élément fini de référence et $\widetilde{\mathcal{T}}_h$ comme maillage. Pour $v \in \mathcal{C}^0(\overline{\Omega}_h)$, on pose

$$\mathcal{I}_h^{\text{Lag}} v = [\widetilde{\mathcal{I}}_h^{\text{Lag}}(v \circ F_h)] \circ F_h^{-1}. \tag{3.66}$$

La fonction $v \circ F_h$ est dans $\mathcal{C}^0(\widetilde{\Omega}_h)$; elle appartient donc au domaine de $\widetilde{\mathcal{I}}_h^{\text{Lag}}$. De plus, on observe que $\mathcal{I}_h^{\text{Lag}} v$ est une fonction définie sur Ω_h. Pour simplifier, on suppose que les degrés de liberté de l'élément fini de Lagrange ont été choisis de sorte que l'interpolé de v soit dans $H^1(\Omega_h)$; voir la section 3.5.

Théorème 3.29. *On suppose que $k + 1 > \frac{d}{2}$ et que la famille de maillages $\{\mathcal{T}_h\}_{h>0}$ est régulière selon la définition* 3.28 *avec $k_{\text{géo}} = k$. Alors, il existe une constante c telle que pour tout h et pour tout $v \in H^{k+1}(\Omega_h)$,*

$$\|v - \mathcal{I}_h^{\text{Lag}} v\|_{0,\Omega_h} + h|v - \mathcal{I}_h^{\text{Lag}} v|_{1,\Omega_h} \leqslant c\, h^{k+1} |v|_{k+1,\Omega_h}. \tag{3.67}$$

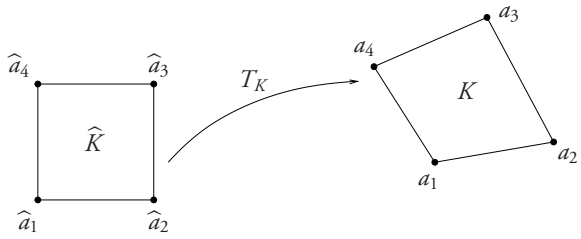


Figure 3.9 – Transformation non-affine du carré de référence en un quadrangle.

On pourra consulter Brenner et Scott [**20**, p. 117] pour la preuve de ce résultat. Voir également Ciarlet [**28**] pour une autre extension de la notion de régularité.

3.7.3 Interpolation sur des quadrangles

On considère un domaine polygonal de $\mathbb{R}^2$ qu'on suppose maillé par des quadrangles. L'élément fini géométrique est donc l'élément fini de Lagrange $\mathbb{Q}_1$. Ceci implique, en particulier, que les transformations T_K ne sont pas nécessairement affines.

Pour un quadrangle $K \in \mathcal{T}_h$, on note h_K son diamètre et on pose $\rho_K = \min_{1\leqslant i\leqslant 4} \rho_i$ où ρ_i est le diamètre du cercle inscrit dans le triangle formé par les trois sommets $(a_j)_{j\neq i}$ de K. On pose

$$\varpi_K = \frac{h_K}{\rho_K}. \tag{3.68}$$

On dit que la famille de maillages $\{\mathcal{T}_h\}_{h>0}$ est régulière s'il existe ϖ_0 tel que pour tout h et pour tout $K \in \mathcal{T}_h$, $\varpi_K \leqslant \varpi_0$.

On suppose que l'élément fini de référence est l'élément fini de Lagrange $\mathbb{Q}_k$ et que $k + 1 > \frac{d}{2}$. Pour tout $K \in \mathcal{T}_h$, on considère l'opérateur d'interpolation local $\mathcal{I}_K^{\mathrm{Lag}}$ défini en (3.36). On considère également l'opérateur d'interpolation global $\mathcal{I}_h^{\mathrm{Lag}}$ défini en (3.59). On a le résultat suivant.

Théorème 3.30. *Il existe une constante c telle que pour tout $K \in \mathcal{T}_h$, pour tout $v \in H^{k+1}(\Omega)$ et pour tout $m \in \{1, \ldots, k+1\}$,*

$$\|v - \mathcal{I}_K^{\mathrm{Lag}} v\|_{0,K} \leqslant c\, \boldsymbol{\varpi}_K h_K^{k+1} |v|_{k+1,K}, \tag{3.69}$$

$$|v - \mathcal{I}_K^{\mathrm{Lag}} v|_{m,K} \leqslant c\, \boldsymbol{\varpi}_K^{4m-1} h_K^{k+1-m} |v|_{k+1,K}. \tag{3.70}$$

De plus, si $\{\mathcal{T}_h\}_{h>0}$ est une famille régulière de maillages conformes, on a pour tout h et pour tout $v \in H^{k+1}(\Omega)$,

$$\|v - \mathcal{I}_h^{\mathrm{Lag}} v\|_{0,\Omega} + h|v - \mathcal{I}_h^{\mathrm{Lag}} v|_{1,\Omega} \leqslant c\, h^{k+1} |v|_{k+1,\Omega}, \tag{3.71}$$

et pour tout $m \in \{2, \ldots, k+1\}$,

$$\left(\sum_{K \in \mathcal{T}_h} \|v - \mathcal{I}_h^{\mathrm{Lag}} v\|_{m,K}^2 \right)^{\frac{1}{2}} \leqslant c\, h^{k+1-m} |v|_{k+1,\Omega}. \tag{3.72}$$

On pourra consulter Girault et Raviart [**46**, p. 104] pour plus de détails.

4 • AUTRES ÉLÉMENTS FINIS

Ce chapitre poursuit l'étude commencée dans le chapitre 3 sur les éléments finis de Lagrange en l'étendant à d'autres types d'éléments finis pour lesquels les degrés de liberté ne sont pas nécessairement définis à partir de valeurs ponctuelles. On donne d'abord une définition générale d'un élément fini et on construit l'opérateur d'interpolation local qui lui est associé. Puis, à l'aide d'un maillage, on construit un opérateur d'interpolation global. Enfin, on étudie l'élément fini de Crouzeix–Raviart, l'élément fini de Raviart–Thomas, l'élément fini de Nédélec et les éléments finis de degré élevé.

4.1 Définition générale d'un élément fini

Conformément à la définition introduite par Ciarlet [**27, 28**], un élément fini est défini comme un triplet $\{K, P, \Sigma\}$ satisfaisant les conditions ci-dessous.

Définition 4.1. *Un élément fini est un triplet* $\{K, P, \Sigma\}$ *où :*

(i) *K est une partie compacte, connexe, d'intérieur non-vide de $\mathbb{R}^d$ dont la frontière est lipschitzienne ; par exemple, un intervalle en dimension 1, un polygone en dimension 2 ou un polyèdre en dimension 3 ;*

(ii) *P est un espace vectoriel de fonctions (en général polynômiales) $p : K \to \mathbb{R}^\mu$ où μ est un entier positif ;*

(iii) *Σ est un ensemble de n_f formes linéaires $\{\sigma_1, \ldots, \sigma_{n_f}\}$ agissant sur les éléments de P et tel que l'application linéaire*

$$P \ni p \longmapsto \big(\sigma_1(p), \ldots, \sigma_{n_f}(p)\big)^T \in \mathbb{R}^{n_f}, \tag{4.1}$$

soit bijective ; ce qui revient à dire que Σ est une base de $\mathcal{L}(P; \mathbb{R})$. Les formes linéaires $\{\sigma_1, \ldots, \sigma_{n_f}\}$ sont appelées les degrés de liberté *de l'élément fini.*

Soit $\{K, P, \Sigma\}$ un élément fini. Puisque Σ est une base de $\mathcal{L}(P;\mathbb{R})$, il existe dans P une base duale de Σ. Cette base de P est notée $\{\theta_1, \ldots, \theta_{n_f}\}$. Par définition, on a

$$\sigma_i(\theta_j) = \delta_{ij}, \qquad i,j \in \{1, \ldots, n_f\}. \tag{4.2}$$

On rappelle que δ_{ij} désigne le symbole de Kronecker tel que $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$. Les fonctions $\{\theta_1, \ldots, \theta_{n_f}\}$ sont appelées les *fonctions de forme* de l'élément fini.

La condition (iii) de la définition 4.1 est équivalente au fait que pour tout $(\alpha_1, \ldots, \alpha_{n_f})^T \in \mathbb{R}^{n_f}$, il existe un et un seul $p \in P$ tel que

$$\sigma_i(p) = \alpha_i, \qquad i \in \{1, \ldots, n_f\}.$$

Cette condition est également équivalente aux propriétés suivantes :

$$\begin{cases} \dim P = \operatorname{card} \Sigma = n_f, \\ \forall p \in P, \ \big(\sigma_i(p) = 0, \ i \in \{1, \ldots, n_f\}\big) \implies (p = 0). \end{cases}$$

Cette propriété est connue sous le nom de *propriété d'unisolvance*. Dans la littérature, on rencontre parfois une définition légèrement différente d'un élément fini où la bijectivité de l'application définie dans (4.1) n'est pas requise pour que le triplet $\{K, P, \Sigma\}$ soit un élément fini et si cette propriété est satisfaite, on parle d'élément fini unisolvant.

Définition 4.2. *Soit $\{K, P, \Sigma\}$ un élément fini. Le plus grand entier k tel que $[\mathbb{P}_k]^\mu \subset P$ est appelé le* degré *de l'élément fini.*

Remarque 4.3

En général, les fonctions de P sont à valeurs scalaires si bien que $\mu = 1$. Deux exemples d'éléments finis à valeurs vectorielles où $\mu = d$ sont l'élément fini de Raviart–Thomas et l'élément fini de Nédélec ; voir les sections 4.5 et 4.6, respectivement.

4.2 Opérateur d'interpolation local

Soit $\{K, P, \Sigma\}$ un élément fini. L'objectif de cette section est de construire à partir de $\{K, P, \Sigma\}$ un opérateur d'interpolation local, c'est-à-dire un opérateur qui permette d'approcher des fonctions définies sur K (et à valeurs dans $\mathbb{R}^\mu$) par des éléments de P dont les degrés de liberté sont convenablement choisis.

Étant donné une fonction v, il semble naturel de définir son interpolé par

$$\sum_{i=1}^{n_f} \sigma_i(v)\theta_i. \tag{4.3}$$

En effet, cet interpolé appartient bien à P et ses degrés de liberté sont les mêmes que ceux de v puisque l'on a pour tout $j \in \{1, \ldots, n_f\}$,

$$\sigma_j\left(\sum_{i=1}^{n_f} \sigma_i(v)\theta_i\right) = \sum_{i=1}^{n_f} \sigma_i(v)\sigma_j(\theta_i) = \sum_{i=1}^{n_f} \sigma_i(v)\delta_{ij} = \sigma_j(v). \tag{4.4}$$

On notera que le terme « interpolation » est utilisé ici dans un sens relativement large puisque l'interpolé n'est pas nécessairement construit en imposant des valeurs ponctuelles en certains nœuds.

Afin de construire l'interpolé de manière rigoureuse, il convient de préciser quelles sont les fonctions que l'on peut interpoler. En d'autres termes, on doit spécifier le domaine de définition de l'opérateur d'interpolation. Ce domaine de définition est noté $V(K)$. Les éléments de $V(K)$ sont des fonctions de K dans $\mathbb{R}^\mu$. L'espace fonctionnel $V(K)$ doit satisfaire certaines propriétés minimales, en l'occurrence d'être un espace vectoriel normé tel que :

(i) $P \subset V(K)$;

(ii) les formes linéaires $\{\sigma_1, \ldots, \sigma_{n_f}\}$ admettent une extension à $V(K)'$.

Cette deuxième condition, qui signifie qu'il existe une constante c telle que pour tout $i \in \{1, \ldots, n_f\}$ et pour tout $v \in V(K)$, $\langle \sigma_i, v \rangle_{V(K)', V(K)} \leqslant c \, \|v\|_{V(K)}$, est importante pour assurer la continuité de l'opérateur d'interpolation local sur son domaine de définition.

Définition 4.4. *L'opérateur d'interpolation local $\mathcal{I}_K$ est défini comme suit :*

$$\mathcal{I}_K : V(K) \ni v \longmapsto \sum_{i=1}^{n_f} \sigma_i(v)\theta_i \in P. \tag{4.5}$$

L'opérateur d'interpolation $\mathcal{I}_K$ est une *projection* de $V(K)$ dans P. En effet, pour tout $p \in P$, en décomposant p dans la base des fonctions de forme selon $p = \sum_{j=1}^{n_f} x_j\theta_j$, on obtient

$$\mathcal{I}_K p = \sum_{i=1}^{n_f} \sigma_i(p)\theta_i = \sum_{i,j=1}^{n_f} x_j\sigma_i(\theta_j)\theta_i = \sum_{i,j=1}^{n_f} x_j\delta_{ij}\theta_i = p. \tag{4.6}$$

Par suite, pour tout $v \in V(K)$, il vient $\mathcal{I}_K(\mathcal{I}_K v) = \mathcal{I}_K v$.

4.3 Opérateur d'interpolation global

Soit Ω un domaine de $\mathbb{R}^d$ et soit $\mathcal{T}_h = \{K_m\}_{1\leqslant m\leqslant N_{ma}}$ un maillage de Ω ; voir la section 3.3. On rappelle que le maillage $\mathcal{T}_h$ ne recouvre pas nécessairement Ω et qu'on désigne par Ω_h l'intérieur de l'ensemble $\bigcup_{m=1}^{N_{ma}} K_m$. La construction d'un opérateur d'interpolation global sur le domaine Ω_h se fait en suivant une démarche analogue à celle adoptée pour les éléments finis de Lagrange : on choisit d'abord un élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ et on génère une famille d'éléments finis sur les mailles de $\mathcal{T}_h$, puis on définit l'opérateur d'interpolation global à partir de l'opérateur d'interpolation local sur chaque maille.

4.3.1 Génération d'éléments finis

On choisit un *élément fini de référence* $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$. On désigne par $\{\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{n_f}\}$ ses degrés de liberté et par $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_f}\}$ ses fonctions de forme (à valeurs dans $\mathbb{R}^\mu$). On note $V(\widehat{K})$ le domaine de l'opérateur d'interpolation local ; on a donc

$$\mathcal{I}_{\widehat{K}} : V(\widehat{K}) \ni \widehat{v} \longmapsto \sum_{i=1}^{n_f} \widehat{\sigma}_i(\widehat{v})\widehat{\theta}_i \in \widehat{P}. \tag{4.7}$$

Afin de générer un élément fini sur une maille $K \in \mathcal{T}_h$, il faut pouvoir passer de fonctions définies sur $\widehat{K}$ à des fonctions définies sur K. Pour cela, on doit d'abord préciser la contre-partie de l'espace fonctionnel $V(\widehat{K})$ sur K, c'est-à-dire un espace vectoriel normé $V(K)$ de fonctions définies sur K et à valeurs dans $\mathbb{R}^\mu$. On suppose qu'il existe un isomorphisme

$$\psi_K : V(K) \longrightarrow V(\widehat{K}). \tag{4.8}$$

Une façon simple de construire cet isomorphisme est de poser

$$\psi_K : V(K) \ni v \longmapsto v \circ T_K \in V(\widehat{K}), \tag{4.9}$$

où T_K est la transformation géométrique envoyant $\widehat{K}$ dans K ; voir la section 3.3.1. La définition (4.9) est pertinente pour générer la plupart des éléments finis mais n'est pas suffisamment générale pour couvrir tous les cas ; voir par exemple l'élément fini de Raviart–Thomas et l'élément fini de Nédélec où on utilise la transformation de Piola.

Proposition 4.5. *Soit $K \in \mathcal{T}_h$. Alors, le triplet $\{K, P_K, \Sigma_K\}$ défini par*

$$\begin{cases} K = T_K(\widehat{K}), \\ P_K = \{\psi_K^{-1}(\widehat{p}) \, ; \, \widehat{p} \in \widehat{P}\}, \\ \Sigma_K = \{\{\sigma_{K,i}\}_{1 \leqslant i \leqslant n_\mathrm{f}} \, ; \, \sigma_{K,i}(p) = \widehat{\sigma}_i(\psi_K(p)), \ \forall p \in P_K\}, \end{cases} \tag{4.10}$$

est un élément fini.

Les *fonctions de forme* de l'élément fini $\{K, P_K, \Sigma_K\}$ sont telles que

$$\theta_{K,i} = \psi_K^{-1}(\widehat{\theta}_i), \qquad i \in \{1, \ldots, n_\mathrm{f}\}, \tag{4.11}$$

et *l'opérateur d'interpolation local* est tel que

$$\mathcal{I}_K : V(K) \ni v \longmapsto \sum_{i=1}^{n_\mathrm{f}} \sigma_{K,i}(v)\theta_{K,i} \in P_K. \tag{4.12}$$

Une propriété importante de $\mathcal{I}_K$ est que pour tout $v \in V(K)$,

$$\mathcal{I}_{\widehat{K}}(\psi_K(v)) = \psi_K(\mathcal{I}_K(v)). \tag{4.13}$$

Cette propriété résulte du fait que

$$\mathcal{I}_{\widehat{K}}(\psi_K(v)) = \sum_{i=1}^{n_{\mathrm{f}}} \widehat{\sigma}_i(\psi_K(v))\,\widehat{\theta}_i = \sum_{i=1}^{n_{\mathrm{f}}} \sigma_{K,i}(v)\,\psi_K(\theta_{K,i}) = \psi_K(\mathcal{I}_K(v)), \tag{4.14}$$

puisque ψ_K est linéaire. En d'autres termes, le schéma suivant est commutatif :

$$\begin{array}{ccc} V(K) & \xrightarrow{\psi_K} & V(\widehat{K}) \\ \Big\downarrow{\scriptstyle \mathcal{I}_K} & & \Big\downarrow{\scriptstyle \mathcal{I}_{\widehat{K}}} \\ P_K & \xrightarrow{\psi_K} & \widehat{P} \end{array}$$

Une deuxième propriété importante de $\mathcal{I}_K$ est la suivante.

Théorème 4.6. *On suppose que :*

(i) *la transformation T_K est affine ;*

(ii) $[\mathbb{P}_k]^\mu \subset P \subset [H^{k+1}(K)]^\mu \subset V(K)$.

Alors, il existe une constante c, indépendante de K, telle que pour tout $v \in [H^{k+1}(K)]^\mu$ et pour tout $m \in \{0, \ldots, k+1\}$,

$$|v - \mathcal{I}_K v|_{m,K} \leqslant c\, h_K^{k+1-m} \varpi_K^m |v|_{k+1,K}, \tag{4.15}$$

où ϖ_K est défini en (3.38).

Remarque 4.7

On peut procéder à des changements d'échelle entre les degrés de liberté de l'élément fini de référence et ceux de l'élément fini $\{K, P_K, \Sigma_K\}$. Une construction possible est la suivante. Pour toute maille $K \in \mathcal{T}_h$, on choisit un vecteur $\alpha_K \in \mathbb{R}^{n_{\mathrm{f}}}$ tel que $\alpha_{K,i} \neq 0$ pour tout $i \in \{1, \ldots, n_{\mathrm{f}}\}$. On définit le triplet $\{K, P_K, \Sigma_K^\alpha\}$ en prenant K et P_K comme dans (4.10) et en choisissant les degrés de liberté $\Sigma_K^\alpha = \{\sigma_{K,1}^\alpha, \ldots, \sigma_{K,n_{\mathrm{f}}}^\alpha\}$ tels que

$$\sigma_{K,i}^\alpha : P_K \ni p \longmapsto \alpha_{K,i}\widehat{\sigma}_i(\psi_K(p)), \qquad i \in \{1, \ldots, n_{\mathrm{f}}\}. \tag{4.16}$$

Dans ces conditions, le triplet $\{K, P_K, \Sigma_K^\alpha\}$ est un élément fini. De plus, les *fonctions de forme* de cet élément fini sont telles que $\theta_{K,i}^\alpha = \alpha_{K,i}^{-1}\psi_K^{-1}(\widehat{\theta}_i)$ pour tout $i \in \{1, \ldots, n_f\}$, et l'opérateur d'interpolation local est tel que

$$\mathcal{I}_K^\alpha : V(K) \ni v \longmapsto \sum_{i=1}^{n_f} \sigma_{K,i}^\alpha(v)\theta_{K,i}^\alpha \in P_K. \tag{4.17}$$

Le théorème d'interpolation local 4.6 s'applique également à l'opérateur d'interpolation $\mathcal{I}_K^\alpha$ défini ci-dessus.

4.3.2 Espaces d'éléments finis totalement discontinus

On pose

$$W_h = \{v_h \in [L^2(\Omega_h)]^\mu \,;\, \forall K \in \mathcal{T}_h,\ v_h|_K \in P_K\}. \tag{4.18}$$

L'espace W_h est souvent appelé un *espace d'éléments finis totalement discontinu*. La terminologie fait référence au fait que les fonctions de W_h peuvent être multi-valuées aux interfaces entre les mailles puisqu'elles ne satisfont aucune condition de raccord. En d'autres termes, pour toute face $F \in \mathcal{F}_h^i$ telle qu'il existe deux mailles K_1 et K_2 dans $\mathcal{T}_h$ telles que $F = K_1 \cap K_2$, on a *a priori* pour tout $x \in F$,

$$\lim_{\substack{y \to x \\ y \in K_1}} v_h(y) \neq \lim_{\substack{y \to x \\ y \in K_2}} v_h(y). \tag{4.19}$$

Ceci ne pose pas de problème théorique puisque l'ensemble $\bigcup_{F \in \mathcal{F}_h} F$ est de mesure nulle.

Pour $\mu = 1$ et lorsque $P_K = \mathbb{P}_k$ pour tout $K \in \mathcal{T}_h$, l'espace totalement discontinu correspondant est noté

$$P_{\mathrm{td},h}^k = \{v_h \in L^2(\Omega_h) \,;\, \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{P}_k\}. \tag{4.20}$$

Par la suite, il sera commode d'introduire les versions discrètes des opérateurs gradient et divergence définis sur $P_{\mathrm{td},h}^k$ de la manière suivante :

$$\nabla_h : P_{\mathrm{td},h}^k \ni v_h \longmapsto \nabla_h v_h \in [P_{\mathrm{td},h}^{k-1}]^d, \quad \forall K \in \mathcal{T}_h,\ (\nabla_h v_h)|_K = \nabla(v_h|_K), \tag{4.21}$$

$$\nabla_h\cdot : [P_{\mathrm{td},h}^k]^d \ni \boldsymbol{\sigma}_h \longmapsto \nabla_h\cdot\boldsymbol{\sigma}_h \in P_{\mathrm{td},h}^{k-1}, \quad \forall K \in \mathcal{T}_h,\ (\nabla_h\cdot\boldsymbol{\sigma}_h)|_K = \nabla\cdot(\boldsymbol{\sigma}_h|_K). \tag{4.22}$$

Ces opérateurs consistent à évaluer localement le gradient ou la divergence sur chaque cellule du maillage sans se préoccuper d'éventuelles discontinuités aux interfaces (qui provoquent l'apparition de masses de Dirac surfaciques si les dérivées sont prises au sens des distributions sur Ω_h).

4.3.3 Construction de l'opérateur d'interpolation global

L'opérateur d'interpolation global est défini comme suit :

$$\mathcal{I}_h : D(\mathcal{I}_h) \ni v \longmapsto \sum_{K\in\mathcal{T}_h} \left(\sum_{i=1}^{n_f} \sigma_{K,i}(v|_K)\theta_{K,i} \right) 1_K \in \mathrm{Im}(\mathcal{I}_h) \subset W_h, \tag{4.23}$$

où 1_K est la fonction indicatrice de K (1_K vaut 1 sur K et 0 ailleurs). Le domaine de définition de $\mathcal{I}_h$, c'est-à-dire l'espace vectoriel des fonctions que l'on peut interpoler, est tel que

$$D(\mathcal{I}_h) = \{v \in [L^2(\Omega_h)]^\mu \,;\, \forall K \in \mathcal{T}_h,\, v|_K \in V(K)\}. \tag{4.24}$$

Pour une fonction $v \in D(\mathcal{I}_h)$, les quantités $\sigma_{K,i}(v|_K)$ ont bien un sens sur chaque maille $K \in \mathcal{T}_h$ et pour chaque indice $i \in \{1, \ldots, n_f\}$.

On s'intéresse enfin à la précision de l'opérateur d'interpolation $\mathcal{I}_h$: étant donné un sous-espace Z de $D(\mathcal{I}_h)$ constitué de fonctions suffisamment régulières, on souhaite estimer l'erreur d'interpolation $v - \mathcal{I}_h v$ pour tout v dans Z.

Théorème 4.8. *On suppose que :*

(i) $\{\mathcal{T}_h\}_{h>0}$ *est une famille régulière de maillages affines ;*

(ii) *pour tout* $K \in \mathcal{T}_h$, $[\mathbb{P}_k]^\mu \subset P \subset [H^{k+1}(K)]^\mu \subset V(K)$.

Alors, il existe une constante c telle que pour tout h et pour tout $v \in [H^{k+1}(\Omega_h)]^\mu$,

$$\|v - \mathcal{I}_h v\|_{0,\Omega_h} + \sum_{m=1}^{k+1} h^m \left(\sum_{K\in\mathcal{T}_h} |v - \mathcal{I}_h v|^2_{m,K} \right)^{\frac{1}{2}} \leqslant c\, h^{k+1} |v|_{k+1,\Omega_h}. \tag{4.25}$$

4.4 Éléments finis de Crouzeix–Raviart

4.4.1 Point de vue local

Soit K un simplexe de $\mathbb{R}^d$. On pose $P = \mathbb{P}_1$ et on prend pour degrés de liberté de $p \in P$ sa valeur moyenne sur les $(d + 1)$ faces de K. On a donc pour $i \in \{0, \ldots, d\}$,

$$\sigma_i(p) = \frac{1}{\mathrm{mes}(F_i)} \int_{F_i} p. \tag{4.26}$$

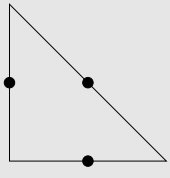
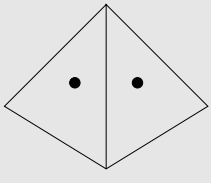
On pose $\Sigma = \{\sigma_i\}_{0 \leqslant i \leqslant d}$.

Proposition 4.9. *Le triplet* $\{K, \mathbb{P}_1, \Sigma\}$ *est un élément fini.*

Cet élément fini a été introduit par Crouzeix et Raviart [**32**] ; voir également Brezzi et Fortin [**22**, pp. 107–109]. Le tableau 4.1 présente les degrés de liberté et les fonctions de forme en dimension 2 et 3. Les cercles noirs sur chaque face indiquent que le degré de liberté consiste à évaluer la valeur moyenne sur cette face. En utilisant les coordonnées barycentriques $(\lambda_0, \ldots, \lambda_d)$ définies en (3.7), les fonctions de forme sont telles que

$$\theta_i(x) = 1 - d\lambda_i(x), \qquad i \in \{0, \ldots, d\}. \tag{4.27}$$

Tableau 4.1 – Degrés de liberté et fonctions de forme de l'élément fini de Crouzeix–Raviart en dimension 2 et 3 ; en dimension 3, seuls les degrés de liberté visibles sont représentés.

$d = 2$	$d = 3$
	
$1 - 2\lambda_i \ \ [0 \leqslant i \leqslant 2]$	$1 - 3\lambda_i \ \ [0 \leqslant i \leqslant 3]$

On observera que $\theta_i|_{F_i} = 1$ et $\theta_i(s_i) = 1 - d$ où s_i est le sommet opposé à la face F_i.

Un choix relativement simple pour le domaine de l'opérateur d'interpolation local consiste à prendre

$$V^{\mathrm{CR}}(K) = \mathcal{C}^0(K). \tag{4.28}$$

L'interpolé local de Crouzeix–Raviart est défini de la manière suivante :

$$\mathcal{I}_K^{\mathrm{CR}} : V^{\mathrm{CR}}(K) \ni v \longmapsto \sum_{i=0}^{d} \left(\tfrac{1}{\mathrm{mes}\, F_i} \int_{F_i} v \right) \theta_i \in \mathbb{P}_1. \tag{4.29}$$

4.4.2 Point de vue global

Pour simplifier, on suppose que le maillage $\mathcal{T}_h$ est affine et conforme. On prend pour élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ l'élément fini de Crouzeix–Raviart. On choisit $V(\widehat{K}) = \mathcal{C}^0(\widehat{K})$, $V(K) = \mathcal{C}^0(K)$ et l'isomorphisme ψ_K défini en (4.9).

On construit la famille $\{K, P_K, \Sigma_K\}_{K \in \mathcal{T}_h}$ comme dans la proposition 4.5. Pour chaque $K \in \mathcal{T}_h$, en posant $F_{K,i} = T_K(\widehat{F}_i)$ pour tout $i \in \{0, \ldots, d\}$, où $\{\widehat{F}_0, \ldots, \widehat{F}_d\}$ sont les faces de $\widehat{K}$, les degrés de liberté s'expriment sous la forme

$$\sigma_{K,i}(v) = \widehat{\sigma}_i(\psi_K(v)) = \tfrac{1}{\mathrm{mes}(\widehat{F}_i)} \int_{\widehat{F}_i} \psi_K(v) = \tfrac{1}{\mathrm{mes}(F_{K,i})} \int_{F_{K,i}} v. \tag{4.30}$$

Par ailleurs, comme le maillage est affine, $P_K = \mathbb{P}_1$. Par conséquent, $\{K, P_K, \Sigma_K\}$ est un élément fini de Crouzeix–Raviart.

L'espace d'éléments finis de Crouzeix–Raviart est défini comme suit :

$$P^1_{\mathrm{pt},h} = \{v_h \in L^2(\Omega_h)\ ;\ \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{P}_1\ ;\ \forall F \in \mathcal{F}_h^{\mathrm{i}},\ \int_F [\![v_h]\!] = 0\}. \tag{4.31}$$

Puisque le saut de v_h est linéaire sur F, la condition $\int_F [\![v_h]\!] = 0$ est équivalente à la continuité de v_h au barycentre de F. L'indice inférieur dans $P^1_{\mathrm{pt},h}$ fait référence au fait que la condition $\int_F [\![v_h]\!] = 0$ est souvent appelée le « patch-test » d'ordre 0.

Pour une face $F \in \mathcal{F}_h$, on définit la fonction $\varphi_F \in P^1_{\mathrm{pt},h}$ dont le support est constitué du ou des deux simplexes contenant F et telle que sur ces simplexes, la fonction φ_F coïncide avec la fonction de forme de l'élément fini de Crouzeix–Raviart associée à F. La figure 4.1 illustre le graphe d'une telle fonction φ_F en dimension 2. La famille $\{\varphi_F\}_{F\in\mathcal{F}_h}$ est une base de $P^1_{\mathrm{pt},h}$. Ceci implique que

$$\dim P^1_{\mathrm{pt},h} = \begin{cases} N_{\mathrm{ar}} & \text{si } d = 2, \\ N_{\mathrm{fa}} & \text{si } d = 3. \end{cases} \tag{4.32}$$

Les fonctions $\{\varphi_F\}_{F\in\mathcal{F}_h}$ sont appelées les *fonctions de forme* dans $P^1_{\mathrm{pt},h}$.

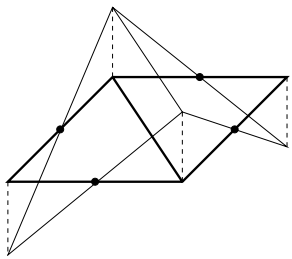


Figure 4.1 – Fonction de forme dans l'espace d'éléments finis de Crouzeix–Raviart en dimension 2. Le support est indiqué en traits gras et le graphe en traits fins.

Pour tout $F \in \mathcal{F}_h$, on définit la forme linéaire $\gamma_F : P^1_{\mathrm{pt},h} \ni v_h \mapsto \frac{1}{\mathrm{mes}(F)} \int_F v_h \in \mathbb{R}$. On observe que même si $v_h \in P^1_{\mathrm{pt},h}$ peut être multi-valuée sur F, la quantité $\gamma_F(v_h)$ est bien définie de manière univoque puisque $\int_F [\![v_h]\!] = 0$. La famille $\{\gamma_F\}_{F\in\mathcal{F}_h}$ est une base de $\mathcal{L}(P^1_{\mathrm{pt},h};\mathbb{R})$. Les formes linéaires $\{\gamma_F\}_{F\in\mathcal{F}_h}$ sont appelées les *degrés de liberté* dans $P^1_{\mathrm{pt},h}$.

L'opérateur d'interpolation de Crouzeix–Raviart est défini comme suit :

$$\mathcal{I}_h^{\mathrm{CR}} : \mathcal{C}^0(\overline{\Omega}_h) \ni v \longmapsto \sum_{F\in\mathcal{F}_h} \left(\frac{1}{\mathrm{mes}(F)} \int_F v \right) \varphi_F \in P^1_{\mathrm{pt},h}. \tag{4.33}$$

L'opérateur $\mathcal{I}_h^{\mathrm{CR}}$ satisfait les conclusions du théorème d'interpolation 4.8 pour $k = 1$.

4.5 Éléments finis de Raviart–Thomas

4.5.1 Point de vue local

Soit K un simplexe de $\mathbb{R}^d$. On considère l'espace vectoriel des polynômes à valeurs dans $\mathbb{R}^d$ défini par

$$\mathbb{RT}_0 = [\mathbb{P}_0]^d \oplus x\,\mathbb{P}_0. \tag{4.34}$$

Il est clair que $\mathbb{RT}_0$ est un espace vectoriel de dimension $(d+1)$. On prend pour degrés de liberté de $p \in \mathbb{RT}_0$ l'intégrale de la composante normale de p sur chacune des $(d+1)$ faces de K. On a donc pour $i \in \{0, \ldots, d\}$,

$$\sigma_i(p) = \int_{F_i} p \cdot n_i. \tag{4.35}$$

On pose $\Sigma = \{\sigma_i\}_{0 \leqslant i \leqslant d}$.

Proposition 4.10. *Le triplet* $\{K, \mathbb{RT}_0, \Sigma\}$ *est un élément fini.*

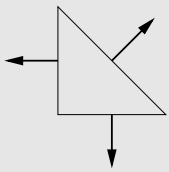
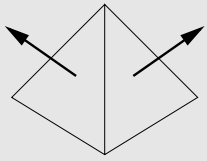
Cet élément fini a été introduit par Raviart et Thomas [**62**] ; voir également Brezzi et Fortin [**22**, p. 113], Quarteroni et Valli [**61**, p. 82]. Il est utilisé, par exemple, dans les problèmes de mécanique des fluides où les fonctions à approcher représentent des vitesses qui doivent satisfaire certaines propriétés de conservation. Le tableau 4.2 présente les degrés de liberté et les fonctions de forme en dimension 2 et 3. Les flèches sur chaque face indiquent que le degré de liberté consiste à évaluer l'intégrale de la composante normale sur cette face. Les fonctions de forme s'écrivent de la manière suivante :

$$\theta_i(x) = \frac{1}{d\,\mathrm{mes}(K)}(x - s_i), \qquad i \in \{0, \ldots, d\}, \tag{4.36}$$

où s_i est le sommet opposé à la face F_i. On observera que la composante normale de θ_i est constante sur la face F_i et est nulle sur les d autres faces de K. La figure 4.2 présente une illustration graphique des fonctions de forme en dimension 2. En dimension 3, pour $i \in \{0, 1, 2, 3\}$, on considère une face F_i dont les sommets s_p, s_q et s_r sont numérotés de sorte que $[(s_p - s_q) \times (s_q - s_r)] \cdot n_i > 0$; dans ces conditions, on a

$$\theta_i = 2(\lambda_p \nabla\lambda_q \times \nabla\lambda_r + \lambda_q \nabla\lambda_r \times \nabla\lambda_p + \lambda_r \nabla\lambda_p \times \nabla\lambda_q). \tag{4.37}$$

Tableau 4.2 – Degrés de liberté et fonctions de forme de l'élément fini de Raviart–Thomas en dimension 2 et 3 ; en dimension 3, seuls les degrés de liberté visibles sont représentés.

$d = 2$	$d = 3$
	
$\frac{1}{2\,\mathrm{mes}(K)}(x - s_i)$ $[0 \leqslant i \leqslant 2]$	$\frac{1}{3\,\mathrm{mes}(K)}(x - s_i)$ $[0 \leqslant i \leqslant 3]$

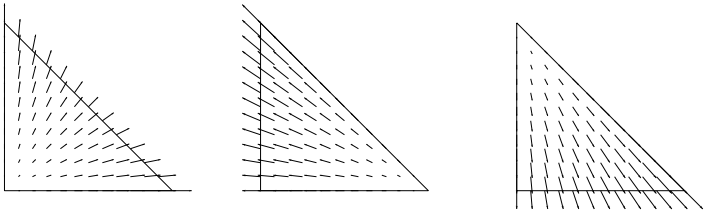


Figure 4.2 – Représentation graphique des fonctions de forme de l'élément fini de Raviart–Thomas en dimension 2.

Un choix relativement simple pour le domaine de l'opérateur d'interpolation local consiste à prendre

$$V^{\mathrm{RT}}(K) = [H^1(K)]^d. \tag{4.38}$$

En effet, la trace d'une fonction de $[H^1(K)]^d$ sur une face de K est bien intégrable sur cette face ; voir la section A.4. L'interpolé local de Raviart–Thomas est défini de la manière suivante :

$$\mathcal{I}_K^{\mathrm{RT}} : V^{\mathrm{RT}}(K) \ni v \longmapsto \sum_{i=0}^{d} \left(\int_{F_i} v \cdot n_i \right) \theta_i \in \mathbb{RT}_0. \tag{4.39}$$

L'opérateur d'interpolation $\mathcal{I}_K^{\mathrm{RT}}$ est tel que

$$\forall v \in V^{\mathrm{RT}}(K), \qquad \pi_K^0(\nabla\cdot v) = \nabla\cdot(\mathcal{I}_K^{\mathrm{RT}} v), \tag{4.40}$$

où π_K^0 désigne le projecteur orthogonal de $L^2(K)$ dans $\mathbb{P}_0$ (pour tout $w \in L^2(K)$, $\pi_K^0 w = \frac{1}{\mathrm{mes}\, K}\int_K w$). En d'autres termes, le diagramme suivant est commutatif :

$$\begin{array}{ccc} V^{\mathrm{RT}}(K) & \xrightarrow{\nabla\cdot} & L^2(K) \\ \downarrow{\scriptstyle \mathcal{I}_K^{\mathrm{RT}}} & & \downarrow{\scriptstyle \pi_K^0} \\ \mathbb{RT}_0 & \xrightarrow{\nabla\cdot} & \mathbb{P}_0 \end{array}$$

4.5.2 Point de vue global

Pour simplifier, on suppose que le maillage $\mathcal{T}_h$ est affine et conforme. On prend pour élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ l'élément fini de Raviart–Thomas. On choisit $V(\widehat{K}) = [H^1(\widehat{K})]^d$ et on définit $V(K)$ de manière analogue. L'isomorphisme entre $V(K)$ et $V(\widehat{K})$ est la *transformation de Piola* définie comme suit :

$$\psi_K : V(K) \ni v \longmapsto \psi_K(v)(\widehat{x}) = \det(J_K) J_K^{-1}\left[v \circ T_K(\widehat{x})\right] \in V(\widehat{K}), \tag{4.41}$$

où J_K est la matrice jacobienne de T_K.

On construit la famille $\{K, P_K, \Sigma_K\}_{K\in\mathcal{T}_h}$ comme dans la proposition 4.5. Pour tout $K \in \mathcal{T}_h$ et pour tout $i \in \{0, \ldots, d\}$, on pose $F_{K,i} = T_K(\widehat{F}_i)$ où $\{\widehat{F}_0, \ldots, \widehat{F}_d\}$ sont les faces de $\widehat{K}$. On vérifie que :

(i) pour tout $i \in \{0, \ldots, d\}$, les degrés de liberté sur K sont tels que $\sigma_{K,i}(v) = \int_{F_{K,i}} v\cdot n_{K,i}$ où $n_{K,i}$ est la normale extérieure à K sur $F_{K,i}$;

(ii) $P_K = \mathbb{RT}_0$.

Par conséquent, $\{K, P_K, \Sigma_K\}$ est un élément fini de Raviart–Thomas.

L'espace d'éléments finis de Raviart–Thomas est défini comme suit :

$$D_h = \{v_h \in [L^2(\Omega_h)]^d \,;\, \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{RT}_0, \forall F \in \mathcal{F}_h^{\mathrm{i}},\ [\![v_h\cdot n]\!]_F = 0\}. \tag{4.42}$$

Cet espace est tel que

$$D_h \subset H(\mathrm{div};\Omega_h) = \{v \in [L^2(\Omega_h)]^d \,;\, \nabla\cdot v \in L^2(\Omega_h)\}. \tag{4.43}$$

On dit que l'espace D_h est $H(\mathrm{div})$-*conforme*.

Pour une face $F \in \mathcal{F}_h$, soit n_F un vecteur unitaire normal à F (sa direction est sans incidence pour la suite). On définit la fonction $\varphi_F \in D_h$ dont le support est constitué du ou des deux simplexes contenant F et telle que sur ces simplexes, la fonction φ_F coïncide (au signe près) avec la fonction de forme de l'élément fini de Raviart–Thomas associée à F. La famille $\{\varphi_F\}_{F\in\mathcal{F}_h}$ est une base de D_h. Ceci implique que

$$\dim D_h = \begin{cases} N_{\mathrm{ar}} & \text{si } d = 2, \\ N_{\mathrm{fa}} & \text{si } d = 3. \end{cases} \tag{4.44}$$

Les fonctions $\{\varphi_F\}_{F\in\mathcal{F}_h}$ sont appelées les *fonctions de forme* dans D_h.

Pour tout $F \in \mathcal{F}_h$, on définit la forme linéaire $\gamma_F : D_h \ni v_h \mapsto \int_F v_h\cdot n_F \in \mathbb{R}$. La famille $\{\gamma_F\}_{F\in\mathcal{F}_h}$ est une base de $\mathcal{L}(D_h;\mathbb{R})$. Les formes linéaires $\{\gamma_F\}_{F\in\mathcal{F}_h}$ sont appelées les *degrés de liberté* dans D_h.

L'opérateur d'interpolation de Raviart–Thomas est défini comme suit :

$$\mathcal{I}_h^{\mathrm{RT}} : [H^1(\Omega_h)]^d \ni v \longmapsto \sum_{F\in\mathcal{F}_h} \left(\int_F v\cdot n_F\right) \varphi_F \in D_h. \tag{4.45}$$

Théorème 4.11. *Avec les hypothèses ci-dessus, il existe une constante c telle que pour tout $K \in \mathcal{T}_h$ et pour tout $v \in [H^1(K)]^d$ tel que $\nabla\cdot v \in H^1(K)$,*

$$\|v - \mathcal{I}_K^{\mathrm{RT}} v\|_{0,K} \leqslant c\,\varpi_K h_K |v|_{1,K}, \tag{4.46}$$

$$\|\nabla\cdot(v - \mathcal{I}_K^{\mathrm{RT}} v)\|_{0,K} \leqslant c\, h_K |\nabla\cdot v|_{1,K}, \tag{4.47}$$

où ϖ_K est défini en (3.38)*. De plus, si la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière, on a pour tout h et pour tout $v \in [H^1(\Omega_h)]^d$ tel que $\nabla\cdot v \in H^1(\Omega_h)$,*

$$\|v - \mathcal{I}_h^{\mathrm{RT}} v\|_{0,\Omega_h} + \|\nabla\cdot(v - \mathcal{I}_h^{\mathrm{RT}} v)\|_{0,\Omega_h} \leqslant c\, h(\|v\|_{1,\Omega_h} + \|\nabla\cdot v\|_{1,\Omega_h}). \tag{4.48}$$

4.6 Éléments finis de Nédélec (ou d'arête)

4.6.1 Point de vue local

Soit K un simplexe de $\mathbb{R}^d$ avec $d = 2$ ou 3. On considère l'espace vectoriel des polynômes à valeurs dans $\mathbb{R}^d$ défini par

$$\mathbb{N}_0 = [\mathbb{P}_0]^2 \oplus \mathbb{P}_0(x_2, -x_1)^T \quad \text{en dimension 2,} \tag{4.49}$$

$$\mathbb{N}_0 = [\mathbb{P}_0]^3 \oplus (x \times [\mathbb{P}_0]^3) \quad \text{en dimension 3.} \tag{4.50}$$

On désigne par n_e le nombre d'arêtes de K ; on a donc $n_e = 3$ si $d = 2$ et $n_e = 6$ si $d = 3$. On désigne par $\{e_i\}_{1 \leqslant i \leqslant n_e}$ l'ensemble des arêtes de K et pour chaque arête e_i, on choisit un vecteur unitaire tangent qu'on note t_i. On prend pour degrés de liberté de $p \in \mathbb{N}_0$ l'intégrale de la composante tangente de p le long des n_e arêtes de K. Pour $i \in \{1, \ldots, n_e\}$, les degrés de liberté sont donc définis de la manière suivante :

$$\sigma_i(p) = \int_{e_i} p \cdot t_i. \tag{4.51}$$

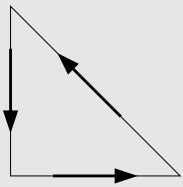
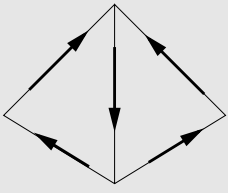
On pose $\Sigma = \{\sigma_i\}_{1 \leqslant i \leqslant n_e}$.

Proposition 4.12. *Le triplet $\{K, \mathbb{N}_0, \Sigma\}$ est un élément fini.*

Cet élément fini a été introduit par Nédélec [**59**]. Il intervient, par exemple, dans les problèmes d'électromagnétisme et de magnéto-hydrodynamique ; voir, par exemple, Bossavit [**16**, Chap. 3]. Le tableau 4.3 présente les degrés de liberté et les fonctions de forme en dimension 2 et 3. Les flèches sur chaque face indiquent que le degré de liberté consiste à évaluer l'intégrale de la composante tangentielle (orientée dans le sens de la flèche) sur cette arête. On observera que la composante tangentielle de θ_i est constante le long de l'arête e_i à laquelle elle est associée et est nulle le long des autres arêtes. La figure 4.3 contient une représentation graphique des fonctions de forme en dimension 2. Enfin, en dimension 3, on peut également écrire les fonctions de forme de la manière suivante. Pour $l, l' \in \{0, 1, 2, 3\}$, on note $e_{l,l'}$ le numéro de l'arête comprise entre les sommets s_l et $s_{l'}$ et pour $l \in \{0, 1, 2, 3\}$, on note λ_l la coordonnée barycentrique associée au sommet s_l ; dans ces conditions, les fonctions de forme $\{\theta_1, \ldots, \theta_6\}$ s'expriment sous la forme

$$\theta_{e(l,l')} = \lambda_l \nabla \lambda_{l'} - \lambda_{l'} \nabla \lambda_l. \tag{4.52}$$

Tableau 4.3 – Degrés de liberté et fonctions de forme de l'élément fini de Nédélec en dimension 2 et 3. En dimension 2, pour $i \in \{1,2,3\}$, s_{i-1} désigne le sommet opposé à l'arête e_i et $\mathcal{R} : \mathbb{R}^2 \to \mathbb{R}^2$ la transformation telle que $\mathcal{R}(x_1,x_2) = (x_2,-x_1)$. En dimension 3, seuls les degrés de liberté visibles sont représentés. De plus, on a introduit l'application $j : \{1,\ldots,6\} \ni i \to j(i) \in \{1,\ldots,6\}$ telle que $j(i)$ soit l'indice de l'arête opposée à e_i, c'est-à-dire de la seule arête de K qui n'intersecte pas e_i ; m_i désigne le point milieu de e_i.

$d = 2$	$d = 3$
	
$\dfrac{\mathcal{R}(x - s_{i-1})}{t_i \cdot \left[\mathcal{R}(\frac{s_{i_1}+s_{i_2}}{2} - s_{i-1})\right] \operatorname{mes}(e_i)}$	$\dfrac{(x - m_{j(i)}) \times t_{j(i)}}{t_i \cdot \left[(m_i - m_{j(i)}) \times t_{j(i)}\right] \operatorname{mes}(e_i)}$
$[i \in \{1,2,3\},\ i_1, i_2 \neq i-1]$	$[i \in \{1,\ldots,6\}]$

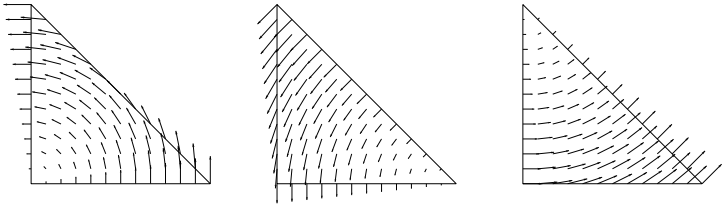


Figure 4.3 – Représentation graphique des fonctions de forme de l'élément fini de Nédélec en dimension 2.

Un choix relativement simple pour le domaine de l'opérateur d'interpolation local consiste à prendre

$$V^{\mathrm{N}}(K) = [H^1(K)]^d. \tag{4.53}$$

L'interpolé local de Nédélec est défini de la manière suivante :

$$\mathcal{I}_K^{\mathrm{N}} : V^{\mathrm{N}}(K) \ni v \longmapsto \sum_{i=1}^{n_e} \left(\int_{e_i} v \cdot t_i \right) \theta_i \in \mathbb{N}_0. \tag{4.54}$$

L'opérateur d'interpolation $\mathcal{I}_K^{\mathrm{N}}$ satisfait les propriétés suivantes.

(i) En dimension 2 ou 3, pour tout $v \in H^2(K)$, $\mathcal{I}_K^{\mathrm{N}}(\nabla v) = \nabla(\mathcal{I}_K^1 v)$ où $\mathcal{I}_K^1 v$ est l'interpolé de Lagrange $\mathbb{P}_1$ de v. En d'autres termes, le diagramme suivant est commutatif :

$$\begin{array}{ccc} H^2(K) & \xrightarrow{\nabla} & [H^1(K)]^d \\ \downarrow{\scriptstyle \mathcal{I}_K^1} & & \downarrow{\scriptstyle \mathcal{I}_K^{\mathrm{N}}} \\ \mathbb{P}_1 & \xrightarrow{\nabla} & \mathbb{N}_0 \end{array}$$

(ii) En dimension 3, pour tout $v \in [H^2(K)]^3$, $\mathcal{I}_K^{\mathrm{RT}}(\nabla \times v) = \nabla \times (\mathcal{I}_K^{\mathrm{N}} v)$. En d'autres termes, le diagramme suivant est commutatif :

$$\begin{array}{ccc} [H^2(K)]^3 & \xrightarrow{\nabla\times} & [H^1(K)]^3 \\ \downarrow{\scriptstyle \mathcal{I}_K^{\mathrm{N}}} & & \downarrow{\scriptstyle \mathcal{I}_K^{\mathrm{RT}}} \\ \mathbb{N}_0 & \xrightarrow{\nabla\times} & \mathbb{RT}_0 \end{array}$$

4.6.2 Point de vue global

Pour simplifier, on suppose que le maillage $\mathcal{T}_h$ est affine et conforme. On se place en dimension 3 (une construction analogue est possible en dimension 2). On prend pour élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ l'élément fini de Nédélec. On choisit $V(\widehat{K}) = [H^1(\widehat{K})]^3$ et on définit $V(K)$ de manière analogue. L'isomorphisme entre $V(K)$ et $V(\widehat{K})$ est la *transformation de Piola*

définie comme suit :

$$\psi_K : V(K) \ni v \longmapsto \psi_K(v)(\widehat{x}) = J_K^T[v \circ T_K(\widehat{x})] \in V(\widehat{K}). \tag{4.55}$$

On construit la famille $\{K, P_K, \Sigma_K\}_{K\in\mathcal{T}_h}$ comme dans la proposition 4.5. On désigne par $\{\widehat{e}_1, \ldots, \widehat{e}_6\}$ les arêtes de $\widehat{K}$ et pour tout $i \in \{1, \ldots, 6\}$, on désigne par $e_{K,i} = T_K(\widehat{e}_i)$ les arêtes correspondantes de K. Soit $\widehat{t}_i$ un des vecteurs unitaires portés par $\widehat{e}_i$; alors, $t_{K,i} = T_K(\widehat{t}_i)$ est un vecteur unitaire porté par $e_{K,i}$. On vérifie que :

(i) pour tout $i \in \{1, \ldots, 6\}$, les degrés de liberté dans K sont tels que $\sigma_i(v) = \int_{e_{K,i}} v \cdot t_{K,i}$;

(ii) $P_K = \mathbb{N}_0$.

Par conséquent, $\{K, P_K, \Sigma_K\}$ est un élément fini de Nédélec.

L'espace d'éléments finis de Nédélec est défini comme suit :

$$R_h = \{v_h \in [L^2(\Omega_h)]^3 \, ; \, \forall K \in \mathcal{T}_h, \, v_h|_K \in \mathbb{N}_0 \, ; \, \forall F \in \mathcal{F}_h^{\mathrm{i}}, \, [\![v_h \times n]\!]_F = 0\}. \tag{4.56}$$

Cet espace est tel que

$$R_h \subset H(\mathrm{rot}; \Omega_h) = \{v \in [L^2(\Omega_h)]^3 \, ; \, \nabla \times v \in [L^2(\Omega_h)]^3\}. \tag{4.57}$$

On dit que l'espace R_h est $H(\mathrm{rot})$-*conforme*.

Pour une arête $e \in \mathcal{E}_h$, on choisit un vecteur unitaire porté par e, que l'on désigne par t_e (sa direction est sans incidence pour la suite). On définit la fonction $\varphi_e \in R_h$ dont le support est constitué du ou des simplexes contenant e et telle que sur ces simplexes, la fonction φ_e coïncide (au signe près) avec la fonction de forme de l'élément fini de Nédélec associée à e. La famille $\{\varphi_e\}_{e\in\mathcal{E}_h}$ est une base de R_h. Ceci implique que

$$\dim R_h = N_{\mathrm{ar}}. \tag{4.58}$$

Les fonctions $\{\varphi_e\}_{e\in\mathcal{E}_h}$ sont appelées les *fonctions de forme* dans R_h.

Pour tout $e \in \mathcal{E}_h$, on définit la forme linéaire $\gamma_e : R_h \ni v_h \mapsto \int_e v_h \cdot t_e \in \mathbb{R}$. La famille $\{\gamma_e\}_{e\in\mathcal{E}_h}$ est une base de $\mathcal{L}(R_h; \mathbb{R})$. Les formes linéaires $\{\gamma_e\}_{e\in\mathcal{E}_h}$ sont appelées les *degrés de liberté* dans R_h.

L'opérateur d'interpolation de Nédélec est défini comme suit :

$$\mathcal{I}_h^{\mathrm{N}} : [H^1(\Omega_h)]^3 \ni v \longmapsto \sum_{e\in\mathcal{E}_h} \left(\int_e v\cdot t_e \right) \varphi_e \in R_h. \tag{4.59}$$

Théorème 4.13. *Avec les hypothèses ci-dessus, il existe une constante c telle que pour tout $K \in \mathcal{T}_h$ et pour tout $v \in [H^1(K)]^3$ avec $\nabla\times v \in [H^1(K)]^3$,*

$$\|v - \mathcal{I}_K^{\mathrm{N}} v\|_{0,p,K} \leqslant c\, \varpi_K h_K |v|_{1,p,K}, \tag{4.60}$$

$$\|\nabla\times(v - \mathcal{I}_K^{\mathrm{N}} v)\|_{0,p,K} \leqslant c\, h_K |\nabla\times v|_{1,p,K}, \tag{4.61}$$

où ϖ_K est défini en (3.38). *De plus, si la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière, on a pour tout h et pour tout $v \in [H^1(\Omega_h)]^3$ avec $\nabla\times v \in [H^1(\Omega_h)]^3$,*

$$\|v - \mathcal{I}_h^{\mathrm{N}} v\|_{0,\Omega_h} + \|\nabla\times(v - \mathcal{I}_h^{\mathrm{N}} v)\|_{0,\Omega_h} \leqslant c\, h(\|v\|_{1,\Omega_h} + \|\nabla\times v\|_{1,\Omega_h}). \tag{4.62}$$

4.7 Éléments finis de degré élevé

Lorsqu'on utilise des éléments finis de degré élevé, il est important de bien choisir la base $\{\theta_1, \ldots, \theta_{n_\mathrm{f}}\}$ des fonctions de forme. Plusieurs critères peuvent fonder ce choix.

(i) La possibilité d'imposer facilement la valeur de l'interpolé sur la frontière de K. Cette propriété est essentielle lorsqu'on souhaite raccorder de façon continue des interpolés sur des mailles adjacentes.

(ii) La possibilité d'inverser facilement la matrice de masse élémentaire $\mathcal{M}^{\text{élé}} \in \mathbb{R}^{n_\mathrm{f}, n_\mathrm{f}}$ dont les composantes sont

$$\mathcal{M}^{\text{élé}}_{mn} = \int_K \theta_m \theta_n, \qquad m, n \in \{1, \ldots, n_\mathrm{f}\}. \tag{4.63}$$

On notera que cette matrice est symétrique définie positive. Dans des algorithmes de marche en temps explicites, la matrice de masse élémentaire doit être inversée à chaque pas de temps ; il est donc important que cette inversion soit la moins coûteuse possible.

(iii) La stabilité de l'opérateur d'interpolation local $\mathcal{I}_K$ défini en (4.5). Cette stabilité peut se mesurer en évaluant la norme

$$\|\mathcal{I}_K\|_{\mathcal{L}(V(K),P)} = \sup_{v\in V(K)} \frac{\|\mathcal{I}_K v\|_P}{\|v\|_{V(K)}}, \tag{4.64}$$

où $\|\cdot\|_P$ et $\|\cdot\|_{V(K)}$ sont des normes sur P et $V(K)$, repsectivement. Pour un élément fini de Lagrange $\{K, P, \Sigma\}$, un choix naturel consiste à équiper $V(K) = \mathcal{C}^0(K)$ et P de la norme canonique $\|\cdot\|_{\mathcal{C}^0(K)}$ définie pour $v \in \mathcal{C}^0(K)$ par $\|v\|_{\mathcal{C}^0(K)} = \sup_{t\in K} |v(t)|$. Un calcul simple montre que pour l'opérateur d'interpolation de Lagrange $\mathcal{I}_K^{\mathrm{Lag}}$, on a

$$\|\mathcal{I}_K^{\mathrm{Lag}}\|_{\mathcal{L}(\mathcal{C}^0(K),P)} = \sup_{v\in\mathcal{C}^0(K)} \frac{\|\mathcal{I}_K^{\mathrm{Lag}} v\|_{\mathcal{C}^0(K)}}{\|v\|_{\mathcal{C}^0(K)}} = \sup_{t\in K} \left(\sum_{n=1}^{n_{\mathrm{f}}} |\theta_n(t)| \right). \tag{4.65}$$

Cette quantité ne dépend que des nœuds $\{a_1, \ldots, a_{n_{\mathrm{f}}}\}$ de l'élément fini de Lagrange et est appelée la *constante de Lebesgue* associée à ces nœuds. Par la suite, cette constante est notée $\lambda(\{a_1, \ldots, a_{n_{\mathrm{f}}}\})$. La constante de Lebesgue permet d'évaluer la qualité de l'opérateur d'interpolation $\mathcal{I}_K^{\mathrm{Lag}}$ dans $\mathcal{C}^0(K)$. En effet, pour tout $f \in \mathcal{C}^0(K)$, on a

$$\|f - \mathcal{I}_K^{\mathrm{Lag}} f\|_{\mathcal{C}^0(K)} \leqslant (1 + \lambda(\{a_1, \ldots, a_{n_{\mathrm{f}}}\})) \inf_{f^*\in P} \|f - f^*\|_{\mathcal{C}^0(K)}. \tag{4.66}$$

Cette estimation montre que si la constante de Lebesgue n'est pas trop grande, l'erreur d'interpolation en norme $\|\cdot\|_{\mathcal{C}^0(K)}$ n'est pas « trop loin » d'être optimale.

Comment satisfaire les critères (i)–(iii) ci-dessus ? Afin d'apprécier la difficulté du problème, on suppose que l'on souhaite utiliser un élément fini de degré k en dimension 1. Si on choisit de travailler avec un élément fini de Lagrange, un choix naturel consiste à prendre pour fonctions de forme les $(k + 1)$ polynômes d'interpolation de Lagrange associés aux nœuds équirépartis sur K ; voir la section 1.4. Ce choix présente l'avantage d'être bien adapté au critère (i) ci-dessus puisque seuls deux des polynômes d'interpolation de Lagrange sont non-nuls aux extrémités de K. Malheureusement, le fait d'avoir pris des nœuds équirépartis sur K conduit à des situations désastreuses quant aux critères (ii) et (iii). On montre en effet que le nombre de

conditionnement[1] de la matrice de masse explose exponentiellement en k. De plus, la constante de Lebesgue associée aux nœuds équirépartis explose, elle aussi, exponentiellement en k.

Cette section présente quelques bases de fonctions de forme permettant de satisfaire (au moins en partie) les critères ci-dessus pour des éléments finis de degré élevé. On abordera d'une part les bases modales, où les fonctions de forme sont définies directement sans la médiation de degrés de liberté, et d'autre part les bases nodales, qui sont associées à des éléments finis de Lagrange pour des nœuds convenablement choisis. La présentation est restreinte aux éléments finis en dimension 1 et à l'utilisation de produits tensoriels sur des hypercubes en dimension supérieure.

4.7.1 Bases modales

On se place d'abord en dimension 1 et on pose $K = [0, 1]$.

Définition 4.14 (Polynômes de Legendre). *Les polynômes de Legendre* $\{\mathcal{E}_m\}_{m\geqslant 0}$ *sur l'intervalle de référence* $K = [0, 1]$ *sont définis par la formule*

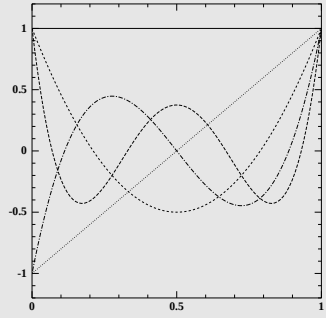
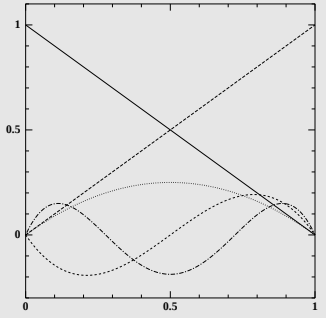
$$\mathcal{E}_m(t) = \frac{1}{m!}\frac{d^m}{dt^m}(t^2 - t)^m. \tag{4.67}$$

Le polynôme de Legendre $\mathcal{E}_m$ est de degré m, il vérifie $\mathcal{E}_m(0) = (-1)^m$, $\mathcal{E}_m(1) = 1$ et ses m racines se trouvent toutes dans K.[2] La colonne de gauche du tableau 4.4 contient une représentation graphique et l'expression analytique des polynômes de Legendre $\mathcal{E}_m$ pour $m \in \{0, 1, 2, 3, 4\}$.

1. Le nombre de conditionnement d'une matrice symétrique $\mathcal{Z}$ est défini comme le rapport entre la plus grande et la plus petite valeur propre de la matrice. Lorsque ce nombre est très grand, les systèmes linéaires de la forme $\mathcal{Z}X = Y$ sont très difficiles à inverser, notamment parce qu'ils deviennent très sensibles aux erreurs d'arrondi ; voir la section 10.1.

2. Dans la littérature, les polynômes de Legendre sont plus fréquemment définis sur l'intervalle de référence $K = [-1, 1]$. En notant $\mathcal{E}^*_m(s)$ les polynômes de Legendre définis sur $[-1, 1]$, on a $\mathcal{E}^*_m(s) = \mathcal{E}_m(\frac{1}{2}(1 + s))$. La même remarque s'applique pour les polynômes de Jacobi définis par la suite.

Tableau 4.4 – Représentation graphique et expression analytique des polynômes de Legendre $\mathcal{E}_m$ (à gauche) et des éléments de la base modale θ_m^{mod} (à droite) pour $m \in \{0, 1, 2, 3, 4\}$.

polynômes de Legendre $\{\mathcal{E}_m\}_{0 \leqslant m \leqslant 4}$	**base modale** $\{\boldsymbol{\theta}_m^{\text{mod}}\}_{0 \leqslant m \leqslant 4}$
	
$\mathcal{E}_0(t) = 1$ $\mathcal{E}_1(t) = 2t - 1$ $\mathcal{E}_2(t) = 6t^2 - 6t + 1$ $\mathcal{E}_3(t) = 20t^3 - 30t^2 + 12t - 1$ $\mathcal{E}_4(t) = 70t^4 - 140t^3 + 90t^2 - 20t + 1$	$\theta_0^{\text{mod}}(t) = 1 - t$ $\theta_1^{\text{mod}}(t) = t$ $\theta_2^{\text{mod}}(t) = t(1 - t)$ $\theta_3^{\text{mod}}(t) = t(1 - t)(4t - 2)$ $\theta_4^{\text{mod}}(t) = t(1 - t)(15t^2 - 15t + 3)$

La propriété fondamentale satisfaite par les polynômes de Legendre est que pour tout $m, n \geqslant 0$,

$$\int_0^1 \mathcal{E}_m(t)\mathcal{E}_n(t)\,\mathrm{d}t = \tfrac{1}{2m+1}\delta_{mn}. \tag{4.68}$$

Par conséquent, la matrice de masse élémentaire associée à la base $\{\mathcal{E}_0, \ldots, \mathcal{E}_k\}$ est diagonale et son nombre de conditionnement vaut $(2k + 1)$, ce qui représente une amélioration considérable par rapport aux polynômes d'interpolation de Lagrange.

Une deuxième propriété intéressante des polynômes de Legendre est qu'ils forment une base hiérarchique de $\mathbb{P}_k$ au sens de la définition suivante.

Définition 4.15 (Base hiérarchique). *Soit un entier $k \geqslant 0$. On dit que la famille de polynômes $\{\mathcal{P}_0, \ldots, \mathcal{P}_k\}$ forme une base hiérarchique de $\mathbb{P}_k$ si pour tout $l \in \{0, \ldots, k\}$, la famille $\{\mathcal{P}_0, \ldots, \mathcal{P}_l\}$ forme une base de $\mathbb{P}_l$.*

L'intérêt des bases hiérarchiques est que si l'on souhaite augmenter le degré de l'élément fini (par exemple pour améliorer la précision de l'interpolé), il suffit de rajouter des nouveaux polynômes à la base sans devoir modifier les anciennes fonctions de base. L'exemple le plus simple de base hiérarchique de $\mathbb{P}_k$ est la famille des monômes $\{1, t, \ldots, t^k\}$. On observera que la famille des polynômes d'interpolation de Lagrange ne constitue pas une base hiérarchique de $\mathbb{P}_k$. En effet, si on rajoute un point d'interpolation, tous les polynômes d'interpolation de Lagrange sont modifiés.

Un défaut des polynômes de Legendre est qu'aucun d'entre eux ne s'annule aux extrémités de K. Afin de pallier cette difficulté, on introduit les polynômes de Jacobi.

Définition 4.16 (Polynômes de Jacobi). *Soient deux entiers*[1] $\alpha > -1$ *et* $\beta > -1$. *Les polynômes de Jacobi $\{\mathcal{J}_m^{\alpha,\beta}\}_{m\geqslant 0}$ sur l'intervalle de référence $K = [0, 1]$ s'expriment sous la forme*

$$\mathcal{J}_m^{\alpha,\beta}(t) = \frac{(-1)^m}{m!} 2^{-\alpha-\beta} (1-t)^{-\alpha} t^{-\beta} \frac{d^m}{dt^m} \left((1-t)^{\alpha+m} t^{\beta+m} \right). \tag{4.69}$$

Les polynômes de Jacobi sont tels que pour tout $m, n \geqslant 0$,

$$\int_0^1 (1-t)^\alpha t^\beta \mathcal{J}_m^{\alpha,\beta}(t) \mathcal{J}_n^{\alpha,\beta}(t)\, \mathrm{d}t = c_{m,\alpha,\beta} \delta_{mn}, \tag{4.70}$$

avec $c_{m,\alpha,\beta} = \frac{1}{2m+\alpha+\beta+1} \frac{(m+\alpha)!(m+\beta)!}{m!(m+\alpha+\beta)!}$. On notera par ailleurs que $\mathcal{J}_m^{0,0} = \mathcal{E}_m$ pour tout $m \geqslant 0$. Pour des compléments sur les polynômes de Legendre et de Jacobi, on pourra consulter Abramowitz et Stegun [**1**, chap. 22] ou Karniadakis et Spencer [**54**, p. 350].

1. Les polynômes de Jacobi peuvent être définis pour des paramètres α et β réels tels que $\alpha > -1$ et $\beta > -1$.

Pour un entier $k \geqslant 2$, on définit la base modale $\{\theta_0^{\text{mod}}, \ldots, \theta_k^{\text{mod}}\}$ de la manière suivante :

$$\theta_m^{\text{mod}}(t) = \begin{cases} 1 - t & \text{si } m = 0, \\ t & \text{si } m = 1, \\ (1-t)t\,\mathcal{J}_{m-2}^{1,1}(t) & \text{si } 1 < m \leqslant k. \end{cases} \tag{4.71}$$

Cette base possède plusieurs propriétés intéressantes :

(i) elle constitue une base hiérarchique de $\mathbb{P}_k$ au sens de la définition 4.15 ;

(ii) la valeur de l'interpolé aux extrémités de K peut être imposée facilement puisque seules les deux premières fonctions de la base sont non-nulles aux extrémités de K ;

(iii) la matrice de masse élémentaire reste relativement bien conditionnée et celle-ci a une structure « presque tridiagonale » puisque $\mathcal{M}_{mn}^{\text{élé}} = 0$ pour $|m - n| > 2$ et $m, n \in \{0, \ldots, k\}$ sauf si $m = k$ et $n \leqslant 2$ ou $n = k$ et $m \leqslant 2$.

La colonne de droite du tableau 4.4 contient une représentation graphique et l'expression analytique des éléments de la base modale pour $m \in \{0, 1, 2, 3, 4\}$.

Lorsque K est un hypercube en dimension $d \geqslant 2$, on peut construire une base modale sur K à partir de produits tensoriels des éléments de la base modale unidimensionelle. Par exemple, en dimension 2, en notant (t_1, t_2) les coordonnées locales du point courant de K, les éléments de la base modale s'expriment sous la forme

$$\theta_{m_1 m_2}^{\text{mod}}(t_1, t_2) = \theta_{m_1}^{\text{mod}}(t_1)\theta_{m_2}^{\text{mod}}(t_2), \qquad m_1, m_2 \in \{0, \ldots, k\}. \tag{4.72}$$

La construction de bases modales sur des simplexes ou des prismes est plus technique car elle utilise des transformations non-linéaires ; voir Karniadakis et Spencer [**54**, pp. 70–94].

4.7.2 Bases nodales et éléments finis spectraux

Les bases nodales sont construites à partir de polynômes d'interpolation de Lagrange, mais les nœuds associés ne sont pas équirépartis. Des répartitions

autres que celle uniforme permettent en effet d'améliorer substantiellement la précision de l'interpolé lorsqu'on utilise des polynômes de degré élevé.
On se place d'abord en dimension 1 et on pose à nouveau $K = [0, 1]$.

Définition 4.17 (Points de Gauß–Lobatto). *Soit un entier $k \geqslant 1$. Les $(k+1)$ points de Gauß–Lobatto $\{g_0^k, \ldots, g_k^k\}$ sur l'intervalle de référence $K = [0,1]$ sont les deux extrémités de l'intervalle, $g_0^k = 0$ et $g_k^k = 1$, et les $(k-1)$ racines de $\mathcal{E}_k'$.*

On notera que comme $\mathcal{E}_k$ admet k racines distinctes sur K, $\mathcal{E}_k'$ admet $(k-1)$ racines distinctes sur K. Les points de Gauß–Lobatto sur K sont portés dans le tableau 4.5 pour $k \in \{1, 2, 3, 4\}$.

	$k=1$	$k=2$	$k=3$	$k=4$
$l=0$	0	0	0	0
$l=1$	1	$\frac{1}{2}$	$\frac{1}{2}(1-(\frac{1}{5})^{\frac{1}{2}})$	$\frac{1}{2}(1-(\frac{3}{7})^{\frac{1}{2}})$
$l=2$		1	$\frac{1}{2}(1+(\frac{1}{5})^{\frac{1}{2}})$	$\frac{1}{2}$
$l=3$			1	$\frac{1}{2}(1+(\frac{3}{7})^{\frac{1}{2}})$
$l=4$				1

Tableau 4.5 – Points de Gauß–Lobatto sur l'intervalle de référence $K = [0, 1]$ pour $k \in \{1, 2, 3, 4\}$.

Proposition 4.18. *Les $(k+1)$ polynômes d'interpolation de Lagrange associés aux $(k+1)$ points de Gauß–Lobatto $\{g_0^k, \ldots, g_k^k\}$ sur l'intervalle de référence $K = [0,1]$ sont notés $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$. Ces polynômes sont tels que*

$$\theta_m^{\mathrm{GL},k}(t) = \frac{(t-1)t\mathcal{E}_k'(t)}{k(k+1)\mathcal{E}_k(g_m^k)(t-g_m^k)}, \qquad m \in \{0, \ldots, k\}. \tag{4.73}$$

La famille $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ forme une base nodale de $\mathbb{P}_k$. Le tableau 4.6 contient une représentation graphique et l'expression analytique des polynômes $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ pour $k \in \{3, 4\}$. Pour $k \in \{1, 2\}$, la base nodale

est constituée des polynômes d'interpolation de Lagrange usuels puisque les points de Gauß–Lobatto sont équirépartis sur $[0, 1]$ (voir le tableau 4.5). La base nodale $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ n'est pas hiérarchique (ce qui explique la présence de l'indice supérieur k) mais elle présente plusieurs avantages à la lumière des critères (i)–(iii) ci-dessus. Le critère (i) est clairement satisfait puisque les deux extrémités de l'intervalle font toujours partie de l'ensemble des points de Gauß–Lobatto. De plus, même si la matrice de masse est dense, elle devient diagonale si ses coefficients sont approchés en utilisant la quadrature de Gauß–Lobatto décrite dans la section 9.2.2. On parle de *condensation statique* de la matrice de masse lorsque les coefficients de cette matrice sont approchés par quadrature. Enfin, un résultat dû à Fejér (1932) est que la base nodale $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ est telle que

$$\sup_{t\in K} \left(\sum_{m=0}^{k} \left(\theta_m^{\mathrm{GL},k}(t) \right)^2 \right) = 1. \tag{4.74}$$

Cette propriété remarquable implique en particulier que

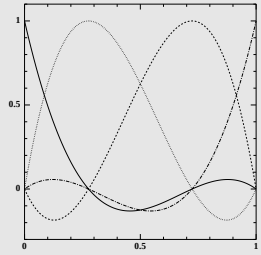
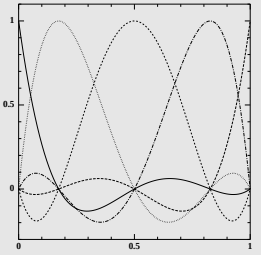
$$\sup_{t\in K} |\theta_m^{\mathrm{GL},k}(t)| = 1, \qquad m \in \{0, \ldots, k\}. \tag{4.75}$$

En d'autres termes, le polynôme $\theta_m^{\mathrm{GL},k}$ atteint son maximum sur K au point de Gauß–Lobatto g_m^k. Cette propriété est clairement visible sur le tableau 4.6. On observera que cette propriété n'est pas satisfaite par les polynômes d'interpolation de Lagrange puisque ceux-ci peuvent prendre des valeurs plus grandes que 1 sur K ; voir le tableau 1.1. Une conséquence importante de la propriété (4.75) est que la constante de Lebesgue associée aux points de Gauß–Lobatto est majorée par $(k + 1)$. En effet, on déduit immédiatement de la formule (4.65) que

$$\lambda(\{g_0^k, \ldots, g_k^k\}) = \sup_{t\in K} \left(\sum_{m=0}^{k} |\theta_m^{\mathrm{GL},k}(t)| \right) \leqslant (k + 1). \tag{4.76}$$

Même si cette estimation est quelque peu pessimiste, elle représente déjà une amélioration considérable par rapport à la constante de Lebesgue associée aux nœuds équirépartis sur K, cette dernière explosant exponentiellement en k.

Tableau 4.6 – Représentation graphique et expression analytique des éléments de la base nodale $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ pour $k \in \{3, 4\}$; les points de Gauß–Lobatto sont indiqués dans le tableau 4.5.

$k = 3$	$k = 4$
	
$\theta_0^{\mathrm{GL},3}(t) = (1-t)(5t^2 - 5t + 1)$ $\theta_1^{\mathrm{GL},3}(t) = 5\frac{1}{2}(t^2 - t)(t - g_2^3)$ $\theta_2^{\mathrm{GL},3}(t) = -5\frac{1}{2}(t^2 - t)(t - g_1^3)$ $\theta_3^{\mathrm{GL},3}(t) = t(5t^2 - 5t + 1)$	$\theta_0^{\mathrm{GL},4}(t) = 2(t-1)(t - \frac{1}{2})(7t^2 - 7t + 1)$ $\theta_1^{\mathrm{GL},4}(t) = \frac{98}{3}t(1-t)(t - \frac{1}{2})(t - g_3^4)$ $\theta_2^{\mathrm{GL},4}(t) = \frac{16}{3}t(t-1)(7t^2 - 7t + 1)$ $\theta_3^{\mathrm{GL},4}(t) = \frac{98}{3}t(1-t)(t - \frac{1}{2})(t - g_1^4)$ $\theta_4^{\mathrm{GL},4}(t) = 2t(t - \frac{1}{2})(7t^2 - 7t + 1)$

En dimension $d \geqslant 2$, lorsque K est un hypercube, on peut construire une base nodale sur K à partir de produits tensoriels des éléments de la base nodale unidimensionelle. Par exemple, en dimension 2, en notant (t_1, t_2) les coordonnées locales du point courant de K, les éléments de la base nodale s'expriment sous la forme

$$\theta_{m_1 m_2}^{\mathrm{GL},k}(t_1, t_2) = \theta_{m_1}^{\mathrm{GL},k}(t_1)\theta_{m_2}^{\mathrm{GL},k}(t_2), \qquad m_1, m_2 \in \{0, \ldots, k\}. \tag{4.77}$$

On observera que la famille de polynômes $\{\theta_{m_1 m_2}^{\mathrm{GL},k}\}_{0 \leqslant m_1, m_2 \leqslant k}$ est constituée des polynômes d'interpolation de Lagrange associés aux $(k+1)^2$ points de coordonnées locales $(g_{m_1}^k, g_{m_2}^k)$. Ces points sont obtenus par produits tensoriels des points de Gauß–Lobatto unidimensionnels.

Définition 4.19. *Un élément fini dont la base nodale est constituée des polynômes d'interpolation de Lagrange associés aux points de Gauß–Lobatto en dimension 1 ou à des produits tensoriels de ceux-ci en dimension $d \geqslant 2$ est appelé un* élément fini spectral.

5 • APPROXIMATION DE PROBLÈMES COERCIFS

Ce chapitre est consacré à l'approximation par éléments finis de problèmes coercifs. Le prototype est le problème de Dirichlet : étant donné une fonction $f : \Omega \to \mathbb{R}$, chercher une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\Delta u = f \quad \text{dans } \Omega, \tag{5.1}$$

$$u = 0 \quad \text{sur } \partial\Omega, \tag{5.2}$$

où $\Delta u = \sum_{i=1}^{d} \partial_{ii}^2 u$ est le *Laplacien* de u. L'équation (5.2), qui impose la nullité de la solution u sur la frontière du domaine, est appelée une *condition aux limites de Dirichlet homogène*. D'autres conditions aux limites (de Dirichlet non-homogène, de Neumann ou de Robin) sont possibles et seront abordées ci-dessous. Le problème (5.1)–(5.2) intervient dans de nombreux modèles physiques, notamment en thermique (u représente une température), en électrostatique (u représente un potentiel électrique), en mécanique (u représente un déplacement vertical de membrane) et en hydraulique (u représente une charge dans un écoulement de Darcy). Un deuxième exemple de problème coercif est celui des équations d'advection–diffusion–réaction avec diffusion dominante. Un troisième exemple est fourni par la mécanique des milieux continus déformables dans le cadre de l'élasticité linéaire. Dans ce cas, l'inconnue est une fonction $u : \Omega \to \mathbb{R}^d$ qui représente un champ de déplacement.

Les problèmes coercifs ont fourni le premier cadre d'application de la méthode des éléments finis, lorsque des ingénieurs ont utilisé cette méthode dans les années 1950 pour approcher la solution de problèmes de mécanique des milieux continus déformables. L'analyse mathématique des problèmes coercifs est relativement simple puisqu'elle repose sur le lemme de Lax–Milgram.

L'approximation par éléments finis se fait en général par une méthode de *Galerkin standard* dans un cadre conforme ou non. L'analyse de convergence est effectuée en s'appuyant d'une part sur les résultats du chapitre 2 concernant l'analyse abstraite de la méthode de Galerkin et d'autre part sur ceux des chapitres 3 et 4 concernant les propriétés interpolantes des espaces d'éléments finis.

5.1 Le Laplacien

Cette section aborde l'approximation par éléments finis du problème de Dirichlet homogène (5.1)–(5.2). Après avoir établi le caractère bien posé du problème continu, on considère la méthode de Galerkin standard dans un cadre conforme puis non-conforme. On étudie également d'autres conditions aux limites pour le Laplacien ainsi que l'approximation par éléments finis des équations d'advection–diffusion–réaction avec diffusion dominante.

5.1.1 Le cadre mathématique

On suppose pour simplifier que la donnée f est dans $L^2(\Omega)$. La formulation faible du problème de Dirichlet homogène (5.1)–(5.2) est la suivante :

$$\left\{\begin{array}{l} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ \int_\Omega \nabla u \cdot \nabla w = \int_\Omega f w, \quad \forall w \in H_0^1(\Omega). \end{array}\right. \tag{5.3}$$

En introduisant l'espace fonctionnel $V = H_0^1(\Omega)$, la forme bilinéaire $a \in \mathcal{L}(V \times V; \mathbb{R})$ définie par $a(v,w) = \int_\Omega \nabla v \cdot \nabla w$ et la forme linéaire[1] $f \in V'$ définie par $f(w) = \int_\Omega f w$, le problème (5.3) s'écrit sous la forme abstraite suivante :

$$\left\{\begin{array}{l} \text{Chercher } u \in V \text{ tel que} \\ a(u,w) = f(w), \quad \forall w \in V. \end{array}\right. \tag{5.4}$$

On retrouve donc le problème (2.6) analysé dans la section 2.1.1.

1. On commet ici un abus de notation en utilisant le même symbole f pour la fonction de $L^2(\Omega)$ et la forme linéaire sur $H_0^1(\Omega)$.

L'espace V équipé de la norme $\|\cdot\|_{1,\Omega}$ (définie par $\|v\|_{1,\Omega} = (\|v\|_{0,\Omega}^2 + \|\nabla v\|_{0,\Omega}^2)^{1/2}$ pour $v \in V$) est un espace de Hilbert et les formes a et f sont continues sur $V \times V$ et V, respectivement. La seule propriété non-triviale pour établir le caractère bien-posé de (5.3) est la *coercivité* de a sur $H_0^1(\Omega)$. Celle-ci résulte de l'inégalité ci-dessous.

Lemme 5.1 (Inégalité de Poincaré). *Soit Ω un ouvert borné de $\mathbb{R}^d$. Il existe une constante ℓ_Ω telle que*

$$\forall v \in H_0^1(\Omega), \qquad \|v\|_{0,\Omega} \leqslant \ell_\Omega \|\nabla v\|_{0,\Omega}. \tag{5.5}$$

On notera que la constante ℓ_Ω a la dimension d'une longueur ; elle peut s'interpréter comme une longueur caractéristique de l'ouvert borné Ω.[1] L'inégalité de Poincaré implique la coercivité de la forme bilinéaire a sur $H_0^1(\Omega)$ puisque

$$\forall v \in H_0^1(\Omega), \qquad a(v,v) = \|\nabla v\|_{0,\Omega}^2 \geqslant \tfrac{1}{1+\ell_\Omega^2} \|v\|_{1,\Omega}^2. \tag{5.6}$$

Le lemme de Lax–Milgram permet de conclure quant au caractère bien posé du problème (5.3).

Afin d'obtenir des estimations d'erreur optimales en norme L^2 pour l'approximation par éléments finis, on utilise par la suite la notion suivante.

Définition 5.2 (Problème régularisant). *On dit que le problème* (5.3) *est* régularisant *s'il existe une constante c_S telle que pour tout $f \in L^2(\Omega)$, la solution unique u de* (5.3) *est telle que*

$$\|u\|_{2,\Omega} \leqslant c_S \|f\|_{0,\Omega}. \tag{5.7}$$

En termes un peu plus abstraits, le problème (5.3) est régularisant si l'opérateur $(-\Delta)^{-1} : L^2(\Omega) \to H^2(\Omega) \cap H_0^1(\Omega)$ est un isomorphisme. Une condition *suffisante* pour que le problème (5.3) soit régularisant est que Ω est un domaine de classe $\mathcal{C}^2$ ou que Ω est un polygone (ou un polyèdre) *convexe* ; voir Grisvard [**48**].

1. L'inégalité de Poincaré s'étend aux ouverts de $\mathbb{R}^d$ bornés dans au moins une direction.

5.1.2 Approximation conforme

On considère une approximation conforme du problème (5.3) par éléments finis de Lagrange. On suppose pour simplifier que Ω est un polygone de $\mathbb{R}^2$ ou un polyèdre de $\mathbb{R}^3$. On considère une famille régulière et conforme de maillages affines de Ω que l'on note $\{\mathcal{T}_h\}_{h>0}$. On choisit comme élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ un élément fini de Lagrange tel que $\mathbb{P}_k \subset \widehat{P}$ et $k+1 > \frac{d}{2}$.[1] On pose

$$L^k_{c,h} = \{v_h \in \mathcal{C}^0(\overline{\Omega})\,;\, \forall K \in \mathcal{T}_h,\ v_h \circ T_K \in \widehat{P}\}, \tag{5.8}$$

où T_K est la transformation géométrique envoyant $\widehat{K}$ dans K. Si on utilise un élément fini de Lagrange $\mathbb{P}_k$ ou $\mathbb{Q}_k$, on obtient les espaces d'approximation $P^k_{c,h}$ et $Q^k_{c,h}$ définis en (3.44) et (3.45), respectivement. Afin de construire un espace d'approximation qui soit inclus dans $V = H^1_0(\Omega)$, on pose

$$V_h = L^k_{c,h} \cap H^1_0(\Omega). \tag{5.9}$$

Les éléments de V_h sont les fonctions de $L^k_{c,h}$ qui s'annulent sur la frontière de Ω. Si on utilise un élément fini de Lagrange $\mathbb{P}_k$ ou $\mathbb{Q}_k$, on obtient les espaces d'approximation $P^k_{c,h,0}$ et $Q^k_{c,h,0}$ définis en (3.46) et (3.47), respectivement.
On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a(u_h, w_h) = f(w_h), \quad \forall w_h \in V_h, \end{cases} \tag{5.10}$$

qui est clairement bien posé puisque a est coercive sur V et que $V_h \subset V$.

Théorème 5.3 (Convergence). *Avec les hypothèses ci-dessus, on suppose que la solution unique u de* (5.3) *est dans* $H^{k+1}(\Omega) \cap H^1_0(\Omega)$. *Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^k |u|_{k+1,\Omega}. \tag{5.11}$$

1. En dimension 2 ou 3, l'hypothèse $k+1 > \frac{d}{2}$ n'est pas restrictive dès que $k \geqslant 1$.

De plus, si le problème (5.3) *est régularisant, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^{k+1} |u|_{k+1,\Omega}. \tag{5.12}$$

L'estimation (5.11) résulte du lemme de Céa 2.12 et du théorème d'interpolation 3.26. En effet, on a

$$\begin{aligned} \|u - u_h\|_{1,\Omega} &\leqslant c \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \\ &\leqslant c\, \|u - \mathcal{I}_h^{\mathrm{Lag}} u\|_{1,\Omega} \\ &\leqslant c\, h^k |u|_{k+1,\Omega}, \end{aligned}$$

où $\mathcal{I}_h^{\mathrm{Lag}}$ est l'opérateur d'interpolation de Lagrange défini dans la section 3.6. Par ailleurs, l'estimation (5.12) résulte du lemme de Aubin–Nitsche 2.21 qui permet d'affirmer que

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h |u - u_h|_{1,\Omega}, \tag{5.13}$$

si bien que (5.12) se déduit de (5.11).

L'exemple canonique d'application du théorème 5.3 en dimension 2 ou 3 est celui de l'approximation par éléments finis de Lagrange $\mathbb{P}_1$ ou $\mathbb{Q}_1$, pour lesquels, si le problème (5.3) est régularisant, on a

$$\|u - u_h\|_{0,\Omega} + h\|u - u_h\|_{1,\Omega} \leqslant c\, h^2 \|f\|_{0,\Omega}. \tag{5.14}$$

On obtient donc une convergence à l'ordre 1 en norme H^1 et une convergence à l'ordre 2 en norme L^2.

On dit que les estimations d'erreur (5.11) et (5.12) sont *optimales* car leur ordre de convergence en h est *le même* que celui de l'erreur d'interpolation. Lorsque la solution exacte u n'est pas suffisamment régulière, l'erreur d'approximation converge bien vers zéro mais son ordre de convergence n'est pas optimal. Par exemple, sous la seule hypothèse que $u \in H_0^1(\Omega)$, on montre que $\lim_{h \to 0} \|u - u_h\|_{1,\Omega} = 0$. Par ailleurs, si $u \in H^{s+1}(\Omega)$ avec $\frac{d}{2} - 1 < s < k$, il existe une constante c telle que pour tout h,

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^s |u|_{s+1,\Omega}.$$

On obtient donc le même ordre de convergence que si on avait utilisé un élément fini de Lagrange de degré $s < k$.

Remarque 5.4

L'hypothèse de régularité de la famille de maillages $\{\mathcal{T}_h\}_{h>0}$ intervient dans le théorème d'interpolation 3.26 pour obtenir des estimations uniformes de l'erreur d'interpolation valables *pour toute fonction* $v \in H^{k+1}(\Omega)$. Or, pour estimer l'erreur d'approximation $\|u - u_h\|_{1,\Omega}$, on ne doit contrôler l'erreur d'interpolation que pour *une seule fonction*, à savoir la solution exacte de (5.3). On peut donc imaginer que si le maillage est « adapté » à la solution exacte, l'erreur d'approximation ne sera pas trop grande. Ces considérations sont à la base des techniques d'adaptation du maillage basées sur la hessienne (la matrice des dérivées secondes) de la solution exacte u ou d'une approximation de celle-ci ; voir la section 8.6 pour plus de détails.

5.1.3 Approximation non-conforme

On considère une approximation non-conforme du problème (5.3) par éléments finis de Crouzeix–Raviart. On suppose à nouveau que Ω est un polyèdre de $\mathbb{R}^d$ et on considère une famille régulière et conforme de maillages affines de Ω que l'on note $\{\mathcal{T}_h\}_{h>0}$. On choisit comme élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ l'élément fini de Crouzeix–Raviart décrit dans la section 4.4. Partant de l'espace d'éléments finis de Crouzeix-Raviart $P^1_{\mathrm{pt},h}$ défini en (4.31), on considère l'espace d'approximation

$$\begin{aligned} P^1_{\mathrm{pt},h,0} &= \left\{ v_h \in P^1_{\mathrm{pt},h} \,;\, \forall F \in \mathcal{F}_h^{\partial},\ \int_F v_h = 0 \right\} \\ &= \{ v_h \in L^2(\Omega_h) \,;\, \forall K \in \mathcal{T}_h,\ v_h|_K \in \mathbb{P}_1 \,;\, \forall F \in \mathcal{F}_h,\ \textstyle\int_F [\![v_h]\!] = 0 \}. \end{aligned} \tag{5.15}$$

On considère le problème approché suivant :

$$\left\{ \begin{array}{l} \text{Chercher } u_h \in P^1_{\mathrm{pt},h,0} \text{ tel que} \\ a_h(u_h, w_h) = f(w_h), \quad \forall v_h \in P^1_{\mathrm{pt},h,0}, \end{array} \right. \tag{5.16}$$

où la forme bilinéaire a_h est telle que pour tout $(v_h, w_h) \in P^1_{\mathrm{pt},h,0} \times P^1_{\mathrm{pt},h,0}$,

$$a_h(v_h, w_h) = \int_\Omega \nabla_h v_h \cdot \nabla_h w_h. \tag{5.17}$$

On rappelle que l'opérateur de gradient discret est défini en (4.21) ; on a donc

$$a_h(v_h, w_h) = \sum_{K\in\mathcal{T}_h} \int_K \nabla v_h \cdot \nabla w_h. \tag{5.18}$$

Le problème discret (5.16) résulte d'une approximation non-conforme de (5.3) puisque $P^1_{\mathrm{pt},h,0} \not\subset H^1_0(\Omega)$. En effet, les gradients des fonctions de $P^1_{\mathrm{pt},h,0}$ ne sont pas nécessairement dans $[L^2(\Omega)]^d$ (c'est pourquoi la forme bilinéaire a a été remplacée par la forme bilinéaire a_h). De plus, les fonctions de $P^1_{\mathrm{pt},h,0}$ ne sont pas nécessairement nulles sur $\partial\Omega$.

On pose $V(h) = P^1_{\mathrm{pt},h,0} + H^1_0(\Omega)$ et pour $v \in V(h)$, on définit la semi-norme H^1-brisée par

$$|v|_{h,1,\Omega} = \|\nabla_h v\|_{0,\Omega} = \left(\sum_{K\in\mathcal{T}_h} \|\nabla v\|^2_{0,K}\right)^{\frac{1}{2}}. \tag{5.19}$$

On munit l'espace $V(h)$ de la norme $\|\cdot\|_{V(h)} = \|\cdot\|_{0,\Omega} + |\cdot|_{h,1,\Omega}$. L'analyse de convergence de l'approximation par éléments finis de Crouzeix–Raviart repose sur les résultats suivants.

- **Stabilité.** La forme bilinéaire a_h définie en (5.17) est $\|\cdot\|_{V(h)}$-coercive sur $P^1_{\mathrm{pt},h,0}$ uniformément en h ; cette propriété implique que le problème approché (5.16) est bien posé. La coercivité de la forme bilinéaire a_h résulte de l'inégalité suivante, qui est une inégalité de Poincaré étendue ; voir, par exemple, Temam [**72**, Prop. 4.13].

  Lemme 5.5 (Inégalité de Poincaré étendue). *Soit Ω un ouvert borné de $\mathbb{R}^d$. On suppose $h \leqslant 1$. Alors, il existe une constante ℓ'_Ω telle que*

  $$\forall v \in V(h), \qquad \|v\|_{0,\Omega} \leqslant \ell'_\Omega |v|_{h,1,\Omega}. \tag{5.20}$$

- **Continuité.** La forme bilinéaire a_h est uniformément continue en h sur $V(h) \times V(h)$: il existe une constante c, indépendante de h, telle que pour tout $(v, w) \in V(h) \times V(h)$,

  $$a_h(v, w) \leqslant c\, \|v\|_{V(h)} \|w\|_{V(h)}. \tag{5.21}$$

- **Consistance asymptotique.** En supposant que la solution exacte u de (5.3) est dans $H^2(\Omega) \cap H^1_0(\Omega)$, il existe une constante c, indépendante de h, telle que pour tout h,

$$\sup_{w_h \in P^1_{\mathrm{pt},h,0}} \frac{|f(w_h) - a_h(u, w_h)|}{\|w_h\|_{V(h)}} \leqslant c\, h |u|_{2,\Omega}. \tag{5.22}$$

- **Densité asymptotique.** Il existe une constante c, indépendante de h, telle que pour tout h et pour tout $v \in H^2(\Omega) \cap H^1_0(\Omega)$,

$$\inf_{v_h \in P^1_{\mathrm{pt},h,0}} \|v - v_h\|_{V(h)} \leqslant c\, h |v|_{2,\Omega}. \tag{5.23}$$

En procédant comme dans la preuve du deuxième lemme de Strang 2.20, on aboutit au théorème de convergence ci-dessous.

Théorème 5.6 (Convergence). *Soit $\{\mathcal{T}_h\}_{h>0}$ une famille régulière de maillages affines et conformes. On suppose que la solution exacte u de* (5.3) *est dans $H^2(\Omega) \cap H^1_0(\Omega)$. Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{V(h)} \leqslant c\, h |u|_{2,\Omega}. \tag{5.24}$$

De plus, si le problème (5.3) *est régularisant, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^2 |u|_{2,\Omega}. \tag{5.25}$$

5.1.4 Variations sur les conditions aux limites

Cette section passe en revue les conditions aux limites, autres que celles de Dirichlet homogènes, que l'on peut considérer pour le Laplacien.

- **Condition aux limites de Dirichlet non-homogène.** Étant donné une fonction $f \in L^2(\Omega)$ et une fonction $g \in \mathcal{C}^{0,1}(\partial\Omega)$ (g est lipschitzienne sur $\partial\Omega$),[1] on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\Delta u = f \quad \text{dans } \Omega, \tag{5.26}$$

$$u = g \quad \text{sur } \partial\Omega. \tag{5.27}$$

1. Plus généralement, on peut prendre g dans l'espace de Sobolev fractionnaire $H^{\frac{1}{2}}(\partial\Omega)$ défini en (A.56).

L'hypothèse $g \in \mathcal{C}^{0,1}(\partial\Omega)$ permet d'affirmer qu'il existe un *relèvement* u_g de g dans $H^1(\Omega)$, c'est-à-dire qu'il existe une fonction u_g dans $H^1(\Omega)$ telle que $u_g|_{\partial\Omega} = g$. Dans ces conditions, on effectue le changement d'inconnue $u_0 = u - u_g$ et on considère la formulation faible suivante :

$$\left\{ \begin{array}{l} \text{Chercher } u_0 \in H^1_0(\Omega) \text{ tel que} \\ a(u_0, w) = f(w) - a(u_g, w), \quad \forall w \in H^1_0(\Omega). \end{array} \right. \tag{5.28}$$

Par le lemme de Lax–Milgram, ce problème est bien posé.
On s'intéresse à une approximation conforme de (5.28) par éléments finis de Lagrange. On reprend le cadre discret de la section 5.1.2. On suppose que la donnée g est suffisamment régulière pour admettre un relèvement u_g dans $\mathcal{C}^0(\overline{\Omega}) \cap H^1(\Omega)$. On note $\mathcal{I}_h^{\mathrm{Lag}}$ l'opérateur d'interpolation associé au maillage $\mathcal{T}_h$ et à l'élément fini de Lagrange de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$; voir la section 3.6. On rappelle que $L^k_{\mathrm{c},h}$ désigne l'espace H^1-conforme basé sur cet élément fini et que V_h est l'espace d'approximation H^1_0-conforme défini en (5.9). On pose $N = \dim L^k_{\mathrm{c},h}$. On désigne par $\{\varphi_1, \ldots, \varphi_N\}$ la base nodale de $L^k_{\mathrm{c},h}$ et par $\{a_1, \ldots, a_N\}$ les nœuds associés. Par définition, pour $u \in \mathcal{C}^0(\overline{\Omega})$, on a

$$\mathcal{I}_h^{\mathrm{Lag}} u = \sum_{i=1}^{N} u(a_i)\varphi_i, \tag{5.29}$$

et on introduit également *l'interpolé de Lagrange surfacique*

$$\mathcal{I}_h^{\mathrm{Lag}\partial}(u|_{\partial\Omega}) = \sum_{a_i \in \partial\Omega} u(a_i)\varphi_i|_{\partial\Omega}. \tag{5.30}$$

On considère le problème approché

$$\left\{ \begin{array}{l} \text{Chercher } u_{0h} \in V_h \text{ tel que} \\ a(u_{0h}, w_h) = f(w_h) - a(\mathcal{I}_h^{\mathrm{Lag}} u_g, w_h), \quad \forall w_h \in V_h, \end{array} \right. \tag{5.31}$$

qui est clairement bien posé. On pose $u_h = u_{0h} + \mathcal{I}_h^{\mathrm{Lag}} u_g$ si bien que $u_h|_{\partial\Omega}$ coïncide avec l'interpolé de Lagrange surfacique de g. En d'autres termes,

u_h est solution du problème

$$\begin{cases} \text{Chercher } u_h \in L^k_{\mathrm{c},h} \text{ tel que} \\ a(u_h, w_h) = f(w_h), \quad \forall w_h \in V_h, \\ u_h|_{\partial\Omega} = \mathcal{I}_h^{\mathrm{Lag}\partial} g. \end{cases} \tag{5.32}$$

Pour tout nœud surfacique $a_i \in \partial\Omega$, on a donc $u_h(a_i) = g(a_i)$, mais en général $u_h|_{\partial\Omega} \neq g$.

Théorème 5.7. *Avec les hypothèses ci-dessus, on suppose que la solution unique u de* (5.28) *est dans $H^{k+1}(\Omega) \cap H^1_0(\Omega)$. Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^k |u|_{k+1,\Omega}. \tag{5.33}$$

De plus, si le problème (5.28) *est régularisant, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^{k+1} |u|_{k+1,\Omega}. \tag{5.34}$$

On peut également considérer une approximation du problème (5.28) par éléments finis de Crouzeix–Raviart. Le problème approché est le suivant :

$$\begin{cases} \text{Chercher } u_h \in P^1_{\mathrm{pt},h} \text{ tel que} \\ a_h(u_h, w_h) = f(w_h), \quad \forall w_h \in P^1_{\mathrm{pt},h,0}, \\ u_h|_{\partial\Omega} = \mathcal{I}_h^{\mathrm{CR}\partial} g, \end{cases} \tag{5.35}$$

où la forme bilinéaire a_h est définie en (5.17), les espaces $P^1_{\mathrm{pt},h}$ et $P^1_{\mathrm{pt},h,0}$ en (4.31) et (5.15), respectivement, et où $\mathcal{I}_h^{\mathrm{CR}\partial}$ désigne l'interpolé de Crouzeix–Raviart surfacique dont la définition se déduit de (4.33) en se restreignant aux faces situées sur $\partial\Omega$. Pour toute face $F \in \mathcal{F}_h^{\partial}$, on a donc $\int_F u_h = \int_F g$, mais en général $u_h|_{\partial\Omega} \neq g$. La solution u_h de (5.35) satisfait des estimations d'erreur analogues à celles du théorème 5.7.

- **Condition aux limites de Neumann.** Étant donné un réel λ strictement positif, une fonction $f \in L^2(\Omega)$ et une fonction $g \in L^2(\partial\Omega)$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\Delta u + \lambda u = f \quad \text{dans } \Omega, \tag{5.36}$$

$$\partial_n u = g \quad \text{sur } \partial\Omega, \tag{5.37}$$

où $\partial_n u$ désigne la *dérivée normale* de u sur la frontière[1]. La formulation faible de (5.36)–(5.37) consiste à

$$\begin{cases} \text{Chercher } u \in H^1(\Omega) \text{ tel que} \\ a_{\mathrm{N}}(u, w) = f_{\mathrm{N}}(w), \quad \forall w \in H^1(\Omega), \end{cases} \tag{5.38}$$

où

$$a_{\mathrm{N}}(v, w) = \int_\Omega \nabla v \cdot \nabla w + \int_\Omega \lambda v w, \tag{5.39}$$

et

$$f_{\mathrm{N}}(w) = \int_\Omega f w + \int_{\partial\Omega} g w. \tag{5.40}$$

Le caractère bien posé de (5.38) résulte de la coercivité de la forme bilinéaire a_{N} sur $H^1(\Omega)$.
Le problème (5.38) peut être approché par des éléments finis de Lagrange. On reprend le cadre discret de la section 5.1.2. On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in L^k_{\mathrm{c},h} \text{ tel que} \\ a_{\mathrm{N}}(u_h, w_h) = f_{\mathrm{N}}(w_h), \quad \forall w_h \in L^k_{\mathrm{c},h}, \end{cases} \tag{5.41}$$

qui est clairement bien posé. Une différence importante entre la condition aux limites de Neumann et celle de Dirichlet est que la première n'est pas imposée explicitement dans l'espace où on cherche la solution mais résulte du fait que les fonctions tests dans (5.38) peuvent prendre des valeurs non-nulles au bord. Cette différence se traduit, au niveau de l'approximation par éléments finis, par le fait que la solution u_h de (5.41) satisfait la condition

1. En notant $n = (n_1, \ldots, n_d)^T$ les coordonnées cartésiennes de la normale extérieure en un point de la frontière, on a par définition $\partial_n u = n \cdot \nabla u = \sum_{i=1}^d n_i \partial_i u$.

aux limites de Neumann de manière approchée et non de manière exacte. Enfin, l'analyse de convergence du problème approché (5.41) conduit, sous les hypothèses du théorème 5.3, aux mêmes estimations que pour le problème de Dirichlet homogène.

On peut également considérer, au prix de quelques difficultés techniques, le problème de Neumann avec $\lambda = 0$. On cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\Delta u = f \quad \text{dans } \Omega, \tag{5.42}$$

$$\partial_n u = g \quad \text{sur } \partial\Omega. \tag{5.43}$$

On observe qu'une condition nécessaire à l'existence d'une solution est que les données f et g satisfassent la condition de compatibilité

$$\int_\Omega f + \int_{\partial\Omega} g = 0, \tag{5.44}$$

puisque $\int_\Omega f + \int_{\partial\Omega} g = -\int_\Omega \Delta u + \int_{\partial\Omega} \partial_n u = 0$ d'après le *théorème de la divergence*[1]. Par ailleurs, une solution de (5.42)–(5.43) n'est déterminée qu'à une constante additive près. On convient donc de chercher la solution dans l'espace fonctionnel

$$H^1_*(\Omega) = \left\{ v \in H^1(\Omega) \, ; \, \int_\Omega v = 0 \right\}. \tag{5.45}$$

La formulation faible de (5.42)–(5.43) est la suivante :

$$\left\{ \begin{array}{l} \text{Chercher } u \in H^1_*(\Omega) \text{ tel que} \\ a(u, w) = f_{\mathrm{N}}(w), \quad \forall w \in H^1_*(\Omega), \end{array} \right. \tag{5.46}$$

1. Le théorème de la divergence exprime le fait que pour tout champ de vecteurs ϕ suffisamment régulier,

$$\int_\Omega \nabla \cdot \phi = \int_{\partial\Omega} \phi \cdot n.$$

En particulier, en prenant $\phi = \nabla u$ où u est une fonction suffisamment régulière, on obtient $\int_\Omega \Delta u = \int_{\partial\Omega} \partial_n u$.

avec la forme bilinéaire a telle que $a(v,w) = \int_\Omega \nabla v \cdot \nabla w$. Le caractère bien posé de (5.46) résulte de la coercivité de la forme bilinéaire a sur $H^1_*(\Omega)$, cette dernière propriété étant elle-même une conséquence de l'inégalité ci-dessous.

Lemme 5.8 (Inégalité de Poincaré–Wirtinger). *Soit Ω un ouvert borné de $\mathbb{R}^d$. Il existe une constante ℓ''_Ω telle que*

$$\forall v \in H^1_*(\Omega), \qquad \|v\|_{0,\Omega} \leqslant \ell''_\Omega \|\nabla v\|_{0,\Omega}. \tag{5.47}$$

Le problème (5.46) peut être approché par des éléments finis de Lagrange, ce qui conduit au problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in L^k_{c,h} \text{ tel que} \\ a(u_h, w_h) = f_N(w_h), \quad \forall w_h \in L^k_{c,h}. \end{cases} \tag{5.48}$$

On observera que l'espace d'approximation $L^k_{c,h}$ n'est pas conforme dans $H^1_*(\Omega)$ et que le système linéaire associé au problème discret (5.48) est singulier. En notant ce système linéaire sous la forme

$$\mathcal{A}^* U = F^*, \tag{5.49}$$

et en introduisant le vecteur $Z = (1, \ldots, 1)^T \in \mathbb{R}^N$ où $N = \dim L^k_{c,h}$, on constate que $\mathrm{Ker}(\mathcal{A}^*) = \mathrm{vect}(Z)$, $\mathrm{Im}(\mathcal{A}^*) = Z^\perp$ et $F \in Z^\perp$. Par conséquent, le système linéaire (5.49) admet une infinité de solutions. L'une d'entre elles peut être approchée à l'aide d'une des méthodes itératives décrites dans le chapitre 11, par exemple la méthode du gradient conjugué. On désigne par U^∞ l'approximation fournie par la méthode itérative. La solution de (5.49) dont les composantes sont de moyenne nulle est obtenue en posant

$$U = U^\infty - \left(\frac{(U^\infty, Z)_{\mathbb{R}^N}}{(Z,Z)_{\mathbb{R}^N}} \right) Z, \tag{5.50}$$

où $(\cdot,\cdot)_{\mathbb{R}^N}$ désigne le produit scalaire euclidien sur $\mathbb{R}^N$.

- **Condition aux limites de Robin.** Étant donné une fonction $f \in L^2(\Omega)$ et une fonction $g \in L^2(\partial\Omega)$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\Delta u = f \quad \text{dans } \Omega, \tag{5.51}$$
$$\gamma u + \partial_n u = g \quad \text{sur } \partial\Omega, \tag{5.52}$$

où $\gamma > 0$ est un paramètre réel[1]. La formulation faible de (5.51)–(5.52) consiste à

$$\begin{cases} \text{Chercher } u \in H^1(\Omega) \text{ tel que} \\ a_{\mathrm{R}}(u, w) = f_{\mathrm{N}}(w), \quad \forall w \in H^1(\Omega), \end{cases} \tag{5.53}$$

où la forme linéaire f_{N} est définie en (5.40) et où la forme bilinéaire a_{R} est telle que

$$a_{\mathrm{R}}(v, w) = \int_\Omega \nabla v \cdot \nabla w + \int_{\partial\Omega} \gamma v w. \tag{5.54}$$

Le caractère bien posé de (5.53) résulte de la coercivité de a_{R} sur $H^1(\Omega)$, cette dernière propriété étant elle-même une conséquence du fait qu'il existe une constante ρ_Ω telle que

$$\forall v \in H^1(\Omega), \qquad \|v\|_{0,\Omega} \leqslant \rho_\Omega(\|\nabla v\|_{0,\Omega} + \|v\|_{0,\partial\Omega}). \tag{5.55}$$

Partant du cadre discret de la section 5.1.2, on s'intéresse à une approximation de (5.53) par éléments finis de Lagrange. On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in L^k_{\mathrm{c},h} \text{ tel que} \\ a_{\mathrm{R}}(u_h, w_h) = f_{\mathrm{N}}(w_h), \quad \forall w_h \in L^k_{\mathrm{c},h}, \end{cases} \tag{5.56}$$

qui est clairement bien posé. Comme pour la condition aux limites de Neumann, celle de Robin résulte du fait que les fonctions tests dans (5.53) peuvent prendre des valeurs non-nulles au bord ; elle n'est donc pas sa-

1. Plus généralement, on peut considérer une fonction $\gamma \in L^\infty(\partial\Omega)$ telle que $\gamma(x) \geqslant \gamma_0 > 0$ pour presque tout $x \in \partial\Omega$.

tisfaite de manière exacte par la solution discrète u_h mais uniquement de manière approchée. Enfin, l'analyse de convergence pour la solution u_h de (5.56) conduit, sous les hypothèses du théorème 5.3, aux mêmes estimations que pour le problème de Dirichlet homogène.

5.1.5 Advection–diffusion–réaction avec diffusion dominante

On considère un opérateur différentiel de la forme

$$\mathcal{L}u = -\nabla\cdot(\sigma\cdot\nabla u) + \beta\cdot\nabla u + \mu u, \tag{5.57}$$

où σ, β et μ sont des fonctions définies sur Ω et à valeurs dans $\mathbb{R}^{d,d}$, $\mathbb{R}^d$ et $\mathbb{R}$ respectivement. L'opérateur (5.57) intervient, par exemple, dans la modélisation des problèmes d'advection–diffusion–réaction ; le premier terme du membre de droite dans (5.57) est le terme diffusif, le deuxième le terme advectif[1] et le troisième le terme réactif. L'opérateur $\mathcal{L}$ intervient également dans des problèmes de finance liés au pricing d'options. Par la suite, on suppose que $\sigma \in [L^\infty(\Omega)]^{d,d}$, $\beta \in [L^\infty(\Omega)]^d$, $\nabla\cdot\beta \in L^\infty(\Omega)$ et $\mu \in L^\infty(\Omega)$. De plus, on suppose que l'opérateur $\mathcal{L}$ est *elliptique* au sens suivant.

Définition 5.9 (Opérateur elliptique). *L'opérateur $\mathcal{L}$ défini en* (5.57) *est dit* elliptique *s'il existe une constante $\sigma_0 > 0$ telle que, presque partout dans Ω,*

$$\forall \xi \in \mathbb{R}^d, \qquad \sum_{i,j=1}^{d} \sigma_{ij}\xi_i\xi_j \geqslant \sigma_0 \left(\sum_{i=1}^{d} \xi_i^2\right). \tag{5.58}$$

Le Laplacien entre dans la catégorie des opérateurs elliptiques puisqu'il est obtenu à partir de $\mathcal{L}$ en prenant $\beta = 0$, $\mu = 0$ et $\sigma = \mathcal{I}_d$, la matrice identité dans $\mathbb{R}^{d,d}$.

1. Ce terme est parfois appelé le terme « convectif » et on parle d'équation de convection–diffusion–réaction.

Étant donné une fonction $f \in L^2(\Omega)$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$\mathcal{L}u = f \quad \text{dans } \Omega, \tag{5.59}$$
$$u = 0 \quad \text{sur } \partial\Omega. \tag{5.60}$$

La formulation faible de (5.59)–(5.60) est la suivante :

$$\begin{cases} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ a_{\sigma\beta\mu}(u, w) = f(w), \quad \forall w \in H_0^1(\Omega), \end{cases} \tag{5.61}$$

avec la forme bilinéaire $a_{\sigma\beta\mu}$ telle que pour tout $(v, w) \in H_0^1(\Omega) \times H_0^1(\Omega)$,

$$a_{\sigma\beta\mu}(v, w) = \int_\Omega \nabla w \cdot \sigma \cdot \nabla v + w(\beta \cdot \nabla v) + \mu w v. \tag{5.62}$$

On pose[1] $p = \operatorname{infess}_{x\in\Omega} \left(\mu - \frac{1}{2}\nabla\cdot\beta\right)$ et on rappelle que ℓ_Ω est la constante intervenant dans l'inégalité de Poincaré (5.5). Alors, on montre que sous la condition

$$\sigma_0 + \min(0, \ell_\Omega p) > 0, \tag{5.63}$$

la forme bilinéaire $a_{\sigma\beta\mu}$ est coercive sur $H_0^1(\Omega)$, si bien que, grâce au lemme de Lax–Milgram, le problème (5.61) est bien posé. La condition (5.63) étant une minoration de σ_0, on retiendra que la coercivité est garantie pourvu que σ_0 soit suffisamment grand, c'est-à-dire en régime de *diffusion dominante*.

L'approximation conforme de (5.61) par éléments finis de Lagrange reprend le cadre discret introduit dans la section 5.1.2 pour le Laplacien. On peut également considérer une approximation non-conforme par éléments finis de Crouzeix–Raviart, la coercivité du problème approché nécessitant en particulier d'utiliser l'inégalité de Poincaré étendue (5.20). Enfin, on peut considérer à la place de (5.60) des conditions aux limites de Dirichlet non-homogène, de Neumann ou de Robin.

1. Pour une fonction $f \in L^\infty(\Omega)$, $\operatorname{infess}_{x\in\Omega} f(x) = \sup\{M \in \mathbb{R}_+ ; f(x) \geqslant M$ presque partout dans $\Omega\}$.

5.1.6 Principe du maximum discret

On considère le problème (5.1)–(5.2) et on suppose que $f \leqslant 0$ dans Ω. On suppose que la solution u est suffisamment régulière, à savoir $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$. Dans ces conditions, $u \leqslant 0$ dans Ω. Cette propriété du problème continu (5.1)–(5.2) porte le nom de *principe du maximum* ; voir, par exemple, Gilbarg et Trudinger [**45**, p. 179]. On considère maintenant une approximation par éléments finis du problème (5.1)–(5.2), par exemple le problème (5.10). Si on a l'implication

$$(f \leqslant 0 \text{ dans } \Omega) \Longrightarrow (u_h \leqslant 0 \text{ dans } \Omega), \tag{5.64}$$

on dit que l'approximation par éléments finis jouit d'un *principe du maximum discret*. Cette propriété n'est pas évidente *a priori*, même pour une approximation aussi simple que celle par éléments finis de Lagrange $\mathbb{P}_1$.

Définition 5.10 (M-matrice). *On dit qu'une matrice $\mathcal{A} \in \mathbb{R}^{N,N}$ est une M-*matrice *si elle est telle que :*

(i) $\mathcal{A}_{ii} > 0$ *pour tout* $i \in \{1, \ldots, N\}$ *;*

(ii) $\mathcal{A}_{ij} \leqslant 0$ *pour tout* $i, j \in \{1, \ldots, N\}$ *et* $i \neq j$ *;*

(iii) *en posant* $Z = (1, \ldots, 1)^T \in \mathbb{R}^N$*, le vecteur* $\mathcal{A}Z$ *a toutes ses composantes positives et au moins une composante strictement positive ; en d'autres termes, la somme des coefficients dans la* i*-ième ligne de la matrice* $\mathcal{A}$ *est positive pour tout* $i \in \{1, \ldots, N\}$ *et cette somme est strictement positive pour au moins un* $i \in \{1, \ldots, N\}$*.*

Une propriété importante des M-matrices est la suivante.

Proposition 5.11. *On suppose que la matrice $\mathcal{A} \in \mathbb{R}^{N,N}$ est une M-matrice. Alors, $\mathcal{A}$ est inversible et tous les coefficients de son inverse sont positifs.*

La proposition 5.11 implique que si le vecteur $F \in \mathbb{R}^N$ a toutes ses composantes négatives, il en est de même du vecteur $U \in \mathbb{R}^N$ qui est solution du système linéaire $\mathcal{A}U = F$. Dans le cadre de la méthode des éléments finis, les composantes du vecteur F s'expriment sous la forme $F_i = f(\varphi_i)$, $i \in \{1, \ldots, N\}$, où $\{\varphi_1, \ldots, \varphi_N\}$ désignent les fonctions de forme dans l'espace d'approximation. Par conséquent, une condition *suffisante* pour que la propriété (5.64) soit satisfaite est que *la matrice de rigidité $\mathcal{A}$ soit une M-matrice*.

On s'intéresse maintenant à l'approximation par éléments finis de Lagrange $\mathbb{P}_1$ du problème de Dirichlet homogène (5.3). Les coefficients de la matrice de rigidité sont tels que

$$\mathcal{A}_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j, \qquad i,j \in \{1,\dots,N\}. \tag{5.65}$$

En dimension 1, la matrice de rigidité est tridiagonale et on vérifie facilement que c'est une M-matrice. En dimension $d \geqslant 2$, la matrice de rigidité est une M-matrice si et seulement si le maillage satisfait certaines conditions géométriques. On considère un maillage de Ω par des simplexes. Pour une maille $K \in \mathcal{T}_h$ et une arête a de K, on désigne par $s_{1,a}$ et $s_{2,a}$ les deux sommets de l'arête et par $F_{1,a}$ et $F_{2,a}$ les faces de K opposées aux sommets $s_{1,a}$ et $s_{2,a}$, respectivement. On désigne par $\mu_{a,K}$ la mesure (en dimension $(d-2)$) de l'intersection $F_{1,a} \cap F_{2,a}$ ($\mu_{a,K} = 1$ en dimension deux) et par $\theta_{a,K}$ l'angle entre les faces $F_{1,a}$ et $F_{2,a}$. On a le résultat suivant ; voir Xu et Zikatanov [77].

Proposition 5.12. *La matrice de rigidité pour le problème de Dirichlet homogène est une M-matrice si et seulement si le maillage satisfait la condition géométrique suivante : pour toute arête a du maillage, on a*

$$\frac{1}{d(d-1)} \sum_{K \in \mathcal{T}(a)} \mu_{a,K} \cot(\theta_{a,K}) \geqslant 0, \tag{5.66}$$

où $\mathcal{T}(a)$ désigne l'ensemble des éléments de $\mathcal{T}_h$ dont a est une arête et cot *la fonction cotangente.*

En deux dimensions d'espace, la condition géométrique de la proposition 5.12 signifie que pour chaque arête intérieure du maillage, la somme des deux angles opposés à cette arête est inférieure ou égale à π. Un tel maillage est appelé une triangulation de *Delaunay* ; voir, par exemple, George et Borouchaki [**44**] ou Frey et George [**41**]. Une condition suffisante pour qu'une triangulation soit de Delaunay est que tous les angles des triangles soient aigus, mais une telle condition est plus restrictive, et donc moins générale, que celle de la proposition 5.12.

On peut également formuler des schémas d'approximation par éléments finis de Lagrange $\mathbb{P}_1$ qui jouissent d'un principe du maximum discret pour les équations d'advection–diffusion–réaction. En particulier, on souhaite que le

principe du maximum discret soit satisfait uniformément pour toutes les valeurs du champ d'advection β, c'est-à-dire aussi bien en régime de diffusion dominante (voir la section 5.1.5) qu'en régime d'advection dominante (voir la section 7.3). Dans ce cas, on ne peut pas imposer des conditions géométriques sur le maillage pour que la matrice de rigidité associée au problème (5.61) soit une M-matrice. Par contre, sous les hypothèses de la proposition 5.12, on montre qu'on peut rajouter un terme de *capture de choc* non-linéaire à la forme bilinéaire $a_{\sigma\beta\mu}$ définie en (5.62) de sorte que le problème approché (5.61) jouisse d'un principe du maximum discret uniformément en la valeur du champ d'advection β ; voir [**26**] pour plus de détails.

5.2 Élasticité linéaire

Cette section aborde l'approximation par éléments finis de problèmes de mécanique des milieux continus déformables en trois dimensions d'espace. Le domaine $\Omega \subset \mathbb{R}^3$ représente un milieu déformable, initialement au repos et auquel on applique un chargement extérieur $f : \Omega \to \mathbb{R}^3$. L'objectif est de déterminer le champ de déplacement $u : \Omega \to \mathbb{R}^3$ induit par ce chargement une fois que le milieu a atteint l'équilibre. On suppose que les déformations sont suffisamment petites pour pouvoir les modéliser dans le cadre de *l'élasticité linéaire*. Pour simplifier, on suppose également que le milieu est isotrope.

On note $\sigma : \Omega \to \mathbb{R}^{3,3}$ le champ des *contraintes*. La condition d'équilibre s'écrit

$$\nabla\cdot\sigma + f = 0 \quad \text{dans } \Omega. \tag{5.67}$$

On note $\varepsilon(u) : \Omega \to \mathbb{R}^{3,3}$ le champ des *déformations linéarisées* défini par[1]

$$\varepsilon(u) = \tfrac{1}{2}(\nabla u + \nabla u^T). \tag{5.68}$$

Dans le cadre de l'élasticité linéaire, le champ des contraintes s'exprime en fonction du champ des déformations linéarisées sous la forme

$$\sigma(u) = \lambda \operatorname{tr}(\varepsilon(u))\mathcal{I}_3 + 2\mu\varepsilon(u), \tag{5.69}$$

1. On rappelle que ∇u est le champ à valeurs dans $\mathbb{R}^{3,3}$ de composantes $(\nabla u)_{ij} = \partial_j u_i$, $i,j \in \{1,2,3\}$.

où λ et μ sont les *coefficients de Lamé* et $\mathcal{I}_3$ la matrice identité dans $\mathbb{R}^{3,3}$. En utilisant (5.68), il vient

$$\sigma(u) = \lambda(\nabla\cdot u)\mathcal{I}_3 + \mu(\nabla u + \nabla u^T). \tag{5.70}$$

Les coefficients de Lamé sont des coefficients phénoménologiques qui, pour des raisons thermodynamiques, sont contraints par les relations $\mu > 0$ et $\lambda + \frac{2}{3}\mu \geqslant 0$. Pour simplifier, on suppose que $\lambda \geqslant 0$ et que le milieu est *homogène* si bien que les coefficients λ et μ sont des constantes. Dans certaines applications, il est plus pratique d'introduire le *module de Young* E et le *coefficient de Poisson* ν tels que

$$E = \mu\frac{3\lambda+2\mu}{\lambda+\mu} \qquad \text{et} \qquad \nu = \tfrac{1}{2}\frac{\lambda}{\lambda+\mu}. \tag{5.71}$$

Le coefficient de Poisson est tel que $-1 \leqslant \nu < \frac{1}{2}$; de plus, $\nu \geqslant 0$ si $\lambda \geqslant 0$. Un coefficient de Poisson très proche de $\frac{1}{2}$ (c'est-à-dire des coefficients de Lamé tels que le rapport $\frac{\lambda}{\mu}$ est très grand) caractérise un matériau pratiquement incompressible.

Le problème modèle (5.67)–(5.70) doit être complété par des conditions aux limites. Par la suite, on considère deux cas.

(i) Le problème de Dirichlet homogène où on impose la condition aux limites

$$u = 0 \quad \text{sur } \partial\Omega. \tag{5.72}$$

(ii) Le problème de traction pure (ou problème de Neumann) où on impose la condition aux limites

$$\sigma(u)\cdot n = g \quad \text{sur } \partial\Omega, \tag{5.73}$$

où $g : \partial\Omega \to \mathbb{R}^3$ représente un champ de forces appliqué sur le bord de Ω et n la normale extérieure à Ω.

D'autres conditions aux limites sont possibles comme, par exemple, des conditions mêlées Dirichlet–Neumann.

5.2.1 Le cadre mathématique

Sur $[H^1(\Omega)]^3 \times [H^1(\Omega)]^3$, on introduit la forme bilinéaire[1]

$$a(v,w) = \int_\Omega \sigma(v){:}\varepsilon(w) = \int_\Omega \lambda(\nabla{\cdot}v)(\nabla{\cdot}w) + \int_\Omega 2\mu\varepsilon(v){:}\varepsilon(w). \quad (5.74)$$

Il est clair que $a \in \mathcal{L}([H^1(\Omega)]^3 \times [H^1(\Omega)]^3; \mathbb{R})$. De plus, la forme bilinéaire a est symétrique et positive. Par contre, la forme bilinéaire a est singulière. En effet, en introduisant l'espace vectoriel des *déplacements rigides*

$$\mathcal{R} = \{z \in [H^1(\Omega)]^3 \,;\, \exists(\alpha,\beta) \in \mathbb{R}^3 \times \mathbb{R}^3,\ \forall x \in \Omega,\ z(x) = \alpha + \beta \times x\}, \quad (5.75)$$

on a l'équivalence

$$(z \in \mathcal{R}) \iff (\forall v \in [H^1(\Omega)]^3,\ a(z,v) = 0). \quad (5.76)$$

On notera qu'un déplacement rigide consiste en la composée d'une translation et d'une rotation et que

$$\mathcal{R} \cap [H_0^1(\Omega)]^3 = \{0\}, \quad (5.77)$$

puisque le seul déplacement rigide qui conserve la frontière est le déplacement nul.

- **Problème de Dirichlet homogène.** Afin d'écrire le problème de Dirichlet homogène sous forme faible, on introduit l'espace fonctionnel

$$V_D = [H_0^1(\Omega)]^3, \quad (5.78)$$

et la forme linéaire $f_D \in V_D'$ telle que pour tout $w \in V_D$,

$$f_D(w) = \int_\Omega f{\cdot}w. \quad (5.79)$$

On obtient le problème suivant :

$$\begin{cases} \text{Chercher } u \in V_D \text{ tel que} \\ a(u,w) = f_D(w), \quad \forall w \in V_D. \end{cases} \quad (5.80)$$

1. Pour deux matrices σ et ε dans $\mathbb{R}^{3,3}$, $\sigma{:}\varepsilon$ désigne leur contraction maximale, qui est égale à $\sum_{i,j=1}^3 \sigma_{ij}\varepsilon_{ij}$.

Le caractère bien posé de (5.80) résulte de l'inégalité ci-dessous.

Lemme 5.13 (Première inégalité de Korn). *Soit* Ω *un domaine de* $\mathbb{R}^3$. *On pose* $\|\varepsilon(v)\|_{0,\Omega} = (\int_\Omega \varepsilon(v){:}\varepsilon(v))^{\frac{1}{2}}$. *Il existe une constante* κ_Ω *telle que*

$$\forall v \in [H_0^1(\Omega)]^3, \qquad \kappa_\Omega \|v\|_{1,\Omega} \leqslant \|\varepsilon(v)\|_{0,\Omega}. \tag{5.81}$$

L'inégalité (5.81) implique la coercivité de la forme bilinéaire a sur $V_{\mathrm{D}} = [H_0^1(\Omega)]^3$ puisque

$$\forall v \in [H_0^1(\Omega)]^3, \qquad a(v,v) \geqslant 2\mu \|\varepsilon(v)\|_{0,\Omega}^2 \geqslant 2\mu\kappa_\Omega^2 \|v\|_{1,\Omega}^2. \tag{5.82}$$

- **Problème de traction pure.** L'étude mathématique du problème de traction pure demande quelques précautions. En effet, on ne peut pas chercher la solution u dans $[H^1(\Omega)]^3$ et demander que pour tout $w \in [H^1(\Omega)]^3$, $a(u,w) = \int_\Omega f\cdot w + \int_{\partial\Omega} g\cdot w$ car la forme bilinéaire a est singulière sur $[H^1(\Omega)]^3 \times [H^1(\Omega)]^3$. Une condition nécessaire pour l'existence d'une solution est que

$$\forall z \in \mathcal{R}, \qquad \int_\Omega f\cdot z + \int_{\partial\Omega} g\cdot z = 0. \tag{5.83}$$

Cette équation exprime le fait que la résultante de l'ensemble des efforts extérieurs et leurs moments sont nuls. De plus, la solution u, si elle existe, n'est déterminée qu'à un déplacement rigide près. On considère donc l'espace fonctionnel

$$V_{\mathrm{N}} = \left\{ v \in [H^1(\Omega)]^3 \,;\, \int_\Omega v = 0 \,;\, \int_\Omega \nabla\times v = 0 \right\}, \tag{5.84}$$

et la forme linéaire $f_{\mathrm{N}} \in V_{\mathrm{N}}'$ telle que pour tout $w \in V_{\mathrm{N}}$,

$$f_{\mathrm{N}}(w) = \int_\Omega f\cdot w + \int_{\partial\Omega} g\cdot w. \tag{5.85}$$

On obtient le problème suivant :

$$\left\{ \begin{array}{l} \text{Chercher } u \in V_{\mathrm{N}} \text{ tel que} \\ a(u,w) = f_{\mathrm{N}}(w), \quad \forall w \in V_{\mathrm{N}}. \end{array} \right. \tag{5.86}$$

Le caractère bien posé de (5.86) résulte de l'inégalité ci-dessous.

Lemme 5.14 (Deuxième inégalité de Korn). *Soit* Ω *un domaine de* $\mathbb{R}^3$. *Il existe une constante* κ'_Ω *telle que*

$$\forall v \in [H^1(\Omega)]^3, \qquad \kappa'_\Omega \|v\|_{1,\Omega} \leqslant \|\varepsilon(v)\|_{0,\Omega} + \|v\|_{0,\Omega}. \tag{5.87}$$

On montre que l'inégalité (5.87) implique la coercivité de la forme bilinéaire a sur V_N.

Remarque 5.15

En mécanique des milieux continus, la fonction test w intervenant dans les formulations faibles (5.80) et (5.86) s'interprète comme un champ de déplacement virtuel admissible et les formulations faibles expriment le *principe des travaux virtuels*. Par ailleurs, la forme bilinéaire a étant symétrique et coercive sur V_D et V_N, la solution unique de (5.80) et (5.86), respectivement, minimise sur V_D et V_N la fonctionnelle d'énergie

$$\mathcal{E}_D(v) = \tfrac{1}{2}\lambda \int_\Omega (\nabla\cdot v)^2 + \tfrac{1}{2}\mu \int_\Omega \varepsilon(v):\varepsilon(v) - f_D(v), \tag{5.88}$$

et

$$\mathcal{E}_N(v) = \tfrac{1}{2}\lambda \int_\Omega (\nabla\cdot v)^2 + \tfrac{1}{2}\mu \int_\Omega \varepsilon(v):\varepsilon(v) - f_N(v). \tag{5.89}$$

On retrouve le *principe de moindre énergie*. Les termes quadratiques en v dans (5.88)–(5.89) représentent l'énergie élastique de déformation et les termes linéaires l'énergie potentielle sous le champ des forces extérieures.

5.2.2 Approximation conforme

On considère une approximation conforme des problèmes (5.80) et (5.86) par éléments finis de Lagrange. On suppose que Ω est un polyèdre de $\mathbb{R}^3$ et on considère une famille régulière et conforme de maillages affines de Ω que l'on note $\{\mathcal{T}_h\}_{h>0}$. On choisit comme élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ un élément fini de Lagrange de degré $k \geqslant 1$.

- **Problème de Dirichlet homogène.** Afin de construire un espace d'approximation V_D-conforme, on pose

$$V_{Dh} = [L^k_{c,h}]^3 \cap [H^1_0(\Omega)]^3, \tag{5.90}$$

où $L^k_{\mathrm{c},h}$ est défini en (5.8). Les éléments de $V_{\mathrm{D}h}$ sont les champs de vecteurs dont chaque composante est dans $L^k_{\mathrm{c},h}$ et qui s'annulent sur la frontière de Ω.
On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_{\mathrm{D}h} \text{ tel que} \\ a(u_h, w_h) = f_{\mathrm{D}}(w_h), \quad \forall w_h \in V_{\mathrm{D}h}, \end{cases} \tag{5.91}$$

qui est clairement bien posé puisque a est coercive sur V_{D} et que $V_{\mathrm{D}h} \subset V_{\mathrm{D}}$.

Théorème 5.16 (Convergence). *Avec les hypothèses ci-dessus, on suppose que la solution unique u de* (5.80) *est dans $[H^{k+1}(\Omega) \cap H^1_0(\Omega)]^3$. Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^k |u|_{k+1,\Omega}. \tag{5.92}$$

De plus, si le problème (5.80) *est* régularisant *(c'est-à-dire s'il existe une constante c_S telle que pour tout $f \in [L^2(\Omega)]^3$, la solution unique de* (5.80) *satisfait $\|u\|_{2,\Omega} \leqslant c_S \|f\|_{0,\Omega}$), il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^{k+1} |u|_{k+1,\Omega}. \tag{5.93}$$

L'estimation (5.92) résulte du lemme de Céa 2.12 et du théorème d'interpolation 3.26 que l'on applique composante par composante. L'estimation (5.93) résulte du lemme de Aubin–Nitsche 2.21.

- **Problème de traction pure.** Pour le problème de traction pure, une manière d'éliminer le déplacement rigide arbitraire au niveau discret consiste à :

  (i) imposer que le déplacement en un nœud du maillage, a_0, est nul ;

  (ii) choisir trois autres nœuds du maillage, a_1, a_2, a_3, et trois vecteurs unitaires, τ_1, τ_2, τ_3, tels que l'ensemble $\{(a_i - a_0) \times \tau_i\}_{1 \leqslant i \leqslant 3}$ forme une base de $\mathbb{R}^3$;

  (iii) imposer que le déplacement au nœud a_i le long de la direction τ_i est nul.

Ceci conduit à l'espace d'approximation

$$V_{\mathrm{N}h} = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^3 \,;\, \forall K \in \mathcal{T}_h,\ v_h \circ T_K \in [\widehat{P}]^3 \,;\ v_h(a_0) = 0 \,;\, v_h(a_i)\cdot\boldsymbol{\tau}_i = 0,\ i \in \{1,2,3\}\}. \tag{5.94}$$

On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_{\mathrm{N}h} \text{ tel que} \\ a(u_h, w_h) = f_{\mathrm{N}}(w_h), \quad \forall w_h \in V_{\mathrm{N}h}. \end{cases} \tag{5.95}$$

En utilisant la deuxième inégalité de Korn, on montre que la forme bilinéaire a est coercive sur $V_{\mathrm{N}h}$ si bien que le problème discret (5.95) est bien posé.

Théorème 5.17 (Convergence). *Avec les hypothèses ci-dessus, on suppose que la solution unique u de* (5.86) *est dans* $[H^{k+1}(\Omega)]^3 \cap V_{\mathrm{N}}$. *Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} \leqslant c\, h^k |u|_{k+1,\Omega}. \tag{5.96}$$

De plus, si $g = 0$ *et si le problème* (5.86) *est* régularisant, *(c'est-à-dire s'il existe une constante* c_S *telle que pour tout* $f \in [L^2(\Omega)]^3$, *la solution unique de* (5.86) *avec* $g = 0$ *satisfait* $\|u\|_{2,\Omega} \leqslant c_S \|f\|_{0,\Omega}$*), il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^{k+1} |u|_{k+1,\Omega}. \tag{5.97}$$

Remarque 5.18

Une condition suffisante pour que les problèmes (5.80) et (5.86) soient régularisants est que le polyèdre Ω soit convexe et que $g = 0$; voir, par exemple, Grisvard [**48**].

5.2.3 Matériaux pratiquement incompressibles : perte de coercivité

Dans cette section, on s'intéresse à des matériaux dont le rapport des coefficients de Lamé est tel que

$$\frac{\lambda}{\mu} \gg 1. \tag{5.98}$$

Cette situation se produit lorsque le coefficient de Poisson est très proche de $\frac{1}{2}$, c'est-à-dire pour des matériaux pratiquement incompressibles.

Pour de tels matériaux, on constate que si on utilise un maillage qui n'est pas suffisamment fin, la solution discrète est polluée par des oscillations parasites. Ce phénomène, qu'on appelle *perte de coercivité*, peut s'expliquer en reprenant l'analyse d'erreur présentée dans la section 5.2.2. Le rapport $\frac{\lambda}{\mu}$ étant très grand, il n'est pas raisonnable de l'absorber dans les constantes génériques c apparaissant dans les estimations d'erreur.

On considère la forme bilinéaire a définie en (5.74). On pose

$$\alpha_a = \inf_{v \in V} \frac{a(v,v)}{\|v\|_{1,\Omega}^2}, \tag{5.99}$$

$$\omega_a = \sup_{v \in V} \sup_{w \in V} \frac{a(v,w)}{\|v\|_{1,\Omega}\|w\|_{1,\Omega}}, \tag{5.100}$$

où V est l'espace fonctionnel sur lequel est posé le problème continu. Sous l'hypothèse (5.98), on montre que le rapport $\frac{\omega_a}{\alpha_a}$ est d'ordre $\frac{\lambda}{\mu}$. En reprenant l'analyse de convergence présentée dans la section 5.2.2, on obtient l'estimation d'erreur

$$\|u - u_h\|_{1,\Omega} \leqslant c\frac{\lambda}{\mu} h^k |u|_{k+1,\Omega}, \tag{5.101}$$

avec une constante c indépendante de h et du rapport $\frac{\lambda}{\mu}$. Cette estimation montre que le maillage doit être suffisamment fin pour que l'erreur soit contrôlée. La figure 5.1 illustre le phénomène de perte de coercivité dans l'approximation par éléments finis ($k = 1$) des déformations d'un barreau élastique percé de deux trous. Lorsque le rapport $\frac{\lambda}{\mu}$ est de l'ordre de l'unité, le maillage considéré conduit à des résultats acceptables. Ce n'est plus le cas lorsque $\frac{\lambda}{\mu} = 10^3$.

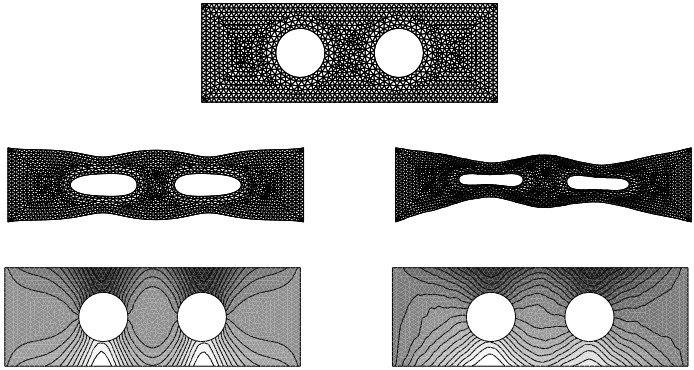


Figure 5.1 – Déformations d'un barreau élastique percé de deux trous ; en haut : maillage et configuration de référence ; au milieu : configurations déformées ; en bas : isovaleurs du déplacement vertical u_2 ; colonne de gauche : $\frac{\lambda}{\mu} = 1$; colonne de droite : $\frac{\lambda}{\mu} = 10^3$.

5.3 Complément : approximation spectrale

Soit Ω un domaine de $\mathbb{R}^d$. On considère le *problème spectral* suivant :

$$\left\{ \begin{array}{l} \text{Chercher } \psi \in H_0^1(\Omega) \setminus \{0\} \text{ et } \lambda \in \mathbb{R} \text{ tels que} \\ -\Delta\psi = \lambda\psi \quad \text{dans } \Omega, \end{array} \right. \tag{5.102}$$

dont la formulation faible consiste à

$$\left\{ \begin{array}{l} \text{Chercher } \psi \in H_0^1(\Omega) \setminus \{0\} \text{ et } \lambda \in \mathbb{R} \text{ tels que} \\ \int_\Omega \nabla\psi\cdot\nabla w = \lambda \int_\Omega \psi w, \quad \forall w \in H_0^1(\Omega). \end{array} \right. \tag{5.103}$$

Lorsque le couple $\{\lambda, \psi\}$ est solution de (5.103), on dit que λ est une *valeur propre du Laplacien* (avec conditions aux limites de Dirichlet homogènes) et que la fonction ψ est une *fonction propre du Laplacien*. On peut montrer (voir Yosida [**78**, p. 284] ou Brezis [**21**, p. 192]) que le problème (5.103) admet

une infinité dénombrable de solutions $\{\lambda_n, \psi_n\}_{n\geqslant 1}$ et que ces solutions sont telles que :

(i) $\{\lambda_n\}_{n\geqslant 1}$ est une suite croissante positive telle que $\lambda_n \to +\infty$;

(ii) $\{\psi_n\}_{n\geqslant 1}$ forme une base hilbertienne orthonormée de $L^2(\Omega)$.

En dimension 1, les valeurs propres du Laplacien peuvent se calculer facilement. Par exemple, pour $\Omega =]0,1[$, ces valeurs propres sont les réels $\lambda_n = n^2\pi^2$ pour $n \geqslant 1$, et les fonctions propres sont les fonctions $\psi_n(x) = \sin(n\pi x)$ pour $n \geqslant 1$. Ces fonctions deviennent de plus en plus oscillantes à mesure que n croît. En dimension $d \geqslant 2$, on dispose d'une formule analytique pour les valeurs propres du Laplacien uniquement lorsque le domaine Ω a une forme géométrique simple. Par exemple, si celui-ci est un hypercube, les valeurs propres s'écrivent comme la somme de d valeurs propres du Laplacien unidimensionnel. Dans le cas général, les valeurs propres et fonctions propres du Lapalcien doivent être approchées, par exemple en utilisant une méthode d'éléments finis.

Pour simplifier, on suppose que Ω est un polygone ou un polyèdre et on considère une famille $\{\mathcal{T}_h\}_{h>0}$ de maillages conformes de Ω. Pour tout $h > 0$, on construit un espace d'approximation H^1-conforme, V_h, en utilisant le maillage $\mathcal{T}_h$ et un élément fini de Lagrange de degré $k \geqslant 1$. Le problème spectral approché est le suivant :

$$\left\{\begin{array}{l} \text{Chercher } \psi_h \in V_h \setminus \{0\} \text{ et } \lambda_h \in \mathbb{R} \text{ tels que} \\ \int_\Omega \nabla\psi_h\cdot\nabla w_h = \lambda_h \int_\Omega \psi_h w_h, \quad \forall w_h \in V_h. \end{array}\right. \tag{5.104}$$

Soit $\{\varphi_1, \ldots, \varphi_N\}$ une base de V_h où $N = \dim V_h$. On désigne par $\Psi_h \in \mathbb{R}^N$ le vecteur formé par les composantes de ψ_h dans cette base. Le problème (5.104) consiste à

$$\left\{\begin{array}{l} \text{Chercher } \Psi_h \in \mathbb{R}^N \setminus \{0\} \text{ et } \lambda_h \in \mathbb{R} \text{ tels que} \\ \mathcal{A}\Psi_h = \lambda_h \mathcal{M}\Psi_h, \end{array}\right. \tag{5.105}$$

où $\mathcal{A}$ est la *matrice de rigidité* et $\mathcal{M}$ la *matrice de masse*. Ces deux matrices ont pour coefficients

$$\mathcal{A}_{ij} = \int_\Omega \nabla\varphi_i\cdot\nabla\varphi_j \qquad \text{et} \qquad \mathcal{M}_{ij} = \int_\Omega \varphi_i\varphi_j. \tag{5.106}$$

Les matrices $\mathcal{A}$ et $\mathcal{M}$ sont clairement symétriques et définies positives. Le problème spectral approché (5.104) admet donc N solutions $\{\lambda_{hn}, \psi_{hn}\}_{1\leqslant n\leqslant N}$. Les fonctions propres approchées peuvent être choisies telles que $\{\psi_{h1}, \ldots, \psi_{hN}\}$ soit une base orthonormée de V_h et telles que $\lambda_{h1} \leqslant \ldots \leqslant \lambda_{hN}$.

On fixe un entier $m \geqslant 1$ et on suppose que le maillage est suffisamment fin pour que $m \leqslant N$. On pose $V_m = \text{vect}\{\psi_1, \ldots, \psi_m\}$ et on note S_m la sphère unité de V_m dans $L^2(\Omega)$. On introduit le *projecteur elliptique* $\Pi_h : H_0^1(\Omega) \to V_h$ tel que pour tout $v \in H_0^1(\Omega)$ et pour tout $v_h \in V_h$,

$$\int_\Omega \nabla(\Pi_h v - v) \cdot \nabla v_h = 0. \tag{5.107}$$

Proposition 5.19. *Soit $m \geqslant 1$ et $N \geqslant m$. On suppose qu'il existe un entier $k \geqslant 1$ et une constante $c_1(m)$ tels que*

$$\inf_{v\in S_m} \left(\|\Pi_h v - v\|_{0,\Omega} + h\|\Pi_h v - v\|_{1,\Omega}\right) \leqslant c_1(m)h^{k+1}. \tag{5.108}$$

Alors, il existe des constantes $c_2(m)$, $c_3(m)$ et $c_4(m)$, indépendantes de h mais qui explosent lorsque $m \to +\infty$, telles que, si h est suffisamment petit, on a

$$\lambda_m \leqslant \lambda_{hm} \leqslant \lambda_m + c_2(m)\, h^{2k}\lambda_m^2. \tag{5.109}$$

De plus, si la valeur propre λ_m est simple, on a

$$\|\psi_m - \psi_{hm}\|_{1,\Omega} \leqslant c_3(m)\, h^k \lambda_m, \tag{5.110}$$

$$\|\psi_m - \psi_{hm}\|_{0,\Omega} \leqslant c_4(m)\, h^{k+1} \lambda_m. \tag{5.111}$$

Si la valeur propre λ_m est multiple, ψ_m peut être choisi de sorte que les estimations (5.110)–(5.111) *sont satisfaites.*

Par exemple, on considère un maillage uniforme de $\Omega =]0,1[$ de pas $h = \frac{1}{N+1}$. Les matrices de rigidité et de masse sont tridiagonales et telles que

$$\mathcal{A} = \frac{1}{h}\text{tridiag}(-1, 2, -1) \qquad \text{et} \qquad \mathcal{M} = \frac{h}{6}\text{tridiag}(1, 4, 1), \tag{5.112}$$

si bien que les valeurs propres du problème spectral approché sont

$$\lambda_{hn} = \frac{6}{h^2}\left(\frac{1 - \cos(n\pi h)}{2 + \cos(n\pi h)}\right), \qquad n \in \{1, \ldots, N\}. \tag{5.113}$$

La figure 5.2 présente les N premières valeurs propres exactes et les N valeurs propres approchées pour $N = 10$ et $N = 100$. On observe que :

(i) les valeurs propres sont approchées par excès, en accord avec l'estimation (5.109) ;

(ii) à N fixé (donc à h fixé), seules les premières valeurs propres sont approchées avec une précision satisfaisante.

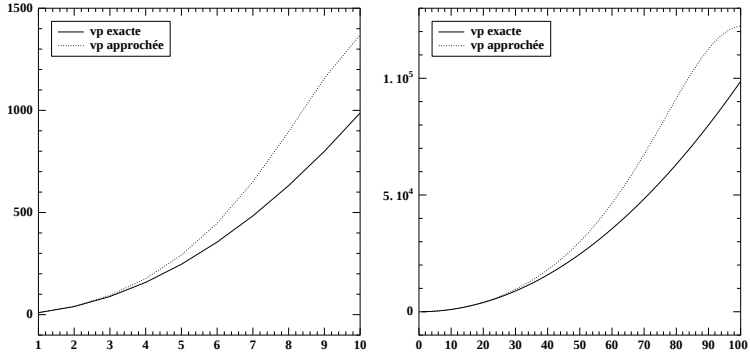


Figure 5.2 – Valeurs propres exactes (trait plein) et approchées par éléments finis (trait pointillé) pour le Laplacien sur l'intervalle $\Omega =]0, 1[$; à gauche : $N = 10$, à droite : $N = 100$.

6 • ÉLÉMENTS FINIS MIXTES

On considère un problème modèle qui s'exprime sous la forme d'un système d'équations aux dérivées partielles où interviennent plusieurs fonctions inconnues qui ne jouent pas le même rôle mathématique et physique. Par exemple, dans le problème de Stokes,

$$-\Delta u + \nabla p = f \quad \text{dans } \Omega, \tag{6.1}$$

$$\nabla \cdot u = g \quad \text{dans } \Omega, \tag{6.2}$$

le champ $u : \Omega \to \mathbb{R}^d$ représente une vitesse et la fonction $p : \Omega \to \mathbb{R}$ une pression. Dans le problème de Darcy,

$$\sigma + \nabla u = f \quad \text{dans } \Omega, \tag{6.3}$$

$$\nabla \cdot \sigma = g \quad \text{dans } \Omega, \tag{6.4}$$

le champ $\sigma : \Omega \to \mathbb{R}^d$ représente une vitesse et la fonction $u : \Omega \to \mathbb{R}$ une charge hydraulique. En général, le caractère bien posé de ces problèmes résulte d'une condition inf-sup qui, comme on l'a vu dans la section 2.3, n'est pas automatiquement transférée au niveau discret. En pratique, cela veut dire que pour que l'approximation par éléments finis soit bien posée, il faut que les espaces d'approximation pour les deux fonctions inconnues satisfassent une *condition de compatibilité* donnant lieu à une *condition inf-sup discrète*. Dans ce cas, on parle *d'éléments finis mixtes*.

Ce chapitre est organisé comme suit. On présente d'abord quelques résultats mathématiques pour l'analyse et l'approximation par éléments finis de problèmes de type point selle. Même si ces problèmes ne recouvrent pas tous les cas rencontrés dans ce chapitre, ils sont historiquement importants car ils ont

fourni le premier cadre théorique pour l'analyse de la méthode des éléments finis ne reposant pas sur le lemme de Lax–Milgram et la notion de coercivité. Puis, on présente divers exemples d'éléments finis mixtes, d'une part pour le problème de Stokes (6.1)–(6.2) et d'autre part pour le problème de Darcy (6.3)–(6.4).

6.1 Problèmes de type point selle

Étant donné deux espaces de Hilbert X et M et deux formes linéaires $f \in X'$ et $g \in M'$, on considère deux formes bilinéaires $a \in \mathcal{L}(X \times X; \mathbb{R})$ et $b \in \mathcal{L}(X \times M; \mathbb{R})$. On s'intéresse au problème abstrait suivant :

$$\left\{ \begin{array}{l} \text{Chercher } (u,p) \in X \times M \text{ tel que} \\ a(u,v) + b(v,p) = f(v), \quad \forall v \in X, \\ b(u,q) \qquad\quad = g(q), \quad \forall q \in M. \end{array} \right. \tag{6.5}$$

Définition 6.1. *Si la forme bilinéaire a est symétrique et positive sur $X \times X$, on dit que* (6.5) *est un problème de type* point selle.

Le problème (6.5) a une structure très particulière puisque :

(i) l'espace solution est le même que l'espace test ;

(ii) la fonction inconnue p n'intervient pas dans la deuxième équation ;

(iii) les fonctions inconnues u et p sont couplées via la même forme bilinéaire dans la première et la deuxième équation.

L'objet de cette section est de présenter les principaux résultats relatifs au caractère bien posé de (6.5) et à son approximation par éléments finis mixtes.

6.1.1 Caractère bien posé

Il est clair que le problème (6.5) est un cas particulier du problème (2.1) dont l'analyse mathématique est présentée dans la section 2.1. En effet, en posant $V = W = X \times M$ et en introduisant la forme bilinéaire $c \in \mathcal{L}(V \times V; \mathbb{R})$ telle que

$$c((u,p),(v,q)) = a(u,v) + b(v,p) + b(u,q), \tag{6.6}$$

et la forme linéaire $h \in \mathcal{L}(V;\mathbb{R})$ telle que

$$h(v,q) = f(v) + g(q), \tag{6.7}$$

le problème (6.5) équivaut à

$$\begin{cases} \text{Chercher } (u,p) \in V \text{ tel que} \\ c((u,p),(v,q)) = h(v,q), \quad \forall (v,q) \in V. \end{cases} \tag{6.8}$$

Ainsi, sur le plan mathématique, tout est dit : le problème (6.5) est bien posé si et seulement si la forme bilinéaire c satisfait les conditions (BNB1) et (BNB2) du théorème BNB.

Lorsque la forme bilinéaire a est coercive, les conditions (BNB1) et (BNB2) sur la forme bilinéaire c peuvent se formuler de manière relativement simple.

Théorème 6.2. *On suppose que la forme bilinéaire a est* coercive *sur X. Alors, le problème* (6.5) *est bien posé si et seulement si la forme bilinéaire b satisfait la* condition inf-sup *suivante : il existe* $\beta > 0$ *tel que*

$$\exists \beta > 0, \qquad \inf_{q \in M} \sup_{v \in X} \frac{b(v,q)}{\|v\|_X \|q\|_M} \geqslant \beta. \tag{6.9}$$

Afin de justifier la terminologie introduite dans la définition 6.1, on introduit le *Lagrangien* $l \in \mathcal{L}(X \times M;\mathbb{R})$ défini par

$$l(v,q) = \tfrac{1}{2} a(v,v) + b(v,q) - f(v) - g(q), \tag{6.10}$$

et la notion de point selle ; voir la figure 6.1.

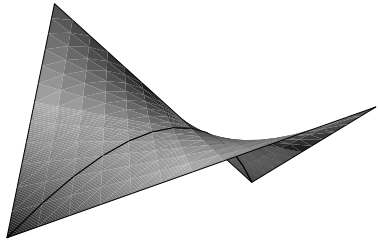


Figure 6.1 – Graphe d'une fonctionnelle présentant un point selle.

Définition 6.3 (Point selle). *On dit que le point* $(u, p) \in X \times M$ *est un* point selle *de l si*

$$\forall (v, q) \in X \times M, \qquad l(u, q) \leqslant l(u, p) \leqslant l(v, p). \tag{6.11}$$

Proposition 6.4. *Sous les hypothèses du théorème* 6.2*, le Lagrangien l défini par* (6.10) *admet un unique point selle. De plus, ce point selle est également la solution unique du problème* (6.5).

6.1.2 Approximation par éléments finis mixtes

On se restreint à une approximation du problème (6.5) par une méthode de Galerkin *standard* dans un cadre *conforme*. Étant donné deux espaces d'approximation $X_h \subset X$ et $M_h \subset M$, on considère le problème approché suivant :

$$\left\{ \begin{array}{l} \text{Chercher } (u_h, p_h) \in X_h \times M_h \text{ tel que} \\ a(u_h, v_h) + b(v_h, p_h) = f(v_h), \quad \forall v_h \in X_h, \\ b(u_h, q_h) \qquad\qquad = g(q_h), \quad \forall q_h \in M_h. \end{array} \right. \tag{6.12}$$

Théorème 6.5. *On suppose que la forme bilinéaire a est coercive sur X. On suppose que* $X_h \subset X$ *et* $M_h \subset M$*. Alors, le problème* (6.12) *est bien posé si et seulement si la* condition inf-sup discrète *suivante est satisfaite : il existe* $\beta_h > 0$ *tel que*

$$\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X \|q_h\|_M} \geqslant \beta_h. \tag{6.13}$$

On retiendra que pour que le problème discret (6.12) soit bien posé, il faut que les éléments finis utilisés pour construire les espaces d'approximation X_h et M_h satisfassent la *condition de compatibilité* (6.13). Dans la littérature, cette condition est souvent appelée la *condition de Babuška–Brezzi* ou la *condition de Ladyshenskaya–Babuška–Brezzi*.

On s'intéresse maintenant à l'analyse d'erreur pour le problème approché (6.12). Le résultat suivant généralise le lemme de Céa 2.12 aux problèmes de type point selle.

Lemme 6.6. *Sous les hypothèses du théorème* 6.5*, la solution unique* (u_h, p_h) *de* (6.12) *satisfait les estimations suivantes :*

$$\|u - u_h\|_X \leqslant c_{1h} \inf_{v_h \in X_h} \|u - v_h\|_X + c_2 \inf_{q_h \in M_h} \|p - q_h\|_M, \tag{6.14}$$

$$\|p - p_h\|_M \leqslant c_{3h} \inf_{v_h \in X_h} \|u - v_h\|_X + c_{4h} \inf_{q_h \in M_h} \|p - q_h\|_M, \tag{6.15}$$

avec $c_{1h} = (1+\frac{\|a\|_{X,X}}{\alpha})(1+\frac{\|b\|_{X,M}}{\beta_h})$, α *est la constante de coercivité de a sur X,* $c_2 = \frac{\|b\|_{X,M}}{\alpha}$, $c_{3h} = c_{1h}\frac{\|a\|_{X,X}}{\beta_h}$ *et* $c_{4h} = 1 + \frac{\|b\|_{X,M}}{\beta_h} + c_2\frac{\|a\|_{X,X}}{\beta_h}$.

On observera que les constantes c_{1h}, c_{3h} et c_{4h} sont d'autant plus grandes que la constante β_h dans la condition inf-sup discrète (6.13) est petite.

6.1.3 Le système linéaire

On s'intéresse à la version matricielle du problème discret (6.12). On pose $N_u = \dim X_h$ et $N_p = \dim M_h$. Soit $\{\phi_i\}_{1\leqslant i\leqslant N_u}$ une base de X_h et $\{\psi_k\}_{1\leqslant k\leqslant N_p}$ une base de M_h. Pour tout $u_h = \sum_{i=1}^{N_u} u_i\phi_i$ dans X_h et pour tout $p_h = \sum_{k=1}^{N_p} p_k\psi_k$ dans M_h, on introduit les vecteurs $U \in \mathbb{R}^{N_u}$ et $P \in \mathbb{R}^{N_p}$ formés par les composantes de u_h et p_h dans ces bases, respectivement. On a donc $U = (u_1, \ldots, u_{N_u})^T$ et $P = (p_1, \ldots, p_{N_p})^T$. Le problème discret (6.12) est équivalent à chercher la solution (U, P) du système linéaire suivant :

$$\left[\begin{array}{c:c} \mathcal{A} & \mathcal{B}^T \\ \hdashline \mathcal{B} & 0 \end{array}\right] \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ G \end{bmatrix}, \tag{6.16}$$

où les matrices $\mathcal{A} \in \mathbb{R}^{N_u,N_u}$ et $\mathcal{B} \in \mathbb{R}^{N_p,N_u}$ ont pour coefficients

$$\mathcal{A}_{ij} = a(\phi_j, \phi_i) \qquad \text{et} \qquad \mathcal{B}_{ki} = b(\phi_i, \psi_k), \tag{6.17}$$

et où les vecteurs $F \in \mathbb{R}^{N_u}$ et $G \in \mathbb{R}^{N_p}$ ont pour composantes $F_i = f(\phi_i)$ et $G_k = g(\psi_k)$. On observe que :

(i) puisque la forme bilinéaire a est symétrique et coercive, la matrice $\mathcal{A}$ est symétrique définie positive ;

(ii) la matrice du système linéaire (6.16) est symétrique mais elle n'est pas positive ;

(iii) la condition inf-sup discrète (6.13) est équivalente au fait que la matrice $\mathcal{B}$ est de rang maximal (ou, en termes équivalents, que $\mathrm{Ker}(\mathcal{B}^T) = \{0\}$) ;

(iv) d'après le théorème 6.5, la matrice du système linéaire (6.16) est inversible.

Les contre-exemples où la condition de compatibilité (6.13) n'est pas vérifiée, et où, donc, la matrice $\mathcal{B}$ n'est pas de rang maximal, relèvent en général de l'une des deux situations suivantes :

(i) $\dim M_h > \dim X_h$: l'espace M_h est trop grand pour que la matrice $\mathcal{B}$ soit de rang maximal ; en général, la matrice $\mathcal{B}$ est injective ($\mathrm{Ker}(\mathcal{B}) = \{0\}$) : on parle de *vérouillage* ;

(ii) il existe un vecteur $Q^* \neq 0$ dans $\mathrm{Ker}(\mathcal{B}^T)$. Le champ discret $q_h^* = \sum_{k=1}^{N_p} Q_k^* \psi_k$ dans M_h est appelé un *mode parasite*. Par construction, ce champ est tel que $b(v_h, q_h^*) = 0$ pour tout $v_h \in X_h$.

Des exemples où ces situations se produisent sont présentés dans la section 6.2.3 pour l'approximation du problème de Stokes.

Remarque 6.7

Des considérations matricielles élémentaires permettent de montrer directement que si $\mathcal{A}$ est définie positive et si $\mathcal{B}$ est de rang maximal, la matrice du système linéaire (6.16) est inversible. Pour cela, on montre simplement que cette matrice est injective. Soit en effet (U^*, P^*) un vecteur du noyau de cette matrice. On a donc $\mathcal{A}U^* + \mathcal{B}^T P^* = 0$ et $\mathcal{B}U^* = 0$. La matrice $\mathcal{A}$ étant inversible, on obtient $\mathcal{B}\mathcal{A}^{-1}\mathcal{B}^T P^* = 0$. En prenant le produit scalaire avec P^* et en utilisant le fait que la matrice $\mathcal{A}^{-1}$ est définie positive, on déduit que $\mathcal{B}^T P^* = 0$, puis que $P^* = 0$ et, enfin, que $U^* = 0$, ce qui termine la preuve.

La matrice

$$\mathcal{U} = \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^T, \tag{6.18}$$

porte le nom de *matrice d'Uzawa*. Cette matrice est clairement symétrique et positive ; de plus, le raisonnement développé dans la remarque 6.7 montre qu'elle est *définie positive*. La matrice d'Uzawa apparaît naturellement lorsqu'on élimine la vitesse U dans le système (6.16), ce qui conduit au système linéaire en P suivant :

$$\mathcal{U}P = \mathcal{B}\mathcal{A}^{-1}F - G. \tag{6.19}$$

L'inversion du système (6.19) (par une méthode itérative) est une des approches possibles pour résoudre le système (6.16). Une autre méthode consiste à utiliser une technique dite de compressibilité artificielle ; voir la section 6.4.

6.2 Éléments finis mixtes pour le problème de Stokes

Cette section passe en revue quelques contre-exemples et de nombreux exemples d'éléments finis mixtes pour l'approximation du problème de Stokes (6.1)–(6.2). Pour simplifier, on se restreint à des conditions aux limites de Dirichlet homogène en la vitesse ; voir, par exemple, [**38**, p. 179] pour d'autres conditions aux limites. Par ailleurs, on suppose que les données f et g sont dans $[L^2(\Omega)]^d$ et $L^2(\Omega)$, respectivement, et que $\int_\Omega g = 0$.[1]

Le problème (6.1)–(6.2) intervient dans la modélisation des écoulements incompressibles à faible nombre de Reynolds. L'équation (6.1) exprime la conservation de la quantité de mouvement (f représente une densité volumique d'efforts extérieurs) et l'équation (6.2) exprime la conservation de la masse (g représente les sources et puits de masse dans l'écoulement). Lorsque $g = 0$, le champ u est à divergence nulle ; on dit que u est *solénoïdal*.

6.2.1 Formulation faible et caractère bien posé

Afin d'écrire le problème de Stokes sous forme faible, on introduit l'espace fonctionnel

$$L^2_*(\Omega) = \left\{ q \in L^2(\Omega) \; ; \; \int_\Omega q = 0 \right\}. \tag{6.20}$$

1. Cette dernière condition s'obtient en intégrant (6.2) sur Ω et en utilisant le théorème de la divergence, ce qui conduit à

$$\int_\Omega g = \int_\Omega \nabla \cdot u = \int_{\partial\Omega} u \cdot n = 0,$$

car on impose une condition aux limites de Dirichlet homogène sur u. La condition $\int_\Omega g = 0$ est donc une condition nécessaire à l'existence d'une solution (u, p) pour le problème de Stokes avec de telles conditions aux limites.

Le problème de Stokes s'écrit sous la forme faible suivante :

$$\left\{\begin{array}{l} \text{Chercher } (u,p)\in [H_0^1(\Omega)]^d\times L^2_*(\Omega) \text{ tel que}\\ \quad \int_\Omega \nabla u{:}\nabla v - \int_\Omega p\nabla{\cdot}v = \int_\Omega f{\cdot}v, \quad \forall v\in [H_0^1(\Omega)]^d,\\ \quad -\int_\Omega q\nabla{\cdot}u = -\int_\Omega gq, \quad \forall q\in L^2_*(\Omega). \end{array}\right. \tag{6.21}$$

On retrouve bien le problème abstrait (6.5) avec $X = [H_0^1(\Omega)]^d$ et $M = L^2_*(\Omega)$, les formes bilinéaires[1]

$$a(u,v)=\int_\Omega \nabla u{:}\nabla v \qquad \text{et} \qquad b(v,p)=-\int_\Omega p\nabla{\cdot}v, \tag{6.22}$$

et les formes linéaires $f(v) = \int_\Omega f{\cdot}v$ et $g(q) = -\int_\Omega gq$. On observera que dans le problème de Stokes (6.1)–(6.2), la pression n'est déterminée qu'à une constante additive près et que dans la formulation faible (6.21), on a convenu de chercher le champ de pression de moyenne nulle sur Ω. De plus, il est inutile de tester la deuxième équation dans (6.21) par les constantes puisque cela conduit à l'équation triviale $\int_\Omega \nabla{\cdot}u = 0 = \int_\Omega g$. On élimine les fonctions tests constantes dans la deuxième équation en prenant les fonctions tests dans $L^2_*(\Omega)$.

Le caractère bien posé du problème de Stokes (6.21) repose de manière fondamentale sur le résultat suivant ; voir Girault et Raviart [**46**, p. 22], Galdi [**42**, lemme 3.1, chap. III] ou Ern et Guermond [**36**, p. 423].

Lemme 6.8. *Soit Ω un domaine de $\mathbb{R}^d$ avec $d\geqslant 2$. Alors, l'opérateur*

$$\nabla{\cdot} : [H_0^1(\Omega)]^d \to L^2_*(\Omega), \tag{6.23}$$

est surjectif.

On déduit du lemme 6.8 et du lemme A.3 qu'il existe une constante $\beta>0$ telle que pour toute fonction $q\in L^2_*(\Omega)$, il existe un champ de vecteurs $v\in[H_0^1(\Omega)]^d$ tel que $\nabla{\cdot}v = q$ et $\|v\|_{1,\Omega}\leqslant \frac{1}{\beta}\|q\|_{0,\Omega}$. En d'autres termes, on a

$$\inf_{q\in L^2_*(\Omega)}\sup_{v\in[H_0^1(\Omega)]^d}\frac{\int_\Omega q\nabla{\cdot}v}{\|v\|_{1,\Omega}\|q\|_{0,\Omega}}\geqslant\beta. \tag{6.24}$$

1. Les champs u et v étant à valeurs vectorielles, on a $a(u,v)=\sum_{i,j=1}^d\int_\Omega \partial_j u_i\partial_j v_i$.

Par conséquent, la *condition inf-sup* (6.9) est satisfaite. Par ailleurs, la forme bilinéaire a est coercive sur $[H_0^1(\Omega)]^d$ grâce à l'inégalité de Poincaré (5.5) que l'on applique composante par composante. Le théorème 6.2 permet de conclure quant au caractère bien posé de (6.21).

6.2.2 Approximation par éléments finis mixtes

On s'intéresse maintenant à une approximation conforme du problème (6.21) par éléments finis mixtes dans le cadre de la méthode de Galerkin standard ; voir la section 6.1.2. Étant donné deux espaces d'approximation $X_h \subset [H_0^1(\Omega)]^d$ et $M_h \subset L_*^2(\Omega)$, on considère le problème approché suivant :

$$\left\{ \begin{array}{l} \text{Chercher } (u_h, p_h) \in X_h \times M_h \text{ tel que} \\ \quad \int_\Omega \nabla u_h : \nabla v_h - \int_\Omega p_h \nabla \cdot v_h = \int_\Omega f \cdot v_h, \quad \forall v_h \in X_h, \\ \quad - \int_\Omega q_h \nabla \cdot u_h = - \int_\Omega g q_h, \quad \forall q_h \in M_h. \end{array} \right. \tag{6.25}$$

D'après le théorème 6.5, ce problème est bien posé si et seulement si les espaces X_h et M_h sont tels qu'il existe une constante $\beta_h > 0$ telle que

$$\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{\int_\Omega q_h \nabla \cdot v_h}{\|q_h\|_{0,\Omega} \|v_h\|_{1,\Omega}} \geqslant \beta_h. \tag{6.26}$$

Lorsque cette condition est satisfaite avec une constante β indépendante de h, on dit que les espaces X_h et M_h satisfont (6.26) *uniformément en h*.

Lorsque la *condition de compatibilité* (6.26) est satisfaite, le lemme 6.6 permet d'obtenir une estimation d'erreur pour la vitesse en norme H^1 et une estimation d'erreur pour la pression en norme L^2. De plus, afin d'obtenir des estimations d'erreur pour la vitesse en norme L^2, on utilise la notion suivante.

Définition 6.9 (Problème de Stokes régularisant). *On dit que le problème de Stokes* (6.21) *est* régularisant *s'il existe une constante c_S telle que pour tout $f \in [L^2(\Omega)]^d$, la solution unique (ξ_u, ζ_p) de* (6.21) *avec $g = 0$ est telle que*

$$\|\xi_u\|_{2,\Omega} + \|\zeta_p\|_{1,\Omega} \leqslant c_S \|f\|_{0,\Omega}. \tag{6.27}$$

Une condition *suffisante* pour que le problème (6.21) soit régularisant est que Ω est un polygone convexe en dimension 2 ou que Ω est un domaine de classe $\mathcal{C}^{1,1}$ dans $\mathbb{R}^d$ avec $d = 2$ ou 3 ; voir, par exemple, Amrouche et Girault [**5**].

Dans la pratique, il n'est pas très commode d'imposer la condition $\int_\Omega q_h = 0$ sur les champs de pression discrets car cela rend difficile la construction d'une base de M_h composée de fonctions dont le support est localisé. On préfère travailler avec un couple d'espaces $X_h \times M_h^*$. L'espace d'approximation en pression M_h^* admet la décomposition $M_h^* = M_h \oplus \mathrm{vect}(Z)$ telle que

(i) le couple d'espaces $X_h \times M_h$ satisfait la condition de compatibilité (6.26) ;

(ii) le vecteur $Z \in \mathbb{R}^{N_p^*}$, où N_p^* désigne la dimension de M_h^*, a toutes ses composantes égales à 1.

La matrice du système linéaire associé au couple d'espaces $X_h \times M_h^*$ est singulière, son noyau étant de dimension 1 et son image étant de co-dimension 1. Plus précisément, le noyau de cette matrice est $\mathbb{R}(0, Z)$ et son image est l'hyperplan $(0, Z)^\perp$. Puisque le membre de droite du système linéaire appartient à cet hyperplan (car $\int_\Omega g = 0$ implique $\sum_{k=1}^{N_p^*} G_k = 0$), le système linéaire admet une infinité de solutions. Si on utilise une méthode itérative pour approcher une des solutions du système linéaire et si la méthode itérative est convergente, on peut soustraire au champ de pression ainsi obtenu une constante de façon à imposer $\int_\Omega p_h = 0$.

6.2.3 Contre-exemples

Cette section présente trois contre-exemples classiques d'éléments finis mixtes pour le problème de Stokes.

- **Élément $\mathbb{P}_1/\mathbb{P}_0$: verrouillage**. On cherche la vitesse discrète dans l'espace X_h construit avec l'élément fini de Lagrange $\mathbb{P}_1$ et la pression dans l'espace M_h constitué des fonctions constantes par morceaux. On a donc $X_h = [P^1_{\mathrm{c},h,0}]^d$ où $P^1_{\mathrm{c},h,0}$ est défini en (3.46) avec $k = 1$ et $M_h = P^0_{\mathrm{td},h}$ où $P^0_{\mathrm{td},h}$ est défini en (4.20) avec $k = 0$.
  Les espaces X_h et M_h ainsi construits ne satisfont pas la condition de compatibilité (6.26). En deux dimensions d'espace sur un domaine Ω qui est simplement connexe, ce fait résulte simplement des relations d'Euler rappelées dans le lemme 3.19. En reprenant les notations du lemme 3.19, on

obtient

$$\dim(X_h) = 2N_{\rm so}^{\rm i} \qquad \text{et} \qquad \dim(M_h) = N_{\rm ma}, \tag{6.28}$$

où $N_{\rm so}^{\rm i} = N_{\rm so} - N_{\rm so}^{\partial}$ désigne le nombre de sommets intérieurs, si bien que

$$\dim(M_h) - \dim(X_h) = N_{\rm ma} - 2N_{\rm so}^{\rm i} = N_{\rm ar}^{\partial} - 2 > 0. \tag{6.29}$$

Un simple argument de dimension montre donc que la matrice $\mathcal{B}$ ne peut pas être de rang maximal. En d'autres termes, l'espace M_h est beaucoup trop riche pour que la condition de compatibilité (6.26) soit satisfaite. Par ailleurs, on a en général $\mathrm{Ker}(\mathcal{B}) = \{0\}$, si bien que le seul champ de vitesse discret u_h^* dont le vecteur composantes U^* satisfait $\mathcal{B}U^* = 0$ est le champ nul.

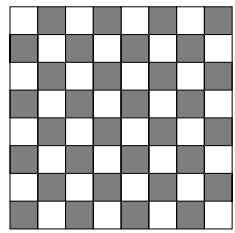


Figure 6.2 – Mode de pression parasite pour les éléments finis mixtes $\mathbb{Q}_1/\mathbb{P}_0$; les mailles grisées correspondent à la valeur $+1$ et les autres mailles à la valeur -1.

- **Élément $\mathbb{Q}_1/\mathbb{P}_0$: mode parasite**. On cherche la vitesse discrète dans l'espace X_h construit avec l'élément fini de Lagrange $\mathbb{Q}_1$ et la pression dans l'espace M_h constitué des fonctions constantes par morceaux. On a donc $X_h = [Q_{{\rm c},h,0}^1]^d$ où $Q_{{\rm c},h,0}^1$ est défini en (3.47) avec $k = 1$ et $M_h = P_{{\rm td},h}^0$. Sur le carré unité maillé avec un maillage uniforme, on peut construire facilement un mode de pression parasite : il s'agit de la fonction q_h^* constante par morceaux qui vaut alternativement $+1$ et -1 comme sur les cases d'un échiquier ; voir la figure 6.2. On vérifie que

$$\forall v_h \in [Q_{{\rm c},h,0}^1]^d, \qquad \int_\Omega q_h^* \nabla\cdot v_h = 0, \tag{6.30}$$

si bien que les espaces X_h et M_h ne satisfont pas la condition de compatibilité (6.26).

- **Élément $\mathbb{P}_1/\mathbb{P}_1$: mode parasite**. On cherche la vitesse discrète dans l'espace X_h construit avec l'élément fini de Lagrange $\mathbb{P}_1$ et la pression dans l'espace M_h construit également avec cet élément fini. On a donc $X_h = [P^1_{c,h,0}]^d$ et $M_h = P^1_{c,h}$.
  Sur le carré unité maillé avec une triangulation structurée et uniforme, on peut construire facilement un mode de pression parasite : il s'agit de la fonction q_h^* qui vaut alternativement $+1$, 0 et -1 sur les sommets du maillage ; voir la figure 6.3. On vérifie que

$$\forall v_h \in [P^1_{c,h,0}]^d, \qquad \int_\Omega q_h^* \nabla\cdot v_h = 0, \tag{6.31}$$

  si bien que les espaces X_h et M_h ne satisfont pas la condition de compatibilité (6.26).

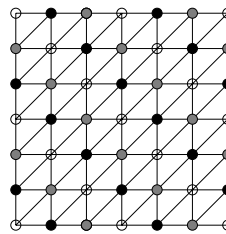


Figure 6.3 – Mode de pression parasite pour les éléments finis mixtes $\mathbb{P}_1/\mathbb{P}_1$; les cercles blancs correspondent à la valeur $+1$, les cercles gris à la valeur 0 et les cercles noirs à la valeur -1.

6.2.4 L'élément mini

La raison pour laquelle l'élément $\mathbb{P}_1/\mathbb{P}_1$ ne conduit pas à la condition inf-sup discrète (6.26) est que l'espace d'approximation des vitesses n'est pas assez riche. L'idée de la construction de l'élément mini est de conserver une approximation de la pression basée sur l'élément fini de Lagrange $\mathbb{P}_1$ et d'enrichir l'espace d'approximation des vitesses afin qu'il existe un champ de vitesse discret v_h^* tel que $\int_\Omega q_h^* \nabla\cdot v_h^* \neq 0$, où q_h^* est le mode de pression parasite illustré dans la figure 6.3.

On suppose que le domaine Ω est un polygone de $\mathbb{R}^2$ ou un polyèdre de $\mathbb{R}^3$ et on considère une famille régulière $\{\mathcal{T}_h\}_{h>0}$ de maillages affines de Ω constitués de simplexes (triangles en dimension 2 ou tétraèdres en dimension 3). On note $\widehat{K}$ le simplexe de référence.

Définition 6.10 (Fonction bulle). *On dit qu'une fonction* $\widehat{b} : \widehat{K} \to \mathbb{R}$ *est une* fonction bulle *si :*

(i) $\widehat{b} \in H_0^1(\widehat{K})$;

(ii) $0 \leqslant \widehat{b}(\widehat{x}) \leqslant 1$ *pour tout* $\widehat{x} \in \widehat{K}$;

(iii) $\widehat{b}(\widehat{G}) = 1$ *où* $\widehat{G}$ *est le barycentre de* $\widehat{K}$.

Un exemple de fonction bulle est la fonction

$$\widehat{b} = (d+1)^{d+1} \prod_{i=0}^{d} \widehat{\lambda}_i, \tag{6.32}$$

où $(\widehat{\lambda}_0, \ldots, \widehat{\lambda}_d)$ désignent les coordonnées barycentriques sur $\widehat{K}$; voir la partie gauche de la figure 6.4. Un autre exemple consiste à construire $(d+1)$ simplexes dans $\widehat{K}$ en reliant le barycentre de $\widehat{K}$ aux $(d+1)$ sommets de $\widehat{K}$, puis à prendre pour $\widehat{b}$ la fonction continue et affine par morceaux dans $\widehat{K}$ qui s'annule aux $(d+1)$ sommets de $\widehat{K}$ et qui vaut un en son barycentre ; voir la partie droite de la figure 6.4.

Figure 6.4 – Lignes de niveau de deux fonctions bulle sur le triangle de référence.

Soit $\widehat{b}$ une fonction bulle sur $\widehat{K}$. On pose

$$\widehat{P} = [\mathbb{P}_1(\widehat{K}) \oplus \mathrm{vect}(\widehat{b})]^d, \tag{6.33}$$

et on introduit les espaces d'approximation

$$X_h = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^d \,;\, \forall K \in \mathcal{T}_h,\ v_h \circ T_K \in \widehat{P} \,;\, v_h|_{\partial\Omega} = 0\}, \tag{6.34}$$

$$M_h = P^1_{c,h}. \tag{6.35}$$

Lemme 6.11. *Les espaces X_h et $M_h \cap L^2_*(\Omega)$ satisfont la condition de compatibilité* (6.26) *uniformément en h.*

Théorème 6.12. *On suppose que la solution (u, p) du problème de Stokes* (6.21) *est suffisamment régulière, à savoir $u \in [H^2(\Omega)]^d \cap [H^1_0(\Omega)]^d$ et $p \in H^1(\Omega) \cap L^2_*(\Omega)$. Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \leqslant c\, h(\|u\|_{2,\Omega} + \|p\|_{1,\Omega}). \tag{6.36}$$

De plus, si le problème de Stokes est régularisant,

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^2(\|u\|_{2,\Omega} + \|p\|_{1,\Omega}). \tag{6.37}$$

L'idée d'utiliser des fonctions bulle dans l'approximation du problème de Stokes remonte aux travaux de Crouzeix et Raviart [**32**]. L'analyse de l'élément $\mathbb{P}_1$-bulle/$\mathbb{P}_1$ est due aux travaux de Arnold, Brezzi et Fortin [**6**]. Dans la littérature, cet élément est souvent appelé *l'élément mini*.

6.2.5 L'élément de Taylor–Hood et variantes

L'élément mini présente l'avantage de satisfaire la condition inf-sup discrète (6.26). Toutefois, l'estimation d'erreur (6.36) n'est pas optimale dans la mesure où l'espace d'approximation en pression est suffisamment riche pour laisser espérer une convergence de $\|p - p_h\|_{0,\Omega}$ à l'ordre deux en h. Par contre, l'espace d'approximation en vitesse n'est pas suffisamment riche pour que $\|u - u_h\|_{1,\Omega}$ converge à l'ordre deux en h.

L'idée de l'élément de Taylor–Hood consiste à enrichir encore davantage l'espace d'approximation de la vitesse. On cherche la vitesse discrète dans l'espace X_h construit avec l'élément fini de Lagrange $\mathbb{P}_2$ et la pression dans l'espace M_h

construit avec l'élément fini de Lagrange $\mathbb{P}_1$. On a donc

$$X_h = [P^2_{c,h,0}]^d, \tag{6.38}$$

$$M_h = P^1_{c,h}. \tag{6.39}$$

Comme dans la section précédente, on suppose que le domaine Ω est un polygone de $\mathbb{R}^2$ ou un polyèdre de $\mathbb{R}^3$ et on considère une famille régulière $\{\mathcal{T}_h\}_{h>0}$ de maillages affines de Ω constitués de simplexes. On suppose que toutes les mailles ont au moins d arêtes dans Ω (ou, en termes équivalents, toute maille a au plus une arête sur la frontière $\partial\Omega$ en dimension 2 et au plus trois arêtes sur la frontière $\partial\Omega$ en dimension 3).

Lemme 6.13. *Les espaces X_h et $M_h \cap L^2_*(\Omega)$ satisfont la condition de compatibilité* (6.26) *uniformément en h.*

Théorème 6.14. *On suppose que la solution (u, p) du problème de Stokes* (6.21) *est suffisamment régulière, à savoir $u \in [H^3(\Omega)]^d \cap [H^1_0(\Omega)]^d$ et $p \in H^2(\Omega) \cap L^2_*(\Omega)$. Alors, il existe une constante c telle que pour tout h,*

$$\|u - u_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \leqslant c\, h^2(\|u\|_{3,\Omega} + \|p\|_{2,\Omega}). \tag{6.40}$$

De plus, si le problème de Stokes est régularisant,

$$\|u - u_h\|_{0,\Omega} \leqslant c\, h^3(\|u\|_{3,\Omega} + \|p\|_{2,\Omega}). \tag{6.41}$$

Pour les preuves des résultats ci-dessus et des compléments, on pourra consulter Bercovier et Pironneau [**15**], Verfürth [**75**], Girault et Raviart [**46**, p. 176] ou Ern et Guermond [**38**, p. 192]. Dans la littérature, l'élément $\mathbb{P}_2/\mathbb{P}_1$ est souvent appelé élément de Taylor–Hood. Plusieurs variantes de cet élément sont possibles.

- **Éléments $\mathbb{P}_k/\mathbb{P}_{k-1}$ et $\mathbb{Q}_k/\mathbb{Q}_{k-1}$.** On considère un entier $k \geqslant 2$. L'élément $\mathbb{P}_k/\mathbb{P}_{k-1}$ consiste à approcher la vitesse et la pression dans les espaces

$$X_h = [P^k_{c,h,0}]^d, \tag{6.42}$$

$$M_h = P^{k-1}_{c,h}. \tag{6.43}$$

  Une autre possibilité consiste à considérer une famille régulière $\{\mathcal{T}_h\}_{h>0}$ de maillages affines constitués de quadrangles en dimension 2 ou de hexaèdres

en dimension 3 et à approcher la vitesse et la pression dans les espaces

$$X_h = [Q^k_{c,h,0}]^d, \tag{6.44}$$

$$M_h = Q^{k-1}_{c,h}. \tag{6.45}$$

Dans les deux cas, les espaces X_h et $M_h \cap L^2_*(\Omega)$ satisfont la condition de compatibilité (6.26) uniformément en h pour tout $k \geqslant 2$. De plus, pourvu que la solution exacte (u, p) soit suffisamment régulière et que le problème de Stokes soit régularisant, on a l'estimation d'erreur [**22**]

$$\|u - u_h\|_{0,\Omega} + h(\|u - u_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega}) \leqslant c\, h^{k+1}(\|u\|_{k+1,\Omega} + \|p\|_{k,\Omega}). \tag{6.46}$$

- **Éléments $\mathbb{P}_1$-iso-$\mathbb{P}_2/\mathbb{P}_1$ et $\mathbb{Q}_1$-iso-$\mathbb{Q}_2/\mathbb{Q}_1$.** On considère d'abord un maillage $\mathcal{T}_h$ constitué de simplexes. En dimension 2, chaque triangle de $\mathcal{T}_h$ est divisé en 4 sous-triangles en joignant les milieux de ses 3 arêtes. En dimension 3, chaque tétraèdre de $\mathcal{T}_h$ est divisé en 8 sous-tétraèdres en divisant chaque face en 4 sous-triangles comme en dimension 2 puis en joignant les milieux d'une paire d'arêtes dont l'intersection est vide ; voir le tableau 6.1. On note $\mathcal{T}_{\frac{h}{2}}$ le nouveau maillage ainsi construit. L'élément $\mathbb{P}_1$-iso-$\mathbb{P}_2/\mathbb{P}_1$ consiste à approcher la vitesse et la pression dans les espaces

$$X_h = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^d \,;\, \forall K \in \mathcal{T}_{\frac{h}{2}},\, v_h \circ T_K \in [\mathbb{P}_1]^d \,;\, v_h|_{\partial\Omega} = 0\}, \tag{6.47}$$

$$M_h = P^1_{c,h}. \tag{6.48}$$

Dans la littérature, cet élément est souvent appelé $4\mathbb{P}_1/\mathbb{P}_1$ en dimension 2 et $8\mathbb{P}_1/\mathbb{P}_1$ en dimension 3. Les espaces X_h et $M_h \cap L^2_*(\Omega)$ satisfont la condition de compatibilité (6.26) uniformément en h. De plus, pourvu que la solution exacte (u, p) soit suffisamment régulière et que le problème de Stokes soit régularisant, on a l'estimation d'erreur

$$\|u - u_h\|_{0,\Omega} + h(\|u - u_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega}) \leqslant c\, h^2(\|u\|_{2,\Omega} + \|p\|_{1,\Omega}). \tag{6.49}$$

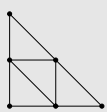
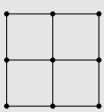
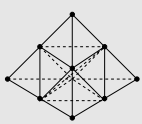
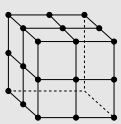
Une construction analogue est possible lorsque le maillage $\mathcal{T}_h$ est constitué de quadrangles en dimension 2 ou de hexaèdres en dimension 3. En dimension 2, chaque quadrangle est divisé en 4 sous-quadrangles en joignant les milieux des arêtes opposées. En dimension 3, chaque hexaèdre est subdivisé en 8 sous-hexaèdres en divisant chaque face en 4 sous-faces comme en dimension 2 puis en joignant les milieux des faces opposées ; voir le tableau 6.1. On note $\mathcal{T}_{\frac{h}{2}}$ le nouveau maillage ainsi construit. L'élément $\mathbb{Q}_1$-iso-$\mathbb{Q}_2/\mathbb{Q}_1$ consiste à approcher la vitesse et la pression dans les espaces

$$X_h = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^d \, ; \, \forall K \in \mathcal{T}_{\frac{h}{2}}, \, v_h \circ T_K \in [\mathbb{Q}_1]^d \, ; \, v_h|_{\partial\Omega} = 0\}, \quad (6.50)$$

$$M_h = Q^1_{c,h}. \quad (6.51)$$

Dans la littérature, cet élément est souvent appelé $4\mathbb{Q}_1/\mathbb{Q}_1$ en dimension 2 et $8\mathbb{Q}_1/\mathbb{Q}_1$ en dimension 3. Les espaces X_h et $M_h \cap L^2_*(\Omega)$ satisfont la condition de compatibilité (6.26) uniformément en h. De plus, pourvu que la solution exacte (u, p) soit suffisamment régulière et que le problème de Stokes soit régularisant, on retrouve l'estimation d'erreur (6.49). L'élément $\mathbb{Q}_1$-iso-$\mathbb{Q}_2/\mathbb{Q}_1$ est souvent utilisé dans les logiciels industriels car il est très simple à programmer.

Tableau 6.1 – Construction de l'espace d'approximation en vitesse pour les éléments $4\mathbb{P}_1/\mathbb{P}_1$ et $4\mathbb{Q}_1/\mathbb{Q}_1$ en dimension 2 et les éléments $8\mathbb{P}_1/\mathbb{P}_1$ et $8\mathbb{Q}_1/\mathbb{Q}_1$ en dimension 3 (seuls les degrés de liberté visibles sont représentés).

dimension 2		dimension 3	
$4\mathbb{P}_1/\mathbb{P}_1$	$4\mathbb{Q}_1/\mathbb{Q}_1$	$8\mathbb{P}_1/\mathbb{P}_1$	$8\mathbb{Q}_1/\mathbb{Q}_1$
			

6.2.6 L'élément Crouzeix–Raviart/$\mathbb{P}_0$

On suppose que le domaine Ω est un polygone de $\mathbb{R}^2$ ou un polyèdre de $\mathbb{R}^3$ et on considère une famille régulière $\{\mathcal{T}_h\}_{h>0}$ de maillages affines de Ω constitués de simplexes. On approche la vitesse dans l'espace X_h construit avec l'élément fini de Crouzeix–Raviart et la pression dans l'espace M_h constitué des fonctions constantes par morceaux. On a donc $X_h = [P^1_{\mathrm{pt},h,0}]^d$ et $M_h = P^0_{\mathrm{td},h}$. On introduit les formes bilinéaires $a_h \in \mathcal{L}(X_h \times X_h; \mathbb{R})$ et $b_h \in \mathcal{L}(X_h \times M_h)$ définies par[1]

$$a_h(v_h, w_h) = \int_\Omega \nabla_h v_h : \nabla_h w_h \qquad \text{et} \qquad b_h(v_h, q_h) = -\int_\Omega q_h \nabla_h \cdot v_h, \quad (6.52)$$

et on considère le problème approché suivant :

$$\left\{ \begin{array}{ll} \text{Chercher } (u_h, p_h) \in X_h \times M_h \text{ tel que} & \\ a_h(u_h, v_h) + b_h(v_h, p_h) = f(v_h), & \forall v_h \in X_h, \\ b_h(u_h, q_h) \qquad\qquad\quad = g(q_h), & \forall q_h \in M_h, \end{array} \right. \quad (6.53)$$

où les formes linéaires f et g sont définies dans la section 6.2.1. On notera que (6.53) réalise une approximation non-conforme (en vitesse) du problème de Stokes du fait que $X_h = [P^1_{\mathrm{pt},h,0}]^d \not\subset [H^1_0(\Omega)]^d$.

Les espaces X_h et M_h définis ci-dessus satisfont la condition inf-sup discrète suivante : il existe $\beta > 0$ tel que pour tout h,

$$\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b_h(v_h, q_h)}{\|v_h\|_{1,h,\Omega} \|q_h\|_{0,\Omega}} \geqslant \beta, \quad (6.54)$$

où la norme $\|\cdot\|_{1,h,\Omega}$ est définie en (6.70). La condition (6.54) implique que le problème discret (6.53) est bien posé. De plus, pourvu que la solution exacte (u, p) soit suffisamment régulière et que le problème de Stokes soit régularisant, on a l'estimation d'erreur

$$\|u - u_h\|_{0,\Omega} + h(\|u - u_h\|_{1,h,\Omega} + \|p - p_h\|_{0,\Omega}) \leqslant c\, h^2 (\|u\|_{2,\Omega} + \|p\|_{1,\Omega}). \quad (6.55)$$

1. Les opérateurs de gradient et de divergence discrets sont définis en (4.21) et (4.22), respectivement.

L'élément Crouzeix–Raviart/$\mathbb{P}_0$ a été introduit dans [**32**]. Un des intérêts de cet élément est que le champ de vitesse approché est tel que

$$\nabla_h \cdot u_h = \Pi_h g. \tag{6.56}$$

En particulier, pour $g = 0$, le champ de vitesse approché est localement solénoïdal (on gardera toutefois à l'esprit que $\nabla_h \cdot u_h = 0$ n'implique pas $\nabla \cdot u_h = 0$ à cause des discontinuités de u_h aux interfaces).

6.3 Éléments finis mixtes pour le problème de Darcy

L'objet de cette section est de présenter quelques exemples d'éléments finis mixtes pour l'approximation du problème de Darcy (6.3)–(6.4). Ce problème intervient dans la modélisation des écoulements stationnaires dans un milieu poreux. L'équation (6.3), en général avec $f = 0$, exprime la loi phénoménologique de Darcy qui stipule que pour de tels écoulements, la vitesse du fluide σ est proportionnelle au gradient de la charge hydraulique u. L'équation (6.4) exprime la conservation de la masse (g représente les sources et puits de masse dans l'écoulement).

Une observation importante est qu'en éliminant σ de (6.4) à l'aide de (6.3), il vient

$$-\Delta u = g - \nabla \cdot f. \tag{6.57}$$

L'inconnue u satisfait donc une équation de Laplace. L'équation (6.57) s'appelle la *formulation primale* du problème de Darcy, alors que le système (6.3)–(6.4) en constitue la *formulation mixte*. L'inconnue u est appelée la *variable primale* et l'inconnue σ le *flux*.

Une première possibilité pour approcher le problème de Darcy consiste à obtenir dans un premier temps une approximation par éléments finis de l'inconnue u en utilisant les techniques d'approximation présentées dans la section 5.1 pour la formulation primale (6.57), puis à reconstruire le flux discret σ_h en prenant le gradient de la solution discrète u_h. Toutefois, dans de nombreuses applications, on souhaite disposer d'une approximation plus précise du flux discret σ_h car celui-ci a vocation à servir, par exemple, pour simuler le

transport de polluants dans le milieu poreux. On préfère donc considérer une approximation de la formulation mixte.

Dans cette section, on considère plusieurs formulations mixtes du problème de Darcy (6.3)–(6.4). Ces formulations diffèrent au niveau de l'espace solution et de l'espace test qu'elles font intervenir. Sur le plan mathématique, toutes ces formulations faibles sont *équivalentes* (pourvu que les données f et g satisfassent des hypothèses de régularité minimales) au sens où elles sont toutes bien posées et leur unique solution (σ, u) est la même. Par contre, chacune d'entre elles conduit à un *problème discret différent* à cause des différentes intégrations par parties qui sont utilisées pour les établir.

Pour simplifier, on se restreint à des conditions aux limites de Dirichlet homogène en u. D'autres conditions aux limites sont possibles, par exemple des conditions portant sur la composante normale du flux ou sur une combinaison linéaire entre la composante normale du flux et la variable primale. On suppose également que les données f et g sont dans $[L^2(\Omega)]^d$ et $L^2(\Omega)$, respectivement, et que le domaine Ω est un polygone de $\mathbb{R}^2$ ou un polyèdre de $\mathbb{R}^3$. On considère une famille régulière $\{\mathcal{T}_h\}_{h>0}$ de maillages affines de Ω constitués de simplexes.

6.3.1 Formulation mixte dans $[L^2(\Omega)]^d \times H_0^1(\Omega)$

On considère une formulation mixte où on prend en compte l'équation de Darcy $\sigma + \nabla u = f$ pour chercher *a priori* le flux dans $[L^2(\Omega)]^d$ et la variable primale dans $H^1(\Omega)$. Les fonctions de $H^1(\Omega)$ étant suffisamment régulières pour admettre une trace sur la frontière $\partial\Omega$, on prend en compte la condition aux limites en cherchant la variable primale dans $H_0^1(\Omega)$. On obtient la formulation mixte suivante du problème de Darcy :

$$\left\{\begin{array}{l} \text{Chercher } (\sigma, u) \in [L^2(\Omega)]^d \times H_0^1(\Omega) \text{ tel que} \\ \int_\Omega \sigma\cdot\tau + \int_\Omega \tau\cdot\nabla u = \int_\Omega f\cdot\tau, \quad \forall \tau \in [L^2(\Omega)]^d, \\ \int_\Omega \sigma\cdot\nabla v = -\int_\Omega g v, \quad \forall v \in H_0^1(\Omega). \end{array}\right. \tag{6.58}$$

On retrouve le problème abstrait (6.5) avec $X = [L^2(\Omega)]^d$ et $M = H_0^1(\Omega)$, les formes bilinéaires

$$a(\sigma, \tau) = \int_\Omega \sigma\cdot\tau \qquad \text{et} \qquad b(\tau, u) = \int_\Omega \tau\cdot\nabla u, \tag{6.59}$$

et les formes linéaires $f(\tau) = \int_\Omega f\cdot\tau$ et $g(v) = -\int_\Omega gv$. Il est clair que

$$\inf_{v\in H_0^1(\Omega)} \sup_{\tau\in[L^2(\Omega)]^d} \frac{\int_\Omega \tau\cdot\nabla v}{\|\tau\|_{0,\Omega}\|v\|_{1,\Omega}} \geqslant \inf_{v\in H_0^1(\Omega)} \frac{\|\nabla v\|_{0,\Omega}}{\|v\|_{1,\Omega}} \geqslant (1+\ell_\Omega^2)^{-\frac{1}{2}}, \quad (6.60)$$

où ℓ_Ω est la constante intervenant dans l'inégalité de Poincaré (5.5). Par conséquent, le problème (6.58) est bien posé.

Afin d'approcher le problème (6.58) dans un cadre *conforme* par des éléments finis mixtes, on considère les espaces d'approximation $S_h = [P^0_{\mathrm{td},h}]^d$ et $V_h = P^1_{\mathrm{c},h,0}$. On approche donc le flux par une fonction constante par morceaux et la variable primale par une fonction continue et affine par morceaux. Ceci conduit au problème discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } (\sigma_h, u_h)\in S_h\times V_h \text{ tel que} \\ \int_\Omega \sigma_h\cdot\tau_h + \int_\Omega \tau_h\cdot\nabla u_h = \int_\Omega f\cdot\tau_h, \quad \forall \tau_h\in S_h, \\ \int_\Omega \sigma_h\cdot\nabla v_h = -\int_\Omega g v_h, \quad \forall v_h\in V_h. \end{array}\right. \quad (6.61)$$

Puisque $\nabla v_h\in S_h$ pour tout $v_h\in V_h$, on a

$$\inf_{v_h\in V_h}\sup_{\tau_h\in S_h} \frac{\int_\Omega \tau_h\cdot\nabla v_h}{\|\tau_h\|_{0,\Omega}\|v_h\|_{1,\Omega}} \geqslant \inf_{v_h\in V_h} \frac{\|\nabla v_h\|_{0,\Omega}}{\|v_h\|_{1,\Omega}} \geqslant (1+\ell_\Omega^2)^{-\frac{1}{2}}, \quad (6.62)$$

si bien que le problème discret (6.61) est bien posé. De plus, si la solution exacte (σ, u) est suffisamment régulière, on a l'estimation d'erreur

$$\|\sigma-\sigma_h\|_{0,\Omega} + \|u-u_h\|_{1,\Omega} \leqslant c\, h(\|\sigma\|_{1,\Omega} + \|u\|_{2,\Omega}). \quad (6.63)$$

En prenant $\tau_h = \nabla v_h$ dans la première équation de (6.61), où v_h est une fonction arbitraire dans V_h, on constate que u_h coïncide avec la solution unique du problème primal discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } u_h\in V_h \text{ tel que} \\ \int_\Omega \nabla u_h\cdot\nabla v_h = \int_\Omega (f\cdot\nabla v_h + g v_h), \quad \forall v_h\in V_h. \end{array}\right. \quad (6.64)$$

La façon pratique de résoudre (6.61) consiste donc :

(i) à résoudre le problème primal discret (6.64) afin d'obtenir u_h ;

(ii) puis, à *reconstruire* le flux discret σ_h en posant

$$\sigma_h = -\nabla u_h + \Pi_h f, \quad (6.65)$$

où Π_h est la projection L^2-orthogonale sur $[P^0_{\mathrm{td},h}]^d$. Le champ $\Pi_h f$ est constant par morceaux et égal, sur chaque maille, à la valeur moyenne de f sur cette maille :

$$\forall K \in \mathcal{T}_h, \qquad \Pi_h f|_K = \Pi_K f = \tfrac{1}{\mathrm{mes}\, K} \int_K f. \tag{6.66}$$

On constate que la solution discrète (σ_h, u_h) satisfait une loi de Darcy *locale*. Par contre, elle ne satisfait l'équation de conservation de la masse que de manière *faible* (au sens des distributions). Plus précisément, pour un sommet intérieur du maillage s, on désigne par $\Omega(s)$ le *macro-élément* formé par la réunion des mailles contenant s et par ω_s la fonction de forme dans $P^1_{\mathrm{c},h,0}$ associée à ce sommet. Le macro-élément $\Omega(s)$ est illustré sur la figure 6.5 à gauche. On note $\mathcal{F}(s)$ l'ensemble des faces du maillage contenant s. Pour une face $F \in \mathcal{F}(s)$, on note $|F|$ sa mesure et $[\sigma_h \cdot n]_F$ le saut de la composante normale de σ_h à-travers F ; voir la section 3.3.2. Avec ces notations, la *conservation de la masse discrète* s'exprime de la manière suivante : pour tout sommet intérieur du maillage s, on a

$$\sum_{F \in \mathcal{F}(s)} \tfrac{1}{d} [\sigma_h \cdot n]_F |F| = -\int_{\Omega(s)} g \omega_s. \tag{6.67}$$

Cette équation s'établit simplement en prenant $v_h = \omega_s$ dans la deuxième équation de (6.61) et en intégrant par parties le membre de gauche.

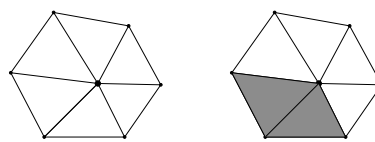


Figure 6.5 – Macro-élément pour la conservation de la masse discrète associé aux problèmes (6.61) (à gauche ; 6 triangles) et (6.68) (à droite ; les 2 triangles en gris).

On peut également approcher le problème (6.3) dans un cadre *non-conforme* par des éléments finis mixtes. On considère les espaces d'approximation $S_h = [P^0_{\mathrm{td},h}]^d$ et $V_h = P^1_{\mathrm{pt},h,0}$, ce dernier espace étant défini en (5.15). On

approche donc le flux par une fonction constante par morceaux et la variable primale par une fonction dans l'espace d'éléments finis de Crouzeix–Raviart. Ceci conduit au problème discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } (\boldsymbol{\sigma}_h, u_h) \in S_h \times V_h \text{ tel que} \\ \int_\Omega \boldsymbol{\sigma}_h \cdot \boldsymbol{\tau}_h + \int_\Omega \boldsymbol{\tau}_h \cdot \nabla_h u_h = \int_\Omega f \cdot \boldsymbol{\tau}_h, \quad \forall \boldsymbol{\tau}_h \in S_h, \\ \int_\Omega \boldsymbol{\sigma}_h \cdot \nabla_h v_h = -\int_\Omega g v_h, \quad \forall v_h \in V_h. \end{array}\right. \tag{6.68}$$

On rappelle que l'opérateur de gradient discret ∇_h est défini en (4.21) ; son utilisation est rendue nécessaire par le fait que les fonctions de V_h peuvent être discontinues aux interfaces entre les mailles. Le problème (6.68) est bien posé et, pourvu que la solution exacte $(\boldsymbol{\sigma}, u)$ soit suffisamment régulière, on a l'estimation d'erreur

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} + \|u - u_h\|_{1,h,\Omega} \leqslant c\, h(\|\boldsymbol{\sigma}\|_{1,\Omega} + \|u\|_{2,\Omega}), \tag{6.69}$$

où la norme H^1-brisée est définie comme suit :

$$\|v\|_{1,h,\Omega} = \left(\|v\|_{0,\Omega}^2 + \|\nabla_h v\|_{0,\Omega}^2\right)^{\frac{1}{2}}. \tag{6.70}$$

Comme pour l'approximation conforme du problème de Darcy (6.61), on constate que u_h coïncide avec la solution unique du problème primal discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } u_h \in V_h \text{ tel que} \\ \int_\Omega \nabla_h u_h \cdot \nabla_h v_h = \int_\Omega (f \cdot \nabla_h v_h + g v_h), \quad \forall v_h \in V_h. \end{array}\right. \tag{6.71}$$

La façon pratique de résoudre la formulation mixte discrète (6.68) consiste donc :

(i) à résoudre le problème primal discret (6.71) afin d'obtenir u_h ;

(ii) puis, à *reconstruire* le flux discret $\boldsymbol{\sigma}_h$ en posant

$$\boldsymbol{\sigma}_h = -\nabla_h u_h + \boldsymbol{\Pi}_h f. \tag{6.72}$$

On constate que la solution discrète $(\boldsymbol{\sigma}_h, u_h)$ satisfait une loi de Darcy *locale*. Par contre, elle ne satisfait l'équation de conservation de la masse que de manière *faible* (au sens des distributions). Plus précisément, pour une face intérieure du maillage F, on désigne par $\boldsymbol{\Omega}(F)$ le macro-élément formé par la réunion des deux mailles partageant F et par $\boldsymbol{\omega}_F$ la fonction de forme dans

$P^1_{\mathrm{pt},b,0}$ associée à cette face. Le *macro-élément* $\Omega(F)$ est illustré sur la figure 6.5 à droite. Avec ces notations, la *conservation de la masse discrète* s'exprime de la manière suivante : pour toute face intérieure dans le maillage F, on a

$$[\sigma_h \cdot n]_F |F| = -\int_{\Omega(F)} g\omega_F. \tag{6.73}$$

Cette équation s'établit simplement en prenant $v_h = \omega_F$ dans la deuxième équation de (6.68) et en intégrant par parties le membre de gauche.

En comparant (6.67) et (6.73), on constate que le macro-élément sur lequel s'exprime la conservation de la masse discrète est plus compact pour l'approximation non-conforme que pour l'approximation conforme. Cette amélioration se traduit par un coût numérique puisque le système linéaire (6.64) est de taille $N_{\mathrm{so}}^{\mathrm{i}}$, le nombre de sommets intérieurs du maillage, alors que le système linéaire (6.71) est de taille $N_{\mathrm{fa}}^{\mathrm{i}}$, le nombre de faces intérieures. En dimension 2 pour des triangles, les relations d'Euler (voir le lemme 3.19) permettent d'affirmer que dans la limite pratique où le nombre de faces sur la frontière est petit devant le nombre de faces intérieures, on a

$$N_{\mathrm{fa}} \simeq N_{\mathrm{fa}}^{\mathrm{i}} \simeq 3N_{\mathrm{so}}^{\mathrm{i}} \qquad \text{et} \qquad N_{\mathrm{ma}} \simeq 2N_{\mathrm{so}}^{\mathrm{i}}. \tag{6.74}$$

Le système linéaire (6.71) est donc approximativement trois fois plus grand que le système linéaire (6.64).

6.3.2 Formulation mixte dans $H(\mathrm{div};\Omega) \times L^2(\Omega)$

On considère une formulation mixte où on prend en compte l'équation de conservation de la masse $\nabla\cdot\sigma = g$ pour chercher *a priori* le flux dans l'espace fonctionnel

$$H(\mathrm{div};\Omega) = \{\sigma \in [L^2(\Omega)]^d \,;\, \nabla\cdot\sigma \in L^2(\Omega)\}. \tag{6.75}$$

On s'intéresse à la formulation faible suivante :

$$\left\{\begin{array}{l} \text{Chercher } (\sigma,u) \in H(\mathrm{div};\Omega) \times L^2(\Omega) \text{ tel que} \\ \quad \int_\Omega \sigma\cdot\tau - \int_\Omega u\nabla\cdot\tau = \int_\Omega f\cdot\tau, \quad \forall \tau \in H(\mathrm{div};\Omega), \\ \quad -\int_\Omega v\nabla\cdot\sigma = -\int_\Omega gv, \quad \forall v \in L^2(\Omega). \end{array}\right. \tag{6.76}$$

On retrouve à nouveau le problème abstrait (6.5) avec $X = H(\mathrm{div};\Omega)$ et $M = L^2(\Omega)$, les formes bilinéaires

$$a(\sigma,\tau) = \int_\Omega \sigma\cdot\tau \qquad \text{et} \qquad b(\tau,u) = -\int_\Omega u\nabla\cdot\tau, \tag{6.77}$$

et les mêmes formes linéaires f et g que pour (6.58). On montre que la forme bilinéaire b satisfait la condition inf-sup (6.9) sur $H(\mathrm{div};\Omega)\times L^2(\Omega)$ si bien que le problème (6.76) est bien posé. De plus, sa solution unique est également celle du problème (6.58), ce qui implique, en particulier, que $u \in H_0^1(\Omega)$. On observera que dans (6.76), la condition aux limites sur la variable primale ne peut pas être imposée *a priori* dans l'espace solution (car les fonctions de $L^2(\Omega)$ n'admettent pas nécessairement de trace sur la frontière $\partial\Omega$).

On s'intéresse à une approximation du problème (6.76) dans un cadre *conforme* par des éléments finis mixtes. On approche le flux dans l'espace d'éléments finis de Raviart–Thomas défini en (4.42) et on approche la variable primale par une fonction constante par morceaux. On pose donc $S_h = D_h$ et $V_h = P^0_{\mathrm{td},h}$, ce qui conduit au problème discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } (\sigma_h,u_h)\in S_h\times V_h \text{ tel que} \\ \quad \int_\Omega \sigma_h\cdot\tau_h - \int_\Omega u_h\nabla\cdot\tau_h = \quad \int_\Omega f\cdot\tau_h, \quad \forall\tau_h\in S_h, \\ -\int_\Omega v_h\nabla\cdot\sigma_h \qquad\qquad = -\int_\Omega g v_h, \quad \forall v_h\in V_h. \end{array}\right. \tag{6.78}$$

Le problème (6.78) est bien posé et, pourvu que la solution exacte (σ,u) soit suffisamment régulière, on a l'estimation d'erreur

$$\|\sigma-\sigma_h\|_{0,\Omega} + \|\nabla\cdot(\sigma-\sigma_h)\|_{0,\Omega} + \|u-u_h\|_{0,\Omega} \leqslant c\,h(\|\sigma\|_{1,\Omega} + \|\nabla\cdot\sigma\|_{1,\Omega} + \|u\|_{1,\Omega}). \tag{6.79}$$

Un des intérêts du problème discret (6.78) est que la deuxième équation implique immédiatement la conservation *locale* de la masse discrète sous la forme

$$\nabla\cdot\sigma_h = \Pi_h g. \tag{6.80}$$

Par contre, la loi de Darcy ne s'exprime plus de manière locale.

Le système linéaire (6.78) est de taille

$$\dim D_h + \dim P^0_{\mathrm{td},h} = N_{\mathrm{ma}} + N_{\mathrm{fa}} \simeq 5N^{\mathrm{i}}_{\mathrm{so}}, \tag{6.81}$$

soit une taille à peu près cinq fois plus grande que celle du système linéaire (6.64). La technique des *éléments finis mixtes hybrides* exposée ci-dessous, qui permet de résoudre le problème (6.78) en se ramenant à un système linéaire de taille plus petite, présente donc un intérêt pratique considérable. L'idée consiste, dans un premier temps, à relâcher la contrainte de continuité de la composante normale du flux aux interfaces et à introduire un multiplicateur de Lagrange associé à cette contrainte sur chaque interface. On introduit les espaces discrets

$$D^*_h = \{\boldsymbol{\tau}_h \in [L^2(\Omega)]^d \,;\, \forall K \in \mathcal{T}_h,\, \boldsymbol{\tau}_h|_K \in \mathbb{RT}_0\}, \tag{6.82}$$

$$F^0_h = \{\lambda_h \in L^2(\mathcal{F}^{\mathrm{i}}_h) \,;\, \forall F \in \mathcal{F}^{\mathrm{i}}_h,\, \lambda_h|_F \in \mathbb{P}_0\}. \tag{6.83}$$

On observera que $D_h = \{\boldsymbol{\tau}_h \in D^*_h \,;\, \forall F \in \mathcal{F}^{\mathrm{i}}_h,\ [\![\boldsymbol{\tau}_h{\cdot}\boldsymbol{n}]\!]_F = 0\}$ et que pour tout $\boldsymbol{\tau}_h \in D^*_h$, la fonction $[\![\boldsymbol{\tau}_h{\cdot}\boldsymbol{n}]\!]_{\mathcal{F}}$ définie pour tout $F \in \mathcal{F}^{\mathrm{i}}_h$ par $[\![\boldsymbol{\tau}_h{\cdot}\boldsymbol{n}]\!]_{\mathcal{F}}|_F = [\![\boldsymbol{\tau}_h{\cdot}\boldsymbol{n}]\!]_F$ est dans F^0_h. On considère le problème approché suivant :

$$\left\{\begin{array}{l} \text{Chercher } (\boldsymbol{\sigma}_h, u_h, \lambda_h) \in D^*_h \times V_h \times F^0_h \text{ tel que} \\ \quad \int_\Omega \boldsymbol{\sigma}_h{\cdot}\boldsymbol{\tau}_h - \int_\Omega u_h \nabla_h{\cdot}\boldsymbol{\tau}_h - \int_{\mathcal{F}^{\mathrm{i}}_h} \lambda_h [\![\boldsymbol{\tau}_h{\cdot}\boldsymbol{n}]\!]_{\mathcal{F}} = \int_\Omega \boldsymbol{f}{\cdot}\boldsymbol{\tau}_h, \quad \forall \boldsymbol{\tau}_h \in D^*_h, \\ -\int_\Omega v_h \nabla_h{\cdot}\boldsymbol{\sigma}_h = -\int_\Omega g v_h, \quad \forall v_h \in V_h, \\ -\int_{\mathcal{F}^{\mathrm{i}}_h} \mu_h [\![\boldsymbol{\sigma}_h{\cdot}\boldsymbol{n}]\!]_{\mathcal{F}} = 0, \quad \forall \mu_h \in F^0_h. \end{array}\right. \tag{6.84}$$

On montre que le problème (6.84) est bien posé et que si $(\boldsymbol{\sigma}_h, u_h, \lambda_h)$ est solution de (6.84), alors $\boldsymbol{\sigma}_h \in D_h$ et $(\boldsymbol{\sigma}_h, u_h)$ est solution de (6.78).

L'intérêt du système (6.84) est que les deux premières équations sont purement locales. Afin d'expliciter ces équations, on se place pour simplifier en dimension 2 et on suppose que $f = 0$. Tout ce qui suit se généralise à la dimension 3 et au cas $f \neq 0$. Soit $K \in \mathcal{T}_h$. On désigne par $U_K \in \mathbb{R}$ la valeur prise par la fonction (constante) $u_h|_K$. On désigne par $\{\theta_{K,1}, \theta_{K,2}, \theta_{K,3}\}$ les fonctions de forme de l'élément fini de Raviart–Thomas sur la maille K et par

B_K la matrice d'ordre 3 de terme générique

$$B_{K,ij} = \int_K \theta_{K,i} \cdot \theta_{K,j}, \qquad i,j \in \{1,2,3\}. \tag{6.85}$$

Soit $S_K \in \mathbb{R}^3$ le vecteur formé par les trois composantes du flux dans la base locale $\{\theta_{K,1}, \theta_{K,2}, \theta_{K,3}\}$. Soit $\Lambda_K \in \mathbb{R}^3$ le vecteur dont les trois composantes sont les multiplicateurs de Lagrange sur les trois faces de K. Avec ces notations, la première équation dans (6.84) est une loi de Darcy locale qui s'écrit dans $\mathbb{R}^3$ sous la forme

$$B_K S_K - U_K Z + \Lambda_K = 0, \tag{6.86}$$

avec $Z = (1,1,1)^T$. La deuxième équation est une équation de conservation de la masse locale qui s'écrit sous la forme

$$(S_K, Z)_{\mathbb{R}^3} = \int_K g, \tag{6.87}$$

où $(\cdot,\cdot)_{\mathbb{R}^3}$ désigne le produit scalaire euclidien dans $\mathbb{R}^3$. Des propriétés élémentaires de la matrice B_K (voir le lemme 6.15 ci-dessous), on déduit

$$U_K = \tfrac{1}{3}(\Lambda_K, Z)_{\mathbb{R}^3} + \tfrac{1}{4}\rho_K^2 \Pi_K g, \tag{6.88}$$

où l'opérateur Π_K est défini en (6.66), puis que

$$S_K = B_K^{-1}\Big([\tfrac{1}{3}(\Lambda_K, Z)_{\mathbb{R}^3} + \tfrac{1}{4}\rho_K^2 \Pi_K g]Z - \Lambda_K\Big). \tag{6.89}$$

La quantité ρ_K (dont les dimensions sont celles d'une longueur) désigne le *rayon de giration* de K et s'évalue de la manière suivante :

$$\rho_K = |K|^{-\frac{1}{2}} \|\pi_K^1\|_{0,K} \qquad \text{avec} \qquad \pi_K^1(x) = G_K x, \tag{6.90}$$

où G_K est le centre de gravité de K et $G_K x$ désigne le vecteur obtenu en joignant G_K au point courant x dans K.

Le principe des éléments finis mixtes hybrides est le suivant : sur chaque face intérieure, on écrit la continuité de la composante normale du flux en utilisant (6.89). Ceci conduit à un système linéaire de taille N_{fa}^{i} où les seules inconnues sont les multiplicateurs de Lagrange. Une fois évalués ces multiplicateurs de Lagrange, on utilise les formules de reconstruction (6.88) et (6.89)

pour évaluer localement le flux et la variable primale. En conclusion, la méthode des éléments finis mixtes hybrides nécessite uniquement la résolution d'un système linéaire de taille $N_{\mathrm{fa}}^{\mathrm{i}} \simeq 3N_{\mathrm{so}}^{\mathrm{i}}$, ce qui représente un gain de 40% par rapport à la taille initiale du problème (6.78).

Lemme 6.15. *Soit K un triangle non dégénéré. Alors, la matrice B_K est inversible et on a*

$$B_K^{-1} = \begin{pmatrix} \gamma_2 + \gamma_3 & -\gamma_3 & -\gamma_2 \\ -\gamma_3 & \gamma_1 + \gamma_3 & -\gamma_1 \\ -\gamma_2 & -\gamma_1 & \gamma_1 + \gamma_2 \end{pmatrix} + \frac{1}{3l_K} Z \otimes Z, \tag{6.91}$$

où $\gamma_i = 2\cot\theta_{K,i}$, $i \in \{1,2,3\}$, $\theta_{K,i}$ étant l'angle entre les deux arêtes au $i$ième sommet de K, et cot *la fonction cotangente. De plus, on a $l_K = \frac{3}{4}\rho_K^2 |K|^{-2}$ où ρ_K est le rayon de giration de K défini en* (6.90).

6.3.3 Formulation mixte dans $H(\mathrm{div};\Omega) \times H_0^1(\Omega)$

On considère la formulation faible suivante :

$$\left\{ \begin{array}{ll} \text{Chercher } (\sigma, u) \in H(\mathrm{div};\Omega) \times H_0^1(\Omega) \text{ tel que} & \\ \int_\Omega \sigma\cdot\tau + \int_\Omega \tau\cdot\nabla u = \int_\Omega f\cdot\tau, & \forall \tau \in [L^2(\Omega)]^d, \\ \int_\Omega v\nabla\cdot\sigma \qquad\quad = \int_\Omega gv, & \forall v \in L^2(\Omega). \end{array} \right. \tag{6.92}$$

Il s'agit d'une formulation mixte de type *non-standard* puisque l'espace solution $H(\mathrm{div};\Omega) \times H_0^1(\Omega)$ est *différent* de l'espace test $[L^2(\Omega)]^d \times L^2(\Omega)$. Cette formulation ne rentre donc pas dans le cadre particulier du problème abstrait (6.5). Son analyse mathématique se fait plus naturellement dans le cadre du théorème BNB. En introduisant les espaces fonctionnels $V = H(\mathrm{div};\Omega) \times H_0^1(\Omega)$ et $W = [L^2(\Omega)]^d \times L^2(\Omega)$, on montre que la forme bilinéaire $c \in \mathcal{L}(V \times W; \mathbb{R})$ définie par

$$c((\sigma,u),(\tau,v)) = \int_\Omega \sigma\cdot\tau + \int_\Omega \tau\cdot\nabla u + \int_\Omega v\nabla\cdot\sigma, \tag{6.93}$$

satisfait les conditions (BNB1) et (BNB2) ; voir, par exemple, Ern et Guermond [**38**, p. 271]. Le problème (6.92) est donc bien posé. De plus, sa solution unique est également celle des problèmes (6.58) et (6.76).

On s'intéresse à une approximation du problème (6.92) dans un cadre *nonconforme* par des éléments finis mixtes [**31**]. On approche le flux dans l'espace

d'éléments finis de Raviart–Thomas et la variable primale dans l'espace d'éléments finis de Crouzeix–Raviart. On considère le problème discret suivant :

$$\left\{\begin{array}{l} \text{Chercher } (\boldsymbol{\sigma}_h, u_h) \in D_h \times P^1_{\mathrm{pt},h,0} \text{ tel que} \\ \int_\Omega \boldsymbol{\sigma}_h\cdot\boldsymbol{\tau}_h + \int_\Omega \boldsymbol{\tau}_h\cdot\nabla_h u_h = \int_\Omega f\cdot\boldsymbol{\tau}_h, \quad \forall \boldsymbol{\tau}_h \in [P^0_{\mathrm{td},h}]^d, \\ \int_\Omega v_h \nabla\cdot\boldsymbol{\sigma}_h \qquad\qquad\quad = \int_\Omega g v_h, \quad \forall v_h \in P^0_{\mathrm{td},h}. \end{array}\right. \tag{6.94}$$

Le problème (6.94) est bien posé et, pourvu que la solution exacte $(\boldsymbol{\sigma}, u)$ soit suffisamment régulière, on a l'estimation d'erreur

$$\begin{aligned} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} + \|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{0,\Omega} + \|u - u_h\|_{1,h,\Omega} \leqslant c\, h(\|\boldsymbol{\sigma}\|_{1,\Omega} \\ + \|\nabla\cdot\boldsymbol{\sigma}\|_{1,\Omega} + \|u\|_{2,\Omega}). \end{aligned} \tag{6.95}$$

On notera, en particulier, que ce sont les relations d'Euler qui permettent d'affirmer que le problème discret (6.94) contient autant d'inconnues que d'équations. On a en effet

$$\begin{aligned} \dim D_h + \dim P^1_{\mathrm{pt},h,0} &= N_{\mathrm{fa}} + N^{\mathrm{i}}_{\mathrm{fa}} = (d+1)N_{\mathrm{ma}} \\ &= \dim[P^0_{\mathrm{td},h}]^d + \dim P^0_{\mathrm{td},h}. \end{aligned} \tag{6.96}$$

Il est possible de considérer dans (6.94) des fonctions tests constantes par morceaux (par exemple, les indicatrices des mailles) aussi bien pour le flux que pour la variable primale. Le problème discret (6.94) s'interprète donc comme un schéma de type « volumes finis ». Dans la littérature, il porte également le nom de *schéma boîte*. Le schéma boîte (6.94) présente deux propriétés intéressantes.

(i) **Conservation locale de la masse.** On déduit immédiatement de la deuxième équation dans (6.94) que

$$\nabla\cdot\boldsymbol{\sigma}_h = \Pi_h g. \tag{6.97}$$

Cette propriété est également satisfaite par le problème discret (6.78) ; voir l'équation (6.80).

(ii) **Reconstruction locale du flux.** Sur toute maille $K \in \mathcal{T}_h$, tout champ discret $\boldsymbol{\sigma}_h \in D_h$ s'écrit sous la forme

$$\boldsymbol{\sigma}_h|_K = \Pi_K \boldsymbol{\sigma}_h + \tfrac{1}{d}(\nabla\cdot\boldsymbol{\sigma}_h)|_K \boldsymbol{\pi}^1_K, \tag{6.98}$$

où $\boldsymbol{\pi}_K^1 \in [\mathbb{P}_1]^d$ est défini en (6.90). On déduit alors de la première équation dans (6.94) et de (6.97) que

$$\boldsymbol{\sigma}_h = -\nabla_h u_h + \Pi_h f + \tfrac{1}{d}(\Pi_h g)\boldsymbol{\pi}_h^1, \tag{6.99}$$

où $\boldsymbol{\pi}_h^1|_K = \boldsymbol{\pi}_K^1$ pour tout $K \in \mathcal{T}_h$.

On observera que ces deux propriétés remarquables résultent du fait que, grâce au choix des espaces d'approximation pour $\boldsymbol{\sigma}_h$ et u_h, aucune intégration par parties n'a été effectuée dans la formulation du problème approché (6.94).

Le système linéaire (6.94) est de taille significativement plus grande que le système linéaire (6.61). Par exemple, en deux dimensions d'espace, ce système est de taille 7 fois supérieure. La technique présentée ci-dessous, qui permet de se ramener à un système de taille plus petite, présente donc un intérêt pratique considérable. L'observation essentielle est que la solution discrète u_h est également l'unique solution du problème approché suivant :

$$\left\{\begin{array}{l} \text{Chercher } u_h \in P^1_{\mathrm{pt},h,0} \text{ tel que} \\ \int_\Omega \nabla_h u_h \cdot \nabla_h v_h = \int_\Omega [f \cdot \nabla_h v_h + (\Pi_h g) v_h], \quad \forall v_h \in P^1_{\mathrm{pt},h,0}. \end{array}\right. \tag{6.100}$$

La façon pratique de résoudre (6.94) consiste donc :

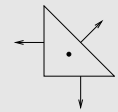
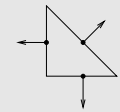
(i) à résoudre le problème primal discret (6.100) afin d'obtenir u_h ;

(ii) puis, à reconstruire le champ discret $\boldsymbol{\sigma}_h$ en utilisant (6.99).

6.3.4 Résumé

Le tableau 6.2 résume les propriétés des formulations mixtes du problème de Darcy étudiées dans cette section et les propriétés de leur approximation par éléments finis. On rappelle l'espace fonctionnel dans lequel est cherché la solution exacte $(\boldsymbol{\sigma}, u)$, le cadre conforme ou non-conforme de l'approximation, les espaces d'éléments finis mixtes, les propriétés qui sont satisfaites localement sur chaque maille (la loi de Darcy (6.3) et/ou la conservation de la masse (6.4)) et enfin, la taille du système linéaire. On rappelle qu'en deux dimensions d'espace, dans la limite pratique où le nombre de faces sur le bord est petit devant le nombre de faces intérieures, on a $N_{\mathrm{fa}}^{\mathrm{i}} \simeq 3N_{\mathrm{so}}^{\mathrm{i}}$.

Tableau 6.2 – Propriétés des formulations mixtes pour le problème de Darcy.

(σ, u)	$[L^2(\Omega)]^d \times H_0^1(\Omega)$		$H(\mathrm{div};\Omega) \times L^2(\Omega)$	$H(\mathrm{div};\Omega) \times H_0^1(\Omega)$
approx.	conforme	non-conforme	conforme	non-conforme
(σ_h, u_h)	$[P^0_{\mathrm{td},h}]^d \times P^1_{\mathrm{c},h,0}$	$[P^0_{\mathrm{td},h}]^d \times P^1_{\mathrm{pt},h,0}$	$D_h \times P^0_{\mathrm{td},h}$	$D_h \times P^1_{\mathrm{pt},h,0}$
				
local	Darcy	Darcy	masse	Darcy+masse
taille	$N^{\mathrm{i}}_{\mathrm{so}}$	$N^{\mathrm{i}}_{\mathrm{fa}}$	$N^{\mathrm{i}}_{\mathrm{fa}}$	$N^{\mathrm{i}}_{\mathrm{fa}}$

6.4 Complément : compressibilité artificielle

Une technique pour résoudre le système linéaire (6.16) consiste à considérer le système linéaire perturbé

$$\begin{bmatrix} \mathcal{A} & \mathcal{B}^T \\ -\mathcal{B} & \epsilon\mathcal{M}_p \end{bmatrix} \begin{bmatrix} U_\epsilon \\ P_\epsilon \end{bmatrix} = \begin{bmatrix} F \\ -G \end{bmatrix}, \tag{6.101}$$

où $\epsilon > 0$ est un coefficient dit de compressibilité artificielle et $\mathcal{M}_p$ la matrice de masse associée à l'espace d'approximation M_h. La terminologie provient du fait qu'en notant $u_{h\epsilon}$ et $p_{h\epsilon}$ les fonctions de X_h et M_h associées à U_ϵ et P_ϵ, respectivement, la deuxième équation dans (6.101) s'écrit sous la forme

$$b(u_{h\epsilon}, q_h) - \epsilon(p_{h\epsilon}, q_h)_{0,\Omega} = g(q_h), \qquad \forall q_h \in M_h. \tag{6.102}$$

Pour le problème de Stokes avec $g = 0$, on constate que l'on a remplacé l'équation d'incompressibilité $\nabla \cdot u = 0$ par une équation de compressibilité artificielle $\nabla \cdot u + \epsilon p = 0$.

En éliminant P_ϵ dans (6.101), il vient

$$\left(\mathcal{A} + \frac{1}{\epsilon} \mathcal{B}^T \mathcal{M}_p^{-1} \mathcal{B} \right) U_\epsilon = F + \frac{1}{\epsilon} \mathcal{B}^T \mathcal{M}_{N_p}^{-1} G. \tag{6.103}$$

L'intérêt du système (6.103) est qu'il fait intervenir une matrice symétrique définie positive. Par conséquent, la solution U_ϵ peut s'obtenir de manière très efficace en mettant en œuvre une méthode du gradient conjugué préconditionné ; voir la section 11.2. De plus, avec les notations du lemme 6.6, l'erreur entre la solution (u_h, p_h) de (6.12) et le couple $(u_{h\epsilon}, p_{h\epsilon})$ résultant de la solution de (6.101) est contrôlée sous la forme

$$\frac{\alpha \beta_h}{\|a\|_{X,X}} \|u_h - u_{h\epsilon}\|_X + \frac{\alpha \beta_h^2}{\|a\|_{X,X}^2} \|p_h - p_{h\epsilon}\|_M \leqslant \epsilon \|p_h\|_M. \tag{6.104}$$

Le conditionnement de la matrice du système (6.101) se dégrade lorsque $\epsilon \to 0$, ce qui rend l'approche par compressibilité artificielle peu attractive lorsqu'on souhaite satisfaire la condition d'incompressibilité avec une bonne précision.

7 • GALERKIN/MOINDRES CARRÉS

Ce chapitre décrit le principe général de la méthode de Galerkin/moindres carrés (en anglais, Galerkin/Least-Squares ou, en abrégé, GaLS) puis en présente des applications à trois problèmes modèles : l'équation d'advection-réaction comme prototype des problèmes d'ordre un ; l'équation d'advection–diffusion avec advection dominante comme exemple de perturbation singulière d'un problème d'ordre un ; enfin, le problème de Stokes approché par des éléments finis pour la vitesse et la pression ne satisfaisant pas la condition de compatibilité (6.26) étudiée au chapitre 6.

7.1 Principe de la méthode

Étant donné deux entiers p et n, une matrice $\mathcal{A} \in \mathbb{R}^{p,n}$ et un vecteur $F \in \mathbb{R}^p$, on considère le système linéaire

$$\mathcal{A}X = F, \tag{7.1}$$

où $X \in \mathbb{R}^n$. On suppose que la matrice $\mathcal{A}$ est injective (c'est-à-dire, $\mathrm{Ker}\,\mathcal{A} = \{0\}$) ce qui implique $p \geqslant n$. Lorsque $p = n$, cette hypothèse équivaut à l'inversibilité de $\mathcal{A}$. Par contre, lorsque $p > n$, la matrice $\mathcal{A}$ n'est pas de rang maximal si bien que le système (7.1) n'admet pas nécessairement de solution. Pour un entier $k \geqslant 1$, on désigne par $(\cdot,\cdot)_{\mathbb{R}^k}$ le produit scalaire euclidien sur $\mathbb{R}^k$ et par $\|\cdot\|_{\mathbb{R}^k}$ la norme induite.

Définition 7.1 (Moindres carrés). *On dit que le vecteur $X \in \mathbb{R}^n$ est solution de* (7.1) au sens des moindres carrés *si U minimise sur $\mathbb{R}^n$ la norme du résidu $\|F - \mathcal{A}X\|_{\mathbb{R}^p}$.*

On vérifie facilement que $U \in \mathbb{R}^n$ est solution de (7.1) au sens des moindres carrés si et seulement si U est solution du système linéaire

$$\mathcal{A}^T\mathcal{A}X = \mathcal{A}^T F. \tag{7.2}$$

La matrice $\mathcal{A}^T\mathcal{A}$ est carrée d'ordre n, elle est clairement symétrique et, de plus, elle est définie positive puisque pour tout $Y \in \mathbb{R}^n$, $(\mathcal{A}^T\mathcal{A}Y, Y)_{\mathbb{R}^n} = \|\mathcal{A}Y\|^2_{\mathbb{R}^n} \geqslant 0$ avec égalité si et seulement si $\mathcal{A}Y = 0$, ce qui équivaut à $Y = 0$ car $\mathcal{A}$ est injective. Par conséquent, le système linéaire (7.2) admet une solution unique.

Par la suite, on s'intéresse aux systèmes linéaires issus d'approximations par éléments finis. Ces systèmes étant carrés, on se restreint au cas $p = n$. On observe qu'une formulation équivalente du système linéaire (7.1) consiste à chercher $X \in \mathbb{R}^n$ tel que

$$(\mathcal{A}X, Y)_{\mathbb{R}^n} = (F, Y)_{\mathbb{R}^n}, \qquad \forall Y \in \mathbb{R}^n. \tag{7.3}$$

De même, le système linéaire (7.2) consiste à chercher $X \in \mathbb{R}^n$ tel que

$$(\mathcal{A}X, \mathcal{A}Y)_{\mathbb{R}^n} = (F, \mathcal{A}Y)_{\mathbb{R}^n}, \qquad \forall Y \in \mathbb{R}^n. \tag{7.4}$$

La matrice $\mathcal{A}$ étant inversible, ces deux problèmes admettent la même solution. Un troisième système linéaire équivalent consiste à choisir un paramètre $\delta > 0$ et à chercher $X \in \mathbb{R}^n$ tel que

$$(\mathcal{A}X, Y)_{\mathbb{R}^n} + \delta(\mathcal{A}X, \mathcal{A}Y)_{\mathbb{R}^n} = (F, Y)_{\mathbb{R}^n} + \delta(F, \mathcal{A}Y)_{\mathbb{R}^n}, \qquad \forall Y \in \mathbb{R}^n. \tag{7.5}$$

Le système linéaire (7.5) est à la base de la méthode de Galerkin/moindres carrés.

Lorsque la matrice $\mathcal{A}$ est *symétrique définie positive*, il est possible de donner une interprétation du système linéaire (7.5) dans un cadre variationnel. On vérifie facilement que $X \in \mathbb{R}^n$ vérifie $\mathcal{A}X = F$ si et seulement si X minimise sur $\mathbb{R}^n$ la fonctionnelle

$$J : \mathbb{R}^n \ni V \longmapsto \tfrac{1}{2}(\mathcal{A}V, V)_{\mathbb{R}^n} - (F, V)_{R^n} \in \mathbb{R}. \tag{7.6}$$

Pour tout $V \in \mathbb{R}^n$, le gradient de J en V, qui est un vecteur de $\mathbb{R}^n$ que l'on note $\nabla J(V)$ (voir la section 11.2.1 pour plus de détails), est tel que

$$\nabla J(V) = \mathcal{A}V - F. \tag{7.7}$$

La formulation (7.3) s'écrit donc sous la forme

$$(\nabla J(X), Y)_{\mathbb{R}^n} = 0, \qquad \forall Y \in \mathbb{R}^n. \tag{7.8}$$

Cette équation est la *condition d'Euler–Lagrange* permettant de caractériser le vecteur $X \in \mathbb{R}^n$ réalisant le minimum de la fonctionnelle J sur $\mathbb{R}^n$. Le système linéaire (7.5) peut également se formuler comme une condition d'Euler–Lagrange associée à la minimisation d'une fonctionnelle sur $\mathbb{R}^n$. On pose

$$J_1 : \mathbb{R}^n \ni V \longmapsto J(V) + \tfrac{\delta}{2}\|F - \mathcal{A}X\|^2_{\mathbb{R}^n} \in \mathbb{R}. \tag{7.9}$$

On constate que (7.5) s'écrit sous la forme

$$(\nabla J_1(X), Y)_{\mathbb{R}^n} = 0, \qquad \forall Y \in \mathbb{R}^n. \tag{7.10}$$

En conclusion, les termes proportionnels à δ dans (7.5) réalisent une *pénalisation* au sens des moindres carrés de la norme euclidienne du résidu.

Afin de donner une interprétation plus spécifique de la méthode de Galerkin/moindres carrés à l'approximation d'équations aux dérivées partielles avec conditions aux limites, on suppose que le système linéaire (7.1) provient de l'approximation par la méthode de Galerkin d'un problème modèle qui consiste à chercher une fonction $u : \Omega \to \mathbb{R}^m$ (où $m \geqslant 1$ est un entier fixé) telle que

$$Au = f \quad \text{dans } \Omega, \tag{7.11}$$

$$Bu = g \quad \text{sur } \partial\Omega, \tag{7.12}$$

où A un opérateur différentiel sur Ω et B un opérateur qui permet d'exprimer les conditions aux limites. On suppose que la formulation faible de (7.11)–(7.12) consiste à

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u, w) = f(w), \quad \forall w \in W, \end{cases} \tag{7.13}$$

où V et W sont des espaces de Hilbert, $a \in \mathcal{L}(V \times W; \mathbb{R})$ et $f \in W'$. En général, on a $V \subset L$ où on a posé $L = [L^2(\Omega)]^m$. On note $(\cdot, \cdot)_L$ le produit scalaire dans L. On distingue deux cas.

(i) L'opérateur A est un isomorphisme de V dans L. Cette situation se présente par exemple pour l'équation d'advection–réaction étudiée dans la section 7.2. Dans ce cas, on peut prendre pour espace test $W = L$ dans (7.13) si bien que pour tout $(v, w) \in V \times L$,

$$a(v, w) = (Av, w)_L \qquad \text{et} \qquad f(w) = (f, w)_L. \tag{7.14}$$

Étant donné un espace d'approximation V_h que l'on suppose V-conforme, la méthode de Galerkin standard conduit au problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ (Au_h, w_h)_L = (f, w_h)_L, \quad \forall w_h \in V_h. \end{cases} \tag{7.15}$$

La méthode des moindres carrés conduit au problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ (Au_h, Aw_h)_L = (f, Aw_h)_L, \quad \forall w_h \in V_h, \end{cases} \tag{7.16}$$

et la méthode de Galerkin/moindres carrés à

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ (Au_h, w_h)_L + \delta(h)(Au_h, Aw_h)_L \\ \qquad = (f, w_h)_L + \delta(h)(f, Aw_h)_L, \quad \forall w_h \in V_h. \end{cases} \tag{7.17}$$

On notera que ces systèmes linéaires sont tous bien posés, mais que leurs solutions respectives sont *a priori* différentes. On notera également que le paramètre de pondération δ dans (7.17) dépend de h. Cette dépendance est en général déterminée par l'analyse de l'erreur $u - u_h$, le but étant d'obtenir une estimation d'erreur en norme $\|\cdot\|_L$ qui soit d'ordre le plus élevé possible en h.

(ii) L'opérateur A est tel que $A(V) \not\subset L$ si bien que A n'est pas un isomorphisme de V dans L. Cette situation se présente par exemple pour l'équation d'advection–diffusion étudiée dans la section 7.3 et pour le problème de Stokes étudié dans la section 7.4. Dans ce cas, le problème modèle est (7.13) et pour simplifier, on se restreint au cas où $W = V$. Soit V_h un espace d'approximation V-conforme. Comme $A(V) \not\subset L$, les produits scalaires $(Av_h, Aw_h)_L$ n'ont *a priori* pas de sens

pour $(v_h, w_h) \in V_h \times V_h$. L'idée consiste alors à localiser ce produit scalaire sur les éléments du maillage $\mathcal{T}_h$ qui a servi à construire l'espace d'approximation V_h. Pour $K \in \mathcal{T}_h$, on note $(\cdot,\cdot)_{L,K}$ le produit scalaire dans $[L^2(K)]^m$. On observe que la restriction à une maille $K \in \mathcal{T}_h$ d'une fonction de V_h est de classe $\mathcal{C}^\infty$ (c'est en général un polynôme), si bien que le produit scalaire $(Av_h, Aw_h)_{L,K}$ a un sens. La méthode de Galerkin/moindres carrés consiste à

$$\left\{\begin{array}{l} \text{Chercher } u_h \in V_h \text{ tel que, } \forall w_h \in V_h, \\ a(u_h, w_h) + \sum_{K\in\mathcal{T}_h} \delta(h_K)(Au_h, Aw_h)_{L,K} \\ \qquad = f(w_h) + \sum_{K\in\mathcal{T}_h} \delta(h_K)(f, Aw_h)_{L,K}. \end{array}\right. \tag{7.18}$$

Enfin, on observera que les deux problèmes discrets obtenus par la méthode de Galerkin/moindres carrés, à savoir (7.17) et (7.18), sont *consistants* puisque la solution exacte u satisfait les équations discrètes. Un exemple de formulation de type Galerkin/moindres carrés non-consistante est étudié dans la section 7.5.

7.2 Advection–réaction

Dans cette section, on considère une équation d'advection–réaction comme prototype d'un problème d'ordre un. On analyse l'approximation de cette équation par la méthode des moindres carrés et par la méthode de Galerkin/moindres carrés.

7.2.1 Un problème modèle unidimensionnel

On pose $\Omega =]0, 1[$. Étant donné une fonction $f \in L^2(\Omega)$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$u' = f \quad \text{dans } \Omega, \tag{7.19}$$

$$u(0) = 0. \tag{7.20}$$

Ce problème rentre dans le cadre du problème modèle (7.11)–(7.12).

Une formulation faible possible de (7.19)–(7.20) consiste à considérer le problème (7.13) avec $V = \{v \in H^1(\Omega)\ ;\ v(0) = 0\}$, $W = L^2(\Omega)$ et les formes

$$a(v, w) = \int_\Omega v'w \qquad \text{et} \qquad f(w) = \int_\Omega fw. \tag{7.21}$$

On montre que la forme bilinéaire a vérifie les conditions (BNB1) et (BNB2) du théorème BNB si bien que le problème (7.13) est bien posé.

On considère le problème discret (7.15) qui résulte d'une approximation de (7.13) par la méthode de Galerkin standard. L'espace d'approximation V_h est construit à l'aide de l'élément fini de Lagrange $\mathbb{P}_1$. Dans ces conditions, on montre qu'il existe des constantes positives c_1 et c_2, indépendantes de h, telles que

$$c_1 h \leqslant \inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{a(v_h, w_h)}{\|v_h\|_{1,\Omega} \|w_h\|_{0,\Omega}} \leqslant c_2 h. \tag{7.22}$$

Ce résultat a des conséquences importantes puisqu'il implique que l'erreur $u - u_h$ n'est pas contrôlée en norme H^1, la dégradation de la condition inf-sup en h étant du même ordre que l'erreur d'interpolation. Dans la pratique, cela se traduit par des oscillations parasites qui polluent la solution discrète u_h.

7.2.2 Un problème d'advection–réaction

Soit Ω un domaine de $\mathbb{R}^d$ de frontière suffisamment régulière. On considère un champ de vecteurs β défini sur Ω et à valeurs dans $\mathbb{R}^d$. On suppose que $\beta \in [L^\infty(\Omega)]^d$ et $\nabla \cdot \beta \in L^\infty(\Omega)$. On pose

$$\partial\Omega^- = \{x \in \partial\Omega\ ;\ \beta(x) \cdot n(x) < 0\}, \tag{7.23}$$

$$\partial\Omega^+ = \{x \in \partial\Omega\ ;\ \beta(x) \cdot n(x) > 0\}, \tag{7.24}$$

où n est la normale extérieure à Ω sur $\partial\Omega$. L'ensemble $\partial\Omega^-$ s'appelle la *frontière entrante* de Ω et l'ensemble $\partial\Omega^+$ s'appelle la *frontière sortante*.

Étant donné des fonctions $\mu \in L^\infty(\Omega)$ et $f \in L^2(\Omega)$, on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$\beta \cdot \nabla u + \mu u = f \quad \text{dans } \Omega, \tag{7.25}$$

$$u = 0 \quad \text{sur } \partial\Omega^-. \tag{7.26}$$

Ce problème rentre dans le cadre du problème modèle (7.11)–(7.12). On introduit les espaces fonctionnels

$$V = \{v \in L^2(\Omega) \,;\, \beta \cdot \nabla v \in L^2(\Omega)\}, \tag{7.27}$$

$$V_* = \{v \in V \,;\, v|_{\partial\Omega^-} = 0\}, \tag{7.28}$$

et on pose $L = L^2(\Omega)$. On équipe V de la norme du graphe,

$$\forall v \in V, \qquad \|v\|_V = \|v\|_{0,\Omega} + \|\beta \cdot \nabla v\|_{0,\Omega}. \tag{7.29}$$

On introduit l'opérateur

$$A : V \ni v \longmapsto \mu v + \beta \cdot \nabla v \in L. \tag{7.30}$$

Clairement, $A \in \mathcal{L}(V; L)$. On introduit également la forme bilinéaire $a \in \mathcal{L}(V \times L; \mathbb{R})$ et la forme linéaire $f \in V'$ définies par

$$\begin{aligned} a(v, w) &= (Av, w)_{0,\Omega} = \textstyle\int_\Omega (\mu v + \beta \cdot \nabla v) w \\ \text{et} \qquad f(w) &= (f, w)_{0,\Omega} = \textstyle\int_\Omega f w. \end{aligned} \tag{7.31}$$

La formulation faible de (7.25)–(7.26) est la suivante :

$$\begin{cases} \text{Chercher } u \in V_* \text{ tel que} \\ a(u, w) = f(w), \quad \forall w \in L. \end{cases} \tag{7.32}$$

Par la suite, on suppose qu'il existe $\mu_0 > 0$ tel que

$$\mu - \tfrac{1}{2} \nabla \cdot \beta \geqslant \mu_0 \quad \text{presque partout dans } \Omega. \tag{7.33}$$

Cette hypothèse implique que la forme bilinéaire a est L-coercive sur V_* puisque l'on a

$$\forall v \in V_*, \qquad a(v, v) \geqslant \mu_0 \|v\|_L^2. \tag{7.34}$$

Bien entendu, cette propriété seule est insuffisante pour prouver le caractère bien posé du problème (7.32). On a le résultat suivant.

Proposition 7.2. *Sous l'hypothèse* (7.33)*, la forme bilinéaire a définie en* (7.31) *satisfait les conditions* (BNB1) *et* (BNB2) *du théorème* BNB*. Par conséquent, le problème* (7.32) *est bien posé (ou, en termes équivalents, l'opérateur* $A : V_* \to L$ *est un isomorphisme).*

Par la suite, on suppose que le problème (7.25)–(7.26) a été adimensionnalisé de sorte que le champ β est d'ordre un et on suppose que la fonction μ est au plus d'ordre un. L'analyse d'erreur pour l'approximation par éléments finis se fera en englobant ces paramètres dans les constantes génériques c.

7.2.3 Approximation par la méthode des moindres carrés

Soit V_h un espace d'approximation H^1-conforme, donc V-conforme ; par exemple, V_h peut être construit à partir d'un élément fini de Lagrange $\mathbb{P}_k$ ou $\mathbb{Q}_k$. En imposant à zéro la valeur aux nœuds du maillage situés sur $\partial\Omega^-$, on peut supposer que V_h est V_*-conforme. En général, l'approximation dans V_h du problème (7.32) par la méthode de Galerkin standard ne conduit pas à des résultats acceptables pour les mêmes raisons que celles évoquées dans la section 7.2.1.

La méthode des moindres carrés consiste à considérer le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ (Au_h, Aw_h)_{0,\Omega} = (f, Aw_h)_{0,\Omega}, \quad \forall w_h \in V_h. \end{cases} \tag{7.35}$$

Puisque l'opérateur $A : V_* \to L$ est un isomorphisme, il existe $\alpha > 0$ tel que pour tout $v \in V_*$, $\|Av\|_L \geqslant \alpha\|v\|_V$; voir la section A.1.4. Cette propriété implique que la forme bilinéaire $a' \in \mathcal{L}(V \times V; \mathbb{R})$ définie par

$$a'(v, w) = (Av, Aw)_{0,\Omega} = \int_\Omega (\mu v + \beta\cdot\nabla v)(\mu w + \beta\cdot\nabla w), \tag{7.36}$$

est V-coercive sur V_*. En effet, pour tout $v \in V_*$, on a

$$a'(v, v) = \|Av\|^2_{0,\Omega} \geqslant \alpha^2\|v\|^2_V. \tag{7.37}$$

Il découle immédiatement du lemme de Lax–Milgram que le problème discret (7.35) est bien posé. Par ailleurs, le lemme de Céa 2.12 conduit au résultat suivant.

Proposition 7.3. *Dans le cadre des hypothèses ci-dessus, il existe une constante c, indépendante de h, telle que*

$$\|u-u_h\|_V = \|u-u_h\|_{0,\Omega} + \|\beta\cdot\nabla(u-u_h)\|_{0,\Omega} \leqslant c \inf_{v_h\in V_h} \|u-v_h\|_V. \quad (7.38)$$

On suppose que l'espace V_h jouit de la propriété d'interpolation suivante : il existe un sous-espace Z dense dans V, un entier $k \geqslant 1$ et une constante c, indépendante de h, tels que pour tout $z \in Z$,

$$\inf_{v_h\in V_h} (\|z-v_h\|_{0,\Omega} + h\|\nabla(z-v_h)\|_{0,\Omega}) \leqslant c\, h^{k+1}\|z\|_Z. \quad (7.39)$$

Par exemple, lorsque V_h est construit à partir d'un élément fini de Lagrange de degré k et d'une famille régulière de maillages affines et conformes, l'estimation (7.39) est satisfaite pour $Z = H^{k+1}(\Omega)$.

Théorème 7.4. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution u de* (7.32) *est dans Z. Alors, il existe une constante c, indépendante de h, telle que*

$$\|u-u_h\|_{0,\Omega} + \|\beta\cdot\nabla(u-u_h)\|_{0,\Omega} \leqslant c\, h^k\|u\|_Z. \quad (7.40)$$

On observe que l'estimation d'erreur est sous-optimale d'un facteur 1 en norme L^2 et qu'elle est optimale dans la semi-norme liée à la dérivée advective[1]. L'estimation en norme L^2 ne peut être améliorée par les techniques de dualité utilisées dans le lemme de Aubin–Nitsche car les problèmes d'advection–réaction ne jouissent pas de propriétés régularisantes.

7.2.4 Approximation par la méthode de Galerkin/moindres carrés

Afin d'obtenir une estimation plus fine en norme L^2, on introduit la méthode de Galerkin/moindres carrés suivante. On choisit une constante $\gamma > 0$ indépendante de h et on pose

$$\delta(h) = \gamma h. \quad (7.41)$$

1. On rappelle que le terme optimal veut dire que l'erreur d'approximation dans une certaine norme est du même ordre en h que l'erreur d'interpolation dans cette même norme.

On introduit la forme bilinéaire $a_h \in \mathcal{L}(V \times V; \mathbb{R})$ définie par

$$a_h(v, w) = (Av, w)_{0,\Omega} + \delta(h)(Av, Aw)_{0,\Omega}, \tag{7.42}$$

et la forme bilinéaire $f_h \in V'$ définie par

$$f_h(w) = (f, w)_{0,\Omega} + \delta(h)(f, Aw)_{0,\Omega}.$$

On considère le problème discret suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{7.43}$$

On introduit les normes suivantes sur V_*,

$$\|v\|_{h,\mathrm{A}} = (Av, v)_{0,\Omega}^{\frac{1}{2}} + h^{\frac{1}{2}}\|Av\|_{0,\Omega}, \tag{7.44}$$

$$\|v\|_{h,\frac{1}{2}} = \|v\|_{h,\mathrm{A}} + h^{-\frac{1}{2}}\|v\|_{0,\Omega}. \tag{7.45}$$

On observera que la définition ci-dessus a bien un sens puisque, grâce à l'hypothèse (7.33), la forme bilinéaire a associée à l'opérateur A est L-coercive sur V_*, ce qui implique $(Av, v)_{0,\Omega} \geqslant 0$ pour $v \in V_*$. On montre les propriétés suivantes.

- **Stabilité.** La forme bilinéaire a_h définie en (7.42) est $\|\cdot\|_{h,\mathrm{A}}$-coercive sur V_h uniformément en h ; cette propriété implique que le problème approché (7.43) est bien posé.
- **Continuité.** Il existe une constante c, indépendante de h, telle que pour tout $(v, w) \in V_* \times V_*$,

$$a_h(v, w) \leqslant c\,\|v\|_{h,\frac{1}{2}}\|w\|_{h,\mathrm{A}}. \tag{7.46}$$

- **Consistance.** Pour tout $w_h \in V_h$,

$$a_h(u - u_h, w_h) = 0. \tag{7.47}$$

En procédant comme dans la preuve du deuxième lemme de Strang 2.20, on déduit l'estimation d'erreur suivante.

Proposition 7.5. *Il existe une constante c, indépendante de h, telle que*

$$\|u - u_h\|_{h,\mathrm{A}} \leqslant c \inf_{w_h \in V_h} \|u - w_h\|_{h,\frac{1}{2}}. \tag{7.48}$$

On suppose que la propriété (7.39) est satisfaite. On en déduit facilement que pour tout $z \in Z$,

$$\inf_{v_h \in V_h} \|z - v_h\|_{h,\frac{1}{2}} \leqslant c\, h^{k+\frac{1}{2}} \|z\|_Z. \tag{7.49}$$

Il en résulte le résultat de convergence suivant.

Théorème 7.6. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution u de* (7.32) *est dans Z. Alors, il existe une constante c, indépendante de h, telle que*

$$\|u - u_h\|_{0,\Omega} + h^{\frac{1}{2}} \|\beta \cdot \nabla (u - u_h)\|_{0,\Omega} \leqslant c\, h^{k+\frac{1}{2}} \|u\|_Z. \tag{7.50}$$

On observe que l'estimation d'erreur est sous-optimale d'un facteur $\frac{1}{2}$ en norme L^2 et qu'elle est optimale dans la semi-norme liée à la dérivée advective. L'amélioration de l'estimation d'erreur en norme L^2 est à l'origine de la popularité de la méthode de Galerkin/moindres carrés.

7.3 Advection–diffusion avec advection dominante

Cette section considère une équation d'advection–diffusion(–réaction) avec advection dominante comme prototype d'une perturbation singulière d'un problème d'ordre un. On analyse l'approximation de cette équation par la méthode de Galerkin/moindres carrés. L'approximation des équations d'advection–diffusion avec advection dominante par une méthode de type Galerkin/moindres carrés a été proposée par Brooks et Hughes [**24**] ; la méthode a été appelée « méthode SUPG » (de l'anglais, Streamline Upwind Petrov–Galerkin). L'analyse de convergence remonte aux travaux de Johnson, Nävert et Pitkäranta [**53**] qui ont proposé le terme de « méthode SD » (de l'anglais, Streamline Diffusion). L'analyse présentée ci-dessous a été introduite par Ern et Guermond [**38**, p. 244]. On pourra également consulter Johnson [**52**], Quarteroni et Valli [**61**] et Codina [**30**] pour des compléments.

7.3.1 Le problème modèle

On reprend les notations et les hypothèses de la section 7.2.2. On se donne en outre un réel $\epsilon > 0$ et on cherche une fonction $u : \Omega \to \mathbb{R}$ telle que

$$-\epsilon\Delta u + \beta\cdot\nabla u + \mu u = f \quad \text{dans } \Omega, \tag{7.51}$$

$$u = 0 \quad \text{sur } \partial\Omega. \tag{7.52}$$

Comme dans la section 7.2.2, l'analyse d'erreur se fera en englobant les paramètres β et μ dans les constantes génériques c. Par contre, on conservera explicitement le paramètre ϵ dans les estimations d'erreur car on souhaite construire une méthode numérique robuste lorsque $\epsilon \to 0$. Pour une maille $K \in \mathcal{T}_h$, le rapport $\frac{h_K}{\epsilon}$ s'interprète comme le *nombre de Péclet local*. L'hypothèse d'advection dominante signifie que sur au moins une partie des mailles, on a $\epsilon \ll h_K$, c'est-à-dire que le nombre de Péclet local est très grand.

La formulation faible du problème (7.51)–(7.52) rentre dans le cadre de (7.13) en posant $V = W = H^1_0(\Omega)$ et en introduisant la forme bilinéaire $a_\epsilon \in \mathcal{L}(V \times V; \mathbb{R})$ telle que

$$a_\epsilon(v, w) = \int_\Omega \epsilon\nabla v\cdot\nabla w + \int_\Omega (\mu v + \beta\cdot\nabla v)w, \tag{7.53}$$

et la forme linéaire $f \in V'$ telle que $f(w) = \int_\Omega fw$. Le problème modèle est le suivant :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a_\epsilon(u, w) = f(w), \quad \forall w \in V. \end{cases} \tag{7.54}$$

Proposition 7.7. *Sous l'hypothèse* (7.33)*, le problème* (7.54) *est bien posé.*

7.3.2 Approximation par la méthode de Galerkin/moindres carrés

Soit V_h un espace d'approximation H^1-conforme ; par exemple, V_h peut être construit à partir d'un élément fini de Lagrange et d'un maillage $\mathcal{T}_h$. Par la suite, on suppose que la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière. De plus, en imposant à zéro la valeur aux nœuds du maillage situés sur $\partial\Omega$, on peut supposer que V_h est V-conforme.

En général, l'approximation dans V_h du problème (7.13) par la méthode de Galerkin standard ne conduit pas à des résultats acceptables lorsque le nombre de Péclet local est trop grand car la solution approchée est polluée par des oscillations parasites. Ces oscillations peuvent être éliminées en raffinant le maillage là où le nombre de Péclet local est élevé. Toutefois, cette approche peut conduire à des coûts de calcul excessifs lorsque le paramètre ϵ est très petit. De plus, dans des problèmes non-linéaires où le nombre de Péclet local dépend de la solution, on ne sait pas *a priori* où celui-ci est grand. Dans ces conditions, il est préférable d'utiliser une méthode de type Galerkin/moindres carrés.

Dans la méthode de Galerkin/moindres carrés appliquée à l'approximation de l'équation d'advection–réaction, le paramètre de pondération δ ne dépend que de la finesse globale du maillage h ; voir (7.41). Pour l'équation d'advection–diffusion, ce paramètre dépend de la taille *locale* des mailles et du coefficient de diffusion ϵ. Par la suite, on considère la fonction suivante :

$$\forall K \in \mathcal{T}_h, \qquad \delta(h_K, \epsilon) = \left(\frac{1}{h_K} + \frac{\epsilon}{h_K^2} \right)^{-1}. \tag{7.55}$$

Plus généralement, l'analyse présentée ci-dessous reste valable pour toute fonction δ de (h_K, ϵ) satisfaisant les propriétés suivantes :

(i) il existe une constante c_1 telle que pour tout (h_K, ϵ), $\delta(h_K, \epsilon) \leqslant c_1 h_K$;

(ii) il existe une constante c_2 telle que pour tout (h_K, ϵ), $\delta(h_K, \epsilon) \leqslant c_2 h_K^2 \epsilon^{-1}$;

(iii) il existe des constantes c_3 et c_4 telles que pour tout (h_K, ϵ), $\epsilon \leqslant c_3 h_K$ implique $\delta(h_K, \epsilon) \geqslant c_4 h_K$.

Il est clair que la fonction δ proposée en (7.55) vérifie ces trois propriétés.

On désigne par $H^2(\mathcal{T}_h)$ le sous-espace de $L^2(\Omega)$ constitué des fonctions dont la restriction à chaque maille $K \in \mathcal{T}_h$ est dans $H^2(K)$. On pose $V(h) = H^2(\mathcal{T}_h) \cap H_0^1(\Omega)$. On introduit la forme bilinéaire $a_{h\epsilon} \in \mathcal{L}(V(h) \times V(h); \mathbb{R})$ définie par

$$a_{h\epsilon}(v, w) = a_\epsilon(v, w) + \sum_{K \in \mathcal{T}_h} \delta(h_K, \epsilon)(A_\epsilon v, A_\epsilon w)_{0,K}, \tag{7.56}$$

où la forme bilinéaire a_ϵ est définie en (7.53) et où pour $V(h)$, on a posé sur chaque maille $K \in \mathcal{T}_h$

$$A_\epsilon v = -\epsilon \Delta v + \beta \cdot \nabla v + \mu v. \tag{7.57}$$

On considère le problème discret suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_{h\epsilon}(u_h, w_h) = f_{h\epsilon}(w_h), \quad \forall w_h \in V_h, \end{cases} \tag{7.58}$$

avec la forme linéaire $f_{h\epsilon} \in V_h'$ définie par

$$f_{h\epsilon}(w_h) = (f, w_h)_{0,\Omega} + \sum_{K \in \mathcal{T}_h} \delta(h_K, \epsilon)(f, A_\epsilon w_h)_{0,K}.$$

On introduit les normes suivantes sur $V(h)$,

$$\|v\|_{h,A_\epsilon} = (Av, v)_{0,\Omega}^{\frac{1}{2}} + \epsilon^{\frac{1}{2}} |v|_{1,\Omega} + \left(\sum_{K \in \mathcal{T}_h} \delta(h_K, \epsilon) \|A_\epsilon v\|_{0,K}^2 \right)^{\frac{1}{2}}, \tag{7.59}$$

$$\|v\|_{h,\frac{1}{2}} = \|v\|_{h,A_\epsilon} + \left(\sum_{K \in \mathcal{T}_h} h_K^{-1} \|v\|_{0,K}^2 \right)^{\frac{1}{2}}. \tag{7.60}$$

On rappelle que l'opérateur A est défini en (7.30) et que compte tenu de l'hypothèse (7.33), on a pour tout $v \in V(h)$, $(Av, v)_{0,\Omega} \geqslant \mu_0 \|v\|_{0,\Omega}^2$. Dans le cadre des hypothèses ci-dessus, on montre les propriétés suivantes.

- **Stabilité.** La forme bilinéaire $a_{h\epsilon}$ définie en (7.56) est $\|\cdot\|_{h,A_\epsilon}$-coercive sur $V(h)$ uniformément en h et en ϵ ; cette propriété implique que le problème approché (7.58) est bien posé.

- **Continuité.** Il existe une constante c, indépendante de h et de ϵ, telle que pour tout $(v, w_h) \in V(h) \times V_h$,

$$a_{h\epsilon}(v, w_h) \leqslant c \, \|v\|_{h,\frac{1}{2}} \|w_h\|_{h,A_\epsilon}. \tag{7.61}$$

- **Consistance.** Pour tout $w_h \in V_h$,

$$a_{h\epsilon}(u - u_h, w_h) = 0. \tag{7.62}$$

En procédant comme dans la preuve du deuxième lemme de Strang 2.20, on déduit l'estimation d'erreur suivante.

Proposition 7.8. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution u de* (7.54) *est dans V(h). Alors, il existe une constante c, indépendante de h et de* ϵ*, telle que*

$$\|u - u_h\|_{h,A_\epsilon} \leqslant c \inf_{w_h \in V_h} \|u - w_h\|_{h,\frac{1}{2}}. \tag{7.63}$$

On suppose que l'espace V_h jouit de la propriété d'interpolation suivante : il existe un sous-espace Z dense dans $H^1(\Omega)$, un entier $k \geqslant 1$ et une constante c indépendante de h tels que pour tout $z \in Z$ et pour tout $K \in \mathcal{T}_h$,

$$\inf_{v_h \in V_h} (\|z - v_h\|_{0,K} + h_K \|\nabla(z - v_h)\|_{0,K} + h_K^2 \|\Delta(z - v_h)\|_{0,K}) \leqslant c\, h_K^{k+1} \|z\|_{Z,K}, \tag{7.64}$$

où $\|\cdot\|_{Z,K}$ désigne la norme de Z localisée sur K. Par exemple, lorsque V_h est construit à partir d'un élément fini de Lagrange de degré k et d'une famille régulière de maillages affines et conformes, l'estimation (7.64) est satisfaite pour $Z = H^{k+1}(\Omega)$ et $\|\cdot\|_{Z,K} = \|\cdot\|_{k+1,K}$. On déduit facilement de (7.64) que pour tout $z \in Z$,

$$\inf_{v_h \in V_h} \|z - v_h\|_{h,\frac{1}{2}} \leqslant c \left(\sum_{K \in \mathcal{T}_h} h_K^{2k} (h_K + \epsilon) \|z\|_{Z,K}^2 \right)^{\frac{1}{2}}. \tag{7.65}$$

Il en résulte le résultat de convergence suivant.

Théorème 7.9. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution u de* (7.54) *est dans Z. Alors, il existe une constante c, indépendante de h et de* ϵ*, telle que*

$$\|u - u_h\|_{h,A_\epsilon} \leqslant c \left(\sum_{K \in \mathcal{T}_h} h_K^{2k} (h_K + \epsilon) \|z\|_{Z,K}^2 \right)^{\frac{1}{2}}. \tag{7.66}$$

Corollaire 7.10. *Pour tout* $\gamma > 0$*, il existe une constante* c_γ*, indépendante de h et de* ϵ*, telle que :*

(i) *si* $\epsilon \geqslant \gamma h$,

$$\|u - u_h\|_{1,\Omega} \leqslant c_\gamma \left(\sum_{K \in \mathcal{T}_h} h_K^{2k} \|u\|_{Z,K}^2 \right)^{\frac{1}{2}}; \tag{7.67}$$

(ii) *si* $\epsilon \leqslant \gamma \min_{K \in \mathcal{T}_h} \{h_K\}$,

$$\left(\sum_{K \in \mathcal{T}_h} h_K \|\mathcal{A}(u - u_h)\|_{0,K}^2 \right)^{\frac{1}{2}} \leqslant c_\gamma \left(\sum_{K \in \mathcal{T}_h} h_K^{2k+1} \|u\|_{Z,K}^2 \right)^{\frac{1}{2}}. \tag{7.68}$$

De plus, si la famille $\{\mathcal{T}_h\}_{h>0}$ *est quasi-uniforme, pour tout* $\gamma > 0$, *il existe une constante* c_γ, *indépendante de* h *et de* ϵ, *telle que :*

(i) *si* $\epsilon \geqslant \gamma h$, $\|\nabla(u - u_h)\|_{0,\Omega} \leqslant c_\gamma h^k \|u\|_Z$;

(ii) *si* $\epsilon \leqslant \gamma h$, $\|\beta \cdot \nabla(u - u_h)\|_{0,\Omega} \leqslant c_\gamma h^k \|u\|_Z$.

Corollaire 7.11. *Il existe une constante* c, *indépendante de* h *et de* ϵ, *telle que*

$$\|u - u_h\|_{0,\Omega} \leqslant c \left(\sum_{K \in \mathcal{T}_h} h_K^{2k} (h_K + \epsilon) \|z\|_{Z,K}^2 \right)^{\frac{1}{2}}. \tag{7.69}$$

De plus, pour tout $\gamma > 0$, *il existe une constante* c_γ, *indépendante de* h *et de* ϵ, *telle que si* $\epsilon \leqslant \gamma h$,

$$\|u - u_h\|_{0,\Omega} \leqslant c_\gamma h^{k+\frac{1}{2}} \|u\|_Z. \tag{7.70}$$

Le corollaire 7.10 signifie que lorsque les nombres de Péclet locaux sont suffisamment petits (régime dit de diffusion dominante), l'approximation par la méthode de Galerkin/moindres carrés fournit un contrôle optimal du gradient de l'erreur. Par contre, lorsque les nombres de Péclet sont grands (régime dit d'advection dominante), seules sont contrôlées les dérivées de l'erreur le long des lignes de courant. Le corollaire 7.11 signifie qu'en régime d'advection dominante, la méthode de Galerkin/moindres carrés reste relativement précise en norme L^2 car elle n'est sous-optimale que d'un facteur $\frac{1}{2}$.

7.4 Problème de Stokes

Cette section est consacrée à l'approximation du problème de Stokes (6.21) par une méthode de Galerkin/moindres carrés. L'intérêt de l'approche réside dans le fait qu'elle permet d'utiliser des éléments finis pour la vitesse et la pression qui ne satisfont pas la condition de compatibilité (6.26). Ces techniques ont été introduites et analysées par Franca et Frey [**40**] et par Tobiska et Verfürth [**73**]. L'analyse présentée ci-dessous a été introduite par Ern et Guermond [**38**, p. 201].

On introduit l'espace fonctionnel $V = [H_0^1(\Omega)]^d \times L_*^2(\Omega)$. On peut formuler le problème de Stokes de la manière suivante :

$$\begin{cases} \text{Chercher } (u,p) \in V \text{ tel que} \\ a((u,p),(v,q)) = \zeta(v,q), \quad \forall (v,q) \in V, \end{cases} \tag{7.71}$$

où la forme bilinéaire $a \in \mathcal{L}(V \times V; \mathbb{R})$ est définie par

$$a((u,p),(v,q)) = (\nabla u, \nabla v)_{0,\Omega} - (\nabla\cdot v, p)_{0,\Omega} + (\nabla\cdot u, q)_{0,\Omega}, \tag{7.72}$$

et la forme linéaire $\zeta \in V'$ par $\zeta(v,q) = \int_\Omega [f\cdot v + gq]$, $f \in [L^2(\Omega)]^d$ et $g \in L^2(\Omega)$ étant les données du problème de Stokes (6.1)–(6.2).

On choisit une constante $\gamma > 0$ indépendante de h. Pour tout $K \in \mathcal{T}_h$, on pose

$$\delta(h_K) = \gamma h_K^2. \tag{7.73}$$

On considère un espace d'approximation pour la vitesse, noté X_h et supposé $[H_0^1(\Omega)]^d$-conforme, et un espace d'approximation pour la pression, noté M_h et supposé $L_*^2(\Omega)$-conforme. On pose $V_h = X_h \times M_h$ et

$$V(h) = ([H^2(\mathcal{T}_h)]^d \cap [H_0^1(\Omega)]^d) \times (H^1(\mathcal{T}_h) \cap L_*^2(\Omega)), \tag{7.74}$$

où $H^m(\mathcal{T}_h)$, $m \in \{1,2\}$, désigne le sous-espace de $L^2(\Omega)$ constitué des fonctions dont la restriction à chaque maille $K \in \mathcal{T}_h$ est dans $H^m(K)$. On introduit la forme bilinéaire $a_h \in \mathcal{L}(V(h) \times V(h); \mathbb{R})$ définie par

$$\begin{aligned} a_h((u_h,p_h),(v_h,q_h)) &= (\nabla u_h, \nabla v_h)_{0,\Omega} - (\nabla\cdot v_h, p_h)_{0,\Omega} + (\nabla\cdot u_h, q_h)_{0,\Omega} \\ &\quad + \sum_{K\in\mathcal{T}_h} \delta(h_K)\big((\nabla p_h - \Delta u_h, \nabla q_h - \Delta v_h)_{0,K} + (\nabla\cdot u_h, \nabla\cdot v_h)_{0,K}\big). \end{aligned} \tag{7.75}$$

On introduit la forme linéaire $\zeta_h \in V(h)'$ définie par

$$\zeta_h(v_h, q_h) = \zeta(v_h, q_h) + \sum_{K \in \mathcal{T}_h} \delta(h_K)((f, g), (\nabla q_h - \Delta v_h, \nabla \cdot v_h))_{0,K}, \quad (7.76)$$

et on considère le problème approché suivant :

$$\begin{cases} \text{Chercher } (u_h, p_h) \in V_h \text{ tel que} \\ a_h((u_h, p_h), (v_h, q_h)) = \zeta_h(v_h, q_h), \quad \forall (v_h, q_h) \in V_h. \end{cases} \quad (7.77)$$

On introduit la norme suivante sur $V(h)$,

$$\|(v, q)\|_{h,\mathrm{A}} = |v|_{1,\Omega} + \|q\|_{0,\Omega} + \left(\sum_{K \in \mathcal{T}_h} \delta(h_K)(\|\nabla q\|_{0,K}^2 + \|\nabla q - \Delta v\|_{0,K}^2 + \|\nabla \cdot v\|_{0,K}^2) \right)^{\frac{1}{2}}. \quad (7.78)$$

On suppose que les inégalités suivantes sont vérifiées : il existe une constante c, indépendante de h, telle que pour tout $v_h \in X_h$ et pour tout $K \in \mathcal{T}_h$,[1]

$$\|\Delta v_h\|_{0,K} \leqslant c\, h_K^{-1} \|\nabla v_h\|_{0,K}, \quad (7.79)$$

$$\|\nabla v_h\|_{0,K} \leqslant c\, h_K^{-1} \|v_h\|_{0,K}. \quad (7.80)$$

Dans le cadre des hypothèses ci-dessus, on montre les propriétés suivantes.

- **Stabilité.** La forme bilinéaire a_h définie en (7.75) satisfait la condition inf-sup suivante : il existe $\alpha > 0$, indépendant de h, tel que

$$\inf_{(v_h, q_h) \in V_h} \sup_{(w_h, r_h) \in V_h} \frac{a_h((v_h, q_h), (w_h, r_h))}{\|(v_h, q_h)\|_{h,\mathrm{A}} \|(w_h, r_h)\|_{h,\mathrm{A}}} \geqslant \alpha. \quad (7.81)$$

  Cette propriété implique que le problème approché (7.77) est bien posé.

1. Les inégalités (7.79) et (7.80) sont satisfaites si l'espace d'approximation X_h est construit à partir d'une famille régulière de maillages.

- **Continuité.** Il existe une constante c, indépendante de h, telle que pour tout $((v,q),(w_h,r_h)) \in V(h) \times V_h$,

$$a_h((v,q),(w_h,r_h)) \leqslant c\,\|(v,q)\|_{h,\mathrm{A}}\|(w_h,r_h)\|_{h,\mathrm{A}}. \tag{7.82}$$

- **Consistance.** Pour tout $(v_h,q_h) \in V_h$,

$$a_h((u-u_h,p-p_h),(v_h,q_h)) = 0. \tag{7.83}$$

On déduit des propriétés ci-dessus l'estimation d'erreur suivante.

Proposition 7.12. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution* (u,p) *de* (7.71) *est dans* $V(h)$*. Alors, il existe une constante* c*, indépendante de* h*, telle que*

$$\|(u-u_h,p-p_h)\|_{h,\mathrm{A}} \leqslant c \inf_{(v_h,q_h)\in V_h} \|(u-v_h,p-q_h)\|_{h,\mathrm{A}}. \tag{7.84}$$

On suppose que les espaces X_h et M_h jouissent des propriétés d'interpolation suivantes : il existe un entier $k \geqslant 1$ et une constante c, indépendante de h, tels que pour tout $(z_u,z_p) \in H^{k+1}(\Omega) \times H^k(\Omega)$, on a pour tout $K \in \mathcal{T}_h$,

$$\begin{aligned} \inf_{v_h\in X_h}(&\|z_u-v_h\|_{0,K} + h_K\|\nabla(z_u-v_h)\|_{0,K} \\ &+ h_K^2\|\Delta(z_u-v_h)\|_{0,K}) \leqslant c\,h_K^{k+1}\|z_u\|_{k+1,K}, \end{aligned} \tag{7.85}$$

$$\inf_{q_h\in M_h}(\|z_p-q_h\|_{0,K} + h_K\|\nabla(z_p-q_h)\|_{0,K}) \leqslant c\,h_K^k\|z_p\|_{k,K}. \tag{7.86}$$

On déduit facilement de (7.85) que pour tout $(z_u,z_p) \in H^{k+1}(\Omega) \times H^k(\Omega)$,

$$\inf_{(v_h,q_h)\in V_h} \|(z_u-v_h,z_p-q_h)\|_{h,\mathrm{A}} \leqslant c\,h^k(\|z_u\|_{k+1,\Omega} + \|z_p\|_{k,\Omega}). \tag{7.87}$$

Il en résulte le résultat de convergence suivant.

Théorème 7.13. *Dans le cadre des hypothèses ci-dessus, on suppose que la solution* (u,p) *de* (7.71) *est dans* $H^{k+1}(\Omega) \times H^k(\Omega)$*. Alors, il existe une constante* c*, indépendante de* h*, telle que*

$$\|u-u_h\|_{1,\Omega} + \|p-p_h\|_{0,\Omega} \leqslant c\,h^k(\|u\|_{k+1,\Omega} + \|p\|_{k,\Omega}). \tag{7.88}$$

De plus, si le problème de Stokes est régularisant (voir la définition 6.9*),*

$$\|u-u_h\|_{0,\Omega} \leqslant c\,h^{k+1}(\|u\|_{k+1,\Omega} + \|p\|_{k,\Omega}). \tag{7.89}$$

7.5 Complément : viscosité de sous-maille

La technique de viscosité de sous-maille a été introduite par Guermond [**49, 50**]. Elle permet de stabiliser la méthode de Galerkin standard lorsque celle-ci est utilisée afin d'approcher par éléments finis des opérateurs linéaires monotones. Pour simplifier, on restreint la présentation à l'opérateur A défini en (7.30) sous l'hypothèse (7.33). Le principe consiste à :

(i) enrichir l'espace d'approximation V_h sous la forme

$$V_h^{\mathrm{e}} = V_h \oplus V_h^{\mathrm{f}}, \tag{7.90}$$

où V_h^{f} est appelé *l'espace des échelles fluctuantes* et V_h^{e} *l'espace d'approximation enrichi* ;

(ii) considérer un problème discret posé sur l'espace d'approximation enrichi V_h^{e} et faisant intervenir une forme bilinéaire qui est la somme de la forme bilinéaire associée à la méthode de Galerkin standard et d'un terme de pénalisation (au sens des moindres carrés) dont le rôle est de contrôler les échelles fluctuantes de la solution discrète.

On reprend le problème modèle (7.32) de la section 7.2. Soit V_h l'espace d'approximation V_*-conforme considéré dans le problème discret (7.43). On suppose qu'on dispose d'un espace des échelles fluctuantes, V_h^{f}, qui est V_*-conforme et en somme directe avec V_h. Par la suite, on décompose les fonctions de V_h^{e} sous la forme $v_h^{\mathrm{e}} = v_h^{\mathrm{f}} + v_h$ avec $v_h^{\mathrm{f}} \in V_h^{\mathrm{f}}$ et $v_h \in V_h$. On fait les hypothèses suivantes.

(i) La décomposition (7.90) est L^2-stable au sens suivant : en notant $P_h \in \mathcal{L}(V_h^{\mathrm{e}}; V_h)$ le projecteur tel que

$$P_h : V_h^{\mathrm{e}} \ni v_h^{\mathrm{e}} \longmapsto v_h \in V_h, \tag{7.91}$$

il existe une constante c_s, indépendante de h, telle que

$$\forall v_h^{\mathrm{e}} \in V_h^{\mathrm{e}}, \qquad \|P_h v_h^{\mathrm{e}}\|_{0,\Omega} \leqslant c_s \|v_h^{\mathrm{e}}\|_{0,\Omega}. \tag{7.92}$$

(ii) La condition inf-sup suivante est satisfaite : il existe une constante $c_a > 0$, indépendante de h, telle que

$$\inf_{v_h \in V_h} \sup_{w_h^e \in V_h^e} \frac{a(v_h, w_h^e)}{\|v_h\|_V \|w_h^e\|_{0,\Omega}} \geqslant c_a, \tag{7.93}$$

où la forme bilinéaire a est définie en (7.31) et la norme $\|\cdot\|_V$ en (7.29).

(iii) L'inégalité suivante est satisfaite sur l'espace d'approximation enrichi V_h^e : il existe une constante c_i, indépendante de h, telle que

$$\forall v_h^e \in V_h^e, \qquad \|v_h^e\|_V \leqslant c_i h^{-1} \|v_h^e\|_{0,\Omega}. \tag{7.94}$$

Par ailleurs, on choisit un réel $c_b > 0$ et on introduit la forme bilinéaire $b_h \in \mathcal{L}(V_h^e \times V_h^e; \mathbb{R})$ telle que

$$b_h(v_h^e, w_h^e) = c_b h (\nabla v_h^f, \nabla w_h^f)_{0,\Omega}. \tag{7.95}$$

On considère le problème discret suivant :

$$\begin{cases} \text{Chercher } u_h^e \in V_h^e \text{ tel que} \\ a(u_h^e, w_h^e) + b_h(u_h^e, w_h^e) = f(w_h^e), \quad \forall w_h^e \in V_h^e. \end{cases} \tag{7.96}$$

On observera que la forme bilinéaire b_h réalise une pénalisation au sens des moindres carrés du gradient de la composante aux échelles fluctuantes de la solution discrète u_h^e. Par la suite, on note θ_h la forme bilinéaire $a + b_h$.

On considère les normes $\|\cdot\|_{h,\mathrm{A}}$ et $\|\cdot\|_{h,\frac{1}{2}}$ définies en (7.44) et (7.45), respectivement. Dans le cadre des hypothèses (7.92), (7.93) et (7.94), on montre les propriétés suivantes.

- **Stabilité.** Il existe une constante α, indépendante de h, telle que

$$\inf_{v_h^e \in V_h^e} \sup_{w_h^e \in V_h^e} \frac{\theta_h(v_h^e, w_h^e)}{\|v_h^e\|_{h,\mathrm{A}} \|w_h^e\|_{h,\mathrm{A}}} \geqslant \alpha. \tag{7.97}$$

  Cette propriété implique que le problème (7.96) est bien posé.

- **Continuité.** Il existe une constante c, indépendante de h, telle que pour tout $(v, w) \in V_* \times V_*$, on a $a_h(v, w) \leqslant c \|v\|_{h,\frac{1}{2}} \|w\|_{h,\mathrm{A}}$.

- **Consistance.** Pour tout $w_h^e \in V_h^e$, on a $\theta_h(u_h^e, w_h^e) = a(u, w_h^e)$.

En procédant comme dans la preuve du deuxième lemme de Strang 2.20, on déduit l'estimation d'erreur suivante.

Proposition 7.14. *Il existe une constante c, indépendante de h, telle que*

$$\|u - u_h^{\mathrm{e}}\|_{h,\mathrm{A}} \leqslant c \inf_{v_h \in V_h} \|u - v_h\|_{h,\frac{1}{2}}. \tag{7.98}$$

En supposant que la propriété (7.39) est satisfaite et que la solution u de (7.32) est dans Z, on en déduit

$$\|u - u_h\|_{0,\Omega} + h^{\frac{1}{2}} \|\beta \cdot \nabla(u - u_h)\|_{0,\Omega} \leqslant c\, h^{k+\frac{1}{2}} \|u\|_Z, \tag{7.99}$$

pour une constante c indépendante de h. On retrouve la même estimation d'erreur que pour la méthode de Galerkin/moindres carrés classique ; voir l'estimation (7.50). La technique de stabilisation par viscosité de sous-maille s'étend aux équations d'advection–diffusion avec advection dominante. On obtient des estimations d'erreur analogues à celles de la section 7.3.2. L'avantage de l'approche par viscosité de sous-maille est que le paramètre c_h est indépendant du coefficient de diffusion ϵ alors que le paramètre δ intervenant dans la méthode de Galerkin/moindres carrés présentée dans la section 7.3 dépend de ϵ. Un autre avantage de l'approche par viscosité de sous-maille est que cette technique s'étend de manière naturelle aux problèmes instationnaires alors que la méthode de Galerkin/moindres carrés nécessite de considérer des éléments finis espace-temps discontinus en temps ; voir, par exemple, [**38**, p. 321] pour l'analyse de la méthode de viscosité de sous-maille instationnaire.

Pour conclure cette section, on décrit quelques exemples d'espaces d'échelles fluctuantes admissibles lorsque l'espace d'approximation V_h est associé à l'élément fini de Lagrange $\mathbb{P}_1$ ou $\mathbb{P}_2$.

(i) Lorsque $V_h = P^1_{\mathrm{c},h} \cap V_*$, l'espace des échelles fluctuantes peut être engendré par des fonctions bulle sur les mailles ; voir la définition 6.10 et la formule (6.32). On pose

$$V_h^{\mathrm{f}} = \{v_h^{\mathrm{f}} \in \mathcal{C}^0(\overline{\Omega}) \,;\, \forall K \in \mathcal{T}_h,\ v_h^{\mathrm{f}} \circ T_K \in \mathrm{vect}(\widehat{b})\}. \tag{7.100}$$

Une autre possibilité pour l'espace des échelles fluctuantes consiste à poser

$$V_h^{\mathrm{f}} = \{v_h^{\mathrm{f}} \in \mathcal{C}^0(\overline{\Omega}) \,;\, \forall K \in \mathcal{T}_h,\ v_h^{\mathrm{f}}|_K \in \mathbb{P}_{1\text{-iso-}2}(K)\}, \tag{7.101}$$

où l'espace $\mathbb{P}_{1\text{-iso-}2}(K)$ est défini de la manière suivante : on divise K en quatre sous-triangles en reliant les milieux de ses trois arêtes ; une fonction est dans $\mathbb{P}_{1\text{-iso-}2}(K)$ si elle est continue et affine par morceaux dans K et si elle s'annule aux trois sommets de K. Avec ce choix, l'espace d'approximation enrichi est celui associé à l'élément fini de Lagrange $\mathbb{P}_1$ sur le maillage $\mathcal{T}_{\frac{h}{2}}$. Les figures 7.1(a) et 7.1(b) contiennent une représentation conventionnelle des degrés de liberté.

(ii) Une construction analogue est possible lorsque $V_h = P^2_{c,h} \cap V_*$. On peut soit utiliser trois bulles par élément pour définir l'espace des échelles fluctuantes soit subdiviser chaque triangle en quatre sous-triangles de sorte que l'espace d'approximation enrichi est celui associé à l'élément fini de Lagrange $\mathbb{P}_2$ sur le maillage $\mathcal{T}_{\frac{h}{2}}$. Les figures 7.1(c) et 7.1(d) contiennent une représentation conventionnelle des degrés de liberté.

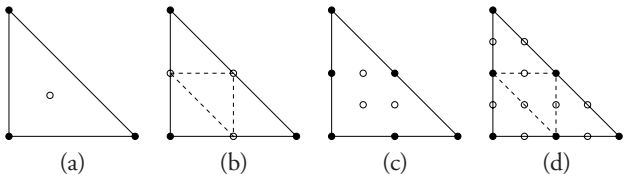


Figure 7.1 – Espaces des échelles fluctuantes permettant d'enrichir les espaces d'approximation associés aux éléments finis de Lagrange : (a) élément fini de Lagrange $\mathbb{P}_1$ enrichi par une bulle sur chaque triangle ; (b) élément fini de Lagrange $\mathbb{P}_1$ enrichi par subdivision de l'élément ; (c) élément fini de Lagrange $\mathbb{P}_2$ enrichi par trois bulles sur chaque triangle ; (d) élément fini de Lagrange $\mathbb{P}_2$ enrichi par subdivision de l'élément. Les degrés de liberté dans l'espace V_h sont représentés par des cercles noirs et ceux associés aux échelles fluctuantes par des cercles blancs.

Pour les choix ci-dessus, on montre, sous des hypothèses raisonnables sur le champ d'advection β, que les hypothèses (7.92), (7.93) et (7.94) sont satisfaites [**49**].

8 • ESTIMATION D'ERREUR *A POSTERIORI*

L'objectif de ce chapitre est d'estimer l'erreur entre la solution exacte d'un problème modèle et son approximation par éléments finis, et ce uniquement en fonction de quantités accessibles au calcul, c'est-à-dire la solution approchée, les données du problème modèle et le maillage. De telles estimations sont appelées estimations d'erreur *a posteriori* et se distinguent des estimations d'erreur *a priori* présentées dans les chapitres précédents par le fait que ces dernières font également intervenir la solution exacte (qui n'est pas, en général, accessible au calcul). Depuis les travaux pionniers de Babuška et Rheinbolt [**10**], l'intérêt pour les estimations d'erreur *a posteriori* s'est considérablement accru, d'une part parce que ces techniques permettent d'estimer explicitement l'erreur d'approximation et d'autre part parce qu'elles sont à l'origine des algorithmes de raffinement adaptatif du maillage qui permettent très souvent de réduire substantiellement le coût des calculs. Trois types d'estimateurs d'erreur *a posteriori* sont présentés dans ce chapitre : les estimateurs par résidu, les estimateurs par dualité et les estimateurs hiérarchiques.

8.1 Cadre général

Cette section précise le cadre dans lequel on se place pour étudier les différentes techniques d'estimation d'erreur *a posteriori*. On introduit également deux notions importantes relatives aux estimations d'erreur *a posteriori* : la fiabilité et l'optimalité.

8.1.1 Le problème modèle

On considère le problème modèle suivant :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u,w) = f(w), \quad \forall w \in W, \end{cases} \tag{8.1}$$

où V et W sont des espaces de Hilbert (dont les éléments sont des fonctions définies sur un domaine Ω de $\mathbb{R}^d$ et à valeurs dans $\mathbb{R}$), a est une forme bilinéaire dans $\mathcal{L}(V \times W; \mathbb{R})$ et f est une forme linéaire dans W'. On suppose que la forme bilinéaire a satisfait les conditions (BNB1) et (BNB2) du théorème BNB si bien que le problème (8.1) est bien posé ; voir le théorème 2.4. De plus, on suppose que le membre de droite dans (8.1) s'écrit sous la forme $\int_\Omega fw$ pour une fonction $f \in L^2(\Omega)$; on abuse des notations en utilisant le même symbole pour la fonction f et la forme linéaire dans W'.

Pour fixer les idées, on pourra supposer que (8.1) est la formulation faible d'un problème d'advection–diffusion–réaction. On a donc :

(i) $V = W = H_0^1(\Omega)$ si on impose des conditions aux limites de Dirichlet homogènes et $V = W = H^1(\Omega)$ si on impose des conditions aux limites de Neumann ;

(ii) pour $(v,w) \in V \times V$, $a(v,w) = \int_\Omega [\nabla w \cdot \sigma \cdot \nabla v + w(\beta \cdot \nabla v) + \mu wv]$ où $\sigma \in [L^\infty(\Omega)]^{d,d}$, $\beta \in [L^\infty(\Omega)]^d$ avec $\nabla \cdot \beta \in L^\infty(\Omega)$, et $\mu \in L^\infty(\Omega)$.

Si on suppose qu'il existe un réel $\sigma_0 > 0$ tel que pour tout $\xi \in \mathbb{R}^d$, $\sum_{i,j=1}^d \sigma_{ij}\xi_i\xi_j \geqslant \sigma_0 \sum_{i=1}^d \xi_i^2$ presque partout dans Ω et qu'il existe un réel $\mu_0 > 0$ tel que $\mu - \frac{1}{2}\nabla \cdot \beta \geqslant \mu_0$ presque partout dans Ω, la forme bilinéaire a est coercive si bien que le problème (8.1) est bien posé. On pourra consulter la section 5.1.5 pour des compléments.

La forme bilinéaire a est associée à un opérateur différentiel $\mathcal{L}$ tel que pour tout $v \in V$ et pour toute fonction $w \in \mathcal{D}(\Omega)$, on a

$$a(v,w) = \langle \mathcal{L}v, w \rangle_{\mathcal{D}',\mathcal{D}}, \tag{8.2}$$

au sens des distributions. Par exemple, pour un problème d'advection–diffusion–réaction, on a

$$\mathcal{L}v = -\nabla \cdot (\sigma \cdot \nabla v) + \beta \cdot \nabla v + \mu v. \tag{8.3}$$

Pour simplifier, on se restreint dans ce chapitre à des approximations *conformes* du problème modèle (8.1) ; voir, par exemple, les références [**35, 68**] pour des estimations d'erreur *a posteriori* dans un cadre non-conforme. On considère donc deux espaces d'approximation V_h et W_h tels que $V_h \subset V$ et $W_h \subset W$. On suppose que les espaces V_h et W_h sont construits à l'aide des techniques présentées dans les chapitres 3 et 4, c'est-à-dire à partir d'une famille régulière de maillages $\{\mathcal{T}_h\}_{h>0}$ et d'un élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$. On considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a(u_h, w_h) = f(w_h), \quad \forall w_h \in W_h. \end{cases} \tag{8.4}$$

Par la suite, on suppose que le problème (8.4) est bien posé. Pour simplifier, on suppose que la même forme bilinéaire a et la même forme linéaire f interviennent dans le problème modèle (8.1) et son approximation (8.4). Ceci implique la propriété de consistance suivante :

$$\forall w_h \in W_h, \qquad a(u - u_h, w_h) = 0. \tag{8.5}$$

Les espaces V et W étant des espaces fonctionnels, leurs normes respectives $\|\cdot\|_V$ et $\|\cdot\|_W$ peuvent être localisées sur les éléments du maillage sous la forme

$$\|\cdot\|_V = \left(\sum_{K \in \mathcal{T}_h} \|\cdot\|_{V,K}^2 \right)^{\frac{1}{2}} \quad \text{et} \quad \|\cdot\|_W = \left(\sum_{K \in \mathcal{T}_h} \|\cdot\|_{W,K}^2 \right)^{\frac{1}{2}}. \tag{8.6}$$

Par exemple, pour $V = H^1(\Omega)$, on a $\|\cdot\|_V = \|\cdot\|_{1,\Omega}$ et $\|\cdot\|_{V,K} = \|\cdot\|_{1,K}$.

8.1.2 Indicateurs d'erreur, fiabilité et optimalité

Soit η une fonctionnelle de $(u_h, \mathcal{T}_h, a, f)$, c'est-à-dire de la solution approchée u_h, du maillage $\mathcal{T}_h$ et des données du problème modèle (représentées par la forme bilinéaire a et la forme linéaire f). On observera que les valeurs de la fonctionnelle η sont accessibles au calcul.

Définition 8.1. *On dit que la fonctionnelle η est un estimateur d'erreur a posteriori s'il existe une constante c, indépendante de h, telle que*

$$\|u - u_h\|_V \leqslant c\, \eta(u_h, \mathcal{T}_h, a, f). \tag{8.7}$$

De plus, si la fonctionnelle η *peut être localisée sous la forme*

$$\eta(u_h, \mathcal{T}_h, a, f) = \left(\sum_{K \in \mathcal{T}_h} \eta_K(u_h, a, f)^2 \right)^{\frac{1}{2}}, \tag{8.8}$$

on dit que les fonctionnelles $\{\eta_K\}_{K \in \mathcal{T}_h}$ *sont des* indicateurs d'erreur locaux.

L'estimation (8.7) s'interprète comme une propriété de *fiabilité* puisqu'elle garantit que l'erreur $\|u - u_h\|_V$ est effectivement contrôlée par l'estimateur d'erreur *a posteriori*.

Les indicateurs d'erreur locaux fournissent l'ingrédient principal dans une stratégie de raffinement adaptatif du maillage. Le principe consiste d'une part à raffiner les mailles $K \in \mathcal{T}_h$ où l'indicateur d'erreur local η_K est grand et d'autre part à regrouper des mailles adjacentes dans lesquelles l'indicateur d'erreur local est petit afin de former une nouvelle maille de taille plus grande. Cette procédure peut être répétée plusieurs fois et conduire (dans une certaine mesure) à optimiser le coût des calculs en fonction du niveau d'erreur atteint. On peut se fixer divers objectifs pour l'adaptation du maillage, par exemple celui d'équirépartir les indicateurs d'erreur locaux sur toutes les mailles (uniformément ou avec une certaine pondération) ou encore celui d'amener l'estimateur d'erreur global η en dessous d'une certaine tolérance.

Pour qu'une procédure de raffinement adaptatif du maillage soit efficace, il faut que les indicateurs d'erreur locaux jouissent d'une certaine propriété d'optimalité. Cette propriété garantit que localement sur chaque maille, l'indicateur d'erreur local est asymptotiquement équivalent à l'erreur réelle. En d'autres termes, on souhaite qu'il existe des constantes $c_1 > 0$ et c_2, indépendantes de h, telles que pour tout $K \in \mathcal{T}_h$,

$$c_1 \eta_K(u_h, a, f) \leqslant \|u - u_h\|_{V,K} \leqslant c_2 \eta_K(u_h, a, f). \tag{8.9}$$

Dans la plupart des problèmes modèles, il est impossible d'obtenir ce type de propriété. On dispose de la majoration globale (8.7) et d'une minoration étendue sous la forme suivante : il existe une constante $c_1 > 0$, indépendante de h, telles que pour tout $K \in \mathcal{T}_h$,

$$c_1 \eta_K(u_h, a, f) \leqslant \|u - u_h\|_{V,\Delta_K} + \Pi(\Delta_K, a, f), \tag{8.10}$$

où Δ_K est un *macro-élément* centré sur la maille K, c'est-à-dire un ensemble de mailles dont l'intersection avec K est non-vide, et $\Pi(\Delta_K, a, f)$ est un terme de perturbation qui est soit négligeable devant l'erreur locale $\|u - u_h\|_{V,\Delta_K}$ soit du même ordre que celle-ci asymptotiquement en h. De plus, on a conventionnellement noté

$$\|\cdot\|_{V,\Delta_K} = \left(\sum_{K' \in \Delta_K} \|\cdot\|^2_{V,K'} \right)^{\frac{1}{2}}. \tag{8.11}$$

L'estimation (8.10) s'interprète comme une propriété *d'optimalité* puisqu'elle garantit que localement sur le macro-élément Δ_K, l'indicateur d'erreur local η_K ne « surévalue pas trop » l'erreur réelle. La figure 8.1 illustre deux exemples de macro-éléments Δ_K en dimension 2 : l'ensemble des mailles (incluant K) partageant un côté avec K et l'ensemble des mailles (incluant K) partageant un sommet avec K.

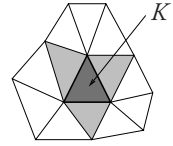


Figure 8.1 – Exemples de macro-éléments Δ_K centrés sur la maille K en dimension 2 ; la maille K est en gris foncé ; l'ensemble des éléments partageant une face avec K est constitué de K et des trois triangles en gris clair ; l'ensemble des éléments partageant un sommet avec K est constitué du macro-élément précédent et des neuf triangles en blanc.

8.2 Estimateurs par résidu

Cette section présente une introduction aux techniques d'estimation d'erreur *a posteriori* par résidu. Ces techniques ont été introduites par Babuška et Rheinbolt [**10**], puis leur analyse a été étendue par Verfürth [**75**] ; voir également [**76**] pour une revue relativement exhaustive.

Définition 8.2. *Soit $w \in W$. On définit le* résidu *de la solution approchée u_h en w par*

$$\rho(u_h; w) = a(u - u_h, w). \tag{8.12}$$

On observera que la propriété de consistance (8.5) s'écrit sous la forme

$$\forall w_h \in W_h, \qquad \rho(u_h; w_h) = 0. \tag{8.13}$$

On suppose que le résidu peut être localisé comme suit :

$$\rho(u_h; w) = \sum_{K \in \mathcal{T}_h} \rho_K(u_h; w), \tag{8.14}$$

$$\rho_K(u_h; w) = (f - \mathcal{L}u_h, w)_{0,K} + ((\Lambda u_h)\cdot n_K, w)_{0,\partial K}, \tag{8.15}$$

où Λ est un opérateur de trace à valeurs dans $\mathbb{R}^d$ et n_K est la normale extérieure à K. On observera qu'on a implicitement supposé que les fonctions de W admettent des traces définies de manière univoque aux interfaces du maillage, ce qui est clairement le cas pour un problème d'advection–diffusion–réaction où $W = H^1(\Omega)$. Pour ce problème, l'opérateur de trace Λ est tel que

$$\Lambda u_h = \sigma\cdot\nabla u_h. \tag{8.16}$$

Soit $F \in \mathcal{F}_h^{\mathrm{i}}$ une face intérieure du maillage, c'est-à-dire telle qu'il existe deux mailles K_1 et K_2 dans $\mathcal{T}_h$ avec $F = K_1 \cap K_2$. On désigne par n_1 et n_2 les normales extérieures à K_1 et K_2, respectivement. On note

$$[\![\Lambda u_h\cdot n]\!] = (\Lambda u_h)|_{K_1}\cdot n_1 + (\Lambda u_h)|_{K_2}\cdot n_2, \tag{8.17}$$

le saut de la composante normale de Λu_h à-travers F. Pour une face $F \in \mathcal{F}_h^{\partial}$ située au bord, on pose simplement $[\![\Lambda u_h\cdot n]\!] = (\Lambda u_h)|_K\cdot n$ où K est l'élément de $\mathcal{T}_h$ dont F est une face et n est la normale extérieure à Ω.

On suppose que l'espace W jouit de la propriété d'interpolation locale suivante : il existe une constante c, indépendante de h, telle que pour tout $w \in W$, il existe $w_h \in W_h$ tel que pour tout $K \in \mathcal{T}_h$,

$$\|w - w_h\|_{0,K} + h_K^{\frac{1}{2}}\|w - w_h\|_{0,\partial K} \leqslant c\, h_K \|w\|_{W,\Delta_K}, \tag{8.18}$$

où h_K est le diamètre de K et Δ_K est un macro-élément autour de K. La propriété d'interpolation (8.18) est satisfaite pour $W = H^1(\Omega)$ si la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière. En dimension $d \geqslant 2$, on ne peut pas prendre pour w_h l'interpolé de Lagrange de w car les fonctions de $H^1(\Omega)$ ne sont pas nécessairement continues. On utilise d'autres techniques d'interpolation où l'interpolé est défini sur chaque maille à partir de moyennes sur un macro-élément centré sur cette maille. La définition précise de ces opérateurs

d'interpolation n'est pas nécessaire pour la suite ; on utilisera simplement le fait que ces opérateurs existent et qu'ils permettent de satisfaire l'hypothèse (8.18). Pour plus de détails, on pourra consulter les travaux de Clément [**29**] ou de Scott et Zhang [**69**].

Théorème 8.3 (Fiabilité). *Pour $K \in \mathcal{T}_h$, on pose*

$$\eta_K(u_h, a, f) = h_K \|f - \mathcal{L}u_h\|_{0,K} + h_K^{\frac{1}{2}} \|[\![\Lambda u_h \cdot n]\!]\|_{0,\partial K}. \tag{8.19}$$

Alors, dans le cadre des hypothèses ci-dessus, η_K est un indicateur d'erreur local ; en d'autres termes, il existe une constante c, indépendante de h, telle que

$$\|u - u_h\|_V \leqslant c \left(\sum_{K \in \mathcal{T}_h} [h_K^2 \|f - \mathcal{L}u_h\|_{0,K}^2 + h_K \|[\![\Lambda u_h \cdot n]\!]\|_{0,\partial K}^2] \right)^{\frac{1}{2}}. \tag{8.20}$$

De plus, lorsque les fonctions de W sont nulles au bord, le terme de saut dans (8.19) *est restreint à la partie de ∂K qui se trouve à l'intérieur de Ω.*

L'estimation (8.20) découle directement des hypothèses ci-dessus. De la condition (BNB1), on déduit

$$\alpha \|u - u_h\|_V \leqslant \sup_{w \in W} \frac{a(u - u_h, w)}{\|w\|_W}.$$

En utilisant la définition (8.12) du résidu et en appliquant la propriété de consistance (8.5) à la fonction w_h résultant de la propriété (8.18), on obtient

$$\begin{aligned}
\alpha \|u - u_h\|_V &\leqslant \sup_{w \in W} \frac{a(u - u_h, w - w_h)}{\|w\|_W} = \sup_{w \in W} \frac{\rho(u_h ; w - w_h)}{\|w\|_W} \\
&\leqslant \sup_{w \in W} \frac{\sum_{K \in \mathcal{T}_h} [\|f - \mathcal{L}u_h\|_{0,K} \|w - w_h\|_{0,K} + \|[\![\Lambda u_h \cdot n]\!]\|_{0,\partial K} \|w - w_h\|_{0,\partial K}]}{\|w\|_W} \\
&\leqslant \left(\sum_{K \in \mathcal{T}_h} [h_K^2 \|f - \mathcal{L}u_h\|_{0,K}^2 + h_K \|[\![\Lambda u_h \cdot n]\!]\|_{0,\partial K}^2] \right)^{\frac{1}{2}} \\
&\times \sup_{w \in W} \frac{\left(\sum_{K \in \mathcal{T}_h} h_K^{-2} \|w - w_h\|_{0,K}^2 + h_K^{-1} \|w - w_h\|_{0,\partial K}^2 \right)^{\frac{1}{2}}}{\|w\|_W} \leqslant c \left(\sum_{K \in \mathcal{T}_h} \eta_K(u_h, a, f)^2 \right)^{\frac{1}{2}}.
\end{aligned}$$

On a utilisé le fait que grâce à l'hypothèse de régularité sur la famille $\{\mathcal{T}_h\}_{h>0}$, le nombre de triangles appartenant à un macro-élément Δ_K peut être majoré uniformément en h. On observera que dans l'analyse d'erreur *a posteriori*, seule intervient la condition inf-sup satisfaite par la forme bilinéaire a sur les espaces fonctionnels V et W et non la condition inf-sup discrète satisfaite sur les espaces d'approximation V_h et W_h.

Par exemple, pour un problème d'advection–diffusion–réaction, l'indicateur d'erreur local par résidu s'exprime sous la forme

$$\begin{aligned}\eta_K(u_h,a,f) = h_K\|f + \nabla\cdot(\sigma\cdot\nabla u_h) - \beta\cdot\nabla u_h - \mu u_h\|_{0,K} \\ + h_K^{\frac{1}{2}}\|[\![(\sigma\cdot\nabla u_h)\cdot n]\!]\|_{0,\partial K}.\end{aligned} \tag{8.21}$$

En particulier, pour le Laplacien où σ est la matrice identité, $\beta = 0$ et $\mu = 0$, on obtient

$$\eta_K(u_h,a,f) = h_K\|f + \Delta u_h\|_{0,K} + h_K^{\frac{1}{2}}\|[\![\nabla u_h\cdot n]\!]\|_{0,\partial K}. \tag{8.22}$$

Si on utilise un élément fini de degré un, le premier terme du membre de droite s'écrit simplement $h_K\|f\|_{0,K}$.

On s'intéresse maintenant à l'optimalité de l'indicateur d'erreur (8.19). Le résultat suivant est dû à Verfürth ; voir [**76**, p. 10].

Théorème 8.4 (Optimalité). *On suppose que l'opérateur $\mathcal{L}$ est défini par* (8.3) *et l'opérateur Λ par* (8.16). *Alors, il existe une constante c, indépendante de h, telle que pour tout $K \in \mathcal{T}_h$,*

$$\eta_K(u_h,a,f) \leqslant c\left(|u-u_h|_{1,\Delta_K} + h_K \inf_{v_h\in P^0_{\mathrm{td},h}} \|f-v_h\|_{0,\Delta_K}\right), \tag{8.23}$$

où Δ_K est le macro-élément constitué de l'ensemble des éléments de $\mathcal{T}_h$ (incluant K) partageant une face avec K.

La conclusion du théorème 8.4 reste vraie pour tout entier $l \geqslant 0$ (avec une constante qui dépend de l) si l'infimum dans (8.23) est pris pour $v_h \in P^l_{\mathrm{td},h}$; voir [**38**, p. 428]. Cette variante permet de montrer que lorsqu'on utilise un élément fini de Lagrange de degré k et que l'indicateur d'erreur local est évalué de manière approchée en utilisant des quadratures (voir la définition 9.1), l'erreur de quadrature reste négligeable devant l'indicateur d'erreur pourvu que la quadrature soit d'ordre $2k$.

8.3 Estimateurs par dualité

Cette section présente une introduction aux techniques d'estimation d'erreur *a posteriori* par dualité. Ces techniques ont été introduites Johnson [**52**]. On pourra consulter Becker et Rannacher [**14**] pour une revue relativement exhaustive.

Un des intérêts des techniques d'estimation d'erreur *a posteriori* par dualité est qu'elles permettent d'estimer d'autres erreurs que celle mesurée dans la norme de stabilité naturelle $\|\cdot\|_V$. Étant donné une fonctionnelle $\Psi : V \to \mathbb{R}$, on souhaite estimer la quantité suivante :

$$\mathcal{E} = \Psi(u) - \Psi(u_h). \tag{8.24}$$

Dans les applications, la fonctionnelle Ψ permet d'extraire une information sur la solution qui intéresse plus directement le numéricien. Par exemple, dans un calcul d'écoulement autour d'un profil d'aile, cette information pourra être la portance ou la trainée de l'aile. Par la suite, on suppose que la fonctionnelle Ψ est suffisamment régulière pour qu'en tout $v \in V$, sa différentielle $\Psi'(v) \in \mathcal{L}(V;\mathbb{R})$ soit définie.

Afin d'estimer la quantité $\mathcal{E}$, on introduit le problème dual suivant :

$$\begin{cases} \text{Chercher } z \in W \text{ tel que} \\ a(v,z) = \int_0^1 \langle \Psi'(u_h + se), v\rangle_{V',V}\, \mathrm{d}s, \quad \forall v \in V, \end{cases} \tag{8.25}$$

où $e = u - u_h$. On rappelle que le résidu $\rho(u_h; w)$ de la solution discrète u_h en $w \in W$ est défini en (8.12).

Proposition 8.5. *On a*

$$\forall z_h \in W_h, \qquad \mathcal{E} = \rho(u_h; z - z_h). \tag{8.26}$$

La formule (8.26) découle simplement du fait que

$$\mathcal{E} = \Psi(u) - \Psi(u_h) = \int_0^1 \langle \Psi'(u_h + se), e\rangle_{V',V}\, \mathrm{d}s = a(e,z) = \rho(u_h; z),$$

et on conclut grâce à la propriété de consistance (8.13). La formule (8.26) s'interprète comme une *formule de représentation de l'erreur* puisqu'elle exprime le

fait que la quantité $\mathcal{E}$ est *égale* au résidu de la solution approchée sur la différence entre la solution du problème dual (8.25) et une fonction arbitraire dans W_h.

Remarque 8.6

On peut également estimer la quantité $\mathcal{E} = \Psi(u - u_h)$. La fonctionnelle Ψ étant *a priori* non-linéaire, cette quantité diffère de celle définie en (8.24). Dans ce cas, on suppose que la fonctionnelle Ψ admet une densité $\psi : V \to V$ telle que pour tout $v \in V$, $\Psi(v) = (\psi(v), v)_V$ où $(\cdot,\cdot)_V$ désigne le produit scalaire dans V. On introduit le problème dual suivant :

$$\begin{cases} \text{Chercher } z \in W \text{ tel que} \\ a(v,z) = (\psi(e), v)_V, \quad \forall v \in V. \end{cases} \tag{8.27}$$

La formule de représentation de l'erreur (8.26) reste valable à condition de l'évaluer sur la solution du problème dual (8.27).

La formule de représentation de l'erreur (8.26) se localise en utilisant (8.15). On suppose que l'espace W jouit d'une propriété d'interpolation locale légèrement plus forte que (8.18), à savoir on suppose qu'il existe un sous-espace Z dense dans W et une constante c, indépendante de h, telle que pour tout $\zeta \in Z$, il existe $w_h \in W_h$ tel que pour tout $K \in \mathcal{T}_h$,

$$\|\zeta - w_h\|_{0,K} + h_K^{\frac{1}{2}} \|\zeta - w_h\|_{0,\partial K} \leqslant c\, h_K^2 \|\zeta\|_{Z,\Delta_K}, \tag{8.28}$$

où h_K est le diamètre de K et Δ_K est un macro-élément autour de K. La propriété (8.28) est satisfaite pour $W = H^1(\Omega)$ et $Z = H^2(\Omega)$ si la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière.

Théorème 8.7 (Fiabilité). *On suppose que la solution z du problème dual* (8.25) *est dans Z. Pour $K \in \mathcal{T}_h$, on pose*

$$\omega_K(z) = h_K \|z\|_{Z,\Delta_K}. \tag{8.29}$$

Alors, dans le cadre des hypothèses ci-dessus, on a

$$|\mathcal{E}| \leqslant c \sum_{K \in \mathcal{T}_h} \eta_K(u_h, a, f)\omega_K(z), \tag{8.30}$$

où $\eta_K(u_h, a, f)$ est défini dans (8.19).

À moins que la fonctionnelle Ψ ne soit linéaire, la solution duale z dépend de la solution exacte u. L'estimation (8.30) n'est donc pas, en toute rigueur, une estimation d'erreur *a posteriori*. Dans la pratique, deux approximations sont nécessaires pour évaluer explicitement l'estimation (8.30).

(i) La première consiste à éliminer la dépendance en u dans le problème dual (8.25). On peut pour cela remplacer le membre de droite par $\langle \Psi'(u_h), v\rangle_{V',V}$. Cette approximation revient à remplacer une intégrale du type $\int_0^1 \Phi(s)\,ds$ par la valeur $\Phi(0)$, ce qui est raisonnable si la fonction Φ varie peu entre 0 et 1, c'est-à-dire si la solution approchée u_h est suffisamment proche de la solution exacte u.

(ii) La deuxième approximation consiste à remplacer le problème dual (8.25) par un problème de dimension finie où une solution duale discrète est cherchée dans un espace d'approximation Z_h. Un choix relativement simple est de prendre $Z_h = W_h$ puis de reconstituer localement une approximation de $\|z\|_{Z,\Delta_K}$ afin d'évaluer $\omega_K(z)$. Par exemple, pour un problème d'advection–diffusion–réaction, la quantité $\|z\|_{Z,\Delta_K}$ fait intervenir les dérivées secondes de z qui peuvent être reconstituées localement à partir des valeurs de la solution duale discrète en utilisant des formules de différences finies ou des projections L^2-orthogonales sur Δ_K.

Malgré les deux difficultés évoquées ci-dessus, les estimations d'erreur *a posteriori* par dualité ont été mises en œuvre avec succès dans de nombreuses applications, notamment afin de générer des maillages adaptatifs pour l'approximation par éléments finis de systèmes d'équations aux dérivées partielles non-linéaires comme les équations de Navier–Stokes ou les équations des gaz réactifs.

Remarque 8.8

En général, il n'est pas facile de prouver une propriété d'optimalité pour les estimateurs d'erreur par dualité car on ne dispose pas d'estimation *a priori* locale sur la solution duale.

8.4 Estimateurs hiérarchiques

Les estimateurs d'erreur *a posteriori* de type hiérarchique ont été introduits par Bank et Weiser en 1975 [**12**]. L'analyse de ces estimateurs a d'abord été effectuée pour des approximations de type Galerkin standard dans un cadre consistant et conforme [**12, 11**], puis étendue aux approximations de type Galerkin non-standard dans un cadre non-consistant et non-conforme [**2**]. Pour simplifier, la présentation ci-dessous est restreinte au cadre consistant et conforme, mais pour des méthodes de Galerkin non-standard. Pour une revue relativement détaillée des estimateurs d'erreur *a posteriori* de type hiérarchique, on pourra consulter Ainsworth et Oden [**4**].

8.4.1 Cadre général

On considère le problème modèle (8.1). On note α la constante intervenant dans la condition inf-sup (BNB1) satisfaite par la forme bilinéaire a et on note ω la norme de a dans $\mathcal{L}(V \times W; \mathbb{R})$; en d'autres termes,

$$\alpha = \inf_{v \in V} \sup_{w \in W} \frac{a(v,w)}{\|v\|_V \|w\|_W}, \tag{8.31}$$

$$\omega = \sup_{v \in V} \sup_{w \in W} \frac{a(v,w)}{\|v\|_V \|w\|_W}. \tag{8.32}$$

On note u_h la solution du problème discret (8.4). On rappelle qu'on a supposé que ce problème est bien posé et pour alléger les notations, on suppose que la forme bilinéaire a satisfait une condition inf-sup discrète sur $V_h \times W_h$ avec la même constante α que dans (8.31).

Le principe des estimateurs d'erreur *a posteriori* de type hiérarchique consiste à enrichir les espaces d'approximation V_h et W_h intervenant dans (8.4) de la manière suivante :

$$V_h^{\mathrm{e}} = V_h \oplus V_h^{\mathrm{f}} \qquad \text{et} \qquad W_h^{\mathrm{e}} = W_h \oplus W_h^{\mathrm{f}}, \tag{8.33}$$

où V_h^{f} et W_h^{f} sont appelés des *espaces d'échelles fluctuantes* et où V_h^{e} et W_h^{e} sont appelés des *espaces d'approximation enrichis*. On considère les problèmes

discrets suivants :

$$\begin{cases} \text{Chercher } u_h^{\mathrm{e}} \in V_h^{\mathrm{e}} \text{ tel que} \\ a(u_h^{\mathrm{e}}, w_h^{\mathrm{e}}) = f(w_h^{\mathrm{e}}), \quad \forall w_h^{\mathrm{e}} \in W_h^{\mathrm{e}}, \end{cases} \tag{8.34}$$

et

$$\begin{cases} \text{Chercher } e_h^{\mathrm{f}} \in V_h^{\mathrm{f}} \text{ tel que} \\ a(e_h^{\mathrm{f}}, w_h^{\mathrm{f}}) = f(w_h^{\mathrm{f}}) - a(u_h, w_h^{\mathrm{f}}), \quad \forall w_h^{\mathrm{f}} \in W_h^{\mathrm{f}}. \end{cases} \tag{8.35}$$

On observera que le membre de droite dans (8.35) est égal à $\rho(u_h; w_h^{\mathrm{f}})$, c'est-à-dire au résidu de la solution approchée u_h sur les fonctions de l'espace des échelles fluctuantes W_h^{f}. On suppose que les problèmes (8.34) et (8.35) sont bien posés, c'est-à-dire que $\dim V_h^{\mathrm{f}} = \dim W_h^{\mathrm{f}}$ et que la forme bilinéaire a satisfait une condition inf-sup discrète sur $V_h^{\mathrm{f}} \times W_h^{\mathrm{f}}$ et sur $V_h^{\mathrm{e}} \times W_h^{\mathrm{e}}$. Pour alléger les notations, on suppose que ces conditions inf-sup discrètes font intervenir la même constante α que dans (8.31)[1].

En pratique, le problème discret (8.34) n'est pas résolu, mais sert uniquement à l'analyse théorique. En revanche, le problème discret (8.35) est résolu et sa solution e_h^{f} permet de calculer l'estimateur d'erreur hiérarchique. L'analyse de l'estimateur hiérarchique repose sur les deux hypothèses suivantes.

- **Inégalité de Cauchy–Buniakowski–Schwarz.** Il existe une constante $\gamma \in [0, 1[$ telle que pour tout h, on a

$$\forall w_h \in W_h,\ \forall w_h^{\mathrm{f}} \in W_h^{\mathrm{f}}, \qquad (w_h, w_h^{\mathrm{f}})_W \leqslant \gamma \|w_h\|_W \|w_h^{\mathrm{f}}\|_W. \tag{8.36}$$

  Dans la littérature, cette inégalité est également appelée *inégalité de Cauchy–Schwarz forte*.

- **Hypothèse de saturation.** Il existe une constante $\beta \in]0, 1[$ telle que, pour tout h, on a

$$\|u - u_h^{\mathrm{e}}\|_V \leqslant \beta \|u - u_h\|_V. \tag{8.37}$$

Ces deux hypothèses sont discutées dans les sections 8.4.2 et 8.4.3. Elles conduisent au résultat suivant.

1. Dans un cadre conforme, c'est clairement le cas si $V = W$, $V_h = W_h$, $V_h^{\mathrm{f}} = W_h^{\mathrm{f}}$ et si la forme bilinéaire a est coercive sur V.

Théorème 8.9. *Dans le cadre des hypothèses ci-dessus, on a*

$$\frac{\alpha}{\omega}(1-\beta)(1-\gamma^2)^{\frac{1}{2}}\|u-u_h\|_V \leqslant \|e_h^{\mathrm{f}}\|_V \leqslant \frac{\omega}{\alpha}\|u-u_h\|_V. \tag{8.38}$$

La quantité $\|e_h^{\mathrm{f}}\|_V$ est appelée un estimateur *a posteriori* de type hiérarchique. Dans la plupart des cas, la norme de V peut se localiser selon $\|\cdot\|_V = (\sum_{K\in\mathcal{T}_h}\|\cdot\|_{V,K}^2)^{\frac{1}{2}}$ si bien que les quantités $\|e_h^{\mathrm{f}}\|_{V,K}$ sont des indicateurs d'erreur locaux.

On observera que les hypothèses (8.36) et (8.37) n'interviennent que pour établir la fiabilité de l'estimateur *a posteriori* $\|e_h^{\mathrm{f}}\|_V$, c'est-à-dire l'inégalité de gauche dans (8.38). Dans un cadre conforme, l'inégalité de droite dans (8.38), qui exprime une propriété d'optimalité globale de l'estimateur *a posteriori* $\|e_h^{\mathrm{f}}\|_V$, ne nécessite pas les hypothèses (8.36) et (8.37).

8.4.2 L'inégalité de Cauchy–Buniakowski–Schwarz

L'inégalité de Cauchy–Buniakowski–Schwarz est une propriété intrinsèque aux espaces tests discrets W_h et W_h^{f}. Elle ne dépend pas du problème modèle dont on cherche à approcher la solution. Une interprétation géométrique de cette inégalité est que l'angle entre les espaces W_h et W_h^{f} est minoré par une quantité strictement positive, et ce *uniformément* en h. Cette propriété s'interprète également en termes de projecteurs de la manière suivante.

Lemme 8.10. *Soit $\Pi_h \in \mathcal{L}(W_h^{\mathrm{e}}; W_h)$ le projecteur (oblique) induit par la décomposition $W_h^{\mathrm{e}} = W_h^{\mathrm{f}} \oplus W_h$. Alors, $\|\Pi_h\|_{\mathcal{L}(W_h^{\mathrm{e}};W_h)}$ est borné uniformément en h par un réel $\varrho < +\infty$ si et seulement si l'inégalité de Cauchy–Buniakowski–Schwarz est satisfaite avec*

$$\gamma = \left(1 - \frac{1}{\varrho^2}\right)^{\frac{1}{2}}. \tag{8.39}$$

En général, la vérification de l'inégalité de Cauchy–Buniakowski–Schwarz se fait localement sur chaque cellule du maillage. Soit $\{\psi_1, \ldots, \psi_N\}$ une base de W_h (les fonctions de forme dans W_h) et soit $\{\psi_1^{\mathrm{f}}, \ldots, \psi_{N^{\mathrm{f}}}^{\mathrm{f}}\}$ une base de W_h^{f}. Soit $K \in \mathcal{T}_h$. Les restrictions à K des fonctions ci-dessus (dont le support a une intersection non-vide avec K) sont, respectivement, les fonctions de forme $\{\theta_{K,1}, \ldots, \theta_{K,n_{\mathrm{f}}}\}$ de l'élément fini sur K et les fluctuations locales

$\{\theta^{\mathrm{f}}_{K,1},\dots,\theta^{\mathrm{f}}_{K,n^{\mathrm{f}}_{\mathrm{f}}}\}$. On définit les matrices de Gram

$$G_K^{11}\in\mathbb{R}^{n_{\mathrm{f}},n_{\mathrm{f}}},\quad (G_K^{21})^T=G_K^{12}\in\mathbb{R}^{n_{\mathrm{f}},n^{\mathrm{f}}_{\mathrm{f}}}\quad\text{et}\quad G_K^{22}\in\mathbb{R}^{n^{\mathrm{f}}_{\mathrm{f}},n^{\mathrm{f}}_{\mathrm{f}}},\tag{8.40}$$

dont les coefficients s'expriment sous la forme

$$G_{K,ij}^{11}=(\theta_i,\theta_j)_W,\qquad i,j\in\{1,\dots,n_{\mathrm{f}}\},\tag{8.41}$$

$$G_{K,ij}^{12}=(\theta_i,\theta^{\mathrm{f}}_j)_W,\qquad i\in\{1,\dots,n_{\mathrm{f}}\},\ j\in\{1,\dots,n^{\mathrm{f}}_{\mathrm{f}}\},\tag{8.42}$$

$$G_{K,ij}^{22}=(\theta^{\mathrm{f}}_i,\theta^{\mathrm{f}}_j)_W,\qquad i,j\in\{1,\dots,n^{\mathrm{f}}_{\mathrm{f}}\}.\tag{8.43}$$

Lemme 8.11. *Pour $K\in\mathcal{T}_h$, on pose*

$$\gamma_K=\left(\max_{X_1\in\mathbb{R}^{n_{\mathrm{f}}}}\frac{X_1^TG_K^{12}(G_K^{22})^{-1}G_K^{21}X_1}{X_1^TG_K^{11}X_1}\right)^{\frac12}.\tag{8.44}$$

Alors, l'inégalité de Cauchy–Buniakowski–Schwarz est satisfaite avec la constante $\gamma=\max_{K\in\mathcal{T}_h}\gamma_K$.

Pour chaque élément $K\in\mathcal{T}_h$, la constante γ_K définie en (8.44) s'obtient en résolvant un problème aux valeurs propres généralisé de petite taille (sa taille est égale au nombre de degrés de liberté de l'élément fini de référence qui a servi à construire l'espace W_h). En général, on ne dispose pas de formule analytique simple pour γ_K, sauf dans certains cas très particuliers comme celui où K est un triangle équilatéral ou celui où K est un triangle rectangle isocèle. Par ailleurs, on peut étudier numériquement la variation de la constante γ_K en fonction de la forme de K. Dans la plupart des cas, on constate qu'une majoration uniforme en K du type $\gamma_K\leqslant\gamma_0<1$ est possible en supposant que la famille $\{\mathcal{T}_h\}_{h>0}$ est régulière. Un exemple est discuté dans la section 8.4.4.

8.4.3 L'hypothèse de saturation

L'hypothèse de saturation signifie que la solution discrète u_h^{e} est une meilleure approximation de la solution exacte que ne l'est la solution discrète u_h, et ce uniformément en h. Contrairement à l'inégalité de Cauchy–Buniakowski–Schwarz, cette hypothèse dépend du problème modèle. Elle ne peut donc être prouvée que dans des cas très particuliers, comme par exemple celui où les espaces enrichis correspondent à une approximation par éléments finis du

même type que celle réalisée avec les espaces V_h et W_h, mais avec un maillage plus fin ou avec un ordre polynômial plus élevé. Par ailleurs, il existe des techniques permettant de s'affranchir de l'hypothèse de saturation ; voir, par exemple, [**35**].

8.4.4 Un exemple simple

On considère un problème d'advection–diffusion–réaction avec des conditions aux limites de Neumann ou de Robin si bien que $V = W = H^1(\Omega)$. Tout ce qui suit s'adapte facilement aux conditions aux limites de Dirichlet. On se place dans le cadre de l'approximation de Galerkin standard, si bien que $V_h = W_h$ et $V_h^{\mathrm{f}} = W_h^{\mathrm{f}}$. On suppose que l'espace V_h est construit à partir d'une famille régulière de maillages affines et de l'élément fini de Lagrange $\mathbb{P}_1$. On a donc $V_h = P^1_{\mathrm{c},h}$; voir l'équation (3.44).

Pour une face $F \in \mathcal{F}_h$, on définit la *bulle conforme* associée à la face F de la manière suivante. On désigne par $\mathcal{T}_F$ l'ensemble des éléments de $\mathcal{T}_h$ dont F est une face (on notera que le cardinal de $\mathcal{T}_F$ est de 2 si $F \in \mathcal{F}_h^{\mathrm{i}}$ et de 1 si $F \in \mathcal{F}_h^{\partial}$). Pour $T \in \mathcal{T}_F$, on désigne par $(\lambda_{0,T}, \ldots, \lambda_{d,T})$ ses coordonnées barycentriques et par $i_{F,T}$ l'indice (compris entre 0 et d) du sommet de T opposé à F. On pose

$$b_F^{\mathrm{c}}|_T = \begin{cases} d^d \displaystyle\prod_{\substack{j=0 \\ j \neq i_{F,T}}}^{d} \lambda_{j,T} & \text{si } T \in \mathcal{T}_F, \\ 0 & \text{si } T \notin \mathcal{T}_F. \end{cases} \tag{8.45}$$

On constate que (i) $b_F^{\mathrm{c}} \in H^1_0(\mathcal{T}_F)$, (ii) $0 \leqslant b_F^{\mathrm{c}}(x) \leqslant 1$ pour tout $x \in \mathcal{T}_F$ et (iii) $b_F^{\mathrm{c}}(x) = 1$ si x est le barycentre de la face F.

On choisit pour espace des échelles fluctuantes l'espace V_h^{f} tel que

$$V_h^{\mathrm{f}} = \mathrm{vect}\{b_F^{\mathrm{c}}\}_{F \in \mathcal{F}_h}. \tag{8.46}$$

Avec ce choix, l'espace d'approximation enrichi est $V_h^{\mathrm{e}} = P^2_{\mathrm{c},h}$. Il est donc raisonnable d'espérer que si h est suffisamment petit et si la solution exacte est au moins dans $H^3(\Omega)$, l'hypothèse de saturation (8.37) est satisfaite.

Il reste à vérifier l'inégalité de Cauchy–Buniakowski–Schwarz (8.36). On présente une vérification numérique en deux dimensions d'espace. Soit $K \in \mathcal{T}_h$.

Grâce à l'isotropie et à l'invariance d'échelle, on peut supposer que deux des sommets du triangle K ont pour coordonnées $(0,0)$ et $(1,0)$, respectivement, et paramétrer le triangle K par ses angles α et β ; voir le dessin de gauche sur la figure 8.2. Pour chaque couple (α,β), on détermine numériquement la quantité γ_K définie en (8.44) en résolvant le problème aux valeurs propres généralisé sur K. Par isotropie et invariance d'échelle, on observe que le même résultat est obtenu pour γ_K si les deux angles α et β sont pris dans l'ensemble $\{\alpha,\beta,\pi-\alpha-\beta\}$. L'étude numérique peut donc être restreinte au domaine

$$D = \{(\alpha,\beta) \in]0,\pi[\times]0,\pi[\ ;\ \alpha \geqslant \beta\ ;\ \alpha + 2\beta \geqslant \pi\ ;\ \alpha + \beta \leqslant \pi\}. \tag{8.47}$$

Ce domaine est illustré par le triangle gris dans la figure 8.2 au centre. On construit ainsi la fonction

$$D \ni (\alpha,\beta) \longmapsto \gamma_K \in \mathbb{R}. \tag{8.48}$$

Les isovaleurs de cette fonction sont présentées sur la figure 8.2 à droite. Les valeurs sont comprises entre $\frac{1}{2}$ (couleur noire) et 1 (couleur blanche). On observe que le coefficient γ_K est minimal lorsque le triangle K est équilatéral. De plus, la valeur 1 n'étant atteinte que lorsqu'un des angles du triangle tend vers π, le coefficient γ_K peut être contrôlé par une quantité strictement plus petite que 1 si la famille de maillages est régulière, c'est-à-dire si les triangles ne sont pas trop aplatis.

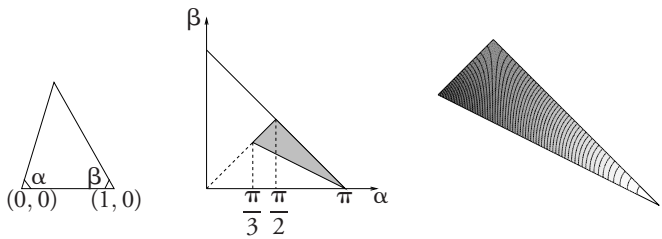


Figure 8.2 – À gauche : triangle K d'angles α et β ; au centre : domaine D (triangle gris) pour les couples (α,β) admissibles ; à droite : isovaleurs de la fonction définie en (8.48) (pour faciliter la visualisation, le domaine D a été agrandi).

8.5 Maillages adaptatifs

Un algorithme relativement général d'adaptation de maillage est le suivant.

(i) On choisit un maillage initial que l'on note $\mathcal{T}_{(0)}$. On pose $i = 0$.

(ii) On résout le problème discret sur $\mathcal{T}_{(i)}$. On note $u_{(i)}$ la solution discrète.

(iii) Sur chaque élément K de $\mathcal{T}_{(i)}$, on calcule un indicateur d'erreur local. On pourra choisir l'un des indicateurs d'erreur présentés dans les sections précédentes.

(iv) Si l'estimation d'erreur globale est inférieure à un seuil de tolérance, on arrête les calculs.

(v) Sinon, sur la base de ces indicateurs d'erreur, on décide de raffiner certaines mailles et d'en déraffiner d'autres. On note $\mathcal{T}_{(i+1)}$ le nouveau maillage ainsi construit.

(vi) On incrémente l'indice i de 1 et on revient à l'étape (ii).

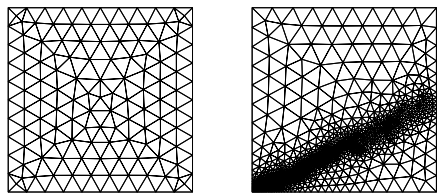


Figure 8.3 – Maillage initial et maillage adaptatif généré à partir d'un indicateur d'erreur *a posteriori* de type résidu pour un problème d'advection–diffusion avec une couche intérieure.

Plusieurs stratégies sont possibles à l'étape (v) lorsque le nouveau maillage est construit à partir des indicateurs d'erreur locaux. Une stratégie fréquemment utilisée dans les applications consiste à *équirépartir l'erreur sur les mailles*. Le critère de raffinement est alors que l'indicateur d'erreur local est supérieur à la moyenne des indicateurs d'erreur sur le maillage. La figure 8.3 présente un exemple de maillage initial et de maillage adaptatif généré à partir d'un indicateur d'erreur *a posteriori* de type résidu pour un problème

d'advection–diffusion. Le domaine Ω est le carré unité, le champ d'advection est $\beta = (2,1)^T$ et le coefficient de diffusion vaut $\epsilon = 10^{-4}$. On impose une valeur unité à la solution sur le côté inférieur du carré, une valeur nulle sur le côté gauche et des conditions aux limites de Neumann homogènes sur les deux autres côtés. La discontinuité de la donnée d'entrée au coin inférieur gauche provoque l'apparition d'une couche intérieure qui est transportée par le champ advectif β et qui reste très localisée du fait de la faible valeur du coefficient de diffusion.

8.6 Compléments

- **Indicateurs d'erreur basés sur une reconstruction locale du gradient.** Le principe de ces indicateurs est d'utiliser la solution discrète u_h afin d'approcher localement le gradient de la solution exacte u. Ces indicateurs sont souvent appelés *indicateurs d'erreur ZZ* en référence aux travaux de Zienkiewicz et Zhu où ils ont été introduits [**80**].

  Pour simplifier, on restreint la présentation aux approximations par des éléments finis de Lagrange $\mathbb{P}_1$. Dans ces conditions, $\nabla u_h \in [P^0_{\mathrm{td},h}]^d$ où $P^0_{\mathrm{td},h}$ est l'espace d'éléments finis totalement discontinu défini en (4.20) pour $k = 0$; il s'agit de l'espace vectoriel constitué des fonctions constantes sur chaque cellule du maillage. On définit un opérateur de reconstruction $R_h : P^0_{\mathrm{td},h} \to P^1_{\mathrm{c},h}$ en utilisant une projection L^2-orthogonale : pour tout $v_h \in P^0_{\mathrm{td},h}$, $R_h v_h$ est l'unique fonction de $P^1_{\mathrm{c},h}$ telle que

  $$\forall w_h \in P^1_{\mathrm{c},h}, \qquad (R_h v_h - v_h, w_h)_{0,\Omega} = 0. \tag{8.49}$$

  L'action de l'opérateur de reconstruction R_h est illustrée en une dimension d'espace sur la figure 8.4 à gauche. Soit $\{\varphi_1, \ldots, \varphi_N\}$ la base nodale des fonctions de forme dans $P^1_{\mathrm{c},h}$. On introduit la *matrice de masse* $\mathcal{M} \in \mathbb{R}^{N,N}$ de terme générique $\mathcal{M} = (\varphi_i, \varphi_j)_{0,\Omega}$ pour tout $i,j \in \{1,\ldots,N\}$. On décompose $R_h v_h$ dans la base nodale selon $R_h v_h = \sum_{i=1}^N V_i^R \varphi_i$ et on pose $V^R = (V_i^R)_{1 \leqslant i \leqslant N}$. On introduit le vecteur $V \in \mathbb{R}^N$ de composantes $V_i = (v_h, \varphi_i)_{0,\Omega}$ pour tout $i \in \{1,\ldots,N\}$. On observe que (8.49) est équivalent à chercher $V^R \in \mathbb{R}^N$ tel que

  $$\mathcal{M} V^R = V. \tag{8.50}$$

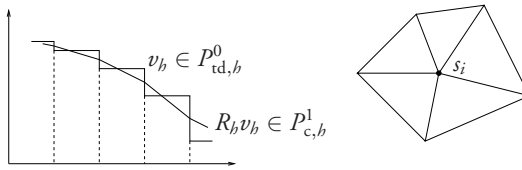


Figure 8.4 – À gauche : action de l'opérateur de reconstruction R_h en une dimension d'espace ; à droite : macro-élément $\Omega(s_i)$ pour approcher les valeurs nodales de $R_h v_h$ en deux dimensions d'espace lorsqu'on utilise une condensation statique de la matrice de masse.

La matrice de masse est symétrique définie positive et elle est bien conditionnée (voir la section 10.1). Par conséquent, le système linéaire (8.50) peut se résoudre facilement à l'aide de la méthode du gradient conjugué décrite dans la section 11.2.2. Une technique encore plus simple consiste à évaluer les coefficients de la matrice de masse de manière approchée en utilisant une quadrature : on remplace l'intégrale d'une fonction sur une cellule du maillage par la moyenne des valeurs que cette fonction prend aux sommets de la maille ; voir la section 9.2. Cette technique d'approximation de la matrice de masse, qu'on appelle *condensation statique*, permet d'approcher cette matrice par une matrice diagonale ; par suite, l'inversion du système linéaire (8.50) devient immédiate. On désigne par $R_h^* v_h$ l'approximation de $R_h v_h$ obtenue par condensation statique de la matrice de masse. Afin d'expliciter $R_h^* v_h$, on désigne par $\{s_1, \ldots, s_N\}$ les sommets du maillage. Pour $i \in \{1, \ldots, N\}$, on désigne par $\mathcal{T}(s_i)$ l'ensemble des mailles dont s_i est un sommet et on introduit le *macro-élément* $\Omega(s_i) = \bigcup_{K \in \mathcal{T}(s_i)} K$. L'ensemble $\Omega(s_i)$ est illustré sur la figure 8.4 à droite. Avec ces notations, on a

$$R_h^* v_h(s_i) = \frac{1}{\mathrm{mes}(\Omega(s_i))} \sum_{K \in \mathcal{T}(s_i)} v_h|_K \,\mathrm{mes}(K), \qquad i \in \{1, \ldots, N\}. \quad (8.51)$$

Une fois défini l'opérateur de reconstruction R_h (ou son approximation R_h^*), on définit l'opérateur de reconstruction locale du gradient $G_h : P^1_{\mathrm{c},h} \to [P^1_{\mathrm{c},h}]^d$ en posant pour tout $v_h \in P^1_{\mathrm{c},h}$ et pour tout $i \in \{1, \ldots, d\}$,

$$(G_h v_h)_i = R_h(\partial_i v_h). \quad (8.52)$$

En d'autres termes, on reconstruit localement le gradient composante par composante en utilisant l'opérateur R_h.
L'opérateur de reconstruction locale G_h défini en (8.52) peut être utilisé comme indicateur d'erreur dans l'approximation du problème modèle (8.1). Pour tout $K \in \mathcal{T}_h$, on pose

$$\eta_K^{\mathrm{rec}} = \|G_h u_h - \nabla u_h\|_{0,K}, \tag{8.53}$$

où u_h est la solution du problème approché par éléments finis (8.4). Les quantités $\{\eta_K^{\mathrm{rec}}\}_{K\in\mathcal{T}_h}$ sont souvent utilisées dans les logiciels d'éléments finis comme indicateurs d'erreur locaux afin de raffiner le maillage de manière adaptative. Même si ces indicateurs d'erreur ne reposent pas sur des bases théoriques aussi solides que les estimateurs d'erreur *a posteriori* présentés dans les sections 8.2, 8.3 et 8.4 (les quantités $\{\eta_K^{\mathrm{rec}}\}_{K\in\mathcal{T}_h}$ ne sont pas, en général, des indicateurs d'erreur locaux selon la définition 8.1), leur utilisation est motivée d'une part par le fait qu'ils sont relativement simples à évaluer et d'autre part par le fait que les maillages basés sur les techniques de reconstruction locale du gradient sont, en général, relativement bien adaptés aux phénomènes physiques que l'on cherche à capter.
Par ailleurs, on dispose d'un certain nombre de résultats qui apportent une caution théorique minimale à l'utilisation des indicateurs d'erreur $\{\eta_K^{\mathrm{rec}}\}_{K\in\mathcal{T}_h}$. D'une part, ces indicateurs sont globalement équivalents à la contribution des sauts dans l'estimateur par résidu défini en (8.19) lorsque la matrice de diffusion est diagonale. Plus précisément, on montre (voir [**64**]) qu'il existe des constantes c_1 et c_2, indépendantes de h, telles que

$$c_1 \sum_{K\in\mathcal{T}_h} h_K \|[\![\nabla u_h\cdot n]\!]\|_{0,\partial K}^2 \leqslant \sum_{K\in\mathcal{T}_h} (\eta_K^{\mathrm{rec}})^2 \leqslant c_2 \sum_{K\in\mathcal{T}_h} h_K \|[\![\nabla u_h\cdot n]\!]\|_{0,\partial K}^2. \tag{8.54}$$

D'autre part, sous certaines hypothèses sur le maillage et la régularité de la solution exacte u, on montre que le gradient reconstruit $G_h u_h$ possède une propriété de superconvergence au sens où il converge plus vite vers ∇u (asymptotiquement en h) que ∇u_h. Dans ces conditions, la décomposition

$$\nabla u - \nabla u_h = (G_h u_h - \nabla u_h) + (\nabla u - G_h u_h) \tag{8.55}$$

montre que le premier terme du membre de droite domine (asymptotiquement en h) le deuxième. En d'autres termes, la différence $(G_h u_h - \nabla u_h)$

fournit une bonne estimation de l'erreur d'approximation en semi-norme H^1, c'est-à-dire de la quantité $\nabla u - \nabla u_h$.
Enfin, le gradient reconstruit $G_h u_h$ peut être utilisé afin d'estimer la hessienne $h(u)$ de la solution exacte u, c'est-à-dire la matrice $h(u) \in \mathbb{R}^{d,d}$ de composantes $h(u)_{ij} = \partial_{ij} u$ pour tout $i, j \in \{1, \ldots, d\}$. Plusieurs techniques sont possibles afin d'estimer ces dérivées secondes, la plus simple consistant à approcher $\partial_{ij} u$ par $\frac{1}{2}(\partial_j (G_h u_h)_i + \partial_i (G_h u_h)_j)$; on peut également considérer des projections locales sur des macro-éléments autour de chaque maille ou des approximations le long des arêtes du maillage. La connaissance d'une approximation de la matrice hessienne de la solution exacte permet d'estimer l'erreur d'interpolation lorsqu'on utilise des éléments finis de degré 1. Dans la mesure où l'analyse d'erreur *a priori* montre que l'erreur d'approximation est contrôlée par l'erreur d'interpolation pour la solution exacte, la matrice hessienne approchée peut être utilisée afin de piloter le raffinement adaptatif du maillage. L'intérêt de ces techniques est qu'elles permettent de construire des maillages dits *anisotropes* où les mailles sont étirées dans les directions où la dérivée seconde de la solution est faible et fines dans les autres directions [**39, 70**]. On observera qu'une caractéristique des estimateurs de l'erreur d'interpolation est qu'ils sont indépendants de la nature du problème modèle que l'on cherche à approcher. Ces estimateurs sont donc facilement portables sur une gamme relativement large de problèmes modèles.

- **Estimation *a posteriori* des erreurs de modélisation.** On s'intéresse à la situation suivante : on suppose qu'on dispose de deux modèles, l'un précis et coûteux, l'autre moins précis et plus économique. On souhaite réaliser un compromis entre la précision et le coût des calculs en n'utilisant le modèle précis que dans une partie du domaine. On souhaite également équilibrer les erreurs de discrétisation et de modélisation afin d'éviter des situations où, par exemple, on utilise un modèle précis, mais coûteux, pour approcher une solution qui est polluée par des erreurs de discrétisation trop importantes. La partie du domaine où le modèle précis doit être utilisé étant *a priori* inconnue, elle est déterminée par une analyse d'erreur *a posteriori*.

On considère le problème suivant :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u,w) + \delta(u,w) = f(w), \quad \forall w \in W, \end{cases} \tag{8.56}$$

où V et W sont deux espaces de Hilbert, $a \in \mathcal{L}(V \times W; \mathbb{R})$, $\delta \in \mathcal{L}(V \times W; \mathbb{R})$ et $f \in W'$. Ici, la forme bilinéaire a représente le modèle grossier, la somme $a + \delta$ le modèle fin et la forme linéaire δ la différence entre le modèle fin et le modèle grossier. L'utilisation du modèle grossier seul conduit au problème suivant :

$$\begin{cases} \text{Chercher } u_m \in V \text{ tel que} \\ a(u_m,w) = f(w), \quad \forall w \in W. \end{cases} \tag{8.57}$$

La différence $u - u_m$ est appelée *l'erreur de modélisation*. On suppose que les problèmes (8.56) et (8.57) sont bien posés. En pratique, on ne résout pas le problème (8.57) mais une approximation de celui-ci, par exemple par la méthode des éléments finis. Ceci conduit à

$$\begin{cases} \text{Chercher } u_{hm} \in V_h \text{ tel que} \\ a(u_{hm},w_h) = f(w_h), \quad \forall w_h \in W_h. \end{cases} \tag{8.58}$$

Pour simplifier, on se place dans le cadre d'une approximation conforme si bien que $V_h \subset V$ et $W_h \subset W$. De plus, on suppose que le problème (8.58) est bien posé.
On considère une fonctionelle $\Psi : V \to \mathbb{R}$; pour simplifier, on suppose que Ψ est linéaire. On souhaite estimer la quantité

$$\mathcal{E} = \Psi(u) - \Psi(u_{hm}), \tag{8.59}$$

c'est-à-dire l'impact sur la fonctionnelle Ψ des erreurs de modélisation et de discrétisation. On introduit le problème dual

$$\begin{cases} \text{Chercher } z \in W \text{ tel que} \\ a(v,z) + \delta(v,z) = \Psi(v), \quad \forall v \in V, \end{cases} \tag{8.60}$$

ainsi que la discrétisation de ce problème avec le modèle grossier, ce qui conduit à

$$\begin{cases} \text{Chercher } z_{hm} \in W_h \text{ tel que} \\ a(v_h, z_{hm}) = \Psi(v_h), \quad \forall v_h \in V_h. \end{cases} \tag{8.61}$$

Proposition 8.12. *Dans le cadre ci-dessus, on a pour tout* $(v_h, w_h) \in V_h \times W_h$,

$$\begin{aligned} \mathcal{E} = & -\delta(u_{hm}, z_{hm}) + \tfrac{1}{2}\big(\rho(u_{hm}; z - w_h) + \rho^*(z_{hm}; u - v_h)\big) \\ & - \tfrac{1}{2}\big(\delta(u_{hm}, z - z_{hm}) + \delta(u - u_{hm}, z_{hm})\big), \end{aligned} \tag{8.62}$$

avec les résidus $\rho(u_{hm}; w) = f(w) - a(u_{hm}, w)$ *pour tout* $w \in W$, *et* $\rho^*(z_{hm}, v) = \Psi(v) - a(v, z_{hm})$ *pour tout* $v \in V$.

La formule de représentation de l'erreur (8.62) peut être utilisée afin de construire un algorithme de raffinement adaptatif du maillage et du modèle. Le principe est que le premier terme du membre de droite dans (8.62), $-\delta(u_{hm}, z_{hm})$, mesure l'erreur de modélisation, le deuxième terme, $\frac{1}{2}(\rho(u_{hm}; z - w_h) + \rho^*(z_{hm}; u - v_h))$, mesure l'erreur de discrétisation et le troisième terme, qui est d'un ordre supérieur en l'erreur de modélisation, est négligé. L'algorithme est le suivant.

(i) Initialisation : on choisit un maillage initial et on divise *a priori* les mailles en deux groupes, celles où on utilise le modèle grossier, $\mathcal{T}_h^{\text{gro}}$, et celles où on utilise le modèle fin, $\mathcal{T}_h^{\text{fin}}$.

(ii) Étant donné un maillage $\mathcal{T}_h$ partitionné selon $\mathcal{T}_h^{\text{gro}} \cup \mathcal{T}_h^{\text{fin}}$, on résout le problème primal (8.58) et le problème dual (8.61) en remplaçant la forme bilinéaire a par la forme bilinéaire

$$a(\cdot, \cdot) + \sum_{K \in \mathcal{T}_h^{\text{gro}}} \delta_K(\cdot, \cdot), \tag{8.63}$$

où $\delta_K(\cdot, \cdot)$ est la localisation de la forme bilinéaire δ à la maille K.

(iii) On utilise les solutions discrètes u_{hm} et z_{hm} obtenues à l'étape (ii) afin d'évaluer les indicateurs locaux d'erreur de modélisation et de discrétisation.

(iv) Sur la base de ces indicateurs, on décide de raffiner certaines mailles, d'en déraffiner d'autres et de modifier la répartition des mailles entre les deux groupes de modélisation.

(v) On revient à l'étape (ii) et on poursuit les itérations jusqu'à ce que l'estimation d'erreur soit inférieure à un seuil de tolérance.

Pour plus de détails sur l'analyse des erreurs de modélisation et de discrétisation et sur la mise en œuvre de l'algorithme adaptatif ci-dessus, on renvoie aux travaux de Braack et Ern [**17**] où ces techniques ont été introduites.

9 • QUADRATURES

Une quadrature est une formule permettant d'évaluer une intégrale de manière approchée. L'utilisation de quadratures est pratiquement incontournable dans la méthode des éléments finis. En effet, une fois construit l'espace d'approximation, la solution discrète s'obtient par la résolution d'un système linéaire dont les coefficients de la matrice et du membre de droite s'évaluent à partir d'intégrales. L'objet de ce chapitre est de décrire le principe général des quadratures, d'en présenter des exemples en une, deux et trois dimensions d'espace et, enfin, d'analyser l'erreur d'approximation due à l'utilisation de quadratures.

9.1 Principe des quadratures

Soit K une partie non-vide, connexe et compacte de $\mathbb{R}^d$ de frontière lipschitzienne. Dans le cadre de la méthode des éléments finis, K est une cellule du maillage.

Définition 9.1. *Soit un entier $l_{\mathrm{q}} \geqslant 1$. Une quadrature sur K à l_{q} points consiste en la donnée de :*

(i) *un ensemble de l_{q} réels $\{\omega_1, \ldots, \omega_{l_{\mathrm{q}}}\}$, ci-après appelés les* poids de la quadrature*, tels que $\sum_{l=1}^{l_{\mathrm{q}}} \omega_l = \mathrm{mes}(K)$;*

(ii) *un ensemble de l_{q} points $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$ dans K, ci-après appelés* points de Gauß *ou* nœuds de la quadrature.

Le plus grand entier k tel que

$$\forall p \in \mathbb{P}_k, \qquad \int_K p(x)\,\mathrm{d}x = \sum_{l=1}^{l_{\mathrm{q}}} \omega_l\, p(\xi_l), \tag{9.1}$$

est appelé l'ordre de la quadrature *et est désigné par k_{q}.*

L'hypothèse $\sum_{l=1}^{l_q} \omega_l = \mathrm{mes}(K)$ permet d'affirmer que la quadrature est au moins d'ordre 0. Par ailleurs, une quadrature sur K permet d'approcher l'intégrale d'une fonction suffisamment régulière sur K. En effet, à l'aide de développements de Taylor, on vérifie aisément le résultat suivant (voir la section A.2 pour les notations).

Proposition 9.2. *Pour tout* $\phi \in \mathcal{C}^{k_q+1}(K)$*, on a*

$$\frac{1}{\mathrm{mes}(K)} \left| \int_K \phi(x)\,\mathrm{d}x - \sum_{l=1}^{l_q} \omega_l\, \phi(\xi_l) \right| \leqslant c\, h_K^{k_q+1} |\phi|_{\mathcal{C}^{k_q+1}(K)}, \tag{9.2}$$

où $h_K = \mathrm{diam}(K)$.

Dans le cadre de la méthode des éléments finis, on souhaite évaluer des intégrales sur le domaine Ω où est posé le problème modèle. Soit ϕ une fonction définie sur Ω que l'on suppose suffisamment régulière. On dispose d'un maillage $\mathcal{T}_h$ de Ω, celui-ci étant construit à partir d'un élément fini géométrique $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$; voir la section 3.3.1. Pour une maille $K \in \mathcal{T}_h$, on note $T_K : \widehat{K} \to K$ la transformation géométrique engendrant K. On rappelle que T_K est un $\mathcal{C}^1$-difféomorphisme et pour $\widehat{x} \in \widehat{K}$, on pose $J_K(\widehat{x}) = \frac{\partial T_K(\widehat{x})}{\partial \widehat{x}} \in \mathbb{R}^{d,d}$ (si le maillage est affine, J_K ne dépend pas de $\widehat{x}$ et son déterminant est égal au rapport entre la mesure de K et celle de $\widehat{K}$). En effectuant le changement de variables $x = T_K(\widehat{x})$, il vient

$$\int_K \phi(x)\,\mathrm{d}x = \int_{\widehat{K}} \phi\big(T_K(\widehat{x})\big)\, \det\big(J_K(\widehat{x})\big)\, \mathrm{d}\widehat{x}. \tag{9.3}$$

On suppose qu'on dispose d'une quadrature avec l_q points de Gauß sur la maille de référence $\widehat{K}$. On désigne par $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_{l_q}\}$ ces points de Gauß et par $\{\widehat{\omega}_1, \ldots, \widehat{\omega}_{l_q}\}$ leurs poids respectifs. Il vient

$$\int_K \phi(x)\,\mathrm{d}x \approx \sum_{l=1}^{l_q} \widehat{\omega}_l\, \det\big(J_K(\widehat{\xi}_l)\big)\, \phi\big(T_K(\widehat{\xi}_l)\big). \tag{9.4}$$

En posant pour $l \in \{1, \ldots, l_{\mathrm{q}}\}$,

$$\omega_{K,l} = \widehat{\omega}_l \det\big(J_K(\widehat{\xi}_l)\big), \tag{9.5}$$

$$\xi_{K,l} = T_K(\widehat{\xi}_l), \tag{9.6}$$

on obtient une quadrature sur K de la forme

$$\int_K \phi(x)\,\mathrm{d}x \approx \sum_{l=1}^{l_{\mathrm{q}}} \omega_{K,l}\,\phi(\xi_{K,l}), \tag{9.7}$$

d'où on déduit, en sommant sur les mailles, une quadrature sur Ω de la forme

$$\int_\Omega \phi(x)\,\mathrm{d}x \approx \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{l_{\mathrm{q}}} \omega_{K,l}\,\phi(\xi_{K,l}). \tag{9.8}$$

Les deux résultats suivants permettent d'estimer l'erreur de quadrature d'abord sur une maille quelconque $K \in \mathcal{T}_h$ puis sur le domaine de calcul Ω.

Lemme 9.3. *On suppose que $\{\mathcal{T}_h\}_{h>0}$ est une famille régulière de maillages affines. On désigne par k_{q} l'ordre de la quadrature sur $\widehat{K}$ et on suppose que $k_{\mathrm{q}} + 1 > \frac{d}{2}$. Soit un entier s tel que $\frac{d}{2} < s \leqslant k_{\mathrm{q}} + 1$. Alors, il existe une constante $c(\widehat{K})$, indépendante de h, telle que, pour tout $K \in \mathcal{T}_h$ et pour tout $\phi \in H^s(K)$,*

$$\left| \int_K \phi(x)\,\mathrm{d}x - \sum_{l=1}^{l_{\mathrm{q}}} \omega_{K,l}\,\phi(\xi_{K,l}) \right| \leqslant c(\widehat{K})\, h_K^s \,\mathrm{mes}(K)^{\frac{1}{2}}\, |\phi|_{s,K}. \tag{9.9}$$

Théorème 9.4. *Dans le cadre des hypothèses du lemme 9.3, il existe une constante c, indépendante de h, telle que, pour tout $\phi \in H^s(\Omega)$,*

$$\left| \int_\Omega \phi(x)\,\mathrm{d}x - \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{l_{\mathrm{q}}} \omega_{K,l}\,\phi(\xi_{K,l}) \right| \leqslant c\, h^s |\phi|_{s,\Omega}. \tag{9.10}$$

Le théorème 9.4 est une conséquence directe du lemme 9.3. En effet, en notant R le membre de gauche dans (9.10), on obtient

$$\begin{aligned} R &\leqslant \sum_{K\in\mathcal{T}_h} \left| \int_K \phi(x)\,\mathrm{d}x - \sum_{l=1}^{l_q} \omega_{K,l}\,\phi(\xi_{K,l}) \right| \\ &\leqslant c(\widehat{K}) \sum_{K\in\mathcal{T}_h} h_K^s \,\mathrm{mes}(K)^{\frac{1}{2}} |\phi|_{s,K} \\ &\leqslant c(\widehat{K})\, h^s \Big(\sum_{K\in\mathcal{T}_h} \mathrm{mes}(K) \Big)^{\frac{1}{2}} \Big(\sum_{K\in\mathcal{T}_h} |\phi|_{s,K}^2 \Big)^{\frac{1}{2}}, \end{aligned}$$

grâce à l'inégalité de Cauchy–Schwarz. On conclut en observant que

$$\sum_{K\in\mathcal{T}_h} \mathrm{mes}(K) = \mathrm{mes}(\Omega).$$

Dans le cadre de la méthode des éléments finis, on a parfois besoin d'évaluer des intégrales sur certaines faces du maillage. C'est le cas par exemple lorsque le problème modèle fait intervenir des conditions aux limites de type Neumann ; voir la section 5.1.4. Pour simplifier, on suppose que toutes les faces de l'élément de référence $\widehat{K}$ sont équivalentes, à une transformation bijective près, à une face de référence $\widehat{F} \subset \mathbb{R}^{d-1}$. Par exemple, si $\widehat{K}$ est un triangle ou un carré de $\mathbb{R}^2$, la face de référence $\widehat{F}$ est le segment unité $[0,1]$. Pour une face $F \in \mathcal{F}_h$, on note $T_F : \widehat{F} \to F$ la transformation géométrique engendrant F. En dimension quelconque, on observera que $\widehat{F} \subset \mathbb{R}^{d-1}$ mais que $F \subset \mathbb{R}^d$. Pour $\widehat{x} \in \widehat{F}$, on pose $J_F(\widehat{x}) = \frac{\partial T_F(\widehat{x})}{\partial \widehat{x}} \in \mathbb{R}^{d,d-1}$ (si le maillage est affine, J_F ne dépend pas de $\widehat{x}$). En effectuant le changement de variables $x = T_F(\widehat{x})$, il vient

$$\int_F \phi(x)\,\mathrm{d}x = \int_{\widehat{F}} \phi\big(T_F(\widehat{x})\big)\,\det(g_F(\widehat{x}))^{\frac{1}{2}}\,\mathrm{d}\widehat{x}, \tag{9.11}$$

avec $g_F(\widehat{x}) = (J_F(\widehat{x}))^T J_F(\widehat{x})$. Si le maillage est affine, la quantité $\det(g_F(\widehat{x}))^{\frac{1}{2}}$ est égale au rapport entre la mesure de F et celle de $\widehat{F}$.

On considère une quadrature sur $\widehat{F}$ définie par les $l_{\mathrm{q}}^{\partial}$ points de Gauß $\{\widehat{\xi}_1^{\partial}, \ldots, \widehat{\xi}_{l_{\mathrm{q}}^{\partial}}^{\partial}\}$ et les $l_{\mathrm{q}}^{\partial}$ poids $\{\widehat{\omega}_1^{\partial}, \ldots, \widehat{\omega}_{l_{\mathrm{q}}^{\partial}}^{\partial}\}$. En posant pour $l \in \{1, \ldots, l_{\mathrm{q}}^{\partial}\}$,

$$\omega_{F,l} = \widehat{\omega}_l^{\partial} \det(g_F(\widehat{\xi}_l^{\partial}))^{\frac{1}{2}}, \tag{9.12}$$

$$\xi_{F,l} = T_F(\widehat{\xi}_l^{\partial}), \tag{9.13}$$

on obtient une *quadrature surfacique* sur F de la forme

$$\int_F \phi(x)\,\mathrm{d}x \approx \sum_{l=1}^{l_{\mathrm{q}}^{\partial}} \omega_{F,l}\,\phi(\xi_{F,l}). \tag{9.14}$$

9.2 Exemples de quadratures

Cette section dresse un catalogue, non-exhaustif, des quadratures les plus fréquemment rencontrées dans le cadre de la méthode des éléments finis. Pour des compléments, on pourra consulter Abramowitz et Stegun [**1**], Davis et Rabinowitz [**34**] ou Stroud [**71**].

9.2.1 Quadratures en dimension un : points de Gauß–Legendre

On considère les polynômes de Legendre introduits dans la définition 4.14. Soit un entier $l_{\mathrm{q}} \geqslant 1$. On rappelle que les l_{q} racines du polynôme de Legendre $\mathcal{E}_{l_{\mathrm{q}}}$ sont toutes dans l'intervalle $[0, 1]$.

Proposition 9.5. *On désigne par* $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$ *les* l_{q} *racines du polynôme de Legendre* $\mathcal{E}_{l_{\mathrm{q}}}$. *On pose pour* $l \in \{1, \ldots, l_{\mathrm{q}}\}$,

$$\omega_l = \int_0^1 \prod_{\substack{j=1 \\ j \neq l}}^{l_{\mathrm{q}}} \frac{t - \xi_j}{\xi_l - \xi_j}\,\mathrm{d}t. \tag{9.15}$$

Alors, la quadrature basée sur les points de Gauß $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$ *avec les poids respectifs* $\{\omega_1, \ldots, \omega_{l_{\mathrm{q}}}\}$ *est d'ordre* $k_{\mathrm{q}} = 2l_{\mathrm{q}} - 1$ *sur l'intervalle de référence* $\widehat{K} = [0, 1]$.

La formule (9.15) pour les poids garantit que la quadrature est au moins d'ordre $(l_{\mathrm{q}} - 1)$. En effet, en notant $\{\mathcal{L}_1^{\xi}, \ldots, \mathcal{L}_{l_{\mathrm{q}}}^{\xi}\}$ les polynômes d'interpolation de Lagrange associés aux points de Gauß $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$, on constate que $\omega_l = \int_0^1 \mathcal{L}_l^{\xi}(t)\,\mathrm{d}t$ pour tout $l \in \{1, \ldots, l_{\mathrm{q}}\}$. Par ailleurs, pour tout $p \in \mathbb{P}_{l_{\mathrm{q}}-1}$, on a

$$p = \sum_{l=1}^{l_{\mathrm{q}}} p(\xi_l)\mathcal{L}_l^{\xi}. \tag{9.16}$$

On en déduit

$$\int_0^1 p(t)\,\mathrm{d}t = \sum_{l=1}^{l_{\mathrm{q}}} p(\xi_l) \int_0^1 \mathcal{L}_l^{\xi}(t)\,\mathrm{d}t = \sum_{l=1}^{l_{\mathrm{q}}} p(\xi_l)\omega_l, \tag{9.17}$$

ce qui montre que la quadrature est au moins d'ordre $(l_{\mathrm{q}} - 1)$. Le fait que la quadrature est d'ordre $(2l_{\mathrm{q}} - 1)$ tient à la propriété d'orthogonalité des polynômes de Legendre ; voir la formule (4.68). Soit $p \in \mathbb{P}_{2l_{\mathrm{q}}-1}$. En effectuant la division euclidienne du polynôme p par $\mathcal{E}_{l_{\mathrm{q}}}$, on obtient $p = q\mathcal{E}_{l_{\mathrm{q}}} + r$ où q et r sont dans $\mathbb{P}_{l_{\mathrm{q}}-1}$. On en déduit

$$\begin{aligned} \int_0^1 p(t)\,\mathrm{d}t &= \int_0^1 [q(t)\mathcal{E}_{l_{\mathrm{q}}}(t) + r(t)]\,\mathrm{d}t \\ &= \int_0^1 r(t)\,\mathrm{d}t = \sum_{l=1}^{l_{\mathrm{q}}} r(\xi_l)\omega_l \\ &= \sum_{l=1}^{l_{\mathrm{q}}} p(\xi_l)\omega_l, \end{aligned} \tag{9.18}$$

car les points de Gauß $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$ sont les racines de $\mathcal{E}_{l_{\mathrm{q}}}$. La quadrature de la proposition 9.5 est donc d'ordre (au moins) $(2l_{\mathrm{q}} - 1)$. Enfin, on peut facilement exhiber un contre-exemple montrant qu'elle n'est pas d'ordre supérieur.

Par un simple changement de variables, on déduit de la quadrature sur $\widehat{K} = [0, 1]$ définie dans la proposition 9.5 une quadrature sur un intervalle quelconque $K = [a, b]$. Les nœuds associés $\{\xi_1, \ldots, \xi_{l_{\mathrm{q}}}\}$ sont appelés les *points de Gauß–Legendre* dans K. Le tableau 9.1 présente les points de Gauß–Legendre et les poids correspondants pour les quadratures d'ordre 1, 3, 5 et 7 sur un intervalle borné $[a, b]$.

Tableau 9.1 – Points de Gauß–Legendre et poids associés pour les quadratures d'ordre 1, 3, 5 et 7 sur l'intervalle $[a, b]$; $m = \frac{1}{2}(a + b)$ et $\delta = b - a$.

k_q	l_q	points de Gauß–Legendre	poids associés
1	1	m	δ
3	2	$m \pm \frac{\delta}{2}\frac{\sqrt{3}}{3}$	$\frac{1}{2}\delta$
5	3	$m \pm \frac{\delta}{2}\sqrt{\frac{3}{5}}$	$\frac{5}{18}\delta$
		m	$\frac{8}{18}\delta$
7	4	$m \pm \frac{\delta}{2}\sqrt{\frac{1}{35}(15 + 2\sqrt{30})}$	$(\frac{1}{4} - \frac{1}{12}\sqrt{\frac{5}{6}})\delta$
		$m \pm \frac{\delta}{2}\sqrt{\frac{1}{35}(15 - 2\sqrt{30})}$	$(\frac{1}{4} + \frac{1}{12}\sqrt{\frac{5}{6}})\delta$

Remarque 9.6

Les racines du polynôme de Legendre $\mathcal{E}_{l_q}$ ne peuvent être déterminées de manière analytique que pour les premières valeurs de l_q. Lorsqu'on souhaite utiliser des quadratures de degré élevé, les points de Gauß–Legendre et leurs poids sont déterminés de manière approchée par une méthode itérative ; on pourra consulter Karniadakis et Spencer [**54**, p. 357] pour plus de détails. La même remarque est valable pour les points de Gauß–Lobatto présentés dans la section 9.2.2.

9.2.2 Quadratures en dimension un : points de Gauß–Lobatto

Soit un entier $k \geqslant 2$. On considère les $(k + 1)$ points de Gauß–Lobatto $\{g_0^k, \ldots, g_k^k\}$ sur l'intervalle de référence $\widehat{K} = [0, 1]$. On désigne par $\{\theta_0^{\mathrm{GL},k}, \ldots, \theta_k^{\mathrm{GL},k}\}$ les $(k + 1)$ polynômes d'interpolation de Lagrange associés aux $(k + 1)$ points de Gauß–Lobatto ; voir la définition 4.17 et la proposition 4.18.

Proposition 9.7. *On pose* $l_{\mathrm{q}} = k + 1$ *et pour* $l \in \{1, \dots, l_{\mathrm{q}}\}$,

$$\omega_l = \frac{1}{k(k+1)} \frac{1}{[\mathcal{E}_k(g_{l-1}^k)]^2}, \tag{9.19}$$

$$\xi_l = g_{l-1}^k. \tag{9.20}$$

Alors, la quadrature basée sur les points de Gauß–Lobatto $\{\xi_1, \dots, \xi_{l_{\mathrm{q}}}\}$ *avec les poids respectifs* $\{\omega_1, \dots, \omega_{l_{\mathrm{q}}}\}$ *est d'ordre* $k_{\mathrm{q}} = 2k - 1 = 2l_{\mathrm{q}} - 3$ *sur l'intervalle de référence* $\widehat{K} = [0, 1]$.

Tableau 9.2 – Points de Gauß–Lobatto et poids associés pour les quadratures d'ordre 3, 5 et 7 sur l'intervalle $[a, b]$; $m = \frac{1}{2}(a + b)$ et $\delta = b - a$.

k_{q}	l_{q}	points de Gauß–Lobatto	poids associés
3	3	$m \pm \frac{\delta}{2}$	$\frac{1}{6}\delta$
		m	$\frac{2}{3}\delta$
5	4	$m \pm \frac{\delta}{2}$	$\frac{1}{12}\delta$
		$m \pm \frac{\delta}{2}\sqrt{\frac{1}{5}}$	$\frac{5}{12}\delta$
7	5	$m \pm \frac{\delta}{2}$	$\frac{1}{20}\delta$
		$m \pm \frac{\delta}{2}\sqrt{\frac{3}{7}}$	$\frac{49}{180}\delta$
		m	$\frac{16}{45}\delta$

Par un simple changement de variables, on déduit de la quadrature sur $\widehat{K} = [0, 1]$ définie dans la proposition 9.7 une quadrature sur un intervalle quelconque $K = [a, b]$. Les nœuds associés $\{\xi_1, \dots, \xi_{l_{\mathrm{q}}}\}$ sont appelés les *points de Gauß–Lobatto* dans K. Le tableau 9.2 présente les points de Gauß–Lobatto et les poids correspondants pour les quadratures d'ordre 3, 5 et 7 sur un intervalle borné $[a, b]$. La quadrature d'ordre 3 est également connue sous le nom de *formule de Simpson*.

9.2.3 Quadratures sur un triangle

Le tableau 9.3 présente quelques quadratures sur un triangle quelconque (non-dégénéré). La position des points de Gauß avec leur poids associé est illustrée dans la figure 9.1 sur un triangle rectangle isocèle.

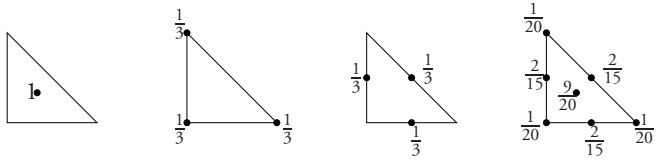


Figure 9.1 – Points de Gauß et poids pour diverses quadratures sur le triangle ; les poids doivent être multipliés par la surface du triangle.

Tableau 9.3 – Points de Gauß et poids pour diverses quadratures sur un triangle non-dégénéré de surface S. La multiplicité indique le nombre de permutations à réaliser sur les coordonnées barycentriques afin d'engendrer tous les points de Gauß.

k_q	l_q	coord. barycentriques	multiplicité	poids
1	1	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	S
1	3	$(1,0,0)$	3	$\frac{1}{3}S$
2	3	$\left(\frac{1}{2},\frac{1}{2},0\right)$	3	$\frac{1}{3}S$
3	7	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	$\frac{9}{20}S$
		$\left(\frac{1}{2},\frac{1}{2},0\right)$	3	$\frac{2}{15}S$
		$(1,0,0)$	3	$\frac{1}{20}S$
5	7	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	$\frac{9}{40}S$
		$(a_i, a_i, 1-2a_i) \quad i \in \{1,2\}$	3	
		$a_1 = \frac{6-\sqrt{15}}{21}$		$\frac{155-\sqrt{15}}{1200}S$
		$a_2 = \frac{6+\sqrt{15}}{21}$		$\frac{155+\sqrt{15}}{1200}S$

Une autre quadrature d'ordre $k_q = 2$ sur le triangle est basée sur les $l_q = 3$ points de Gauß de coordonnées barycentriques $(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$, $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$ et $(\frac{1}{6}, \frac{1}{6}, \frac{2}{3})$, chacun avec le poids $\frac{1}{3}S$. Une autre quadrature d'ordre $k_q = 3$ est basée sur les $l_q = 4$ points de Gauß de coordonnées barycentriques $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $(\frac{3}{5}, \frac{1}{5}, \frac{1}{5})$, $(\frac{1}{5}, \frac{3}{5}, \frac{1}{5})$ et $(\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$, le premier avec le poids $-\frac{9}{16}S$ et les trois autres avec le poids $\frac{25}{48}S$. L'avantage de cette quadrature d'ordre 3 par rapport à celle présentée dans le tableau 9.3 est qu'elle emploie 4 points de Gauß au lieu de 7 ; son coût de calcul est donc moindre.

9.2.4 Quadratures sur un tétraèdre

Le tableau 9.4 présente quelques quadratures sur un tétraèdre quelconque (non-dégénéré).

Tableau 9.4 – Points de Gauß et poids pour diverses quadratures sur un tétraèdre non-dégénéré de volume V. La multiplicité indique le nombre de permutations à réaliser sur les coordonnées barycentriques afin d'engendrer tous les points de Gauß.

k_q	l_q	coord. barycentriques	multiplicité	poids
1	1	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1	V
1	4	$(1, 0, 0, 0)$	4	$\frac{1}{4}V$
2	4	$(a, a, a, 1-3a)$ $a = \frac{5-\sqrt{5}}{20}$	4	$\frac{1}{4}V$
2	10	$(\frac{1}{2}, \frac{1}{2}, 0, 0)$ $(1, 0, 0, 0)$	6 4	$\frac{1}{5}V$ $-\frac{1}{20}V$
3	5	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2})$	1 4	$-\frac{4}{5}V$ $\frac{9}{20}V$
5	15	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ $(a_i, a_i, a_i, 1-2a_i) \quad i \in \{1, 2\}$ $a_1 = \frac{7-\sqrt{15}}{34}$ $a_2 = \frac{7+\sqrt{15}}{34}$ $(a, a, \frac{1}{2}-a, \frac{1}{2}-a)$ $a = \frac{10-2\sqrt{5}}{40}$	1 4 6 	$\frac{16}{135}V$ $\frac{2665+14\sqrt{15}}{37800}V$ $\frac{2665-14\sqrt{15}}{37800}V$ $\frac{10}{189}V$

9.2.5 Quadratures sur un hypercube

Lorsque K est un hypercube (un carré en dimension 2 ou un cube en dimension 3), la manière la plus simple de construire une quadrature sur K consiste à considérer des produits tensoriels de quadratures en dimension 1. On considère une quadrature d'ordre k_q sur l'intervalle de référence $[0, 1]$. On désigne par $\{\xi_1, \ldots, \xi_{l_q}\}$ les l_q points de Gauß de cette quadrature et par $\{\omega_1, \ldots, \omega_{l_q}\}$ leurs poids respectifs.

Proposition 9.8. *On considère un hypercube de* $\mathbb{R}^d$ *de la forme* $K = \prod_{j=1}^{d} [a_j, b_j]$. *La quadrature à* $(l_q)^d$ *points de Gauß* $\{\xi_{i_1 \ldots i_d}\}_{1 \leqslant i_1, \ldots, i_d \leqslant l_q}$ *de coordonnées cartésiennes*

$$\xi_{i_1 \ldots i_d} = (a_j + (b_j - a_j)\xi_{i_j})_{1 \leqslant j \leqslant d}, \tag{9.21}$$

et de poids associés

$$\omega_{i_1 \ldots i_d} = \prod_{j=1}^{d} (b_j - a_j)\omega_{i_j}, \tag{9.22}$$

est exacte pour les polynômes de $\mathbb{Q}_{k_q}$ *sur* K.

La figure 9.2 présente trois exemples de quadratures sur un carré construites selon la proposition 9.8 à partir des quadratures de Gauß–Lobatto d'ordre 3, 4 et 5 en dimension un ; les quadratures sur le carré font intervenir 9, 16 et 25 points de Gauß, respectivement.

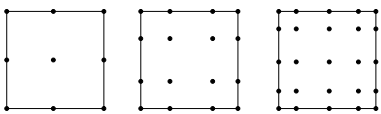


Figure 9.2 – Position des points de Gauß pour une quadrature sur le carré construite à partir des quadratures de Gauß–Lobatto d'ordre 3 (gauche), 4 (centre) et 5 (droite) sur l'intervalle de référence $[0, 1]$.

Remarque 9.9

Les polynômes de $\mathbb{Q}_{k_q}$ peuvent contenir des monômes de degré total dk_q ; toutefois, la quadrature construite dans la proposition 9.8 n'est pas d'ordre dk_q. En d'autres termes, il existe des polynômes de $\mathbb{P}_{dk_q}$ pour lesquels cette quadrature n'est pas exacte.

9.3 Erreurs de quadrature dans la méthode des éléments finis

L'objet de cette section est d'étudier l'impact des quadratures sur la méthode des éléments finis. L'utilisation de quadratures induit des perturbations au niveau de la forme bilinéaire et du membre de droite intervenant dans le problème approché. Deux questions se posent :

(i) le problème perturbé par l'utilisation de quadratures reste-t-il bien posé ?

(ii) en cas de réponse positive, comment choisir l'ordre des quadratures pour que la solution de ce problème perturbé satisfasse une estimation d'erreur optimale en h.

9.3.1 Le problème discret avec quadratures

Afin de répondre aux questions ci-dessus, on considère le problème discret suivant :

$$\left\{ \begin{array}{l} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in W_h, \end{array} \right. \tag{9.23}$$

où V_h et W_h sont des espaces d'approximation, $a_h \in \mathcal{L}(V_h \times W_h; \mathbb{R})$ et $f_h \in \mathcal{L}(W_h; \mathbb{R})$. Les espaces d'approximation V_h et W_h sont construits par la méthode des éléments finis à partir d'un maillage $\mathcal{T}_h$ de Ω et d'un élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$.

On suppose que la forme bilinéaire a_h et la forme linéaire f_h s'écrivent sous la forme

$$a_h(v_h, w_h) = \int_\Omega A_h(x, v_h, w_h)\, \mathrm{d}x, \tag{9.24}$$

$$f_h(w_h) = \int_\Omega F_h(x, w_h)\, \mathrm{d}x + \int_{\partial\Omega_N} G_h(x, w_h)\, \mathrm{d}s, \tag{9.25}$$

où $A_h : \Omega \times V_h \times W_h \to \mathbb{R}$, $F_h : \Omega \times W_h \to \mathbb{R}$ et $G_h : \partial\Omega_N \times W_h \to \mathbb{R}$ sont des opérateurs et où $\partial\Omega_N$ est une partie de la frontière de Ω (dans ce qui suit, $\partial\Omega_N$ peut éventuellement être de mesure nulle).

Pour fixer les idées, on considère un problème d'advection–diffusion–réaction avec des conditions aux limites mêlées Dirichlet–Neumann. On introduit l'opérateur différentiel

$$\mathcal{L}v = -\nabla\cdot(\sigma\cdot\nabla v) + \beta\cdot\nabla v + \mu v, \tag{9.26}$$

avec $\sigma \in [L^\infty(\Omega)]^{d,d}$, $\beta \in [L^\infty(\Omega)]^d$, $\nabla\cdot\beta \in L^\infty(\Omega)$ et $\mu \in L^\infty(\Omega)$. On suppose que l'opérateur $\mathcal{L}$ est elliptique au sens de la définition 5.9. On considère le problème modèle suivant : étant donné une partition $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ où $\partial\Omega_D$ est de mesure strictement positive et des fonctions $f \in L^2(\Omega)$ et $g \in L^2(\partial\Omega_N)$,

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a_{\sigma\beta\mu}(u, w) = f(w), \quad \forall w \in V, \end{cases} \tag{9.27}$$

où $V = \{v \in H^1(\Omega)\ ;\ v|_{\partial\Omega_D} = 0\}$ et pour $(v, w) \in V \times V$,

$$a_{\sigma\beta\mu}(v, w) = \int_\Omega \nabla w\cdot\sigma\cdot\nabla v + w(\beta\cdot\nabla v) + \mu wv, \tag{9.28}$$

$$f(w) = \int_\Omega fw + \int_{\partial\Omega_N} gw. \tag{9.29}$$

On suppose que le problème modèle (9.27) est bien posé[1]. La solution unique de (9.27) satisfait $\mathcal{L}u = f$ dans $L^2(\Omega)$, $u|_{\partial\Omega_D} = 0$ et $n\cdot\sigma\cdot\nabla u|_{\partial\Omega_N} = g$. Enfin, en utilisant une approximation conforme de (9.27), on obtient le problème discret (9.23) avec

$$A_h(x, v_h, w_h) = \nabla w_h\cdot\sigma(x)\cdot\nabla v_h + w_h(\beta(x)\cdot\nabla v_h) + \mu(x)v_h w_h, \tag{9.30}$$

$$F_h(x, w_h) = f(x)w_h, \tag{9.31}$$

$$G_h(x, w_h) = g(x)w_h. \tag{9.32}$$

1. Une condition suffisante pour que le problème modèle (9.27) soit bien posé est que la constante de diffusion σ_0 intervenant dans (5.58) soit suffisamment grande. Cette condition est donnée en (5.63) lorsque $\partial\Omega = \partial\Omega_D$. Lorsque $\partial\Omega_N$ est de mesure strictement positive, la condition est analogue mais fait intervenir une constante de Poincaré sur V au lieu de la constante de Poincaré classique sur $H^1_0(\Omega)$.

L'utilisation de quadratures dans l'évaluation des intégrales volumiques et surfaciques intervenant dans la forme bilinéaire a_h et dans la forme linéaire f_h induit une perturbation du problème discret (9.23). En effet, ces formes sont approchées de la manière suivante :

$$a_h(v_h, w_h) \approx a_{hQ}(v_h, w_h), \tag{9.33}$$

$$f_h(w_h) \approx f_{hQ}(w_h), \tag{9.34}$$

avec

$$a_{hQ}(v_h, w_h) = \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{l_q} \omega_{K,l} A_h(\xi_{K,l}, v_h(\xi_{K,l}), w_h(\xi_{K,l})), \tag{9.35}$$

$$f_{hQ}(w_h) = \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{l_q} \omega_{K,l} F_h(\xi_{K,l}, w_h(\xi_{K,l})) \tag{9.36}$$

$$+ \sum_{F\in\mathcal{N}_h} \sum_{l=1}^{l_q^\partial} \omega_{F,l} G_h(\xi_{F,l}, w_h(\xi_{F,l})). \tag{9.37}$$

Pour simplifier, on a supposé que le maillage $\mathcal{T}_h$ est tel que $\partial\Omega_N$ est une union de faces d'éléments de $\mathcal{T}_h$ et on a désigné par $\mathcal{N}_h$ l'union de ces faces. L'utilisation de quadratures conduit donc au problème approché suivant :

$$\begin{cases} \text{Chercher } u_{hQ} \in V_h \text{ tel que} \\ a_{hQ}(u_{hQ}, w_h) = f_{hQ}(w_h), \quad \forall w_h \in W_h. \end{cases} \tag{9.38}$$

9.3.2 Analyse d'erreur

Les deux résultats suivants permettent de déterminer l'ordre minimal des quadratures volumique et surfacique afin que d'une part le problème perturbé (9.38) soit bien posé et d'autre part qu'il conduise aux mêmes estimations d'erreur qu'en l'absence de quadratures. On pourra consulter Ciarlet [**27**, p. 178], Dautray et Lions [**33**, Chap. 12] ou Raviart et Thomas [**63**, p. 123] pour des compléments sur les résultats ci-dessous.

Théorème 9.10. *On suppose que* $\{\mathcal{T}_h\}_{h>0}$ *est une famille régulière de maillages affines. On suppose par ailleurs que* $\mathbb{P}_1 \subset \widehat{P}$ *et qu'il existe un entier* $l \geqslant 1$ *tel que :*

(i) $\widehat{P} \subset \mathbb{P}_l$ *;*

(ii) *la quadrature sur* $\widehat{K}$ *est d'ordre* $2l-2$.

Alors, le problème (9.38) *est bien posé si h est suffisamment petit.*

Théorème 9.11. *Avec les hypothèses du théorème* 9.10*, on suppose qu'il existe un entier* $k \geqslant 1$ *tel que :*

(i) $\mathbb{P}_k \subset \widehat{P}$ *;*

(ii) *la quadrature surfacique est d'ordre* $k+l-1$.

On suppose par ailleurs que $f \in \mathcal{C}^k(\Omega)$, $g \in H^{k+1}(\partial\Omega_N)$ *et* $u \in H^{k+1}(\Omega)$. *Alors, il existe* $h_0 > 0$ *et c tels que, pour tout* $h \leqslant h_0$,

$$\|u - u_{hQ}\|_{1,\Omega} \leqslant ch^k(\|u\|_{k+1,\Omega} + \|f\|_{\mathcal{C}^k(\Omega)} + h^{\frac{1}{2}}\|g\|_{H^{k+1}(\partial\Omega_N)}). \qquad (9.39)$$

Par exemple, si on considère une approximation par éléments finis de Lagrange $\mathbb{P}_1$, la quadrature volumique doit être au moins d'ordre 0 et la quadrature surfacique au moins d'ordre 1 pour que l'erreur $u - u_{hQ}$ soit d'ordre 1 en norme H^1. La régle générale pour déterminer l'ordre de la quadrature volumique est que les dérivées d'ordre le plus élevé dans la forme bilinéaire doivent être évaluées exactement lorsque les coefficients de l'équation aux dérivées partielles (ici, σ, β et μ) sont constants.

10 • MATRICES D'ÉLÉMENTS FINIS

La méthode des éléments finis permet d'approcher la solution d'un problème modèle en résolvant un problème discret formulé à l'aide d'un espace d'approximation. En cherchant les composantes de la solution discrète dans une base particulière de l'espace d'approximation constituée des fonctions de forme associées à l'élément fini, on se ramène à la résolution d'un système linéaire

$$\mathcal{A}U = F, \tag{10.1}$$

où $\mathcal{A} \in \mathbb{R}^{N,N}$ est appelée la *matrice de rigidité*, $F \in \mathbb{R}^N$ dépend des données du problème et N est la dimension de l'espace d'approximation ; voir la section 2.2.

L'objet de ce chapitre est d'étudier quelques propriétés importantes de la matrice de rigidité. La première est son *conditionnement*. Il s'agit d'évaluer un réel permettant de quantifier la difficulté numérique à résoudre le système (10.1) ou à en construire une solution approchée par une méthode itérative. Plus le nombre de conditionnement de $\mathcal{A}$ est élevé, plus ces difficultés sont grandes. Ces difficultés se déclinent d'une part en termes de stabilité de la solution de (10.1) par rapport à des perturbations des coefficients de $\mathcal{A}$ et de F et d'autre part en termes de taux de convergence si on utilise une méthode itérative pour approcher U. La deuxième section de ce chapitre présente quelques rappels sur les techniques de factorisation LU (de l'anglais Lower–Upper Factorization). La présentation de cette section est relativement générale et ne tient pas compte des spécificités de la structure des matrices d'éléments finis. La troisième section est centrée autour d'une propriété spécifique des matrices d'éléments finis, à savoir leur *structure creuse* : la matrice de rigidité contient

très peu d'éléments non-nuls et la disposition de ces coefficients dans la matrice dépend directement de la numérotation choisie pour les fonctions de forme dans l'espace d'approximation. On présente quelques techniques de renumérotation permettant de regrouper autant que possible ces coefficients autour de la diagonale.

10.1 Conditionnement

L'objet de cette section est d'étudier d'une part le conditionnement de la matrice de rigidité $\mathcal{A}$ et d'autre part la stabilité de la solution U de (10.1) par rapport à des perturbations des coefficients de $\mathcal{A}$ et de F.

10.1.1 Préliminaires

On rappelle que la norme euclidienne sur $\mathbb{R}^N$ est définie pour tout $X = (X_i)_{1 \leqslant i \leqslant N} \in \mathbb{R}^N$ par

$$\|X\|_{\mathbb{R}^N} = \left(\sum_{i=1}^{N} |X_i|^2 \right)^{\frac{1}{2}}. \tag{10.2}$$

On utilise la même notation pour la norme matricielle induite : pour $\mathcal{Z} \in \mathbb{R}^{N,N}$, on pose

$$\|\mathcal{Z}\|_{\mathbb{R}^N} = \sup_{X \in \mathbb{R}^N} \frac{\|\mathcal{Z}X\|_{\mathbb{R}^N}}{\|X\|_{\mathbb{R}^N}}. \tag{10.3}$$

Définition 10.1. *Le* nombre de conditionnement *d'une matrice inversible* $\mathcal{Z} \in \mathbb{R}^{N,N}$ *est le réel*

$$\kappa(\mathcal{Z}) = \|\mathcal{Z}\|_{\mathbb{R}^N} \|\mathcal{Z}^{-1}\|_{\mathbb{R}^N}. \tag{10.4}$$

Par construction, on a $\kappa(\mathcal{Z}) \geqslant 1$. Par ailleurs, si la matrice $\mathcal{Z}$ est *symétrique*, on montre que le nombre de conditionnement de $\mathcal{Z}$ est tel que

$$\kappa(\mathcal{Z}) = \frac{\lambda_{\max}(\mathcal{Z})}{\lambda_{\min}(\mathcal{Z})}, \tag{10.5}$$

où $\lambda_{\min}(\mathcal{Z})$ et $\lambda_{\max}(\mathcal{Z})$ désignent, respectivement, la plus petite et la plus grande valeur propre de $\mathcal{Z}$ en valeur absolue.

Remarque 10.2

On peut également définir le nombre de conditionnement d'une matrice inversible en utilisant d'autres normes matricielles, par exemple celles induites par les normes vectorielles $\|X\|_p = (\sum_{i=1}^N |X_i|^p)^{\frac{1}{p}}$ pour $p \in [1, +\infty[$ ou $\|X\|_\infty = \max_{1\leqslant i\leqslant N} |X_i|$. Le nombre de conditionnement défini en (10.4) est alors appelé le nombre de conditionnement euclidien de la matrice $\mathcal{Z}$.

Définition 10.3. *On dit qu'une matrice inversible* $\mathcal{Z} \in \mathbb{R}^{N,N}$ *est* mal conditionnée *si on a*

$$\kappa(\mathcal{Z}) \gg 1. \tag{10.6}$$

Par la suite, on considère le problème approché suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a_h(u_h, w_h) = f_h(w_h), \quad \forall w_h \in W_h, \end{cases} \tag{10.7}$$

où V_h et W_h sont des espaces d'approximation, a_h est une forme bilinéaire sur $V_h \times W_h$ et f_h une forme linéaire sur W_h. Les espaces V_h et W_h sont construits en utilisant les techniques d'éléments finis présentées dans le chapitres 3 et 4 ; en particulier, on note $\mathcal{T}_h$ le maillage du domaine Ω. On introduit les deux constantes suivantes :

$$\alpha_h = \inf_{v_h\in V_h} \sup_{w_h\in W_h} \frac{a_h(v_h, w_h)}{\|v_h\|_{V_h}\|w_h\|_{W_h}}, \tag{10.8}$$

$$\omega_h = \sup_{v_h\in V_h} \sup_{w_h\in W_h} \frac{a_h(v_h, w_h)}{\|v_h\|_{V_h}\|w_h\|_{W_h}}. \tag{10.9}$$

Par la suite, on suppose que

$$\alpha_h > 0 \qquad \text{et} \qquad \dim V_h = \dim W_h. \tag{10.10}$$

Ces deux conditions impliquent que le problème (10.7) est bien posé.

On pose $N = \dim V_h = \dim W_h$ et on désigne par $\{\varphi_1, \ldots, \varphi_N\}$ et $\{\psi_1, \ldots, \psi_N\}$ la base de V_h et de W_h constituée des fonctions de forme dans V_h et dans W_h, respectivement. Pour une fonction $v_h \in V_h$, on note $X \in \mathbb{R}^N$ le vecteur formé par les coordonnées de v_h dans la base $\{\varphi_1, \ldots, \varphi_N\}$. De même, pour une fonction $w_h \in W_h$, on note $Y \in \mathbb{R}^N$ le vecteur formé par les

coordonnées de w_h dans la base $\{\psi_1, \ldots, \psi_N\}$. On a donc

$$v_h = \sum_{i=1}^{N} X_i \varphi_i \qquad \text{et} \qquad w_h = \sum_{i=1}^{N} Y_i \psi_i. \tag{10.11}$$

On introduit les isomorphismes suivants :

$$C_{V_h} : \quad V_h \ni v_h \longmapsto X \in \mathbb{R}^N, \tag{10.12}$$

$$C_{W_h} : \; W_h \ni w_h \longmapsto Y \in \mathbb{R}^N. \tag{10.13}$$

Lemme 10.4. *On suppose que la famille $\{\mathcal{T}_h\}_{h>0}$ est quasi-uniforme. Alors, il existe des constantes positives c_1, c_2, c_3 et c_4, indépendantes de h, telles que*

$$\forall v_h \in V_h, \qquad c_1 h^{\frac{d}{2}} \leqslant \frac{\|v_h\|_{0,\Omega}}{\|C_{V_h} v_h\|_{\mathbb{R}^N}} \leqslant c_2 h^{\frac{d}{2}}, \tag{10.14}$$

$$\forall w_h \in W_h, \qquad c_3 h^{\frac{d}{2}} \leqslant \frac{\|w_h\|_{0,\Omega}}{\|C_{W_h} w_h\|_{\mathbb{R}^N}} \leqslant c_4 h^{\frac{d}{2}}. \tag{10.15}$$

On introduit les *matrices de masse* $\mathcal{M}_s \in \mathbb{R}^{N,N}$ et $\mathcal{M}_t \in \mathbb{R}^{N,N}$ de terme générique

$$\mathcal{M}_{s,ij} = \int_\Omega \varphi_i \varphi_j \qquad \text{et} \qquad \mathcal{M}_{t,ij} = \int_\Omega \psi_i \psi_j, \tag{10.16}$$

pour $i, j \in \{1, \ldots, N\}$. Les indices s et t font référence au fait que la matrice $\mathcal{M}_s$ est relative à l'espace solution V_h et que la matrice $\mathcal{M}_t$ est relative à l'espace test discret W_h. Les matrices $\mathcal{M}_s$ et $\mathcal{M}_t$ sont symétriques définies positives. On note $\{\mu_{s,1}, \ldots, \mu_{s,N}\}$ et $\{\mu_{t,1}, \ldots, \mu_{t,N}\}$ les valeurs propres (classées par valeurs croissantes) de $\mathcal{M}_s$ et $\mathcal{M}_t$, respectivement. On pose

$$\kappa_s = \kappa(\mathcal{M}_s)^{\frac{1}{2}} = \left(\frac{\mu_{s,N}}{\mu_{s,1}}\right)^{\frac{1}{2}} \qquad \text{et} \qquad \kappa_t = \kappa(\mathcal{M}_t)^{\frac{1}{2}} = \left(\frac{\mu_{t,N}}{\mu_{t,1}}\right)^{\frac{1}{2}}. \tag{10.17}$$

Une conséquence intéressante du lemme 10.4 est la suivante.

Corollaire 10.5. *On suppose que la famille $\{\mathcal{T}_h\}_{h>0}$ est quasi-uniforme. Alors, pour tout $i \in \{1, \ldots, N\}$, on a*

$$c_1^2 h^d \leqslant \mu_{s,i} \leqslant c_2^2 h^d \qquad \text{et} \qquad c_3^2 h^d \leqslant \mu_{t,i} \leqslant c_4^2 h^d, \tag{10.18}$$

si bien que

$$\kappa_s \leqslant \frac{c_2}{c_1} \qquad \text{et} \qquad \kappa_t \leqslant \frac{c_4}{c_3}. \tag{10.19}$$

À titre d'exemple, la matrice de masse associée à l'élément fini de Lagrange $\mathbb{P}_1$ en une dimension d'espace sur un maillage uniforme de pas h est telle que

$$\mathcal{M} = \tfrac{h}{6}\text{tridiag}(1, 4, 1). \tag{10.20}$$

Un calcul direct montre que ses valeurs propres sont les N réels $\{\frac{h}{3}(2 + \cos(ih\pi))\}_{1\leqslant i\leqslant N}$. Toutes ces valeurs propres sont équivalentes à h en accord avec l'estimation (10.18). De plus, le nombre de conditionnement de $\mathcal{M}$, qui est égal au rapport entre sa plus grande et sa plus petite valeur propre, est bien borné uniformément en h en accord avec l'estimation (10.19).

10.1.2 Conditionnement de la matrice de rigidité

L'objectif de cette section est de présenter des bornes inférieures et supérieures pour le nombre de conditionnement $\kappa(\mathcal{A})$ de la matrice de rigidité. Les résultats ci-dessous ont été établis dans [**37**]. On pose

$$\alpha_{0,h} = \inf_{v_h\in V_h}\sup_{w_h\in W_h} \frac{a_h(v_h, w_h)}{\|v_h\|_{0,\Omega}\|w_h\|_{0,\Omega}}, \tag{10.21}$$

$$\omega_{0,h} = \sup_{v_h\in V_h}\sup_{w_h\in W_h} \frac{a_h(v_h, w_h)}{\|v_h\|_{0,\Omega}\|w_h\|_{0,\Omega}}. \tag{10.22}$$

Théorème 10.6. *Pour tout h, on a*

$$\kappa_s^{-1}\kappa_t^{-1}\frac{\omega_{0,h}}{\alpha_{0,h}} \leqslant \kappa(\mathcal{A}) \leqslant \kappa_s\kappa_t\frac{\omega_{0,h}}{\alpha_{0,h}}. \tag{10.23}$$

On souhaite affiner le résultat ci-dessus dans le cas où on connaît le comportement asymptotique en h des constantes $\alpha_{0,h}$ et $\omega_{0,h}$. On suppose qu'il existe

des réels γ et δ tels que

$$0 < c^{\alpha}_{\inf} = \liminf_{h\to 0} \alpha_{0,h} h^{-\gamma} < +\infty, \tag{10.24}$$

$$0 < c^{\alpha}_{\sup} = \limsup_{h\to 0} \alpha_{0,h} h^{-\gamma} < +\infty, \tag{10.25}$$

$$0 < c^{\omega}_{\inf} = \liminf_{h\to 0} \omega_{0,h} h^{-\delta} < +\infty, \tag{10.26}$$

$$0 < c^{\omega}_{\sup} = \limsup_{h\to 0} \omega_{0,h} h^{-\delta} < +\infty. \tag{10.27}$$

Théorème 10.7. *Dans le cadre des hypothèses ci-dessus, pour tout* $\epsilon \in]0,1[$, *il existe* h_ϵ *tel que pour tout* $h \leqslant h_\epsilon$,

$$(1-\epsilon)\frac{c^{\omega}_{\inf}}{c^{\alpha}_{\sup}} \kappa_s^{-1} \kappa_t^{-1} h^{-\gamma-\delta} \leqslant \kappa(\mathcal{A}) \leqslant (1+\epsilon)\frac{c^{\omega}_{\sup}}{c^{\alpha}_{\inf}} \kappa_s \kappa_t h^{-\gamma-\delta}. \tag{10.28}$$

Le théorème 10.7 montre que le nombre de conditionnement de $\mathcal{A}$ est asymptotiquement de l'ordre de $h^{-\gamma-\delta}$. À titre d'illustration, on considère les exemples suivants. On suppose que la famille $\{\mathcal{T}_h\}_{h>0}$ est quasi-uniforme si bien que grâce au corollaire 10.5, les constantes κ_s et κ_t sont bornées inférieurement et supérieurement par des constantes indépendantes de h.

- **Laplacien en formulation primale.** On considère le problème discret (5.3). On montre que $\gamma = 0$ et $\delta = 2$ si bien qu'il existe des constantes positives c_1 et c_2, indépendantes de h, telles que

$$c_1 h^{-2} \leqslant \kappa(\mathcal{A}) \leqslant c_2 h^{-2}. \tag{10.29}$$

  En une dimension d'espace avec des éléments finis de Lagrange $\mathbb{P}_1$ sur un maillage uniforme de pas h, la matrice de rigidité est telle que

$$\mathcal{A} = \tfrac{1}{h}\text{tridiag}(-1, 2, -1). \tag{10.30}$$

  Un calcul direct montre que ses valeurs propres sont les N réels $\{\frac{2}{h}(1 - \cos(ih\pi))\}_{1\leqslant i\leqslant N}$. On vérifie que le nombre de conditionnement de $\mathcal{A}$, qui est égal au rapport entre sa plus grande et sa plus petite valeur propre, explose bien en h^{-2} en accord avec l'estimation (10.29).
- **Laplacien en formulation mixte.** On considère le problème discret (6.94). On montre que $\gamma = 0$ et $\delta = 1$ si bien qu'il existe des constantes positives

c_1 et c_2, indépendantes de h, telles que

$$c_1 h^{-1} \leqslant \kappa(\mathcal{A}) \leqslant c_2 h^{-1}. \tag{10.31}$$

On obtient la même estimation pour les problèmes (6.68) et (6.78).

- **Advection–réaction et Galerkin/moindres carrés.** On considère le problème discret (7.43). On montre que $\gamma = 0$ et $\delta = 1$ si bien qu'il existe des constantes positives c_1 et c_2, indépendantes de h, telles que

$$c_1 h^{-1} \leqslant \kappa(\mathcal{A}) \leqslant c_2 h^{-1}. \tag{10.32}$$

10.1.3 Stabilité discrète

L'objet de cette section est d'étudier la sensibilité de la solution U du système linéaire (10.1) par rapport à des perturbations dans les coefficients de la matrice $\mathcal{A}$ et du membre de droite F. Pour $\delta F \in \mathbb{R}^N$ et $\delta\mathcal{A} \in \mathbb{R}^{N,N}$, on note $U + \delta U$ la solution du système perturbé

$$(\mathcal{A} + \delta\mathcal{A})(U + \delta U) = F + \delta F. \tag{10.33}$$

Proposition 10.8. *On suppose que $F \neq 0$ et que la perturbation $\delta\mathcal{A}$ est suffisamment petite pour que $\|\delta\mathcal{A}\|_{\mathbb{R}^N} \leqslant \frac{1}{2}\|\mathcal{A}^{-1}\|_{\mathbb{R}^N}^{-1}$. Alors, on a*

$$\frac{\|\delta U\|_{\mathbb{R}^N}}{\|U\|_{\mathbb{R}^N}} \leqslant 2\kappa(\mathcal{A}) \left(\frac{\|\delta F\|_{\mathbb{R}^N}}{\|F\|_{\mathbb{R}^N}} + \frac{\|\delta\mathcal{A}\|_{\mathbb{R}^N}}{\|\mathcal{A}\|_{\mathbb{R}^N}} \right). \tag{10.34}$$

La preuve de (10.34) est élémentaire. Un calcul direct montre que

$$\delta U = (\mathcal{I}_N + \mathcal{A}^{-1}\delta\mathcal{A})^{-1}\mathcal{A}^{-1}(\delta F - \delta\mathcal{A}\, U), \tag{10.35}$$

où $\mathcal{I}_N$ est la matrice identité d'ordre N. Puisque $\|\delta\mathcal{A}\|_{\mathbb{R}^N} \leqslant \frac{1}{2}\|\mathcal{A}^{-1}\|_{\mathbb{R}^N}^{-1}$, on a $\|(\mathcal{I}_N + \mathcal{A}^{-1}\delta\mathcal{A})^{-1}\|_{\mathbb{R}^N} \leqslant 2$. On en déduit

$$\|\delta U\|_{\mathbb{R}^N} \leqslant 2\kappa(\mathcal{A})\|\mathcal{A}\|_{\mathbb{R}^N}^{-1}(\|\delta F\|_{\mathbb{R}^N} + \|\delta\mathcal{A}\|_{\mathbb{R}^N}\|U\|_{\mathbb{R}^N}), \tag{10.36}$$

et on conclut en observant que $\|F\|_{\mathbb{R}^N} \leqslant \|\mathcal{A}\|_{\mathbb{R}^N}\|U\|_{\mathbb{R}^N}$. Par ailleurs, il existe des perturbations δF et $\delta\mathcal{A}$ pour lesquelles la borne supérieure dans l'estimation (10.34) est atteinte ; on dit que cette estimation est *optimale*.

Un autre point de vue intéressant sur la stabilité d'un système linéaire consiste à faire le lien entre résidu et erreur. On suppose qu'on a déterminé (par

exemple à l'aide d'une des méthodes itératives décrites dans le chapitre 11) une approximation U^δ de la solution U de (10.1). On définit l'erreur et le résidu associés à U^δ de la manière suivante :

$$E^\delta = U - U^\delta \qquad \text{et} \qquad R^\delta = F - \mathcal{A}U^\delta. \tag{10.37}$$

On observe que

$$\mathcal{A}E^\delta = R^\delta. \tag{10.38}$$

Si on connaît explicitement U^δ mais pas U, on peut évaluer le résidu mais pas l'erreur. La question naturelle qui se pose est de déterminer un critère sur la taille du résidu permettant d'affirmer que l'erreur est suffisamment petite.

Proposition 10.9. *On suppose que $F \neq 0$. Alors, on a*

$$\frac{\|E^\delta\|_{\mathbb{R}^N}}{\|U\|_{\mathbb{R}^N}} \leqslant \kappa(\mathcal{A}) \frac{\|R^\delta\|_{\mathbb{R}^N}}{\|F\|_{\mathbb{R}^N}}. \tag{10.39}$$

De plus, cette inégalité est optimale.

En combinant les estimations (10.34) et (10.39) avec le théorème 10.7, on aboutit à des résultats très pessimistes quant à la stabilité discrète de (10.1) puisque $\kappa(\mathcal{A})$ explose en $h^{-\gamma-\delta}$ (par exemple, en h^{-2} pour le Laplacien en formulation primale). En fait, les estimations ci-dessus peuvent être nettement améliorées lorsque le système linéaire résulte d'une approximation par éléments finis. On utilise pour cela les propriétés de stabilité du problème approché (10.7) et, notamment, la condition inf-sup discrète $\alpha_h > 0$ où α_h est défini en (10.8).

Soit U^δ une approximation de la solution U du système linéaire (10.1). On pose $u_h^\delta = C_{V_h}^{-1} U^\delta$. Par ailleurs, on introduit la norme suivante sur $\mathbb{R}^N$,

$$\|X\|_* = \sup_{Y \in \mathbb{R}^N} \frac{(X, Y)_{\mathbb{R}^N}}{\|C_{W_h}^{-1} Y\|_{W_h}}. \tag{10.40}$$

On observe que pour tout $v_h \in V_h$,

$$\alpha_h \|v_h\|_{V_h} \leqslant \sup_{w_h \in W_h} \frac{a_h(v_h, w_h)}{\|w_h\|_{W_h}} = \|\mathcal{A} C_{V_h} v_h\|_*. \tag{10.41}$$

De plus,

$$\|F\|_* = \sup_{W\in\mathbb{R}^N} \frac{(F,W)_{\mathbb{R}^N}}{\|C_{W_h}^{-1}W\|_{W_h}} = \sup_{w_h\in W_h} \frac{f_h(w_h)}{\|w_h\|_{W_h}} = \sup_{w_h\in W_h} \frac{a_h(u_h,w_h)}{\|w_h\|_{W_h}} \leqslant \omega_h \|u_h\|_{V_h}. \tag{10.42}$$

Il en résulte le résultat suivant.

Proposition 10.10. *On a*

$$\frac{\|u_h - u_h^\delta\|_{V_h}}{\|u_h\|_{V_h}} \leqslant \frac{\omega_h}{\alpha_h} \frac{\|R^\delta\|_*}{\|F\|_*}. \tag{10.43}$$

En général, l'approximation par éléments finis est construite de sorte qu'il existe des constantes positives c_1 et c_2, indépendantes de h, telles que

$$c_1 \leqslant \alpha_h \leqslant \omega_h \leqslant c_2. \tag{10.44}$$

Dans ces conditions, on déduit de l'estimation (10.43) que l'erreur relative $\frac{\|u_h-u_h^\delta\|_{V_h}}{\|u_h\|_{V_h}}$ est bien contrôlée par le résidu relatif $\frac{\|R^\delta\|_*}{\|F\|_*}$. En d'autres termes, le problème discret (10.7) étant bien posé, le système linéaire (10.1) jouit de propriétés de stabilité *même si la matrice de rigidité est mal conditionnée*. La norme $\|\cdot\|_*$ étant coûteuse à évaluer, on préfère en pratique évaluer le résidu relatif en norme euclidienne.

10.2 Factorisation LU et variantes

Cette section expose quelques notions relatives à la factorisation LU des matrices inversibles. La présentation est limitée aux résultats essentiels ; on renvoie, par exemple, à Golub et van Loan [**47**], Lascaux et Théodor [**56**], Ortega [**60**] ou Saad [**66**] pour plus de détails et des nombreux compléments. On oublie provisoirement que la matrice provient de l'approximation d'un problème modèle par éléments finis ; on y reviendra dans la section 10.3. Dans cette section, $\mathcal{A}$ désigne donc simplement une matrice *inversible* d'ordre N,

$$\mathcal{A} = \begin{pmatrix} \mathcal{A}_{11} & \dots & \mathcal{A}_{1N} \\ \vdots & & \vdots \\ \mathcal{A}_{N1} & \dots & \mathcal{A}_{NN} \end{pmatrix}. \tag{10.45}$$

10.2.1 Méthodes d'inversion directe et coût asymptotique

On considère le problème suivant : étant donné une matrice inversible $\mathcal{A} \in \mathbb{R}^{N,N}$ et un vecteur $F \in \mathbb{R}^N$, évaluer le vecteur $U \in \mathbb{R}^N$ tel que $\mathcal{A}U = F$. On s'intéresse à une classe particulière de méthodes pour résoudre ce problème, la classe dite des *méthodes d'inversion directe* au sens de la définition suivante.

Définition 10.11. *Une méthode d'inversion directe fournit, en l'absence d'erreurs d'arrondi, la solution du système linéaire $\mathcal{A}U = F$ en un nombre fini d'opérations.*

On souhaite estimer le coût numérique d'une méthode d'inversion directe. On s'intéresse à la dépendance de ce coût en N et, en particulier, à son comportement asymptotique lorsque le paramètre N est grand. Afin de quantifier ce coût, on effectue un décompte des opérations élémentaires qui sont réalisées par la méthode d'inversion directe.

Définition 10.12. *On convient de définir une* opération élémentaire *comme une multiplication et une addition entre deux réels. On dit qu'une méthode d'inversion directe a un* coût asymptotique *de $\varphi(N)$ pour une certaine fonction $\varphi : \mathbb{N} \to \mathbb{N}$, si le nombre d'opérations réalisées par cet algorithme, que l'on note $\omega(N)$, est tel que*

$$\lim_{N \to +\infty} \frac{\omega(N)}{\varphi(N)} = 1. \tag{10.46}$$

Dans la pratique, une méthode d'inversion directe n'est viable que si son coût asymptotique est polynômial en N, par exemple, $\varphi(N) = N^2$ ou N^3. Une méthode dont le coût asymptotique est exponentiel en N ne peut être utilisée que pour des valeurs de N petites, ce qui limite considérablement son champ d'application. Pour mémoire, une méthode d'inversion directe dont le coût asymptotique est exponentiel en N est celle basée sur le calcul des déterminants de Cramer : on évalue d'abord la matrice $\mathcal{A}^{-1}$ en calculant son déterminant ainsi que celui de tous ses cofacteurs puis on effectue le produit matrice–vecteur $U = \mathcal{A}^{-1}F$.

10.2.2 L'algorithme du pivot de Gauß

La méthode d'inversion directe la plus couramment utilisée pour résoudre le système linéaire $\mathcal{A}U = F$ est celle basée sur l'algorithme du pivot de Gauß. Cet algorithme construit une matrice triangulaire inférieure $\mathcal{T}^{\text{inf}}$ et une matrice triangulaire supérieure $\mathcal{T}^{\text{sup}}$ telles que[1]

$$\mathcal{A} = \mathcal{T}^{\text{inf}}\mathcal{T}^{\text{sup}}. \tag{10.47}$$

De plus, une des deux matrices, par exemple $\mathcal{T}^{\text{sup}}$, peut être construite de façon à ce que ses éléments diagonaux soient tous égaux à 1. Lorsque la matrice $\mathcal{A}$ est écrite sous la forme (10.47), on parle de *factorisation* LU.

Le principe de l'algorithme du pivot de Gauß consiste à construire une suite de matrices $(\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(N)})$ telles que $\mathcal{A}^{(1)} = \mathcal{A}$ et pour $k \in \{2, \ldots, N\}$, la matrice $\mathcal{A}^{(k)}$ possède la structure suivante :

$$\mathcal{A}^{(k)} = \left(\begin{array}{ccccccc} 1 & \mathcal{A}^{(k)}_{12} & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & & & & & \\ \vdots & & \ddots & & & & \\ \vdots & & & 1 & \mathcal{A}^{(k)}_{k-1,k} & \cdots & \cdots \\ \hline 0 & \cdots & \cdots & 0 & \mathcal{A}^{(k)}_{kk} & \cdots & \mathcal{A}^{(k)}_{kN} \\ \vdots & & & \vdots & \mathcal{A}^{(k)}_{k+1,k} & & \mathcal{A}^{(k)}_{k+1,N} \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \mathcal{A}^{(k)}_{Nk} & \cdots & \mathcal{A}^{(k)}_{NN} \end{array}\right). \tag{10.48}$$

En supposant que $\mathcal{A}^{(k)}_{kk} \neq 0$, la matrice $\mathcal{A}^{(k+1)}$ est construite de la manière suivante :

(i) on divise la k-ième ligne de $\mathcal{A}^{(k)}$ par $\mathcal{A}^{(k)}_{kk}$; on note L'_k la ligne ainsi obtenue ;

1. On rappelle qu'une matrice triangulaire inférieure $\mathcal{T}^{\text{inf}}$ est telle que $\mathcal{T}^{\text{inf}}_{ij} = 0$ pour $j > i$; de même, une matrice triangulaire supérieure $\mathcal{T}^{\text{sup}}$ est telle que $\mathcal{T}^{\text{sup}}_{ij} = 0$ pour $j < i$.

(ii) pour tout $i \in \{k+1, \ldots, N\}$, on retranche à la i-ième ligne de $\mathcal{A}^{(k)}$ la ligne L'_k multipliée par $\mathcal{A}^{(k)}_{ik}$.

Ces opérations permettent de remplacer la colonne

$$\begin{pmatrix} \mathcal{A}^{(k)}_{kk} \\ \mathcal{A}^{(k)}_{k+1,k} \\ \vdots \\ \mathcal{A}^{(k)}_{Nk} \end{pmatrix} \quad \text{par} \quad \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{10.49}$$

et de former ainsi la nouvelle matrice $\mathcal{A}^{(k+1)}$. On observera que l'on a $\mathcal{T}^{\inf(k)}\mathcal{A}^{(k)} = \mathcal{A}^{(k+1)}$ où $\mathcal{T}^{\inf(k)}$ est une matrice triangulaire inférieure telle que $\mathcal{T}^{\inf(k)}_{ij} = \delta_{ij}$ pour tout $i,j \in \{1, \ldots, N\}$ sauf pour $j = k$ et $i \in \{k, \ldots, N\}$ où on a

$$\mathcal{T}^{\inf(k)}_{kk} = \frac{1}{\mathcal{A}^{(k)}_{kk}} \qquad \text{et} \qquad \mathcal{T}^{\inf(k)}_{ik} = -\frac{\mathcal{A}^{(k)}_{ik}}{\mathcal{A}^{(k)}_{kk}},\ i \in \{k+1, \ldots, N\}. \tag{10.50}$$

Une fois engendrée la matrice $\mathcal{A}^{(N)}$, les matrices $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$ dans (10.47) sont obtenues en posant $\mathcal{T}^{\sup} = \mathcal{A}^{(N)}$ et $\mathcal{T}^{\inf} = \prod_{l=1}^{N} \mathcal{T}^{\inf(l)}$. On vérifie à l'aide d'un décompte d'opérations que le coût asymptotique d'évaluation des matrices $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$ est de $\frac{N^3}{3}$ opérations. Dans la pratique, il est souvent commode de stocker les matrices $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$ en une seule matrice $\mathcal{T} \in \mathbb{R}^{N,N}$ telle que

$$\mathcal{T}_{ij} = \begin{cases} \mathcal{T}^{\inf}_{ij} & \text{si } j \leqslant i, \\ \mathcal{T}^{\sup}_{ij} & \text{si } j > i. \end{cases} \tag{10.51}$$

La matrice $\mathcal{T}$ recueille toute l'information contenue dans les matrices $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$ puisque les coefficients diagonaux de $\mathcal{T}^{\sup}$ sont égaux à 1.

L'algorithme décrit ci-dessus peut se terminer prématurément lorsque $\mathcal{A}^{(k)}_{kk} = 0$; or, cette éventualité peut se produire même si la matrice $\mathcal{A}$ est inversible. De plus, les erreurs d'arrondi dans les calculs ont des effets d'autant plus prononcés que l'on est amené à effectuer des divisions par des nombres petits en valeur absolue. Il est donc souhaitable que dans

l'algorithme ci-dessus, le coefficient $\mathcal{A}_{kk}^{(k)}$ soit, en valeur absolue, le plus grand possible. On peut envisager deux stratégies.

(i) Permutation de lignes : on recherche l'indice $i_0 \in \{k, \dots, N\}$ tel que

$$|\mathcal{A}_{i_0 k}^{(k)}| = \max_{i \in \{k,\dots,N\}} |\mathcal{A}_{ik}^{(k)}|, \tag{10.52}$$

et on permute les lignes i_0 et k dans $\mathcal{A}^{(k)}$ avant de former la matrice $\mathcal{A}^{(k+1)}$. La matrice $\mathcal{A}$ étant inversible, on peut montrer que $\mathcal{A}_{i_0 k}^{(k)} \neq 0$. Une variante de cette stratégie consiste à adimensionaliser la colonne avant d'effectuer la recherche du coefficient maximal : on évalue d'abord $s_i = \max_{j \in \{k,\dots,N\}} |\mathcal{A}_{ij}^{(k)}|$ pour tout $i \in \{k, \dots, N\}$, puis on recherche l'indice $i_0 \in \{k, \dots, N\}$ tel que

$$\frac{|\mathcal{A}_{i_0 k}^{(k)}|}{s_{i_0}} = \max_{i \in \{k,\dots,N\}} \frac{|\mathcal{A}_{ik}^{(k)}|}{s_i}. \tag{10.53}$$

(ii) Permutation de lignes et de colonnes : on recherche les indices $i_0, j_0 \in \{k, \dots, N\}$ tels que

$$|\mathcal{A}_{i_0 j_0}^{(k)}| = \max_{i,j \in \{k,\dots,N\}} |\mathcal{A}_{ij}^{(k)}|, \tag{10.54}$$

et on permute les lignes i_0 et k ainsi que les colonnes j_0 et k dans $\mathcal{A}^{(k)}$ avant d'effectuer les opérations ci-dessus. On peut également considérer une variante adimensionnée de cette stratégie.

Le coût de la stratégie par permutation de lignes est de $\mathcal{O}(n^2)$ comparaisons[1] alors que celui de la stratégie par permutation de lignes et de colonnes est de $\mathcal{O}(n^3)$ comparaisons[2]. Dans la pratique, on utilise de préférence la stratégie par permutation de lignes car elle permet d'éviter à moindre coût la plupart

1. Pour deux fonctions $\psi_1 : \mathbb{N} \to \mathbb{N}$ et $\psi_2 : \mathbb{N} \to \mathbb{N}$, la notation $\psi_1 = \mathcal{O}(\psi_2)$ signifie qu'il existe des constantes c_1 et c_2 telles que pour tout n suffisamment grand, $c_1\psi_1(n) \leqslant \psi_2(n) \leqslant c_2\psi_1(n)$.

2. Une comparaison entre deux réels est, en général, plus coûteuse que les opérations élémentaires considérées dans la définition 10.12 ; ce surcoût est négligé en première approximation.

des problèmes liés aux erreurs d'arrondi et aux singularités des coefficients diagonaux.

L'algorithme du pivot de Gauß avec permutation de lignes conduit à la décomposition suivante :

$$\Pi^{\sigma}\mathcal{A} = \mathcal{T}^{\inf}\mathcal{T}^{\sup}, \tag{10.55}$$

où Π^{σ} est une matrice de permutation. Étant donné une permutation $\sigma : \{1, \ldots, N\} \to \{1, \ldots, N\}$, les coefficients de la matrice Π^{σ} s'expriment à l'aide du symbole de Kronecker sous la forme

$$\Pi^{\sigma}_{ij} = \delta_{\sigma(i),j}, \qquad i,j \in \{1, \ldots, N\}. \tag{10.56}$$

On vérifie facilement que la matrice Π^{σ} est inversible et que

$$(\Pi^{\sigma})^{-1} = \Pi^{(\sigma^{-1})} = (\Pi^{\sigma})^{T}. \tag{10.57}$$

Une fois obtenue la décomposition (10.55), la solution du système linéaire $\mathcal{A}U = F$ s'obtient de la manière suivante :

(i) résoudre le système triangulaire inférieur $\mathcal{T}^{\inf}V = \Pi^{\sigma}F$;

(ii) résoudre le système triangulaire supérieur $\mathcal{T}^{\sup}U = V$.

On vérifie que

$$\mathcal{A}U = (\Pi^{\sigma})^{-1}\mathcal{T}^{\inf}\mathcal{T}^{\sup}U = (\Pi^{\sigma})^{-1}\mathcal{T}^{\inf}V = (\Pi^{\sigma})^{-1}\Pi^{\sigma}F = F. \tag{10.58}$$

Par ailleurs, le coût asymptotique de la résolution des deux systèmes linéaires ci-dessus est de $\frac{N^2}{2}$ opérations. En conclusion, on retiendra que lorsque N est grand, le coût total de la résolution du système linéaire $\mathcal{A}U = F$ par l'algorithme du pivot de Gauß (avec permutation de lignes ou non) est dominé par le coût de la décomposition (10.55) ; il est donc de l'ordre de $\frac{N^3}{3}$ opérations.

Remarque 10.13

L'algorithme du pivot de Gauß avec permutation de lignes et de colonnes conduit à la décomposition suivante :

$$\Pi^{\sigma}\mathcal{A}\Pi^{\tau} = \mathcal{T}^{\inf}\mathcal{T}^{\sup}, \tag{10.59}$$

où Π^{σ} et Π^{τ} sont deux matrices de permutation.

10.2.3 Variantes dans le cas symétrique

Lorsque la matrice $\mathcal{A}$ est symétrique, on considère généralement des algorithmes de factorisation qui exploitent cette symétrie. En l'absence de permutations de lignes ou de colonnes, on peut décomposer une matrice symétrique $\mathcal{A}$ de la manière suivante :

$$\mathcal{A} = \mathcal{T}^{\text{inf}}\mathcal{D}(\mathcal{T}^{\text{inf}})^T, \tag{10.60}$$

où $\mathcal{D}$ est une matrice diagonale et où $\mathcal{T}^{\text{inf}}$ est une matrice triangulaire inférieure dont les coefficients diagonaux sont tous égaux à 1. Dans la littérature, la décomposition (10.60) est connue sous le nom de *factorisation* LDL^T. L'algorithme de factorisation est décrit dans Golub et van Loan [**47**, p. 137] ou dans Burden et Faires [**25**, p. 376]. L'avantage de la factorisation LDL^T par rapport à l'algorithme du pivot de Gauß est qu'en exploitant la symétrie de la matrice $\mathcal{A}$, elle permet de diviser par deux le coût asymptotique de la factorisation ; celui-ci est donc de l'ordre de $\frac{N^3}{6}$ opérations.

Une variante de la factorisation LDL^T, connue sous le nom *d'algorithme de Choleski*, conduit à la décomposition $\mathcal{A} = \mathcal{T}^{\text{inf}}(\mathcal{T}^{\text{inf}})^T$ où $\mathcal{T}^{\text{inf}}$ est toujours une matrice triangulaire inférieure mais dont les coefficients diagonaux ne sont pas nécessairement égaux à 1. L'algorithme de Choleski nécessite l'évaluation de N racines carrées, ce qui peut être pénalisant en termes de coût de calcul. On pourra consulter Golub et van Loan [**47**, p. 141] pour plus de détails sur l'algorithme de Choleski.

10.2.4 Factorisation QR

On considère une matrice $\mathcal{A} \in \mathbb{R}^{M,N}$ (à M lignes et N colonnes). On suppose que $M \geqslant N$ et que la matrice $\mathcal{A}$ est de rang N. Le principe de l'algorithme de factorisation QR consiste à construire une matrice $\mathcal{Q} \in \mathbb{R}^{M,N}$ et une matrice $\mathcal{R} \in \mathbb{R}^{N,N}$ telles que :

(i) $\mathcal{Q}^T\mathcal{Q}$ est une matrice diagonale que l'on note $\mathcal{D}$ et dont les coefficients sont strictement positifs ;

(ii) $\mathcal{R}$ est une matrice triangulaire supérieure dont les coefficients diagonaux sont tous égaux à 1 ;

(iii) la matrice $\mathcal{A}$ s'écrit sous la forme

$$\mathcal{A} = \mathcal{Q}\mathcal{R}. \tag{10.61}$$

Les matrices $\mathcal{Q}$ et $\mathcal{R}$ peuvent être construites grâce à divers algorithmes, par exemple le procédé d'orthogonalisation de Gram–Schmidt, la méthode de Householder (basée sur l'utilisation de matrices de symétrie) ou encore la méthode de Givens (basée sur l'utilisation de matrices de rotation). Pour plus de détails, on pourra consulter Golub et van Loan [**47**, p. 211]. La méthode de Givens est brièvement décrite dans la section 11.2.3 ci-après dans le cadre de la méthode GMRes. Les méthodes de factorisation QR sont, en général, plus coûteuses que l'algorithme du pivot de Gauß mais elles présentent l'avantage d'être parfois plus robustes vis-à-vis des erreurs d'arrondi. La factorisation QR est également utile dans la résolution de problèmes de minimisation au sens des moindres carrés. Par exemple, pour $F \in \mathbb{R}^M$, la solution $X \in \mathbb{R}^N$ du problème

$$\inf_{Y \in \mathbb{R}^N} \|F - \mathcal{A}Y\|_{\mathbb{R}^N}, \tag{10.62}$$

satisfait les équations, dites *normales*, $\mathcal{A}^T\mathcal{A}X = \mathcal{A}^T F$. La solution X s'obtient en posant $X' = \mathcal{D}^{-1}\mathcal{Q}^T F$ puis en résolvant le système triangulaire supérieur $\mathcal{R}X = X'$.

10.3 Matrices creuses et renumérotation

Dans cette section, on introduit les notions de matrice creuse et de graphe associé à de telles matrices. Puis, on présente une technique de renumérotation des matrices creuses qui est souvent employée pour les matrices d'éléments finis. On pourra consulter Saad [**66**] pour une présentation détaillée des techniques de renumérotation dans le cadre de la résolution des systèmes linéaires par méthodes itératives et George et Liu [**43**] lorsque cette résolution se fait par une méthode d'inversion directe.

10.3.1 Notion de matrice creuse

Les matrices provenant de l'approximation par éléments finis d'équations aux dérivées partielles sont *creuses* au sens de la définition suivante.

Définition 10.14. *On dit que la matrice $\mathcal{A} \in \mathbb{R}^{N,N}$ est* creuse *si dans chaque ligne (et chaque colonne) de $\mathcal{A}$, le nombre d'éléments non-nuls est majoré par une quantité indépendante de N. Par la suite, on désigne par χ_{lin} et χ_{col}, respectivement, le nombre maximum d'éléments non-nuls dans une ligne et une colonne de $\mathcal{A}$.*

Un exemple simple de matrice creuse est la matrice bloc-tridiagonale

$$\mathcal{A} = \begin{pmatrix} A_1 & B_1 & 0 & \dots & \dots & 0 \\ C_2 & A_2 & B_2 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & C_{n-1} & A_{n-1} & B_{n-1} \\ 0 & \dots & \dots & 0 & C_n & A_n \end{pmatrix}, \tag{10.63}$$

où les matrices $\{A_i\}_{1 \leqslant i \leqslant n}$, $\{B_i\}_{1 \leqslant i \leqslant n-1}$ et $\{C_i\}_{2 \leqslant i \leqslant n}$ sont carrées d'ordre n et tridiagonales. La matrice $\mathcal{A}$ est carrée d'ordre $N = n^2$. Cette matrice provient, par exemple, de l'approximation par éléments finis de Lagrange $\mathbb{P}_1$ du problème de Poisson avec conditions aux limites de Dirichlet homogènes sur le carré unité avec un maillage structuré uniforme où chaque cellule carrée est subdivisée en deux mailles triangulaires selon la première diagonale ; voir la figure 10.1 à gauche. Dans cet exemple, on a

$$\chi_{\text{lin}} = \chi_{\text{col}} = 9. \tag{10.64}$$

Une observation importante est que la *structure creuse* de la matrice $\mathcal{A}$, c'est-à-dire la disposition des éléments non-nuls de $\mathcal{A}$, dépend de la numérotation des degrés de liberté dans l'espace d'approximation. Dans l'exemple ci-dessus, les degrés de liberté sont associés aux nœuds du maillage et ceux-ci ont été numérotés en balayant les lignes (par exemple, de la gauche vers la droite et du bas vers le haut).

Lorsqu'on utilise des maillages non-structurés (voir la figure 10.1 à droite), la disposition des éléments non-nuls dans $\mathcal{A}$ ne s'organise pas de manière simple. La structure creuse de $\mathcal{A}$ se déduit de la numérotation des fonctions de forme dans l'espace d'approximation à partir de l'observation suivante : pour deux

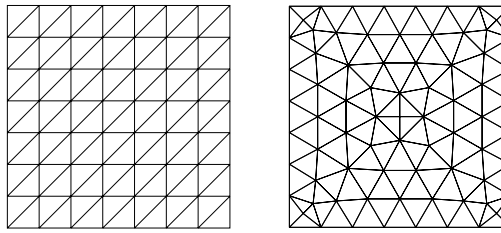


Figure 10.1 – À gauche : maillage structuré uniforme du carré unité, où chaque cellule carrée est subdivisée en deux mailles triangulaires selon la première diagonale ; à droite : maillage non-structuré du carré unité par des triangles.

fonctions de forme φ_i et φ_j avec $i, j \in \{1, \ldots, N\}$ et $i \neq j$, on a

$$(\text{support}(\varphi_i) \cap \text{support}(\varphi_j) = \varnothing) \Longrightarrow (\mathcal{A}_{ij} = 0). \tag{10.65}$$

Dans la pratique, on souhaite numéroter les degrés de liberté de manière à regrouper autant que possible les éléments non-nuls de $\mathcal{A}$ autour de la diagonale. Une des motivations est que cela permet de réduire le *niveau de remplissage* dans une factorisation LU. Pour préciser cette notion, on pose pour $i \in \{1, \ldots, N\}$,

$$l_i = \min\{j \leqslant i \,;\, \mathcal{A}_{ij} \neq 0\}, \tag{10.66}$$

$$u_i = \max\{j \geqslant i \,;\, \mathcal{A}_{ij} \neq 0\}. \tag{10.67}$$

La quantité

$$\delta = \max_{1 \leqslant i \leqslant N} (u_i - l_i), \tag{10.68}$$

s'appelle la *largeur de bande* de $\mathcal{A}$. Pour la matrice $\mathcal{A}$ définie en (10.63), on a $\delta = \mathcal{O}(n) = \mathcal{O}(N^{\frac{1}{2}})$.

On considère maintenant la décomposition LU de la matrice $\mathcal{A}$ sous la forme (10.47). On peut montrer que la matrice triangulaire inférieure $\mathcal{T}^{\text{inf}}$ est telle que $\mathcal{T}^{\text{inf}}_{ij} = 0$ pour $j < l_i$, mais $\mathcal{T}^{\text{inf}}_{ij} \neq 0$ pour la plupart des $j \in \{l_i, \ldots, i\}$. De même, la matrice triangulaire supérieure $\mathcal{T}^{\text{sup}}$ est telle que $\mathcal{T}^{\text{sup}}_{ij} = 0$ pour $j > u_i$, mais $\mathcal{T}^{\text{sup}}_{ij} \neq 0$ pour la plupart des $j \in \{i, \ldots, u_i\}$.

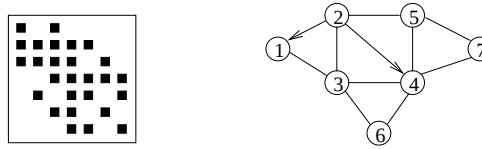


Figure 10.2 – Matrice creuse d'ordre 7 et graphe associé. Un carré noir dans la matrice indique que l'élément correspondant est non-nul.

En d'autres termes, la matrice $\mathcal{T}$ définie dans (10.51) a une largeur de bande δ (qui est donc la même que celle de la matrice $\mathcal{A}$), mais le nombre d'éléments non-nuls par ligne dans $\mathcal{T}$ est de l'ordre de δ et non plus de χ_{lin}. On observera que le rapport entre le nombre d'éléments non-nuls de $\mathcal{T}$ et ceux de $\mathcal{A}$ explose en n.

10.3.2 Graphes et matrices creuses

Les algorithmes de renumérotation des matrices creuses sont basés sur la notion de graphe. Soit V un ensemble (fini) et soit $\mathcal{R}$ une relation binaire sur V. On pose $E = \{(x, y) \in V \times V \,;\, x\mathcal{R}y\}$. Le couple $G = (V, E)$ est appelé un *graphe*. Les éléments de V s'appellent les *nœuds* du graphe et les éléments de E s'appellent les *arêtes* du graphe. Lorsque la relation binaire $\mathcal{R}$ est *symétrique*, on dit que le graphe est *non-orienté*. Pour un sous-ensemble $X \subset V$, on pose

$$\text{Adj}(X) = \{y \in V \backslash X \,;\, \exists x \in X,\, (x, y) \in E\}. \tag{10.69}$$

Pour $x \in V$, l'ensemble $\text{Adj}(\{x\})$, qui est simplement noté $\text{Adj}(x)$, s'appelle le *voisinage* de x et le cardinal de $\text{Adj}(x)$ s'appelle le *degré* de x.

On peut associer à une *matrice creuse* $\mathcal{A} \in \mathbb{R}^{N,N}$ un graphe de la manière suivante : on pose $V = \{1, \ldots, N\}$ et $E = \{(x, y) \in V \times V \,;\, \mathcal{A}_{xy} \neq 0\}$. Ce graphe est non-orienté si $(\mathcal{A}_{xy} \neq 0) \iff (\mathcal{A}_{yx} \neq 0)$. Une façon simple de représenter un graphe consiste à associer à chaque nœud un point du plan et à tracer une flèche de x vers y lorsque $x\mathcal{R}y$. Lorsque $x\mathcal{R}y$ et $y\mathcal{R}x$, on trace une ligne plutôt que deux flèches de sens contraire. Enfin, lorsque $x\mathcal{R}x$, on trace un cercle autour de x. La figure 10.2 présente un exemple de matrice creuse d'ordre 7 et son graphe associé.

10.3.3 Renumérotation par ensembles de niveau

Soit $G = (V, E)$ un graphe et soit $x \in V$. On associe au nœud x une suite (finie) d'ensembles de niveau $(N_1, N_2, \ldots)$ de la manière suivante :

(i) on pose $N_1 = \{x\}$;

(ii) pour $k \geqslant 1$, on pose $N_{k+1} = \mathrm{Adj}(N_k) \backslash (\bigcup_{l=1}^{k} N_l)$.

Si le graphe G est tel qu'il existe un chemin reliant deux nœuds quelconques, la suite finie $(N_1, N_2, \ldots)$ forme une partition de V.

La figure 10.3 à gauche présente un exemple de matrice creuse d'ordre 15. La figure 10.4 illustre le graphe associé à cette matrice. Les ensembles de niveau associés au nœud 2 sont les suivants :

$$N_1 = \{2\}, \quad N_2 = \{5, 7\}, \quad N_3 = \{9, 11, 14\}, \quad N_4 = \{1, 3, 12, 15\},$$
$$N_5 = \{8, 10, 13\}, \quad N_6 = \{4, 6\}. \tag{10.70}$$

En concaténant ces ensembles, on forme un tableau `perm` qui induit une permutation de l'ensemble $V = \{1, \ldots, 15\}$. On obtient

$$\texttt{perm} = (2, 5, 7, 9, 11, 14, 1, 3, 12, 15, 8, 10, 13, 4, 6). \tag{10.71}$$

Le graphe de la matrice considérée dans cet exemple est tel que les ensembles de niveau forment une partition de V. Si tel n'est pas le cas, on considère un nouveau nœud y qui n'est pas relié à x et on construit les ensembles de niveau associés à y. Un nouveau tableau `perm` est ensuite formé en concaténant les tableaux `perm` associés à x et à y. Ce processus est répété jusqu'à ce que tous les nœuds de V figurent dans le tableau `perm`.

Une fois construit le tableau `perm`, on pose

$$\mathcal{B}_{ij} = \mathcal{A}_{\texttt{perm}[i],\texttt{perm}[j]}, \tag{10.72}$$

où `perm`$[i]$ désigne le i-ième élément du tableau `perm`. La technique de renumérotation décrite ci-dessus porte le nom de *renumérotation BFS* (de l'anglais Breadth-First-Search). La matrice $\mathcal{B}$ jouit de la propriété remarquable suivante.

Proposition 10.15. *On suppose que le graphe G associé à la matrice $\mathcal{A}$ est non-orienté. Alors, la matrice $\mathcal{B}$ est bloc-tridiagonale ; plus précisément, pour*

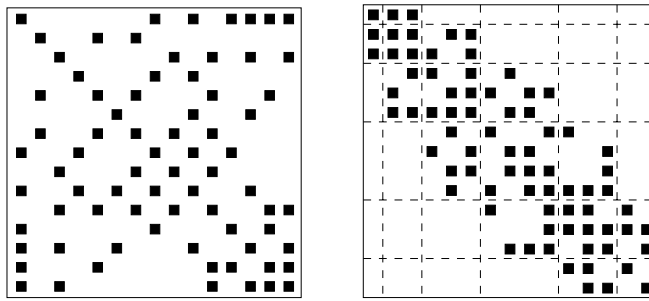


Figure 10.3 – Matrice creuse avant renumérotation (à gauche) et après renumérotation BFS (à droite).

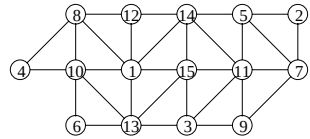


Figure 10.4 – Graphe de la matrice creuse illustrée sur la figure 10.3 à gauche.

$i, j \in \{1, \ldots, N\}$ *tels qu'il existe un chemin reliant* $\texttt{perm}(i)$ *à* $\texttt{perm}(j)$*, on a* $\mathcal{B}_{ij} = 0$ *si* $\texttt{perm}(i) \in N_k$ *et* $\texttt{perm}(j) \in N_{k'}$ *avec* $|k - k'| \geqslant 2$.

La figure 10.3 à droite présente la matrice $\mathcal{B}$ obtenue par renumérotation BFS. La structure bloc-tridiagonale est clairement visible.

Remarque 10.16

Dans l'exemple traité ci-dessus, les nœuds sont classés au sein de chaque ensemble de niveau selon l'ordre naturel (croissant). Cette stratégie peut être améliorée en choisissant d'ordonner les nœuds d'une autre façon, par exemple par degré croissant ou décroissant. On obtient ainsi l'algorithme de renumérotation de Cuthill–McKee (CMK).

11 • SOLVEURS ITÉRATIFS

Le chapitre précédent a montré que l'approximation par éléments finis d'un problème modèle fait intervenir un système linéaire, que l'on note

$$\mathcal{A}U = F, \tag{11.1}$$

dont la matrice associée est, en général, de très grande taille et creuse. Pour de tels systèmes, les méthodes de résolution directe (basées sur une factorisation LU ou des variantes de celle-ci) ne sont pas bien adaptées car, même en employant une technique de renumérotation performante, on aboutit souvent à des matrices à structure bande pour lesquelles le taux de remplissage lors d'une factorisation LU reste élevé. On préfère alors utiliser une méthode itérative. Le principe consiste à approcher (plutôt qu'à calculer exactement) la solution U de (11.1) à l'aide d'une suite de vecteurs de $\mathbb{R}^N$. Étant donné $U^0 \in \mathbb{R}^N$, on génère une suite $(U^k)_{k\geqslant 1}$ de vecteurs de $\mathbb{R}^N$ jusqu'à satisfaire un certain critère de convergence. L'erreur $U - U^k$ ne pouvant être évaluée explicitement, ce critère fait, en général, intervenir le résidu $R^k = F - \mathcal{A}U^k$.

Deux considérations interviennent dans l'évaluation d'une méthode itérative. Le premier est la vitesse de convergence ; le deuxième est le coût par itération. En général, l'amélioration de la vitesse de convergence induit une augmentation du coût par itération. Le coût total étant égal au produit du nombre d'itérations effectuées par le coût d'une itération, la sélection d'une méthode de résolution itérative relève souvent d'un compromis entre ces deux considérations.

Ce chapitre présente quelques solveurs itératifs employés dans le cadre de la méthode des éléments finis. On considère les méthodes dites de relaxation,

puis la méthode du gradient conjugué et ses variantes, et enfin, les méthodes multi-échelles, dont le prototype est la méthode multi-grille. Afin d'alléger les notations dans ce chapitre, on désigne par $(\cdot,\cdot)_N$ le produit scalaire euclidien sur $\mathbb{R}^N$ et par $\|\cdot\|_N$ la norme induite, où N est la taille du système linéaire (11.1).

11.1 Méthodes de relaxation

Cette section décrit le principe général des méthodes de relaxation et en présente quelques exemples classiques : la méthode de Jacobi, la méthode de Gauß–Seidel et la méthode SOR.

11.1.1 Principe général

Le principe d'une méthode de relaxation consiste à décomposer la matrice $\mathcal{A}$ sous la forme

$$\mathcal{A} = \mathcal{P} - \mathcal{Z}, \tag{11.2}$$

où la matrice $\mathcal{P} \in \mathbb{R}^{N,N}$ est inversible (la matrice $\mathcal{Z}$ n'est pas nécessairement inversible). Étant donné $U^0 \in \mathbb{R}^N$, on génère la suite $(U^k)_{k\geqslant 1}$ en résolvant, pour tout $k \geqslant 0$, le système linéaire

$$\mathcal{P}U^{k+1} = \mathcal{Z}U^k + F. \tag{11.3}$$

Il est clair que si la suite $(U^k)_{k\geqslant 1}$ converge, sa limite est la solution du système linéaire $\mathcal{A}U = F$.

Afin de mettre en œuvre de manière efficace la méthode de relaxation basée sur la décomposition (11.2), il convient d'effectuer une sélection judicieuse de la matrice $\mathcal{P}$. Celle-ci doit être telle que :

(i) la suite $(U^k)_{k\geqslant 1}$ définie en (11.3) converge (de préférence quel que soit le vecteur initial $U^0 \in \mathbb{R}^N$) ;

(ii) la matrice $\mathcal{P}$ est relativement simple à inverser si bien que le coût de calcul de U^{k+1} à partir de U^k reste modéré ;

(iii) la suite $(U^k)_{k\geqslant 1}$ converge « rapidement » vers sa limite.

Afin d'analyser la convergence de la suite $(U^k)_{k\geqslant 1}$, on introduit la *matrice d'itération* $\mathcal{T}$ telle que

$$\mathcal{T} = \mathcal{P}^{-1}\mathcal{Z}. \tag{11.4}$$

On note $\mathcal{T}^k$ la k-ième puissance de la matrice $\mathcal{T}$.[1]

Définition 11.1. *On dit qu'une matrice $\mathcal{T} \in \mathbb{R}^{N,N}$ est* convergente *si*

$$\lim_{k\to+\infty} \mathcal{T}^k = 0. \tag{11.5}$$

En notant $E^k = U - U^k$ l'erreur à la k-ième itération, on observe que

$$E^{k+1} = \mathcal{T}E^k, \tag{11.6}$$

si bien que, par récurrence, on déduit immédiatement de (11.3) que

$$E^k = \mathcal{T}^k E^0, \tag{11.7}$$

où $E^0 = U - U^0$ est l'erreur initiale.

Proposition 11.2. *On suppose que la matrice $\mathcal{T}$ est convergente. Alors, quel que soit $U^0 \in \mathbb{R}^N$, la suite $(U^k)_{k\geqslant 1}$ définie par* (11.3) *converge.*

Un critère important dans la sélection de la matrice $\mathcal{P}$ est donc de garantir que la matrice $\mathcal{T}$ est convergente. On désigne par $\sigma(\mathcal{T})$ le *spectre* de $\mathcal{T}$, c'est-à-dire l'ensemble des valeurs propres de $\mathcal{T}$, et par

$$\rho(\mathcal{T}) = \max_{\lambda\in\sigma(\mathcal{T})} |\lambda|, \tag{11.8}$$

le *rayon spectral* de $\mathcal{T}$. On a les résultats suivants.

Proposition 11.3. *Une matrice $\mathcal{T} \in \mathbb{R}^{N,N}$ est convergente si et seulement si $\rho(\mathcal{T}) < 1$.*

Proposition 11.4. *On suppose que la matrice $\mathcal{A}$ est symétrique définie positive. Alors, si la matrice $\mathcal{P}$ est telle que $\mathcal{P} + \mathcal{P}^T - \mathcal{A}$ est définie positive, la matrice $\mathcal{T}$ est convergente.*

1. Il y a une collision de notations avec l'exposant utilisé pour la suite U^k. L'exposant k indique donc une puissance pour les matrices (en lettres calligraphiées) et l'indice d'une suite pour les vecteurs (en lettres majuscules).

Afin de quantifier la vitesse de convergence de la suite $(U^k)_{k\geqslant 1}$, on introduit la notion de *taux de convergence*.

Définition 11.5. *Soit $\|\cdot\|$ une norme sur $\mathbb{R}^N$ (en général, on choisit la norme euclidienne). On suppose qu'il existe un réel $\sigma \in]0,1[$ tel que pour tout $k \geqslant 1$,*

$$\|U - U^k\| \leqslant \sigma^k \|U - U^0\|. \tag{11.9}$$

Le plus petit réel $\sigma \in]0,1[$ satisfaisant la condition ci-dessus est appelé le taux de convergence *de la méthode itérative en norme $\|\cdot\|$.*

L'intérêt pratique de la notion de taux de convergence est le suivant. On se fixe une tolérance, par exemple de la forme 10^{-t} où t est un entier positif, et on souhaite arrêter les itérations lorsque la norme de l'erreur initiale aura été réduite du facteur 10^t. La relation (11.9) montre que ce critère de convergence est satisfait lorsque

$$k \geqslant \frac{t}{-\log_{10}(\sigma)}. \tag{11.10}$$

Plus la quantité σ est proche de 1, plus la convergence de la suite $(U^k)_{k\geqslant 1}$ est lente ; plus la quantité σ est proche de 0, plus la convergence de la suite $(U^k)_{k\geqslant 1}$ est rapide. La relation (11.7) permet d'estimer le taux de convergence d'une méthode de relaxation. On obtient en première approximation

$$\sigma = \rho(\mathcal{T}). \tag{11.11}$$

En conclusion, une méthode de relaxation est utilisable si le rayon spectral de $\mathcal{T}$ n'est pas « trop proche » de 1.

11.1.2 Exemples classiques

Les méthodes de relaxation présentées dans cette section sont basées sur la décomposition

$$\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{U}, \tag{11.12}$$

où la matrice $\mathcal{D} \in \mathbb{R}^{N,N}$ est diagonale, la matrice $\mathcal{L} \in \mathbb{R}^{N,N}$ est triangulaire inférieure stricte ($\mathcal{L}_{ij} = 0$ si $i \leqslant j$) et la matrice $\mathcal{U} \in \mathbb{R}^{N,N}$ est triangulaire supérieure stricte ($\mathcal{U}_{ij} = 0$ si $i \geqslant j$). Clairement, la relation (11.12) détermine de manière univoque les matrices $\mathcal{D}$, $\mathcal{L}$ et $\mathcal{U}$ à partir de la matrice $\mathcal{A}$.

La *méthode de Jacobi* consiste à prendre

$$\mathcal{P}_{\mathrm{J}} = \mathcal{D}, \tag{11.13}$$

$$\mathcal{T}_{\mathrm{J}} = \mathcal{D}^{-1}(\mathcal{L} + \mathcal{U}). \tag{11.14}$$

La méthode de Jacobi est particulièrement simple à mettre en œuvre puisque la construction de la suite $(U^k)_{k\geqslant 1}$ ne demande que l'inversion d'une matrice diagonale à chaque itération. En notant $(U_i^k)_{1\leqslant i\leqslant N}$ les composantes de U^k, la relation (11.3) s'écrit sous la forme

$$U_i^{k+1} = \frac{1}{\mathcal{A}_{ii}}\left(F_i - \sum_{j=1}^{i-1}\mathcal{A}_{ij}U_j^k - \sum_{j=i+1}^{N}\mathcal{A}_{ij}U_j^k\right). \tag{11.15}$$

La *méthode de Gauß–Seidel* consiste à prendre

$$\mathcal{P}_{\mathrm{GS}} = \mathcal{D} - \mathcal{L}, \tag{11.16}$$

$$\mathcal{T}_{\mathrm{GS}} = (\mathcal{D} - \mathcal{L})^{-1}\mathcal{U}. \tag{11.17}$$

Le coût par itération dans la méthode de Gauß–Seidel reste relativement modéré puisque la construction de la suite $(U^k)_{k\geqslant 1}$ ne demande que l'inversion d'une matrice triangulaire inférieure à chaque itération. La relation (11.3) s'écrit sous la forme

$$U_i^{k+1} = \frac{1}{\mathcal{A}_{ii}}\left(F_i - \sum_{j=1}^{i-1}\mathcal{A}_{ij}U_j^{k+1} - \sum_{j=i+1}^{N}\mathcal{A}_{ij}U_j^k\right). \tag{11.18}$$

On constate que, contrairement à la méthode de Jacobi, la méthode de Gauß–Seidel utilise les composantes de U^{k+1} qui ont déjà été évaluées. Par ailleurs, les expressions (11.15) et (11.18) montrent que le coût asymptotique par itération dans les méthodes de Jacobi et de Gauß–Seidel est le même. Il reste à étudier la convergence de ces deux méthodes.

Proposition 11.6. *On suppose que la matrice $\mathcal{A}$ est telle que $\mathcal{A}_{ii} > 0$ pour tout $i \in \{1,\dots,N\}$ et $\mathcal{A}_{ij} \leqslant 0$ pour tout $i,j \in \{1,\dots,N\}$ et $i \neq j$. Alors, l'un des quatre points suivants (mutuellement exclusifs) est vrai :*

(i) $0 \leqslant \rho(\mathcal{T}_{\mathrm{GS}}) < \rho(\mathcal{T}_{\mathrm{J}}) < 1$;

(ii) $1 < \rho(\mathcal{T}_J) < \rho(\mathcal{T}_{GS})$;

(iii) $\rho(\mathcal{T}_{GS}) = \rho(\mathcal{T}_J) = 1$;

(iv) $\rho(\mathcal{T}_{GS}) = \rho(\mathcal{T}_J) = 0$.

La proposition 11.6 montre que, pour la classe de matrices considérées dans l'énoncé, les méthodes de Jacobi et de Gauß–Seidel convergent ou divergent ensemble et que si elles convergent, le taux de convergence de la méthode de Gauß–Seidel est meilleur. Par ailleurs, les points (iii) et (iv) ne se produisent que dans des cas très particuliers. Par exemple, le point (iii) ne peut se réaliser que si la matrice $\mathcal{A}$ est singulière ; en effet, un vecteur $X \in \mathbb{R}^N$ est dans le noyau de $\mathcal{A}$ si et seulement si X est vecteur propre de la matrice d'itération avec la valeur propre 1.

Un deuxième résultat de convergence, spécifique à la méthode de Gauß–Seidel, est le suivant.

Proposition 11.7. *On suppose que la matrice $\mathcal{A}$ est symétrique définie positive. Alors, la méthode de Gauß–Seidel converge.*

On observera que le caractère symétrique défini positif de $\mathcal{A}$ ne suffit pas à garantir la convergence de la méthode de Jacobi.

Une variante de la méthode de Gauß–Seidel, qui, en général, donne lieu à un meilleur taux de convergence, est la *méthode SOR* (de l'anglais, Successive Over-Relaxation). La méthode SOR consiste à prendre

$$\mathcal{P}_{SOR} = \tfrac{1}{\omega}\mathcal{D} - \mathcal{L}, \tag{11.19}$$

$$\mathcal{T}_{SOR} = (\mathcal{D} - \omega\mathcal{L})^{-1}((1-\omega)\mathcal{D} + \omega\mathcal{U}), \tag{11.20}$$

où ω est un paramètre réel. Pour $\omega = 1$, on retrouve la méthode de Gauß–Seidel. La relation (11.3) s'écrit sous la forme

$$U_i^{k+1} = U_i^k + \frac{\omega}{\mathcal{A}_{ii}}\left(F_i - \sum_{j=1}^{i-1}\mathcal{A}_{ij}U_j^{k+1} - \sum_{j=i}^{N}\mathcal{A}_{ij}U_j^k\right). \tag{11.21}$$

Cette relation montre que le coût asymptotique par itération de la méthode SOR est le même que celui de la méthode de Gauß–Seidel. Il reste à étudier la convergence de la méthode SOR.

Proposition 11.8. *La matrice $\mathcal{T}_{\mathrm{SOR}}$ est convergente seulement si $0 < \omega < 2$. Réciproquement, si la matrice $\mathcal{A}$ est symétrique définie positive et si $0 < \omega < 2$, la matrice $\mathcal{T}_{\mathrm{SOR}}$ est convergente.*

Dans la pratique, on souhaite optimiser la valeur du paramètre ω afin d'obtenir le meilleur taux de convergence possible. La difficulté est que cette valeur optimale ne peut être déterminée explicitement que dans certains cas particuliers car elle dépend du spectre de la matrice $\mathcal{A}$. Par exemple, si la matrice $\mathcal{A}$ est tridiagonale, symétrique et définie positive, on montre que

$$\rho(\mathcal{T}_{\mathrm{GS}}) = \rho(\mathcal{T}_{\mathrm{J}})^2 < 1, \tag{11.22}$$

et la valeur optimale du paramètre ω est telle que

$$\omega_{\mathrm{opt}} = \frac{2}{1 + (1 - \rho(\mathcal{T}_{\mathrm{GS}}))^{\frac{1}{2}}}. \tag{11.23}$$

Pour cette valeur optimale, on a

$$\rho(\mathcal{T}_{\mathrm{SOR}}) = \omega_{\mathrm{opt}} - 1. \tag{11.24}$$

Afin d'illustrer la sensibilité de la méthode SOR au choix du paramètre ω, la figure 11.1 présente le nombre d'itérations effectuées par cette méthode en fonction du paramètre ω afin de résoudre un système linéaire avec une matrice tridiagonale d'ordre $N = 20$. La matrice $\mathcal{A}$ est celle résultant de la discrétisation par éléments finis de Lagrange $\mathbb{P}_1$ sur maillage uniforme du Laplacien en une dimension d'espace. Le membre de droite F a toutes ses

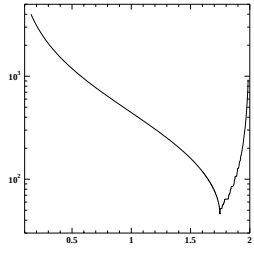


Figure 11.1 – Nombre d'itérations effectuées par la méthode SOR en fonction du paramètre ω.

composantes nulles sauf la première et la dernière qui sont égales à 1. Le critère de convergence est $\|U^k - U^{k-1}\|_N \leqslant 10^{-6}\|U^k\|_N$.

11.2 Gradient conjugué et variantes

L'objet de cette section est de présenter une méthode classique pour la résolution itérative des grands systèmes linéaires dans le cadre de la méthode des éléments finis : la méthode du gradient conjugué. On introduit d'abord les idées essentielles qui sous-tendent cette méthode à partir de l'étude d'une classe de méthodes plus simples, à savoir les méthodes de type gradient. Puis, on présente la méthode du gradient conjugué pour les systèmes symétriques définis positifs. On en décrit ensuite une généralisation, connue sous le nom de méthode GMRes, qui est pertinente pour des systèmes linéaires où la seule propriété de la matrice est le fait d'être inversible. Enfin, on aborde la question, très importante, du préconditionnement de ces méthodes.

11.2.1 Les méthodes de type gradient

On considère la fonctionnelle

$$J : \mathbb{R}^N \ni V \longmapsto \tfrac{1}{2}(\mathcal{A}V, V)_N - (F, V)_N \in \mathbb{R}. \tag{11.25}$$

On suppose que la matrice $\mathcal{A}$ est *symétrique définie positive*. La symétrie de $\mathcal{A}$ implique que pour tout $V, W \in \mathbb{R}^N$ et pour tout $t \in \mathbb{R}$, on a

$$J(V + tW) = J(V) + t(\mathcal{A}V - F, W)_N + \tfrac{1}{2}t^2(\mathcal{A}W, W)_N. \tag{11.26}$$

L'expression (11.26) est un développement de Taylor de J en V à l'ordre deux ; ce développement est exact car J est quadratique en V. De plus, on observe que le gradient de J en V est égal à l'opposé du résidu de V, ce que l'on écrit sous la forme

$$\nabla J(V) = \mathcal{A}V - F = -R(V), \tag{11.27}$$

et que la matrice hessienne de J en V est égale à la matrice $\mathcal{A}$. La matrice $\mathcal{A}$ étant définie positive, la fonctionnelle J est convexe sur $\mathbb{R}^N$. Enfin, des expressions (11.26) et (11.27) on déduit aisément le résultat suivant.

Proposition 11.9. *Le vecteur $U \in \mathbb{R}^N$ est solution du système linéaire $\mathcal{A}U = F$ si et seulement si $\nabla J(U) = 0$, c'est-à-dire si et seulement si U réalise le minimum de la fonctionnelle J sur $\mathbb{R}^N$.*

La proposition 11.9 est à la base de la construction des méthodes itératives de type gradient. En effet, le principe de ces méthodes est de chercher à minimiser la fonctionnelle J sur $\mathbb{R}^N$. Une façon simple de procéder consiste à choisir un vecteur initial $U^0 \in \mathbb{R}^N$, puis pour tout $k \geqslant 0$, à réaliser les opérations suivantes :

(i) calculer le gradient de la fonctionnelle J en U^k,

$$\nabla J(U^k) = -R^k = \mathcal{A}U^k - F. \tag{11.28}$$

(ii) sélectionner une direction de descente D^k ; on prend

$$D^k = -\nabla J(U^k) = R^k. \tag{11.29}$$

Ce choix est guidé par le fait que localement au voisinage de U^k, D^k est la direction selon laquelle J décroît le plus rapidement.

(iii) choisir U^{k+1} sur la droite $\{U^k + tD^k\,;\, t \in \mathbb{R}\}$,

$$U^{k+1} = U^k + t^k D^k. \tag{11.30}$$

Dans la méthode du *gradient à pas fixe*, on se fixe *a priori* une valeur du pas $t^k = t^*$ qui n'est pas modifiée au fil des itérations. Dans la méthode du *gradient à pas optimal*, on effectue une recherche de minimum sur la droite $\{U^k + tD^k\,;\, t \in \mathbb{R}\}$. On a donc

$$J(U^{k+1}) = \inf_{t\in\mathbb{R}} J(U^k + tD^k). \tag{11.31}$$

En posant $\psi_k : \mathbb{R} \ni t \longmapsto J(U^k + tD^k) \in \mathbb{R}$, on obtient $\psi_k'(t) = t(\mathcal{A}D^k, D^k)_N - (R^k, D^k)_N$, si bien que

$$t^k = \frac{(R^k, D^k)_N}{(\mathcal{A}D^k, D^k)_N}. \tag{11.32}$$

En utilisant (11.29), il vient

$$t^k = \frac{(R^k, R^k)_N}{(\mathcal{A}R^k, R^k)_N}. \tag{11.33}$$

On observe que le rapport ci-dessus est bien défini tant que $R^k \neq 0$; or, $R^k = 0$ implique que U^k est la solution cherchée.

On peut envisager plusieurs tests de convergence pour les itérations (i)–(iii). Le plus couramment utilisé est basé sur la norme euclidienne du résidu R^k : on arrête les itérations lorsque $\|R^k\|_N \leqslant \texttt{tol}$. En général, la tolérance $\texttt{tol}$ est évaluée à partir d'une tolérance absolue, une tolérance relative et la norme euclidienne du résidu initial selon l'expression $\texttt{tol} = \texttt{tol}_{\text{abs}} + \texttt{tol}_{\text{rel}}\|R^0\|_N$. L'algorithme 11.1 récapitule la mise en œuvre de la méthode du gradient à pas optimal. On observera que le coût asymptotique par itération est dominé par le produit matrice–vecteur $Z^k = \mathcal{A}R^k$; ce coût est de l'ordre de N^2 opérations si la matrice $\mathcal{A}$ est dense, mais il est proportionnel à N si la matrice $\mathcal{A}$ est creuse ; voir la définition 10.14. La méthode du gradient à pas fixe a clairement le même coût asymptotique par itération. Il reste à étudier la convergence de ces deux méthodes.

Algorithme 11.1 Méthode du gradient à pas optimal

choisir $U^0 \in \mathbb{R}^N$, poser $R^0 = F - \mathcal{A}U^0$
choisir une tolérance $\texttt{tol}$
poser $k = 0$
while $\|R^k\|_N > \texttt{tol}$ **do**
  calculer le vecteur $Z^k = \mathcal{A}R^k$
  $t^k = (R^k, R^k)_N / (Z^k, R^k)_N$
  $U^{k+1} = U^k + t^k R^k$
  $R^{k+1} = R^k - t^k Z^k$
  $k \leftarrow k + 1$
end while

Proposition 11.10. *On suppose que la matrice $\mathcal{A}$ est symétrique définie positive. Alors,*

(i) *la méthode du gradient à pas optimal converge quel que soit $U^0 \in \mathbb{R}^N$;*

(ii) *si le paramètre t^* est suffisamment petit, la méthode du gradient à pas fixe converge quel que soit $U^0 \in \mathbb{R}^N$.*

Dans la pratique, la méthode du gradient à pas optimal est bien plus efficace que la méthode du gradient à pas fixe : les deux méthodes ont le même coût asymptotique par itération et la première offre un meilleur taux de convergence que la seconde.

11.2.2 La méthode du gradient conjugué

La méthode du gradient conjugué, introduite par Hestenes et Stiefel en 1952, est une généralisation de la méthode du gradient à pas optimal présentée dans la section précédente. Assortie d'un bon préconditionneur (voir la section 11.2.4 ci-dessous), la méthode du gradient conjugué est une des méthodes les plus efficaces pour la résolution itérative des systèmes linéaires obtenus dans le cadre de la méthode des éléments finis et dont la matrice associée est *symétrique définie positive*. L'objet de cette section est de présenter le principe général de la méthode du gradient conjugué et ses propriétés remarquables en évitant de trop rentrer dans les détails techniques ; on renvoie à Saad [**66**, p. 193] ou à Golub et van Loan [**47**, p. 525] pour des compléments. Soit $U^0 \in \mathbb{R}^N$; on pose $R^0 = F - \mathcal{A}U^0$. Pour un entier $k \geqslant 1$, on introduit *l'espace de Krylov*

$$\mathbb{K}_k = \text{vect}\{R^0, \mathcal{A}R^0, \ldots, \mathcal{A}^{k-1}R^0\}. \tag{11.34}$$

Par récurrence, il est clair que la suite $(U^k)_{k\geqslant 0}$ générée par la méthode du gradient à pas optimal (voir l'algorithme 11.1 ci-dessus) est telle que

$$U^k \in U^0 + \mathbb{K}_k. \tag{11.35}$$

Il est également clair que

$$\mathbb{K}_k = \text{vect}\{R^0, \ldots, R^{k-1}\}. \tag{11.36}$$

De plus, par construction, le vecteur U^k fourni par la méthode du gradient à pas optimal jouit d'une propriété d'optimalité unidimensionnelle dans $U^0 + \mathbb{K}_k$ puisqu'il réalise le minimum de la fonctionnelle J sur la droite $\{U^{k-1} + tD^{k-1} ; t \in \mathbb{R}\}$ (qui est incluse dans $U^0 + \mathbb{K}_k$). On souhaiterait modifier l'algorithme afin que U^k satisfasse une propriété d'optimalité sur tout le sous-espace affine $U^0 + \mathbb{K}_k$.

Définition 11.11. *Soit un entier $k \geqslant 0$. Une famille de vecteurs $\{P^0, \ldots, P^k\}$ est dite $\mathcal{A}$-*conjuguée *si pour tout $m, n \in \{0, \ldots, k\}$ avec $m \neq n$, on a*

$$(\mathcal{A}P^m, P^n)_N = 0. \tag{11.37}$$

Le principe de la méthode du gradient conjugué consiste à construire trois suites de vecteurs, $(U^k)_{k\geqslant 0}$, $(R^k)_{k\geqslant 0}$ et $(P^k)_{k\geqslant 0}$, telles que $U^0 \in \mathbb{R}^N$, $P^0 = R^0 = F - \mathcal{A}U^0$ et pour tout $k \geqslant 1$,

(i) R^k est le résidu de U^k ;

(ii) $\mathbb{K}_k = \text{vect}\{R^0, \ldots, R^{k-1}\} = \text{vect}\{P^0, \ldots, P^{k-1}\}$;

(iii) la famille $\{R^0, \ldots, R^{k-1}\}$ forme une base *orthogonale* de $\mathbb{K}_k$;

(iv) la famille $\{P^0, \ldots, P^{k-1}\}$ est $\mathcal{A}$-conjuguée ;

La construction de ces suites est basée sur une forme astucieuse du procédé d'orthogonalisation de Gram-Schmidt. Elle repose sur les récurrences suivantes :

$$U^{k+1} = U^k + s^k P^k, \tag{11.38}$$

$$R^{k+1} = R^k - s^k \mathcal{A}P^k, \tag{11.39}$$

$$P^{k+1} = R^{k+1} + t^k P^k, \tag{11.40}$$

où s^k et t^k sont des réels donnés par les formules

$$s^k = \frac{(R^k, R^k)_N}{(\mathcal{A}P^k, P^k)_N} \qquad \text{et} \qquad t^k = -\frac{(R^{k+1}, \mathcal{A}P^k)_N}{(\mathcal{A}P^k, P^k)_N}. \tag{11.41}$$

Les relations (11.38) et (11.39) impliquent clairement que R^k est le résidu de U^k. De plus, on constate que

$$\begin{aligned}(R^{k+1}, R^k)_N &= (R^k, R^k)_N - s^k(\mathcal{A}P^k, R^k)_N \\ &= (R^k, R^k)_N - s^k(\mathcal{A}P^k, P^k - t^{k-1}P^{k-1})_N \\ &= (R^k, R^k)_N - s^k(\mathcal{A}P^k, P^k)_N = 0,\end{aligned} \tag{11.42}$$

car la famille $\{P^0, \ldots, P^{k-1}\}$ est $\mathcal{A}$-conjuguée. Ce calcul justifie le choix de s^k comme celui permettant d'orthogonaliser R^{k+1} et R^k. De même,

$$(P^{k+1}, \mathcal{A}P^k)_N = (R^{k+1}, \mathcal{A}P^k)_N + t^k(P^k, \mathcal{A}P^k)_N = 0, \tag{11.43}$$

ce qui justifie le choix de t^k comme celui assurant la $\mathcal{A}$-conjugaison de P^{k+1} et de P^k. La propriété remarquable est que les récurrences d'ordre 1 induites

par les formules (11.39) et (11.40) permettent d'assurer que le nouveau résidu R^{k+1} est orthogonal à *tous* les vecteurs $\{R^0, \ldots, R^k\}$ et que le nouveau vecteur P^{k+1} est $\mathcal{A}$-conjugué à *tous* les vecteurs $\{P^0, \ldots, P^k\}$. La preuve, de nature technique, est omise ; on retiendra simplement le fait qu'elle repose fondamentalement sur la symétrie de la matrice $\mathcal{A}$.

La construction des suites $(U^k)_{k\geqslant 0}$, $(R^k)_{k\geqslant 0}$ et $(P^k)_{k\geqslant 0}$ selon les récurrences (11.38)–(11.40) est possible tant que $(\mathcal{A}P^k, P^k)_N \neq 0$. Or, ceci ne peut se produire que si U^k est la solution du système linéaire $\mathcal{A}U = F$, c'est-à-dire si $R^k = 0$. En effet, $(\mathcal{A}P^k, P^k)_N = 0$ implique $P^k = 0$ si bien qu'à l'itération précédente, on a obtenu $0 = R^k + t^{k-1}P^{k-1}$. Comme $P^{k-1} \in \mathbb{K}_k$ et que R^k est orthogonal à cet espace, on en déduit $R^k = 0$. En conclusion, tant que $R^k \neq 0$, les récurrences (11.38)–(11.40) sont bien définies.

Une autre propriété remarquable de la méthode du gradient conjugué est la suivante.

Proposition 11.12. *Le vecteur U^k réalise le minimum de la fonctionnelle J sur le sous-espace affine $U^0 + \mathbb{K}_k$.*

En effet, on déduit de (11.38) que

$$U^k = U^0 + \sum_{l=0}^{k-1} s^l P^l. \tag{11.44}$$

Or, un vecteur $V \in \mathbb{R}^N$ qui s'écrit sous la forme $V = U^0 + \sum_{l=0}^{k-1} \gamma^l P^l$ réalise le minimum de J sur le sous-espace affine $U^0 + \mathbb{K}_k$ si et seulement si pour tout $l \in \{0, \ldots, k-1\}$,

$$(\nabla J(V), P^l)_N = 0. \tag{11.45}$$

Un calcul direct montre que $\nabla J(V) = -R^0 + \sum_{l=0}^{k-1} \gamma^l \mathcal{A}P^l$ et en utilisant le fait que la famille $\{P^0, \ldots, P^{k-1}\}$ est $\mathcal{A}$-conjuguée, on en déduit

$$\gamma_l = \frac{(R^0, P^l)_N}{(\mathcal{A}P^l, P^l)_N}, \qquad l \in \{0, \ldots, k-1\}. \tag{11.46}$$

Il est clair que pour tout $l \in \{0, \ldots, k-1\}$,

$$\begin{aligned}(R^0, P^l)_N &= (R^l + \textstyle\sum_{m=0}^{l-1} s^m \mathcal{A} P^m, P^l)_N \\ &= (R^l, P^l)_N = (R^l, R^l)_N + t^{l-1}(R^l, P^{l-1})_N = (R^l, R^l)_N.\end{aligned} \tag{11.47}$$

Par conséquent, $\gamma_l = s_l$ pour tout $l \in \{0, \ldots, k-1\}$, ce qui complète la preuve de la proposition 11.12.

Algorithme 11.2 Méthode du gradient conjugué

choisir $U^0 \in \mathbb{R}^N$, poser $R^0 = F - \mathcal{A}U^0$ et $P^0 = R^0$
choisir une tolérance `tol`
poser $k = 0$
while $\|R^k\|_N >$ `tol` **do**
  calculer le vecteur $Z^k = \mathcal{A}P^k$
  $s^k = (R^k, R^k)_N / (Z^k, P^k)_N$
  $U^{k+1} = U^k + s^k P^k$
  $R^{k+1} = R^k - s^k Z^k$
  $t^k = (R^{k+1}, R^{k+1})_N / (R^k, R^k)_N$
  $P^{k+1} = R^{k+1} + t^k P^k$
  $k \leftarrow k + 1$
end while

L'algorithme 11.2 décrit la mise en œuvre de la méthode du gradient conjugué. Les coefficients s^k et t^k ont été réécrits de manière équivalente à (11.41). Le coût asymptotique par itération est dominé par le produit matrice–vecteur $Z^k = \mathcal{A}P^k$. Lorsque la matrice $\mathcal{A}$ est creuse, celui-ci est proportionnel à N. À ce coût s'ajoutent celui des deux produits scalaires pour calculer le coefficient s^k et celui des trois mises à jour vectorielles dans (11.38)–(11.40). Tous ces coûts sont proportionnels à N.

Il reste à étudier la convergence de la méthode du gradient conjugué. La proposition 11.12 implique que la méthode du gradient conjugué converge en, au plus, N itérations ; en effet, si la convergence ne s'est pas produite pour $k < N$, on constate que pour $k = N$, on a $U^0 + \mathbb{K}_k = \mathbb{R}^N$ si bien que U^N est la solution cherchée d'après la proposition 11.9. Dans ces conditions, la

méthode du gradient conjugué peut être vue comme une méthode d'inversion directe plutôt que comme une méthode itérative. Toutefois, dans la pratique, la méthode du gradient conjugué est employée comme méthode itérative car le nombre d'itérations nécessaires afin d'obtenir une précision suffisante est souvent bien inférieur à N.

Le résultat suivant permet d'estimer le taux de convergence de la méthode du gradient conjugué en fonction du conditionnement de la matrice $\mathcal{A}$. Pour $X \in \mathbb{R}^N$, on considère la norme $\|X\|_{\mathcal{A}} = (\mathcal{A}X, X)_N^{\frac{1}{2}}$. On rappelle que $\kappa(\mathcal{A})$ désigne le nombre de conditionnement de la matrice $\mathcal{A}$; voir la section 10.1.

Proposition 11.13. *Pour tout $k \geqslant 1$, on a*

$$\|U - U^k\|_{\mathcal{A}} \leqslant 2 \left(\frac{\kappa(\mathcal{A})^{\frac{1}{2}} - 1}{\kappa(\mathcal{A})^{\frac{1}{2}} + 1} \right)^k \|U - U^0\|_{\mathcal{A}}. \tag{11.48}$$

La proposition 11.13 montre que le taux de convergence de la méthode du gradient conjugué se détériore lorsque $\kappa(\mathcal{A}) \gg 1$, c'est-à-dire lorsque la matrice $\mathcal{A}$ est mal conditionnée. Dans ce cas, il convient de modifier la méthode du gradient conjugué à l'aide d'un préconditionneur ; voir la section 11.2.4 pour plus de détails.

11.2.3 La méthode GMRes

On souhaite généraliser la méthode du gradient conjugué à la résolution itérative de systèmes linéaires dont la matrice n'est plus symétrique. Or, en l'absence de propriété de symétrie, on peut montrer que de telles généralisations ne peuvent préserver qu'une seule des deux propriétés remarquables de la méthode du gradient conjugué :

(i) soit le fait que la k-ième itérée U^k jouit d'une propriété d'optimalité sur un sous-espace affine de dimension k (voir la proposition 11.12) ;

(ii) soit le fait que la méthode n'emploie que des récurrences d'ordre 1 si bien que le coût par itération n'augmente pas au fil des itérations.

La méthode GMRes, qui est présentée dans cette section, préserve la propriété (i). La méthode Bi-CGStab, qui est présentée dans la section 11.4, préserve la propriété (ii). Ces deux méthodes sont souvent employées pour

la résolution itérative de systèmes linéaires non-symétriques dans le cadre de la méthode des éléments finis.

La méthode GMRes (de l'anglais, Generalized Minimal Residual Method) a été introduite par Saad et Schultz en 1986 [**67**]. Soit $U^0 \in \mathbb{R}^N$; on pose $R^0 = F - \mathcal{A}U^0$ et on considère à nouveau l'espace de Krylov $\mathbb{K}_k$ défini en (11.34). Le principe de la méthode GMRes consiste à construire la k-ième itérée U^k de sorte que celle-ci minimise la norme euclidienne du résidu R^k sur le sous-espace affine $U^0 + \mathbb{K}_k$.

Afin d'écrire les conditions d'optimalité permettant de déterminer U^k, on considère une base judicieusement choisie de $\mathbb{K}_k$. La base en question, $\{V^1, \ldots, V^k\}$, est telle que :

(i) la famille $\{V^1, \ldots, V^k\}$ est orthonormée ;

(ii) $R^0 = \beta V^1$ avec $\beta = \|R^0\|_N$;

(iii) pour tout $l < k$, $\mathcal{A}V^l \in \text{vect}\{V^1, \ldots, V^{l+1}\}$.

La construction de cette base se fait selon une variante du procédé d'orthogonalisation de Gram–Schmidt connue sous le nom d'algorithme d'Arnoldi ; voir l'algorithme 11.3. L'éventualité où $h_{j+1,j} = 0$ se produit lorsque $\mathbb{K}_j$ est invariant par multiplication par $\mathcal{A}$. Dans ce cas, on montre que la solution du système linéaire $\mathcal{A}U = F$ est dans $U^0 + \mathbb{K}_j$.

Algorithme 11.3 Algorithme d'Arnoldi

poser $\beta = \|R^0\|_N$ et $V^1 = R^0/\beta$
poser $j = 1$
while $j \leqslant k$ **do**
  $h_{i,j} = (\mathcal{A}V^j, V^i)_N$ pour $i \in \{1, \ldots, j\}$
  $\widehat{V}^{j+1} = \mathcal{A}V^j - \sum_{i=1}^{j} h_{i,j} V^i$
  $h_{j+1,j} = \|\widehat{V}^{j+1}\|_N$
  si $h_{j+1,j} = 0$ **stop**
  $V^{j+1} = \widehat{V}^{j+1}/h_{j+1,j}$
  $j \leftarrow j + 1$
end while

L'algorithme d'Arnoldi génère $(k+1)$ vecteurs $\{V^1,\dots,V^{k+1}\}$ tels que la famille $\{V^1,\dots,V^k\}$ satisfait les trois propriétés (i)–(iii) ci-dessus. On note $\mathcal{V}^k \in \mathbb{R}^{N,k}$ la matrice dont les colonnes sont les vecteurs $\{V^1,\dots,V^k\}$ et on utilise une notation analogue pour $\mathcal{V}^{k+1} \in \mathbb{R}^{N,k+1}$. Les vecteurs $\{V^1,\dots,V^{k+1}\}$ étant orthonormés, on a $(\mathcal{V}^{k+1})^T\mathcal{V}^{k+1} = \mathcal{I}_{k+1}$ où $\mathcal{I}_{k+1}$ désigne la matrice identité dans $\mathbb{R}^{k+1,k+1}$. On introduit également la matrice $\mathcal{H}^k \in \mathbb{R}^{k+1,k}$ de terme générique $\mathcal{H}^k_{ij} = h_{i,j}$ pour $j \in \{1,\dots,k\}$ et $i \in \{1,\dots,j+1\}$ et $\mathcal{H}^k_{ij} = 0$ pour $j \in \{1,\dots,k-1\}$ et $i \in \{j+2,\dots,k+1\}$; la matrice $\mathcal{H}^k$ est rectangulaire et telle que

$$\mathcal{H}^k = \begin{pmatrix} \mathcal{H}^k_{11} & \dots & \dots & \mathcal{H}^k_{1k} \\ \mathcal{H}^k_{21} & \mathcal{H}^k_{22} & & \vdots \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{H}^k_{kk} \\ 0 & \dots & 0 & \mathcal{H}^k_{k+1,k} \end{pmatrix}. \tag{11.49}$$

Une matrice avec une telle structure est appelée une *matrice de Hessenberg*. Enfin, d'après l'algorithme 11.3, on a

$$\mathcal{A}\mathcal{V}^k = \mathcal{V}^{k+1}\mathcal{H}^k. \tag{11.50}$$

Un vecteur $U^k \in U^0 + \mathbb{K}_k$ peut s'écrire sous la forme $U^k = U^0 + \sum_{l=1}^k Y_l V^l$ avec $Y_l \in \mathbb{R}$ pour tout $l \in \{1,\dots,k\}$. En posant $Y = (Y_l)_{1\leqslant l\leqslant k} \in \mathbb{R}^k$, on a donc $U^k = U^0 + \mathcal{V}^k Y$. Par ailleurs, en utilisant le fait que $R^0 = \beta V^1$ et $(\mathcal{V}^{k+1})^T\mathcal{V}^{k+1} = \mathcal{I}_{k+1}$, il vient

$$\|R^k\|_N = \|R^0 - \mathcal{A}\mathcal{V}^k Y\|_N = \|\beta V^1 - \mathcal{V}^{k+1}\mathcal{H}^k Y\|_N = \|\beta e_1 - \mathcal{H}^k Y\|_{\mathbb{R}^{k+1}}, \tag{11.51}$$

où e_1 est le premier vecteur de la base canonique de $\mathbb{R}^{k+1}$. Chercher le vecteur U^k minimisant la norme euclidienne du résidu sur le sous-espace affine $U^0 + \mathbb{K}_k$ équivaut donc à chercher $Y^* \in \mathbb{R}^k$ réalisant le minimum de $\|\beta e_1 - \mathcal{H}^k Y\|_{\mathbb{R}^{k+1}}$ sur $\mathbb{R}^k$. Il s'agit d'un problème d'optimisation standard. La matrice $\mathcal{H}^k$ étant de rang maximal, la solution Y^* est unique. Par ailleurs, celle-ci peut être évaluée en effectuant une factorisation QR de la matrice $\mathcal{H}^k$: il existe une matrice unitaire $\mathcal{Q}^k \in \mathbb{R}^{k+1,k+1}$ telle que la matrice

$\mathcal{R}^k = \mathcal{Q}^k\mathcal{H}^k \in \mathbb{R}^{k+1,k}$ admet la structure bloc suivante :

$$\mathcal{R}^k = \begin{pmatrix} \mathcal{U}^k \\ \hdashline 0 \ldots 0 \end{pmatrix}, \tag{11.52}$$

où la matrice $\mathcal{U}^k \in \mathbb{R}^{k,k}$ est triangulaire supérieure et inversible. La construction pratique des matrices $\mathcal{Q}^k$ et $\mathcal{R}^k$ est détaillée ci-dessous. Par conséquent, on obtient

$$\|\beta e_1 - \mathcal{H}^k Y\|_{\mathbb{R}^{k+1}} = \|\mathcal{Q}^k(\beta e_1 - \mathcal{H}^k Y)\|_{\mathbb{R}^{k+1}} = \|G^k - \mathcal{R}^k Y\|_{\mathbb{R}^{k+1}}, \tag{11.53}$$

où $G^k = \beta\mathcal{Q}^k e_1 \in \mathbb{R}^{k+1}$. Le vecteur $Y^* \in \mathbb{R}^k$ réalisant le minimum de $\|\beta e_1 - \mathcal{H}^k Y\|_{\mathbb{R}^{k+1}}$ sur $\mathbb{R}^k$ s'obtient donc en inversant le système triangulaire supérieur

$$\mathcal{U}^k \begin{pmatrix} Y_1^* \\ \vdots \\ Y_k^* \end{pmatrix} = \begin{pmatrix} G_1^k \\ \vdots \\ G_k^k \end{pmatrix}. \tag{11.54}$$

La k-ième itérée de la méthode GMRes s'écrit sous la forme

$$U^k = U^0 + \sum_{l=1}^{k} Y_l^* V^l, \tag{11.55}$$

et le résidu correspondant est

$$\|R^k\|_N = |G_{k+1}^k|. \tag{11.56}$$

Une propriété agréable de la méthode GMRes est qu'il n'est pas nécessaire de reconstruire le vecteur U^k afin d'évaluer son résidu. Cette idée est reprise dans l'algorithme 11.4.

Afin de construire les matrices $\mathcal{Q}^k$ et $\mathcal{R}^k$ au fil des itérations de l'algorithme 11.4, on procède de la manière suivante. La matrice $\mathcal{Q}^k$ est décomposée en un produit de matrices de rotation élémentaires (dites matrices de Givens) sous la forme $\mathcal{Q}^k = \prod_{l=1}^{k} \mathcal{F}^l$. Les matrices de Givens

Algorithme 11.4 Méthode GMRes

choisir $U^0 \in \mathbb{R}^N$ et une tolérance `tol`
poser $k = 0$, $R^0 = F - \mathcal{A}U^0$, $\beta = \|R^0\|_N$ et $V^1 = R^0/\beta$
while $\|R^k\|_N >$ `tol` **do**
  $k \leftarrow k + 1$
  $h_{i,k} = (\mathcal{A}V^k, V^i)_N$ pour $i \in \{1, \ldots, k\}$
  $\widehat{V}^{k+1} = \mathcal{A}V^k - \sum_{i=1}^k h_{i,k} V^i$
  $h_{k+1,k} = \|\widehat{V}^{k+1}\|_N$
  $V^{k+1} = \widehat{V}^{k+1}/h_{k+1,k}$
  évaluer les matrices $\mathcal{H}^k$, $\mathcal{Q}^k$ et $\mathcal{R}^k$
  évaluer le vecteur $G^k = \mathcal{Q}^k \beta e_1$
  poser $\|R^k\|_N = |G^k_{k+1}|$
end while
résoudre le système triangulaire supérieur (11.54)
poser $U^k = U^0 + \sum_{l=1}^k Y_l^* V^l$

ont la structure suivante :

$$\mathcal{F}^l = \begin{pmatrix} \mathcal{I}_{l-1} & 0 & 0 \\ 0 & \begin{matrix} c_l & s_l \\ -s_l & c_l \end{matrix} & 0 \\ 0 & 0 & \mathcal{I}_{k-l} \end{pmatrix}, \qquad l \in \{1, \ldots, k\},$$

où, pour un entier $m \geqslant 1$, $\mathcal{I}_m$ désigne la matrice identité dans $\mathbb{R}^{m,m}$ (lorsque $m = 0$, la matrice est omise de la structure bloc de $\mathcal{F}^l$). Les coefficients c_l et s_l sont calculés au fil des itérations. En désignant par $h_{ij}^{(l)}$ les coefficients génériques de la matrice $(\prod_{l'=1}^l \mathcal{F}^{l'})\mathcal{H}^k$, on a

$$c_l = \frac{\alpha_l}{(\alpha_l^2 + \beta_l^2)^{\frac{1}{2}}} \qquad \text{et} \qquad s_l = \frac{\beta_l}{(\alpha_l^2 + \beta_l^2)^{\frac{1}{2}}}, \tag{11.57}$$

avec $\alpha_l = h_{ll}^{(l-1)}$ et $\beta_l = h_{l+1,l}^{(l-1)}$.
Contrairement à la méthode du gradient conjugué, la méthode GMRes nécessite la connaissance de tous les vecteurs $\{V^1, \ldots, V^k\}$ constituant la base

d'Arnoldi de l'espace de Krylov $\mathbb{K}_k$ afin de construire le nouveau vecteur V^{k+1}. Pour cette raison, le coût par itération croît linéairement au fil des itérations ; il en va de même de la place mémoire nécessaire pour stocker la base d'Arnoldi. Dans la pratique, on utilise souvent une variante de la méthode GMRes avec réinitialisation. On se fixe *a priori* une dimension maximale n pour l'espace de Krylov. Si la convergence n'a pas été obtenue au bout de n itérations, on évalue l'itérée courante U^n et on réinitialise l'algorithme en prenant $U^0 = U^n$. Cette variante de la méthode de GMRes est connue sous le nom de GMRes(n).

Il reste à étudier la convergence des méthodes GMRes et GMRes(n). Les principaux résultats sont rappelés ci-dessous ; voir, par exemple, Saad [**66**, p. 195] pour la preuve et des compléments.

Proposition 11.14. *On suppose que la matrice $\mathcal{A} \in \mathbb{R}^{N,N}$ est inversible. Alors,*

(i) *l'algorithme* 11.4 *converge en au plus N itérations quel que soit $U^0 \in \mathbb{R}^N$;*

(ii) *si la matrice $\mathcal{A}$ est définie positive, l'algorithme* GMRes(n) *converge pour tout $n \geqslant 1$ quel que soit $U^0 \in \mathbb{R}^N$.*

Afin de quantifier le taux de convergence de la méthode GMRes, on suppose que la matrice $\mathcal{A}$ est diagonalisable dans $\mathbb{C}$ sous la forme $\mathcal{A} = \Pi\Lambda\Pi^{-1}$ où Λ est une matrice diagonale. On note $E(c, d, a)$ l'ellipse du plan complexe de centre $c \in \mathbb{R}$, de distance focale d et de demi-grand axe a (les paramètres a et d sont soit réels soit imaginaires purs). On suppose que toutes les valeurs propres de $\mathcal{A}$ sont contenues dans l'ellipse $E(c, d, a)$ et que $0 \notin E(c, d, a)$.

Proposition 11.15. *Dans le cadre des hypothèses ci-dessus, on a pour tout $k \geqslant 1$,*

$$\|R^k\|_N \leqslant \kappa(\Pi) \frac{\Xi_k(\frac{a}{d})}{\Xi_k(\frac{c}{d})} \|R^0\|_N, \tag{11.58}$$

où $\Xi_k(z) = \xi(z)^k + \xi(z)^{-k}$ et $\xi(z) = z + (z^2 - 1)^{\frac{1}{2}}$.

Lorsque la matrice $\mathcal{A}$ est mal conditionnée, on obtient l'estimation

$$\|R^k\|_N \leqslant c \left(1 - \frac{2}{\kappa(\mathcal{A})^{\frac{1}{2}}}\right)^k \|R^0\|_N. \tag{11.59}$$

Le taux de convergence a donc un comportement analogue à celui observé pour la méthode du gradient conjugué.

11.2.4 Préconditionnement

L'objectif d'un préconditionneur est d'améliorer le taux de convergence du solveur itératif lorsque la matrice associée au système linéaire est mal conditionnée. Le principe général consiste à remplacer le système linéaire $\mathcal{A}U = F$ par un nouveau système, $\widetilde{\mathcal{A}}\widetilde{U} = \widetilde{F}$, dont la matrice $\widetilde{\mathcal{A}}$ est mieux conditionnée que $\mathcal{A}$.

On considère deux matrices $\mathcal{P}_{\mathrm{g}} \in \mathbb{R}^{N,N}$ et $\mathcal{P}_{\mathrm{d}} \in \mathbb{R}^{N,N}$ telles que les systèmes linéaires $\mathcal{P}_{\mathrm{g}}X = Y$ et $\mathcal{P}_{\mathrm{d}}X' = Y'$ sont relativement simples à inverser. Par exemple, les matrices $\mathcal{P}_{\mathrm{g}}$ et $\mathcal{P}_{\mathrm{d}}$ peuvent être diagonales ou triangulaires inférieure ou supérieure. On considère le nouveau système linéaire

$$\underbrace{(\mathcal{P}_{\mathrm{g}}^{-1}\mathcal{A}\mathcal{P}_{\mathrm{d}}^{-1})}_{\widetilde{\mathcal{A}}}\underbrace{(\mathcal{P}_{\mathrm{d}}U)}_{\widetilde{U}} = \underbrace{\mathcal{P}_{\mathrm{g}}^{-1}F}_{\widetilde{F}}. \tag{11.60}$$

La matrice $\mathcal{P}_{\mathrm{g}}$ est appelée le *préconditionneur à gauche* et la matrice $\mathcal{P}_{\mathrm{d}}$ est appelée le *préconditionneur à droite*. Des cas particuliers de (11.60) sont celui où le système linéaire est uniquement préconditionné à gauche ($\mathcal{P}_{\mathrm{d}} = \mathcal{I}_N$) et celui où le système linéaire est uniquement préconditionné à droite ($\mathcal{P}_{\mathrm{g}} = \mathcal{I}_N$).

Afin de mettre en œuvre une version préconditionnée de la méthode du gradient conjugué, on doit préserver la symétrie du système linéaire. On choisit donc $\mathcal{P}_{\mathrm{d}} = \mathcal{P}_{\mathrm{g}}^T$ et on pose $\mathcal{P} = \mathcal{P}_{\mathrm{g}}\mathcal{P}_{\mathrm{g}}^T$. La version préconditionnée de la méthode du gradient conjugué est présentée dans l'algorithme 11.5. On observera que seule la matrice $\mathcal{P}$ intervient dans cet algorithme. En posant $\widetilde{U}^k = \mathcal{P}_{\mathrm{g}}^T U^k$, $\widetilde{P}^k = \mathcal{P}_{\mathrm{g}}^T P^k$ et $\widetilde{R}^k = \mathcal{P}_{\mathrm{g}}^{-1} R^k$, on vérifie que $\widetilde{U}^k$, $\widetilde{P}^k$ et $\widetilde{R}^k$ sont les itérées générées par la méthode du gradient conjugué appliquée au système linéaire (11.60). Par suite, le taux de convergence de l'algorithme 11.5 dépend du nombre de conditionnement de la nouvelle matrice $\widetilde{\mathcal{A}}$.

Choisir un préconditionneur relève d'un compromis entre le coût par itération et le taux de convergence. Dans la méthode préconditionnée, le coût par itération dépend de la simplicité avec laquelle on peut inverser la matrice $\mathcal{P}$ alors que le taux de convergence dépend du nombre de conditionnement de la nouvelle matrice $\widetilde{\mathcal{A}}$. Si on choisit un préconditionneur relativement

Algorithme 11.5 Méthode du gradient conjugué préconditionnée

choisir $U^0 \in \mathbb{R}^N$, poser $R^0 = F - \mathcal{A}U^0$ et $P^0 = \mathcal{P}^{-1}R^0$
choisir une tolérance `tol`
poser $k = 0$
while $\|R^k\|_N >$ `tol` **do**
  calculer le vecteur $Z^k = \mathcal{A}P^k$
  $s^k = (\mathcal{P}^{-1}R^k, R^k)_N/(Z^k, P^k)_N$
  $U^{k+1} = U^k + s^k P^k$
  $R^{k+1} = R^k - s^k Z^k$
  $t^k = (\mathcal{P}^{-1}R^{k+1}, R^{k+1})_N/(\mathcal{P}^{-1}R^k, R^k)_N$
  $P^{k+1} = \mathcal{P}^{-1}R^{k+1} + t^k P^k$
  $k \leftarrow k + 1$
end while

simple (par exemple, la matrice diagonale de terme générique $\mathcal{P}_{ij} = \delta_{ij}\mathcal{A}_{ij}$, $i,j \in \{1,\ldots,N\}$), le coût par itération reste modéré mais, en général, le taux de convergence n'est guère amélioré. Deux exemples de préconditionneurs pour la méthode du gradient conjugué sont les suivants :

(i) **Gauß–Seidel symétrique (SGS).** On considère la décomposition (11.12). La matrice $\mathcal{A}$ étant symétrique, on a

$$\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{L}^T. \tag{11.61}$$

Le préconditionneur SGS est défini comme suit :

$$\mathcal{P}_{\mathrm{SGS}} = (\mathcal{D} - \mathcal{L})\mathcal{D}^{-1}(\mathcal{D} - \mathcal{L})^T. \tag{11.62}$$

L'emploi de ce préconditionneur nécessite la résolution d'un système triangulaire inférieur et d'un système triangulaire supérieur à chaque itération.

(ii) **Factorisation LDL^T incomplète.** Le principe de ce préconditionneur est de former la factorisation LDL^T de la matrice $\mathcal{A}$ (voir la section 10.2.3) mais de ne stocker les coefficients de $\mathcal{T}^{\mathrm{inf}}$ que s'ils sont inclus dans la structure creuse de $\mathcal{A}$. En d'autres termes, on ne stocke pas $\mathcal{T}^{\mathrm{inf}}_{ij}$ si $\mathcal{A}_{ij} = 0$. L'avantage de ce préconditionneur est qu'il

contient plus d'informations sur la matrice $\mathcal{A}$ que le préconditionneur SGS si bien qu'en pratique la matrice $\widetilde{\mathcal{A}}$ est souvent relativement bien conditionnée. Le préconditionneur LDL^T est néanmoins plus cher, tant sur le plan du coût des calculs que de la place mémoire nécessaire pour le stocker.

La méthode GMRes ne nécessitant pas de propriété de symétrie, elle peut être préconditionnée à gauche ou à droite uniquement. Un préconditionnement à gauche conduit à considérer l'espace de Krylov préconditionné

$$\mathbb{K}_k^{\mathcal{P}} = \mathrm{vect}\{R^0, (\mathcal{P}^{-1}\mathcal{A})R^0, \ldots, (\mathcal{P}^{-1}\mathcal{A})^{k-1}R^0\}. \tag{11.63}$$

La *méthode GMRes préconditionnée à gauche* se déduit de l'algorithme 11.4 en remplaçant $\mathcal{A}$ par $\mathcal{P}^{-1}\mathcal{A}$ et F par $\mathcal{P}^{-1}F$. Deux exemples de préconditionneurs pour la méthode GMRes sont le préconditionneur SGS décrit ci-dessus et la *factorisation ILU* (de l'anglais, Incomplete LU Factorization). Le principe de la factorisation LU incomplète est de former la décomposition LU de la matrice $\mathcal{A}$ sous la forme (10.47) mais de ne conserver que les coefficients $\mathcal{T}_{ij}^{\inf}$ et $\mathcal{T}_{ij}^{\sup}$ pour lesquels $\mathcal{A}_{ij} \neq 0$. On peut également modifier le préconditionneur ILU en compensant la perte des coefficients de $\mathcal{T}^{\inf}$ et $\mathcal{T}^{\sup}$ qui n'ont pas été stockés. Le préconditionneur correspondant porte le nom de préconditionneur MILU (de l'anglais, Modified ILU). Pour approfondir l'étude des techniques de préconditionnement, on pourra consulter Saad [**66**, p. 286].

11.3 Méthodes multi-échelles

Cette section contient une introduction aux méthodes multi-échelles pour la résolution itérative des systèmes linéaires. L'exemple prototype de ces méthodes est fourni par les *méthodes multi-grilles*. On présente d'abord quelques éléments relatifs à l'analyse en fréquence des méthodes de relaxation en une dimension d'espace. Puis, on étudie le principe général des méthodes multi-échelles et on décrit l'exemple classique du V-cycle. Pour une introduction détaillée aux fondements spectraux des méthodes multi-grilles, on pourra se référer à Briggs [**23**]. Pour approfondir l'étude des méthodes multi-grilles et multi-échelles, on pourra consulter, entre autres, Brandt [**19**], Hackbusch [**51**], Knabner et Angermann [**55**] ou McCormick [**57**].

11.3.1 Analyse en fréquence des méthodes de relaxation

Pour simplifier, on se place dans cette section en une dimension d'espace et on considère l'approximation par éléments finis de Lagrange $\mathbb{P}_1$ du problème de Dirichlet homogène (1.15) avec $\alpha = 1$; voir la section 1.3. On rappelle que cette approximation par éléments finis conduit à la résolution d'un système linéaire de la forme $\mathcal{A}U = F$ où la matrice $\mathcal{A}$ est d'ordre N ($N = N_{\text{so}}$, où N_{so} désigne le nombre de sommets intérieurs du maillage) et tridiagonale :

$$\mathcal{A} = \frac{1}{h}\,\text{tridiag}(-1, 2, -1), \tag{11.64}$$

où h est le pas du maillage supposé uniforme. La matrice $\mathcal{A}$ est symétrique définie positive. On vérifie que ses valeurs propres s'expriment sous la forme

$$\lambda_m = \frac{4}{h}\sin^2\left(\frac{m\pi h}{2}\right), \qquad m \in \{1, \ldots, N\}, \tag{11.65}$$

et sont associées aux vecteurs propres $\{S_1, \ldots, S_N\}$ de composantes

$$S_{m,n} = \sin(mn\pi h), \qquad m, n \in \{1, \ldots, N\}. \tag{11.66}$$

À chaque vecteur propre S_m, $m \in \{1, \ldots, N\}$, on associe la fonction propre $s_m \in P^1_{\text{c},h,0}$ définie par

$$s_m = \sum_{n=1}^{N} S_{m,n}\varphi_n, \qquad m \in \{1, \ldots, N\}, \tag{11.67}$$

où $\{\varphi_1, \ldots, \varphi_N\}$ sont les fonctions de forme dans l'espace d'approximation $P^1_{\text{c},h,0}$. La fonction propre s_m est affine par morceaux et ses valeurs aux sommets du maillage coïncident avec les composantes correspondantes du vecteur propre S_m. La figure 11.2 présente les fonctions propres s_1, s_4, s_{16} et s_{128} pour $N = 256$. On observera que la fonction propre s_m est de plus en plus oscillante à mesure que m croît. Par ailleurs, on notera que, par construction, la famille $\{s_1, \ldots, s_N\}$ forme une base de $P^1_{\text{c},h,0}$ et la famille $\{S_1, \ldots, S_N\}$ forme une base de $\mathbb{R}^N$. On pose

$$Z_{\text{BF}} = \text{vect}\{S_m\}_{1 \leqslant m \leqslant \frac{N}{2}}, \tag{11.68}$$

$$Z_{\text{HF}} = \text{vect}\{S_m\}_{\frac{N}{2} < m \leqslant N}, \tag{11.69}$$

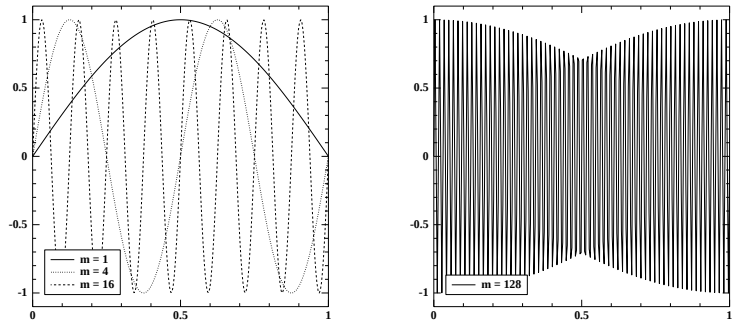


Figure 11.2 – Fonctions propres s_m pour $m \in \{1, 4, 16, 128\}$ et $N = 256$.

si bien que
$$\mathbb{R}^N = Z_{\mathrm{BF}} \oplus Z_{\mathrm{HF}}. \tag{11.70}$$

Le sous-espace Z_{BF} est appelé l'espace des *modes basse-fréquence* et le sous-espace Z_{HF} est appelé l'espace des *modes haute-fréquence*. Pour un vecteur $X \in \mathbb{R}^N$, on note $X = X_{\mathrm{BF}} + X_{\mathrm{HF}}$ la décomposition de X induite par (11.70).

On considère une méthode de relaxation pour la résolution itérative du système linéaire $\mathcal{A}U = F$; voir la section 11.1. On note $\mathcal{T}$ la matrice d'itération définie en (11.4). Étant donné un vecteur initial $U^0 \in \mathbb{R}^N$, on désigne par $(U^k)_{k\geqslant 1}$ la suite de vecteurs générée par la méthode de relaxation. On rappelle qu'en notant $E^0 = U - U^0$ l'erreur initiale et $E^k = U - U^k$ l'erreur à la k-ième itération, on a $E^k = \mathcal{T}^k E^0$. On décompose l'erreur initiale en ses composantes basse-fréquence et haute-fréquence sous la forme $E^0 = E^0_{\mathrm{BF}} + E^0_{\mathrm{HF}}$. On déduit que pour tout $k \geqslant 1$,
$$E^k = \mathcal{T}^k E^0_{\mathrm{BF}} + \mathcal{T}^k E^0_{\mathrm{HF}}. \tag{11.71}$$

Définition 11.16. *On dit que la matrice d'itération $\mathcal{T} \in \mathbb{R}^{N,N}$ est un* lisseur *pour la matrice $\mathcal{A}$ s'il existe un réel $\sigma \in]0,1[$ tel que pour tout $X = (X_{\mathrm{BF}}, X_{\mathrm{HF}}) \in \mathbb{R}^N$, on a pour tout $k \geqslant 0$,*
$$\|\mathcal{T}^k X_{\mathrm{HF}}\|_N \leqslant c\sigma^k \|\mathcal{T}^k X_{\mathrm{BF}}\|_N, \tag{11.72}$$
la constante c pouvant dépendre de X mais pas de k.

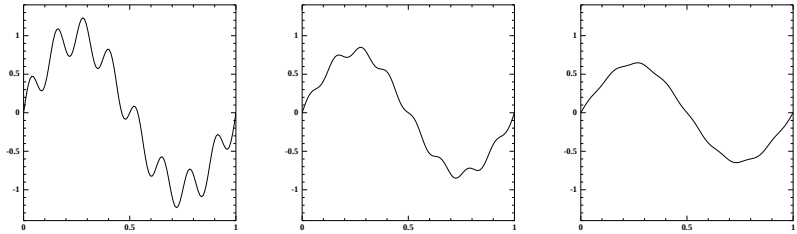


Figure 11.3 – Évolution des composantes haute-fréquence et basse-fréquence de l'erreur dans une méthode de relaxation dont la matrice d'itération $\mathcal{T}$ est un lisseur pour la matrice $\mathcal{A}$. À gauche : erreur initiale ; au milieu : erreur après une itération ; à droite : erreur après deux itérations. La composante haute-fréquence de l'erreur est amortie très rapidement alors que la composante basse-fréquence décroît plus lentement.

La signification de la définition 11.16 est claire à la lumière de la formule (11.71) : les composantes haute-fréquence de l'erreur initiale sont amorties plus rapidement que les composantes basse-fréquence. La figure 11.3 illustre ce phénomène. L'erreur initiale est prise égale à $S_2 + \frac{1}{4}S_{16}$ avec $N = 128$. Les erreurs E^1 et E^2 ont été évaluées en supposant que $\mathcal{T}S_2 = 0.8S_2$ et $\mathcal{T}S_{16} = 0.25S_{16}$.

Une classe particulière de matrices d'itération $\mathcal{T}$ satisfaisant la définition 11.16 est celle où les matrices $\mathcal{T}$ et $\mathcal{A}$ ont les mêmes vecteurs propres et où en notant $\lambda_m(\mathcal{T})$ la m-ième valeur propre de $\mathcal{T}$ telle que $\mathcal{T}S_m = \lambda_m(\mathcal{T})S_m$, on a

$$\inf_{1 \leqslant m \leqslant \frac{N}{2}} |\lambda_m(\mathcal{T})| > \sup_{\frac{N}{2} < m \leqslant N} |\lambda_m(\mathcal{T})|. \tag{11.73}$$

Dans ces conditions, en décomposant l'erreur initiale dans la base de vecteurs propres $\{S_1, \ldots, S_m\}$ selon

$$E^0 = \sum_{m=1}^{N} \eta_m S_m, \tag{11.74}$$

on obtient pour tout $k \geqslant 1$,

$$E^k = \sum_{m=1}^{N} \eta_m \lambda_m(\mathcal{T})^k S_m, \tag{11.75}$$

ce qui, compte tenu de (11.73), montre que les composantes haute-fréquence de l'erreur sont amorties plus rapidement que les composantes basse-fréquence. De plus, lorsque N est grand, on a en général

$$\sup_{1 \leqslant m \leqslant \frac{N}{2}} |\lambda_m(\mathcal{T})| \simeq 1, \tag{11.76}$$

si bien que les composantes basse-fréquence de l'erreur décroissent très lentement.

Pour conclure cette section, on présente deux exemples de méthodes de relaxation dont la matrice d'itération $\mathcal{T}$ est un lisseur pour la matrice tridiagonale $\mathcal{A}$ définie en (11.64). Dans la première méthode, les matrices $\mathcal{A}$ et $\mathcal{T}$ ont les mêmes vecteurs propres, mais ce n'est pas le cas dans la deuxième méthode.

(i) **La méthode de Richardson.** Étant donné un paramètre réel $\theta \in]0, 1[$ et un vecteur initial $U^0 \in \mathbb{R}^N$, la méthode de Richardson consiste à générer la suite $(U^k)_{k \geqslant 1}$ selon

$$U^{k+1} = U^k + \theta h(F - \mathcal{A}U^k). \tag{11.77}$$

Soit la matrice d'itération

$$\mathcal{T}_{\text{Rich}} = \mathcal{I}_N - \theta h \mathcal{A}. \tag{11.78}$$

Il est clair que les matrices $\mathcal{T}_{\text{Rich}}$ et $\mathcal{A}$ ont les mêmes vecteurs propres et que les valeurs propres de $\mathcal{T}_{\text{Rich}}$ sont telles que

$$\lambda_m(\mathcal{T}_{\text{Rich}}) = 1 - 4\theta \sin^2\left(\frac{m\pi h}{2}\right), \qquad m \in \{1, \ldots, N\}. \tag{11.79}$$

On vérifie facilement que pour $\theta = \frac{1}{4}$, la propriété (11.73) est satisfaite. On a en fait la propriété plus forte suivante :

$$1 > \lambda_1(\mathcal{T}_{\text{Rich}}) > \ldots > \lambda_N(\mathcal{T}_{\text{Rich}}) > 0, \tag{11.80}$$

avec $\lambda_1(\mathcal{T}_{\text{Rich}}) \simeq 1$ et $\lambda_N(\mathcal{T}_{\text{Rich}}) \simeq 0$ lorsque N est grand. On observera que pour $\theta = \frac{1}{2}$, on obtient la méthode de Jacobi mais que la matrice d'itération associée n'est pas un lisseur pour la matrice $\mathcal{A}$.

(ii) **La méthode de Gauß–Seidel.** La matrice d'itération associée à la méthode de Gauß–Seidel s'écrit sous la forme $\mathcal{T}_{\text{GS}} = (\mathcal{D} - \mathcal{L})^{-1}\mathcal{U}$ où les matrices $\mathcal{D}$, $\mathcal{L}$ et $\mathcal{U}$ sont associées, respectivement, à la partie diagonale, triangulaire inférieure et triangulaire supérieure de la matrice $\mathcal{A}$; voir la formule (11.12). On vérifie que les valeurs propres de la matrice $\mathcal{T}_{\text{GS}}$ sont telles que

$$\lambda_m(\mathcal{T}_{\text{GS}}) = \cos^2(m\pi h), \qquad m \in \{1, \ldots, N\}, \tag{11.81}$$

et que les vecteurs propres associés $\{S'_1, \ldots, S'_N\}$ ont pour composantes

$$S'_{m,n} = [\cos(m\pi h)]^n \sin(mn\pi h), \qquad m, n \in \{1, \ldots, N\}. \tag{11.82}$$

On peut montrer qu'il existe des réels $\{\sigma_1, \ldots, \sigma_N\}$ tels que $1 > \sigma_1 > \ldots > \sigma_N > 0$ et tels que pour $m \in \{1, \ldots, N\}$, on a pour tout $k \geqslant 1$,

$$\|\mathcal{T}_{\text{GS}}^k S_m\|_N \leqslant c_m \sigma_m^k. \tag{11.83}$$

La matrice d'itération $\mathcal{T}_{\text{GS}}$ est donc un lisseur pour la matrice $\mathcal{A}$.

11.3.2 Principe des méthodes multi-échelles

Cette section présente le principe des méthodes multi-échelles. On commence par présenter ces méthodes dans un cadre simplifié où n'interviennent que deux échelles. Plus précisément, le point de départ est constitué d'un problème à l'échelle fine

$$\mathcal{A}_1 U_1 = F_1, \tag{11.84}$$

où $\mathcal{A}_1$ est une matrice d'ordre N_1 et F_1 est un vecteur de $\mathbb{R}^{N_1}$. On suppose que la matrice $\mathcal{A}_1$ est symétrique définie positive. Le problème (11.84) correspond au système linéaire $\mathcal{A}U = F$ que l'on obtient dans le cadre de la méthode des éléments finis. L'espace $\mathbb{R}^{N_1}$ représente l'espace des échelles fines. Étant donné une approximation $U_1^* \in \mathbb{R}^{N_1}$ de la solution exacte de (11.84), on

note $E_1 = U_1 - U_1^*$ l'erreur correspondante. Le principe des méthodes multi-échelles présentées ci-dessous consiste à formuler un problème à une échelle plus grossière dont la solution peut être utilisée afin de réduire l'erreur E_1.

On considère un entier $N_2 < N_1$ et deux matrices

$$\Pi_{1\to 2} \in \mathbb{R}^{N_2,N_1} \qquad \text{et} \qquad \Pi_{2\to 1} \in \mathbb{R}^{N_1,N_2}. \tag{11.85}$$

L'espace $\mathbb{R}^{N_2}$ représente l'espace des échelles grossières. La matrice $\Pi_{1\to 2}$ permet de passer des échelles fines aux échelles grossières ; l'opérateur linéaire associé est généralement appelé *opérateur de restriction*. La matrice $\Pi_{2\to 1}$ permet de passer des échelles grossières aux échelles fines ; l'opérateur linéaire associé est généralement appelé *opérateur de prolongement*. On suppose que la matrice $\Pi_{1\to 2}$ est de *rang maximal*, c'est-à-dire que tout vecteur de $\mathbb{R}^{N_2}$ est la restriction d'au moins un vecteur de $\mathbb{R}^{N_1}$. On suppose aussi que la matrice $\Pi_{2\to 1}$ est *injective*, c'est-à-dire que si X_2 et Y_2 sont des vecteurs de $\mathbb{R}^{N_2}$ distincts, leurs prolongements $\Pi_{2\to 1}X_2$ et $\Pi_{2\to 1}Y_2$ dans $\mathbb{R}^{N_1}$ sont distincts.

On définit la matrice $\mathcal{A}_2$ d'ordre N_2 par

$$\mathcal{A}_2 = \Pi_{1\to 2}\mathcal{A}_1\Pi_{2\to 1}. \tag{11.86}$$

Par la suite, on suppose qu'il existe une constante γ telle que

$$\Pi_{1\to 2} = \gamma\Pi_{2\to 1}^T. \tag{11.87}$$

Cette propriété implique que $\mathbb{R}^{N_1} = \mathrm{Ker}(\Pi_{1\to 2}) \oplus \mathrm{Im}(\Pi_{2\to 1})$. La matrice $\mathcal{A}_1$ étant symétrique définie positive, on en déduit

$$\mathbb{R}^{N_1} = \mathrm{Ker}(\Pi_{1\to 2}\mathcal{A}_1) \oplus \mathrm{Im}(\Pi_{2\to 1}). \tag{11.88}$$

On observera que cette décomposition est orthogonale par rapport au produit scalaire $(\cdot,\cdot)_{\mathcal{A}_1}$ induit par la matrice $\mathcal{A}_1$, c'est-à-dire tel que pour tout $X_1, Y_1 \in \mathbb{R}^{N_1}$, $(X_1, Y_1)_{\mathcal{A}_1} = (\mathcal{A}_1X_1, Y_1)_{N_1}$ où $(\cdot,\cdot)_{N_1}$ désigne le produit scalaire euclidien dans $\mathbb{R}^{N_1}$. En effet, pour $X_1 \in \mathrm{Ker}(\Pi_{1\to 2}\mathcal{A}_1)$ et $Y_1 \in \mathrm{Im}(\Pi_{2\to 1})$, il existe $Y_2 \in \mathbb{R}^{N_2}$ tel que $Y_1 = \Pi_{2\to 1}Y_2$ si bien que

$$\begin{aligned}(X_1, Y_1)_{\mathcal{A}_1} &= (X_1, \Pi_{2\to 1}Y_2)_{\mathcal{A}_1} = (\mathcal{A}_1X_1, \Pi_{2\to 1}Y_2)_{N_1} \\ &= \tfrac{1}{\gamma}(\Pi_{1\to 2}\mathcal{A}_1X_1, Y_2)_{N_2} = 0.\end{aligned} \tag{11.89}$$

L'algorithme 11.6 présente la méthode de correction d'erreur à deux échelles. Même si cet algorithme n'est pas applicable directement car on ne connaît pas

Algorithme 11.6 Correction d'erreur à deux échelles

Initialisation : erreur $E_1 \in \mathbb{R}^{N_1}$ à l'échelle fine
======former le résidu à l'échelle grossière
$R_2 = \Pi_{1\to 2} R_1$ avec $R_1 = \mathcal{A}_1 E_1$
======évaluer l'erreur à l'échelle grossière
résoudre le système linéaire $\mathcal{A}_2 E_2 = R_2$
======correction d'erreur à l'échelle fine
$E_1 \leftarrow E_1 - \Pi_{2\to 1} E_2$

l'erreur E_1, celui-ci constitue le noyau des méthodes multi-échelles présentées dans cette section. On constate que l'algorithme 11.6 revient à effectuer l'opération $E_1 \leftarrow \mathcal{T}_{\text{C2E}} E_1$ avec la matrice $\mathcal{T}_{\text{C2E}}$ d'ordre N_1 telle que

$$\mathcal{T}_{\text{C2E}} = \mathcal{I}_{N_1} - \Pi_{2\to 1} \mathcal{A}_2^{-1} \Pi_{1\to 2} \mathcal{A}_1. \tag{11.90}$$

Cette matrice a deux propriétés remarquables, à savoir

$$\forall X_1 \in \text{Ker}(\Pi_{1\to 2}\mathcal{A}_1), \quad \mathcal{T}_{\text{C2E}} X_1 = X_1, \tag{11.91}$$

$$\forall Y_1 \in \text{Im}(\Pi_{2\to 1}), \quad \mathcal{T}_{\text{C2E}} Y_1 = 0. \tag{11.92}$$

En d'autres termes, l'algorithme 11.6 consiste à projeter l'erreur E_1 sur $\text{Ker}(\Pi_{1\to 2}\mathcal{A}_1)$ parallèlement à $\text{Im}(\Pi_{2\to 1})$ selon la décomposition (11.88). La propriété (11.91) est de vérification immédiate. Quant à la propriété (11.92), pour $Y_1 \in \text{Im}(\Pi_{2\to 1})$, on a $Y_1 = \Pi_{2\to 1} Y_2$ avec $Y_2 \in \mathbb{R}^{N_2}$ si bien que

$$\begin{aligned} \mathcal{T}_{\text{C2E}} Y_1 &= Y_1 - \Pi_{2\to 1} \mathcal{A}_2^{-1} \Pi_{1\to 2} \mathcal{A}_1 \Pi_{2\to 1} Y_2 \\ &= Y_1 - \Pi_{2\to 1} \mathcal{A}_2^{-1} \mathcal{A}_2 Y_2 = Y_1 - Y_1 = 0. \end{aligned} \tag{11.93}$$

Dans le cadre de l'approximation par éléments finis, une façon naturelle de mettre en place une formulation à deux échelles consiste à introduire un espace d'approximation basé sur un maillage dont les mailles sont deux fois plus grandes que celles du maillage considéré initialement. En dimension 1, on a donc $N_2 \simeq \frac{1}{2} N_1$ et, plus généralement, en dimension d, on a $N_2 \simeq 2^{-d} N_1$. Lorsque la formulation multi-échelles est obtenue à partir d'une hiérarchie de maillages, on parle de *méthode multi-grilles*.

Afin d'illustrer les notions introduites ci-dessus, on revient à l'exemple de l'approximation par éléments finis de Lagrange $\mathbb{P}_1$ du problème de Dirichlet homogène considéré dans la section 11.3.1. On suppose que l'entier N est impair, on pose $N_1 = N$ et $N_2 = \frac{1}{2}(N-1)$ et on introduit *l'opérateur de restriction* dont la matrice est

$$\Pi_{1\to 2} = \frac{1}{4}\begin{pmatrix} 1 & 2 & 1 & & & & & \\ & & 1 & 2 & 1 & & & \\ & & & & 1 & 2 & 1 & \\ & & & & & & \dots \end{pmatrix}, \tag{11.94}$$

ainsi que *l'opérateur de prolongement* dont la matrice est

$$\Pi_{2\to 1} = 2\Pi_{1\to 2}^T = \frac{1}{2}\begin{pmatrix} 1 & & & \\ 2 & & & \\ 1 & 1 & & \\ & 2 & & \\ & 1 & 1 & \\ & & 2 & \\ & & 1 & \dots \end{pmatrix}. \tag{11.95}$$

On vérifie que la matrice $\Pi_{1\to 2}$ est de rang maximal et que la matrice $\Pi_{2\to 1}$ est injective. La restriction d'un vecteur $X_1 \in \mathbb{R}^{N_1}$ est le vecteur $X_2 = \Pi_{1\to 2}X_1 \in \mathbb{R}^{N_2}$ de composantes

$$X_{2,i} = \tfrac{1}{4}(X_{1,2i-1} + 2X_{1,2i} + X_{1,2i+1}), \qquad i \in \{1,\dots,N_2\}. \tag{11.96}$$

Le prolongement d'un vecteur $X_2 \in \mathbb{R}^{N_2}$ est le vecteur $X_1 = \Pi_{2\to 1}X_2 \in \mathbb{R}^{N_1}$ de composantes $X_{1,i} = X_{2,\frac{i}{2}}$ si i est pair et $X_{1,i} = \frac{1}{2}(X_{2,\frac{i-1}{2}} + X_{2,\frac{i+1}{2}})$ si i est impair. La figure 11.4 illustre l'action des opérateurs de restriction et de prolongement. On observera qu'en désignant par $\{\varphi_{2,1},\dots,\varphi_{2,N_2}\}$ les fonctions de forme dans $P^1_{c,2h,0}$ et par $\{\varphi_{1,1},\dots,\varphi_{1,N_1}\}$ les fonctions de forme dans $P^1_{c,h,0}$, l'opérateur de prolongement est tel que si $X_1 = \Pi_{2\to 1}X_2$, on a

$$\sum_{i=1}^{N_2} X_{2,i}\varphi_{2,i} = \sum_{i=1}^{N_1} X_{1,i}\varphi_{1,i}. \tag{11.97}$$

Figure 11.4 – Illustration de l'action des opérateurs de restriction (à gauche) et de prolongement (à droite) définis par les formules (11.94) et (11.95) respectivement.

Enfin, un calcul direct montre que

$$\mathcal{A}_2 = \Pi_{1\to 2}\mathcal{A}_1\Pi_{2\to 1} = \frac{1}{2h}\,\mathrm{tridiag}(-1,2,-1). \tag{11.98}$$

La matrice $\mathcal{A}_2$ est donc exactement la matrice de rigidité obtenue en discrétisant le problème de Dirichlet homogène par l'élément fini de Lagrange $\mathbb{P}_1$ sur un maillage uniforme de pas $2h$.

On revient au cas général. Tous les éléments du puzzle sont maintenant en place pour présenter l'algorithme itératif à deux échelles pour la résolution du problème (11.84). On rappelle qu'on dispose de deux décompositions en somme directe de l'espace des échelles fines, à savoir

$$\mathbb{R}^{N_1} = Z_{\mathrm{BF}} \oplus Z_{\mathrm{HF}}, \tag{11.99}$$

$$\mathbb{R}^{N_1} = \mathrm{Ker}(\Pi_{1\to 2}\mathcal{A}_1) \oplus \mathrm{Im}(\Pi_{2\to 1}). \tag{11.100}$$

La première décomposition résulte du spectre de la matrice $\mathcal{A}_1$ et la deuxième du cadre multi-échelles. On rappelle également que ces deux décompositions sont orthogonales par rapport au produit scalaire associé à la matrice $\mathcal{A}_1$. On note $\|\cdot\|_{\mathcal{A}_1}$ la norme induite par ce produit scalaire. On fait les deux hypothèses suivantes.

(i) Il existe une constante $c \ll 1$ telle que pour tout $X \in Z_{\mathrm{BF}}$, en décomposant X sous la forme $X = X_{\mathrm{Ker}} + X_{\mathrm{Im}}$ avec $X_{\mathrm{Ker}} \in \mathrm{Ker}(\Pi_{1\to 2}\mathcal{A}_1)$ et $X_{\mathrm{Im}} \in \mathrm{Im}(\Pi_{2\to 1})$, on a

$$\|X_{\mathrm{Ker}}\|_{\mathcal{A}_1} \leqslant c\,\|X_{\mathrm{Im}}\|_{\mathcal{A}_1}. \tag{11.101}$$

Cette hypothèse signifie que l'opérateur de prolongement excite peu les hautes fréquences, ou, en d'autres termes, que son image est proche du

sous-espace Z_{BF}. Pour l'exemple uni-dimensionnel présenté ci-dessus, on a pour $m \in \{1, \ldots, N_2\}$,

$$\Pi_{2\to 1} S_{2,m} = (1 - \epsilon_m) S_{1,m} + \epsilon_m S_{1,N_1-m}, \tag{11.102}$$

avec $\epsilon_m = \sin^2(2m\pi h)$ et où $\{S_{2,1}, \ldots, S_{2,N_2}\}$ et $\{S_{1,1}, \ldots, S_{1,N_1}\}$ désignent les vecteurs propres associés aux matrices $\mathcal{A}_2$ et $\mathcal{A}_1$ respectivement. On observe que si $m \ll N_2$, on a $\epsilon_m \ll 1$ si bien que la composante haute-fréquence de $\Pi_{2\to 1} S_{2,m}$ est pratiquement nulle.

(ii) On dispose d'une matrice d'itération $\mathcal{T}_1 \in \mathbb{R}^{N_1,N_1}$ qui est un lisseur pour la matrice $\mathcal{A}_1$. La matrice $\mathcal{T}_1$ amortissant rapidement les composantes haute-fréquence, on peut supposer qu'on peut choisir un entier n_1 tel que pour tout $X_1 \in \mathbb{R}^{N_1}$, le vecteur $\mathcal{T}_1^{n_1} X_1$ est pratiquement dans Z_{BF}.

L'algorithme 11.7 présente la méthode itérative à deux échelles pour la résolution du système linéaire (11.84). À l'échelle fine, on applique une méthode de relaxation au système linéaire $\mathcal{A}_1 U_1 = F_1$, ce qui conduit à une récurrence de la forme

$$U_1^{k+1} = \mathcal{T}_1 U_1^k + G_1, \tag{11.103}$$

où $\mathcal{A}_1 = \mathcal{P}_1 - \mathcal{Z}_1$, $\mathcal{T}_1 = \mathcal{P}_1^{-1}\mathcal{Z}_1$ et $G_1 = \mathcal{P}_1^{-1} F_1$.

La figure 11.5 illustre le déroulement de l'algorithme 11.7. On a représenté les deux décompositions $\mathcal{A}_1$-orthogonales de $\mathbb{R}^{N_1}$ définies en (11.99) et (11.100). L'hypothèse (i) ci-dessus se traduit géométriquement par le fait que l'angle entre les sous-espaces Z_{BF} et $\mathrm{Im}(\Pi_{2\to 1})$ est petit. Étant donné un vecteur $U_1^0 \in \mathbb{R}^{N_1}$, l'erreur initiale E_1^0 est représentée par un point du plan. Après avoir effectué n_1 pas de la méthode de relaxation avec la matrice d'itération $\mathcal{T}_1$, l'hypothèse (ii) ci-dessus permet d'affirmer que l'erreur E_1^{0,n_1} est pratiquement dans Z_{BF}. Dans la mesure où les composantes basse-fréquence de l'erreur ont été peu amorties, l'erreur E_1^{0,n_1} est proche de la projection de E_1^0 sur Z_{BF} parallèlement à Z_{HF}. Puis, l'étape de correction à l'échelle grossière, qui reprend l'algorithme 11.6, permet d'affirmer que l'erreur E_1^1 est la projection de E_1^{0,n_1} sur $\mathrm{Ker}(\Pi_{1\to 2}\mathcal{A}_1)$ parallèlement à $\mathrm{Im}(\Pi_{2\to 1})$. L'algorithme 11.7 se poursuit en zigzag comme illustré sur la figure 11.5. Son efficacité est d'autant plus grande que l'angle entre les sous-espaces Z_{BF} et $\mathrm{Im}(\Pi_{2\to 1})$ est petit.

L'étape la plus coûteuse de l'algorithme 11.7 est la résolution du système linéaire $\mathcal{A}_2 E_2 = R_2$ à l'échelle grossière ; ce coût reste inférieur à celui de la

Algorithme 11.7 Algorithme itératif à deux échelles

choisir $U_1^0 \in \mathbb{R}^{N_1}$ et poser $R_1^0 = F_1 - \mathcal{A}_1 U_1^0$
choisir une tolérance `tol`
poser $k = 0$
while $\|R_1^k\|_{N_1} >$ `tol` **do**
  ======n_1 pas de relaxation à l'échelle fine
  poser $U_1^{k,0} = U_1^k$
  for $m \in \{1, \ldots, n_1\}$ **do**
    $U_1^{k,m} = \mathcal{T}_1 U_1^{k,m-1} + G_1$
  end for
  ======correction à l'échelle grossière
  $R_2 = \Pi_{1\to 2}(F_1 - \mathcal{A}_1 U_1^{k,n_1})$
  résoudre le système linéaire $\mathcal{A}_2 E_2 = R_2$
  $U_1^{k+1} = U_1^{k,n_1} + \Pi_{2\to 1} E_2$
end while

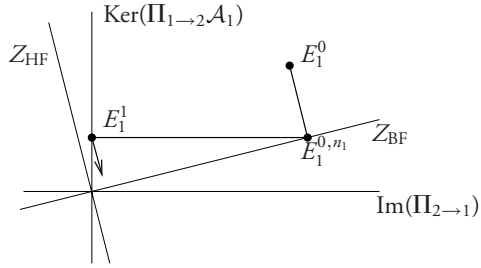


Figure 11.5 – Illustration du déroulement de l'algorithme 11.7.

résolution du système linéaire (11.84) à l'échelle fine puisque $N_2 < N_1$. Toutefois, ce coût reste encore relativement élevé et l'algorithme 11.7 ne prend toute son efficacité que lorsqu'il est mis en œuvre sur une hiérarchie de problèmes.

Étant donné un entier $M > 0$ et une famille d'entiers $\{N_1, \ldots, N_M\}$ telle que $N_1 > \ldots > N_M$, on considère une famille d'opérateurs de restriction

associés à des matrices $\{\Pi_{m\to m+1}\}_{1\leqslant m\leqslant M-1}$ et une famille d'opérateurs de prolongement associés à des matrices $\{\Pi_{m+1\to m}\}_{1\leqslant m\leqslant M-1}$. On suppose que pour tout $m \in \{1,\ldots,M-1\}$, les hypothèses formulées ci-dessus pour le couple $(\Pi_{1\to 2},\Pi_{2\to 1})$ sont valables pour le couple $(\Pi_{m\to m+1},\Pi_{m+1\to m})$. Étant donné une matrice $\mathcal{A}_1$ d'ordre N_1 (qui correspond à la matrice obtenue par la méthode des éléments finis), on construit une famille de matrices $\{\mathcal{A}_m\}_{1\leqslant m\leqslant M}$ en posant pour $m \in \{1,\ldots,M-1\}$,

$$\mathcal{A}_{m+1} = \Pi_{m\to m+1}\mathcal{A}_m\Pi_{m+1\to m}. \tag{11.104}$$

Algorithme 11.8 Correction d'erreur multi-échelles

```
=======Algorithme Multi−Echelles(m, A_m, F_m, X_m)
Input : m ∈ {1,...,M}, A_m ∈ R^{N_m,N_m}, F_m ∈ R^{N_m} et X_m ∈ R^{N_m}
if m < M then
   X_m ← Relax(n, m, A_m, F_m, X_m)
   F_{m+1} ← Π_{m→m+1}(F_m − A_m X_m)
   for p ∈ {1,...,p_cycl} do
      Y_{m+1} ← Multi−Echelles(m + 1, A_{m+1}, F_{m+1}, 0)
   end for
   X_m ← X_m + Π_{m+1→m} Y_{m+1}
else
   résoudre le système linéaire A_M X_M = F_M
end if
Output : X_m
```

L'algorithme 11.8 présente la méthode de correction multi-échelles. La notation

$$X_m \leftarrow \texttt{Relax}(n, m, \mathcal{A}_m, F'_m, X_m), \tag{11.105}$$

signifie qu'étant donné un entier $n \geqslant 1$, un entier $m \in \{1,\ldots,M\}$, une matrice $\mathcal{A}_m$ d'ordre N_m et deux vecteurs $F'_m \in \mathbb{R}^{N_m}$ et $X_m \in \mathbb{R}^{N_m}$, on effectue n pas d'une méthode de relaxation sur le système linéaire $\mathcal{A}_m U'_m = F'_m$ en initialisant les itérations avec le vecteur X_m. À l'issue des itérations, le vecteur X_m contient la solution approchée du système linéaire en question. On observera que l'algorithme 11.8 est formulé de manière récursive. Le paramètre entier

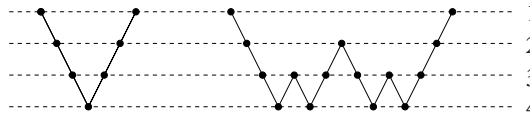


Figure 11.6 – Déroulement de l'algorithme 11.8 dans le cas $M = 4$: V-cycle (à gauche) et W-cycle (à droite) ; en ordonnée, l'indice $m \in \{1, 2, 3, 4\}$.

p_{cycl} détermine la forme du cycle multi-échelles. En pratique, on considère les valeurs $p_{\text{cycl}} = 1$, qui donne lieu au V-cycle, et $p_{\text{cycl}} = 2$, qui donne lieu au W-cycle. La forme de ces cycles est illustrée sur la figure 11.6. À l'issue du V-cycle ou du W-cycle, la norme du résidu sur l'échelle fine est évaluée et un test de convergence est effectué. Les cycles se poursuivent jusqu'à ce que la norme du résidu soit inférieure à un seuil de tolérance fixé par le numéricien. L'algorithme 11.8 remplace donc le contenu de la boucle **while** dans l'algorithme 11.7.

11.4 Compléments

- **Méthode des directions alternées (ADI).** La méthode ADI (de l'anglais, Alternating Direction Iterative Method) est basée sur la décomposition

$$\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2. \tag{11.106}$$

On suppose que la matrice $\mathcal{A}$ est définie positive, que les matrices $\mathcal{A}_1$ et $\mathcal{A}_2$ sont positives et que l'une d'entre elles est définie positive. La terminologie adoptée provient du fait que les méthodes ADI ont été, à l'origine, développées pour des matrices $\mathcal{A}$ issues de la discrétisation d'équations aux dérivées partielles en dimension deux à l'aide de la méthode des différences finies. La matrice $\mathcal{A}_1$ représente les contributions des dérivées selon la première direction spatiale et la matrice $\mathcal{A}_2$ la contribution des dérivées selon la deuxième direction spatiale.

Étant donné un vecteur initial $U^0 \in \mathbb{R}^N$, on génère deux suites de vecteurs, $(U^{k+\frac{1}{2}})_{k\geqslant 0}$ et $(U^k)_{k\geqslant 1}$, de la manière suivante :

$$\gamma(U^{k+\frac{1}{2}} - U^k) + \mathcal{A}_1 U^{k+\frac{1}{2}} = F - \mathcal{A}_2 U^k, \tag{11.107}$$

$$\gamma(U^{k+1} - U^{k+\frac{1}{2}}) + \mathcal{A}_2 U^{k+1} = \mathcal{A}_2 U^k + \rho\gamma(U^{k+\frac{1}{2}} - U^k). \tag{11.108}$$

La méthode ADI fait intervenir deux paramètres réels, γ et ρ. On suppose que $\gamma > 0$ et que $\rho \neq -1$. Pour $\rho = 0$, on obtient le *schéma de Douglas–Rachford* et pour $\rho = 1$, le *schéma de Peaceman–Rachford*. On observera que les matrices $(\gamma\mathcal{I}_N + \mathcal{A}_1)$ et $(\gamma\mathcal{I}_N + \mathcal{A}_2)$ sont inversibles car définies positives ; les suites $(U^{k+\frac{1}{2}})_{k\geqslant 0}$ et $(U^k)_{k\geqslant 1}$ sont donc bien définies. De plus, le schéma (11.107)–(11.108) est consistant avec le système linéaire $\mathcal{A}U = F$ puisque si les suites $(U^{k+\frac{1}{2}})_{k\geqslant 0}$ et $(U^k)_{k\geqslant 1}$ convergent vers des limites notées V et W, respectivement, on déduit aisément de (11.107)–(11.108) que $V = W = \mathcal{A}^{-1}F = U$.
Le schéma (11.107)–(11.108) rentre dans la catégorie des méthodes de relaxation étudiées dans la section 11.1. En effet, en éliminant $U^{k+\frac{1}{2}}$ dans (11.107)–(11.108), on montre facilement que U^{k+1} se déduit de U^k à l'aide d'une relation de la forme (11.3) avec la matrice d'itération

$$\mathcal{T} = (\gamma\mathcal{I}_N + \mathcal{A}_2)^{-1}(\gamma\mathcal{I}_N + \mathcal{A}_1)^{-1}(\gamma^2\mathcal{I}_N - \rho\gamma\mathcal{A} + \mathcal{A}_1\mathcal{A}_2). \quad (11.109)$$

Pour le schéma de Peaceman–Rachford, on peut montrer assez simplement que $\rho(\mathcal{T}) < 1$; voir, par exemple, Quarteroni et Valli [**61**, p. 38].

- **La méthode Bi-CGStab.** La méthode Bi-CGStab (de l'anglais, Bi-Conjugate Gradient Stabilized Method) a été proposée par van der Vorst en 1992 [**74**]. Cette méthode offre une alternative intéressante à la méthode GMRes pour la résolution itérative des systèmes linéaires dont la matrice n'est pas symétrique (ni définie positive). Contrairement à la méthode GMRes qui assure une propriété d'optimalité de l'itérée U^k sur un sous-espace affine de dimension k mais au prix d'une augmentation linéaire du coût des calculs au fil des itérations, la méthode Bi-CGStab renonce à garantir une propriété d'optimalité mais propose des récurrences d'ordre 1. La méthode Bi-CGStab est présentée dans l'algorithme 11.9. On constate que le coût par itération est de deux produits matrice–vecteur et de quatre produits scalaires (pour $\overline{R}^0 = R^0$). Lorsque la matrice $\mathcal{A}$ est creuse, le coût par itération est donc proportionnel à N. En raison de l'absence de propriété d'optimalité, il n'est pas certain que l'algorithme 11.9 converge (on peut facilement construire des contre-exemples). L'expérience montre qu'en pratique, les méthodes Bi-CGStab et GMRes se comportent de manière comparable lorsqu'elles sont appliquées à la résolution itérative

de grands systèmes linéaires provenant de l'approximation par éléments finis de problèmes modèles.

Algorithme 11.9 Méthode Bi-CGStab

choisir $U^0 \in \mathbb{R}^N$, poser $R^0 = F - \mathcal{A}U^0$ et $P^0 = R^0$
choisir $\overline{R}^0 \in \mathbb{R}^N$ (par exemple, $\overline{R}^0 = R^0$)
choisir une tolérance `tol`
poser $k = 0$
while $\|R^k\|_N >$ `tol` **do**
  calculer le vecteur $Z^k = \mathcal{A}P^k$
  $\alpha^k = (R^k, \overline{R}^0)_N / (Z^k, \overline{R}^0)_N$
  $S^k = R^k - \alpha^k Z^k$
  calculer le vecteur $T^k = \mathcal{A}S^k$
  $\omega^k = (T^k, S^k)_N / (T^k, T^k)_N$
  $U^{k+1} = U^k + \alpha^k P^k + \omega^k S^k$
  $R^{k+1} = S^k - \omega^k T^k$
  $\beta^k = (\alpha^k/\omega^k) \times (R^{k+1}, \overline{R}^0)_N / (R^k, \overline{R}^0)_N$
  $P^{k+1} = R^{k+1} + \beta^k (P^k - \omega^k Z^k)$
  $k \leftarrow k + 1$
end while

- **Autres solveurs itératifs.** Les dernières décennies ont connu un développement très important des solveurs itératifs et des techniques de préconditionnement pour les grands systèmes linéaires creux, tels ceux issus de la méthode des éléments finis. Une liste relativement exhaustive des solveurs disponibles sur Internet se trouve à l'adresse suivante : `http://www.netlib.org/utk/people/JackDongarra/la-sw.html`
  Un autre axe de recherches qui a connu un développement considérable est celui de l'implémentation de solveurs itératifs et de préconditionneurs sur des ordinateurs à architecture parallèle. La *parallélisation* d'un logiciel d'éléments finis pouvant être assez technique, des équipes de recherche ont développé des outils de parallélisation qui peuvent être utilisés en boîte noire.

Parmi les exemples de librairies disponibles sur Internet, on peut citer la librairie Aztec

`http://www.cs.sandia.gov/CRF/aztec1.html`

et la boîte à outils PETSc

`http://www-unix.mcs.anl.gov/petsc/petsc-2/`

12 • PROGRAMMER LES ÉLÉMENTS FINIS

L'objectif de ce chapitre est de fournir quelques éléments relatifs à l'implémentation de la méthode des éléments finis afin que le lecteur puisse se familiariser avec l'organisation générale d'un logiciel d'éléments finis. Il s'agit d'une part de présenter les structures de données relatives au maillage et aux quadratures et d'autre part de décrire brièvement les opérations relatives à l'assemblage et au stockage de la matrice de rigidité. Enfin, on présente le fonctionnement général d'un mailleur et quelques techniques permettant d'implémenter les conditions aux limites de Dirichlet non-homogènes.

12.1 Structure de données pour le maillage

Un maillage $\mathcal{T}_h$ est un nuage de points qui sont numérotés et connectés entre eux. On rappelle que ces points, appelés *nœuds géométriques*, sont générés à partir d'un élément fini géométrique de référence $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$; voir la section 3.3.1. On suppose que $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$ est un élément fini de Lagrange. On désigne par $\{\widehat{g}_1, \ldots, \widehat{g}_{n_{\text{géo}}}\}$ les nœuds de $\widehat{K}$, où $n_{\text{géo}}$ est le nombre de nœuds, et par $\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{n_{\text{géo}}}\}$ les fonctions de forme de l'élément fini géométrique.

On désigne par $N_{\text{géo}}$ le nombre de nœuds géométriques dans le maillage. Par exemple, $N_{\text{géo}} = N_{\text{so}}$ si l'élément fini géométrique de référence est de degré 1. Les nœuds géométriques sont numérotés de 1 à $N_{\text{géo}}$; cette numérotation est dite *globale* car elle concerne l'ensemble des nœuds géométriques

indépendamment des cellules du maillage auxquelles ils appartiennent. Les nœuds géométriques sont identifiés par leurs coordonnées ; celles-ci sont stockées dans un tableau de taille $d \times N_{\text{géo}}$ où d est la dimension d'espace. Ce tableau est noté

$$\texttt{coord}(1\!:\!d, 1\!:\!N_{\text{géo}}). \tag{12.1}$$

Pour $k \in \{1, \ldots, d\}$ et $n \in \{1, \ldots, N_{\text{géo}}\}$, $\texttt{coord}(k, n)$ est la k-ième coordonnée du n-ième nœud géométrique.

Les nœuds géométriques sont regroupés en éléments (ou mailles) par le biais d'un tableau de connectivité géométrique. On désigne par N_{ma} le nombre de mailles. Les mailles sont numérotées de 1 à N_{ma} ; cette numérotation est *globale*. On rappelle qu'une maille $K \in \mathcal{T}_h$ est l'image de $\widehat{K}$ par une transformation géométrique $T_K : \widehat{K} \to K$. L'image par T_K des nœuds géométriques de $\widehat{K}$ définit localement les nœuds géométriques dans K. La numérotation des nœuds dans K est celle induite par la numérotation dans $\widehat{K}$. Cette numérotation est *locale* car elle concerne uniquement une maille. Le tableau de connectivité géométrique est de taille $n_{\text{géo}} \times N_{\text{ma}}$ et est noté

$$\texttt{connect\_géo}(1\!:\!n_{\text{géo}}, 1\!:\!N_{\text{ma}}). \tag{12.2}$$

Pour $n \in \{1, \ldots, n_{\text{géo}}\}$ et $m \in \{1, \ldots, N_{\text{ma}}\}$, $\texttt{connect\_géo}(n, m)$ est le numéro global du n-ième nœud local dans la m-ième maille. La figure 12.1 présente un exemple de numérotation locale et de numérotation globale des nœuds dans un maillage composé de triangles et où chaque triangle contient trois nœuds géométriques ($n_{\text{géo}} = 3$) qui sont ses trois sommets. Trois mailles, de numéros 67, 148 et 472, sont représentées sur la figure. Le tableau de connectivité géométrique prend les valeurs suivantes :

$$\texttt{connect\_géo}(1, 67) = 13,\ \texttt{connect\_géo}(1, 148) = 77,\ \texttt{connect\_géo}(1, 472) = 250,$$
$$\texttt{connect\_géo}(2, 67) = 37,\ \texttt{connect\_géo}(2, 148) = 53,\ \texttt{connect\_géo}(2, 472) = 37,$$
$$\texttt{connect\_géo}(3, 67) = 53,\ \texttt{connect\_géo}(3, 148) = 37,\ \texttt{connect\_géo}(3, 472) = 13.$$

Dans certaines situations, il est commode de pouvoir identifier les éléments qui partagent une face avec une maille donnée. Pour cela, on introduit un tableau de voisinage. Par exemple, dans un maillage composé de simplexes, chaque maille partage une face avec au plus $(d + 1)$ mailles voisines (ce nombre est effectivement égal à $(d + 1)$ si la maille en question ne contient pas de face située sur le bord). On numérote les voisins localement. Une façon

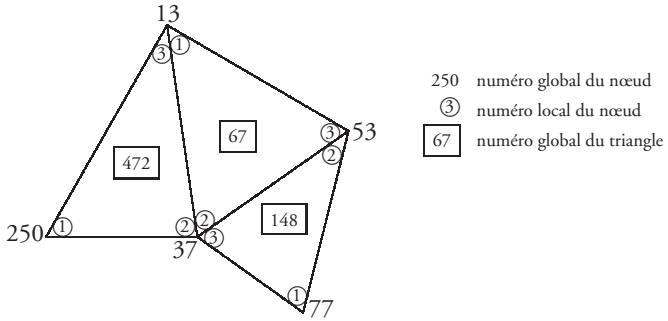


Figure 12.1 – Numérotations locale et globale des nœuds dans un maillage composé de triangles.

simple de procéder lorsque chaque simplexe contient $(d+1)$ nœuds géométriques qui sont ses $(d+1)$ sommets, consiste à affecter à chaque voisin le numéro local du sommet opposé à la face en commun. Le tableau de voisinage est de taille $(d+1) \times N_{\text{ma}}$ et est noté

$$\texttt{vois}(1\!:\!d\!+\!1, 1\!:\!N_{\text{ma}}). \tag{12.3}$$

Pour $n \in \{1, \ldots, d+1\}$ et $m \in \{1, \ldots, N_{\text{ma}}\}$, $\texttt{vois}(n, m)$ est le numéro global du n-ième voisin de la m-ième maille. Si la face opposée au n-ième nœud de cette maille est située sur le bord du domaine, on peut poser conventionnellement $\texttt{vois}(n, m) = 0$. La figure 12.2 présente un exemple de numérotation locale et de numérotation globale des voisins d'un triangle. Le tableau de voisinage prend les valeurs suivantes :

$$\texttt{vois}(1, 326) = 450, \qquad \texttt{vois}(2, 326) = 128, \qquad \texttt{vois}(3, 326) = 29.$$

Afin de pouvoir imposer des conditions aux limites, il est commode de disposer d'un tableau de connectivité à la frontière. Si l'intersection d'une maille avec la frontière du domaine est une variété de dimension $(d-1)$, cette intersection est appelée une face de frontière. Les faces de frontière sont numérotées de 1 à N_{fa}^{∂}. Pour simplifier, on suppose que l'élément fini géométrique de référence a été choisi de sorte que le nombre de nœuds sur chaque face de $\widehat{K}$ est le même ; on note $n_{\text{géo}}^{\partial}$ ce nombre. Par conséquent, toutes les faces de

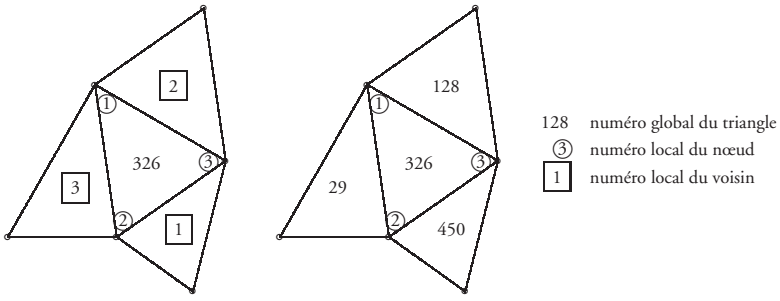


Figure 12.2 – Voisinage du triangle de numéro 326.

frontière contiennent $n^{\partial}_{\text{géo}}$ nœuds. Le tableau de connectivité à la frontière est de taille $n^{\partial}_{\text{géo}} \times N^{\partial}_{\text{fa}}$ et est noté

$$\texttt{connect\_g\_front}(1{:}n^{\partial}_{\text{géo}}, 1{:}N^{\partial}_{\text{fa}}). \tag{12.4}$$

Pour $n \in \{1, \ldots, n^{\partial}_{\text{géo}}\}$ et pour $m \in \{1, \ldots, N^{\partial}_{\text{fa}}\}$, `connect_g_front`$(n, m)$ est l'indice global du n-ième nœud local dans la m-ième face de frontière. Dans certaines situations, on souhaite imposer plusieurs types de conditions aux limites. La frontière du domaine est donc partitionnée en $I^{\partial\Omega}$ composantes sous la forme $\partial\Omega = \bigcup_{i=1}^{I^{\partial\Omega}} \partial\Omega_i$ et une condition aux limites fixée (par exemple, de type Dirichlet, Neumann ou Robin) est imposée sur chaque $\partial\Omega_i$. Afin de repérer le type de condition aux limites que l'on souhaite imposer, on peut introduire un tableau de taille N^{∂}_{fa} noté

$$\texttt{i\_cond\_lim}(1{:}N^{\partial}_{\text{fa}}). \tag{12.5}$$

Pour $m \in \{1, \ldots, N^{\partial}_{\text{fa}}\}$, `i_cond_lim`$(m)$ indique le type de condition aux limites qui est imposée sur la m-ième face de frontière.

Enfin, certains problèmes font intervenir une décomposition du domaine en I^{Ω} composantes sous la forme $\Omega = \bigcup_{i=1}^{I^{\Omega}} \Omega_i$. Sur chaque sous-domaine, certains paramètres du modèle peuvent prendre des valeurs différentes ; on peut également considérer des modèles différents par sous-domaine. Afin d'identifier quelle est la valeur du paramètre ou quel modèle doit être considéré sur

chaque maille, on peut introduire un tableau de taille N_{ma} noté

$$\texttt{i\_dom}(1{:}\,N_{\text{ma}}). \tag{12.6}$$

Pour $m \in \{1, \ldots, N_{\text{ma}}\}$, $\texttt{i\_dom}(m)$ indique la valeur du paramètre ou le modèle qui doit être considéré sur la m-ième maille.

12.2 Structure de données pour les quadratures

L'utilisation de quadratures est pratiquement incontournable dans la méthode des éléments finis puisque la solution discrète s'obtient en résolvant un système linéaire dont la matrice et le membre de droite s'évaluent à partir d'intégrales. Le principe général des quadratures est décrit dans le chapitre 9. Dans le cadre de la méthode des éléments finis, on rappelle qu'on choisit d'abord une quadrature sur l'élément de référence $\widehat{K}$, c'est-à-dire un ensemble de points de Gauß $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_{l_{\text{q}}}\}$ et de poids associés $\{\widehat{\omega}_1, \ldots, \widehat{\omega}_{l_{\text{q}}}\}$, puis on génère une quadrature sur un maille $K \in \mathcal{T}_h$ par l'intermédiaire de la transformation géométrique T_K ; voir les formules (9.5) et (9.6).

On rappelle que $\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{n_{\text{géo}}}\}$ désignent les fonctions de forme de l'élément fini géométrique $\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$. On stocke les valeurs des fonctions de forme aux points de Gauß $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_{l_{\text{q}}}\}$ dans un tableau de taille $n_{\text{géo}} \times l_{\text{q}}$ noté

$$\texttt{psi}(1{:}\,n_{\text{géo}}, 1{:}\,l_{\text{q}}). \tag{12.7}$$

Pour $n \in \{1, \ldots, n_{\text{géo}}\}$ et $l \in \{1, \ldots, l_{\text{q}}\}$, on a

$$\texttt{psi}(n, l) = \widehat{\psi}_n(\widehat{\xi}_l). \tag{12.8}$$

De même, on stocke les valeurs des dérivées des fonctions de forme aux points de Gauß dans un tableau de taille $d \times n_{\text{géo}} \times l_{\text{q}}$ noté

$$\texttt{dpsi\_dx}(1{:}\,d, 1{:}\,n_{\text{géo}}, 1{:}\,l_{\text{q}}). \tag{12.9}$$

Pour $k \in \{1, \ldots, d\}$, $n \in \{1, \ldots, n_{\text{géo}}\}$ et $l \in \{1, \ldots, l_{\text{q}}\}$, on a

$$\texttt{dpsi\_dx}(k, n, l) = \frac{\partial \widehat{\psi}_n}{\partial \widehat{x}_k}(\widehat{\xi}_l). \tag{12.10}$$

Lorsque le maillage est affine, les dérivées des fonctions de forme sont constantes sur $\widehat{K}$ et ne dépendent donc pas du point de Gauß considéré. Par conséquent, la taille du tableau `dpsi_dx` peut être réduite à $d \times n_{\text{géo}}$.

Pour $k \in \{1, \ldots, d\}$, $l \in \{1, \ldots, l_{\text{q}}\}$ et $m \in \{1, \ldots, N_{\text{ma}}\}$, on déduit de (3.23) et (9.6) que la k-ième coordonnée du l-ième point de Gauß dans la m-ième maille s'écrit sous la forme

$$\left(\xi_{K_m,l}\right)_k = \sum_{n=1}^{n_{\text{géo}}} \texttt{coord}(k, \texttt{connect\_géo}(n, m))\ \texttt{psi}(n, l). \tag{12.11}$$

Cette formule permet de positionner les points de Gauß $\{\xi_{K,1}, \ldots, \xi_{K,l_{\text{q}}}\}$ définis en (9.6). Par ailleurs, on stocke les valeurs des poids $\{\omega_{K,1}, \ldots, \omega_{K,l_{\text{q}}}\}$ définis en (9.5) dans un tableau de taille $l_{\text{q}} \times N_{\text{ma}}$ noté

$$\texttt{poids}(1\colon l_{\text{q}}, 1\colon N_{\text{ma}}). \tag{12.12}$$

Pour $l \in \{1, \ldots, l_{\text{q}}\}$ et $m \in \{1, \ldots, N_{\text{ma}}\}$, `poids`$(l, m)$ est le poids du l-ième point de Gauß dans la m-ième maille. On a

$$\texttt{poids}(l, m) = \widehat{\omega}_l \det\left(J_{K_m}(\widehat{\xi}_l)\right), \tag{12.13}$$

où $J_{K_m} \in \mathbb{R}^{d,d}$ est la matrice jacobienne de la transformation géométrique T_{K_m} associée à la m-ième maille. Les coefficients de la matrice jacobienne s'évaluent sous la forme suivante :

$$\left(J_{K_m}(\widehat{\xi}_l)\right)_{k_1,k_2} = \sum_{n=1}^{n_{\text{géo}}} \texttt{coord}(k_1, \texttt{connect\_géo}(n, m))\ \texttt{dpsi\_dx}(k_2, n, l). \tag{12.14}$$

Lorsque le maillage est affine, la matrice J_{K_m} est constante sur $\widehat{K}$ et ne dépend donc pas du point de Gauß considéré. Son déterminant est égal au rapport de la mesure de K_m sur celle de $\widehat{K}$. Afin de réduire la taille des tableaux à stocker, on peut ne pas stocker le tableau `poids` mais réévaluer les poids $\{\omega_{K,1}, \ldots, \omega_{K,l_{\text{q}}}\}$ à partir de la formule (9.5) chaque fois que cela s'avère nécessaire.

On considère maintenant un espace d'approximation V_h construit à partir du maillage $\mathcal{T}_h$ et d'un élément fini de référence $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$. On désigne par

$\{\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_f}\}$ les fonctions de forme de l'élément fini $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$, où n_f désigne le nombre de degrés de liberté de l'élément fini de référence. Il est commode de stocker les valeurs des fonctions $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_f}\}$ aux points de Gauß $\{\widehat{\xi}_1, \ldots, \widehat{\xi}_{l_q}\}$ dans un tableau de taille $n_f \times l_q$ noté

$$\texttt{theta}(1\!:\!n_f, 1\!:\!l_q). \tag{12.15}$$

Pour $n \in \{1, \ldots, n_f\}$ et $l \in \{1, \ldots, l_q\}$, on a

$$\texttt{theta}(n, l) = \widehat{\theta}_n(\widehat{\xi}_l). \tag{12.16}$$

De même, on stocke les valeurs des dérivées de ces fonctions de forme aux points de Gauß dans un tableau de taille $d \times n_f \times l_q$ noté

$$\texttt{dtheta\_dx}(1\!:\!d, 1\!:\!n_f, 1\!:\!l_q). \tag{12.17}$$

Pour $k \in \{1, \ldots, d\}$, $n \in \{1, \ldots, n_f\}$ et $l \in \{1, \ldots, l_q\}$, on a

$$\texttt{dtheta\_dx}(k, n, l) = \frac{\partial \widehat{\theta}_n}{\partial \widehat{x}_k}(\widehat{\xi}_l). \tag{12.18}$$

On s'intéresse maintenant aux fonctions de forme dans V_h et à leurs dérivées. On désigne par N le nombre de fonctions de forme dans V_h (N est égal à la dimension de V_h). Les fonctions de forme sont numérotées de 1 à N et notées $\{\varphi_1, \ldots, \varphi_N\}$; cette numérotation est *globale*. La restriction d'une fonction de forme dans V_h à une maille $K \in \mathcal{T}_h$ est, par construction, une des fonctions de forme $\{\theta_{K,1}, \ldots, \theta_{K,n_f}\}$. On suppose que ces fonctions sont générées à partir des fonctions de forme de l'élément fini de référence selon la formule

$$\theta_{K,n} = \widehat{\theta}_n \circ T_K^{-1}, \qquad n \in \{1, \ldots, n_f\}. \tag{12.19}$$

Afin de faire le lien entre fonctions de forme sur K et les fonctions de forme dans V_h, on introduit un tableau de taille $n_f \times N_{ma}$, appelé tableau de connectivité des fonctions de forme et noté

$$\texttt{connect\_forme}(1\!:\!n_f, 1\!:\!N_{ma}). \tag{12.20}$$

Pour $n \in \{1, \ldots, n_f\}$ et $m \in \{1, \ldots, N_{ma}\}$, $\texttt{connect\_forme}(n, m)$ est le numéro de la fonction de forme dans V_h dont la restriction à la m-ième maille

est la n-ième fonction de forme sur cette maille. On a donc

$$\forall x \in K_m, \qquad \varphi_{\texttt{connect\_forme}(n,m)}(x) = \widehat{\theta}_n \circ T_{K_m}^{-1}(x), \tag{12.21}$$

si bien que pour $l \in \{1, \ldots, l_q\}$, on obtient

$$\varphi_{\texttt{connect\_forme}(n,m)}(\xi_{K_m,l}) = \texttt{theta}(n,l). \tag{12.22}$$

Afin d'évaluer les dérivées des fonctions de forme dans V_h, on utilise la formule (12.19) et la règle de la dérivation composée. Pour $k_1 \in \{1, \ldots, d\}$, on obtient

$$\begin{aligned} \frac{\partial \varphi_{\texttt{connect\_forme}(n,m)}}{\partial x_{k_1}}(\xi_{K_m,l}) &= \frac{\partial \widehat{\theta}_n \circ T_{K_m}^{-1}}{\partial x_{k_1}}(\xi_{K_m,l}) \\ &= \sum_{k_2=1}^{d} \frac{\partial \widehat{\theta}_n}{\partial \widehat{x}_{k_2}}(\widehat{\xi}_l) \left(\left[J_{K_m}(\widehat{\xi}_l) \right]^{-1} \right)_{k_2,k_1}. \end{aligned} \tag{12.23}$$

Une première stratégie consiste à stocker ces dérivées dans un tableau de taille $d \times n_f \times l_q \times N_{ma}$ noté

$$\texttt{dphi\_dx}(1\!:\!d, 1\!:\!n_f, 1\!:\!l_q, 1\!:\!N_{ma}). \tag{12.24}$$

Pour $k \in \{1, \ldots, d\}$, $n \in \{1, \ldots, n_f\}$, $l \in \{1, \ldots, l_q\}$ et $m \in \{1, \ldots, N_{ma}\}$, on a

$$\texttt{dphi\_dx}(k,n,l,m) = \frac{\partial \varphi_{\texttt{connect\_forme}(n,m)}}{\partial x_k}(\xi_{K_m,l}). \tag{12.25}$$

La taille du tableau `dphi_dx` pouvant être très grande, on peut concevoir des stratégies alternatives qui allégent les besoins en place mémoire au prix d'une (légère) augmentation du coût des calculs. Par exemple, lorsque la transformation T_{K_m} est affine (ce qui est souvent le cas pour des transformations subparamétriques), la matrice jacobienne J_{K_m} est constante sur K_m et ne dépend donc pas du point de Gauß considéré. Dans ce cas, on peut opter pour la stratégie suivante : on stocke l'inverse de la matrice jacobienne dans un tableau de taille $d \times d \times N_{ma}$ noté

$$\texttt{inv\_jac}(1\!:\!d, 1\!:\!d, 1\!:\!N_{ma}), \tag{12.26}$$

tel que pour $k_1, k_2 \in \{1, \ldots, d\}$ et $m \in \{1, \ldots, N_{ma}\}$, on a

$$\texttt{inv\_jac}(k_1,k_2,m) = \left(\left[J_{K_m} \right]^{-1} \right)_{k_1,k_2}, \tag{12.27}$$

puis, chaque fois que l'on a besoin des dérivées des fonctions de forme dans V_h, on effectue les opérations suivantes :

$$\frac{\partial \varphi_{\texttt{connect\_forme}(n,m)}}{\partial x_{k_1}}(\xi_{K_m,l}) = \sum_{k_2=1}^{d} \texttt{dtheta\_dx}(k_2,n,l)\,\texttt{inv\_jac}(k_2,k_1,m). \tag{12.28}$$

On observera que le rapport entre la taille des tableaux `dphi_dx` et `inv_jac` est de $\frac{n_l k_q}{d}$. On peut pousser la stratégie encore plus loin en ne stockant pas le tableau `inv_jac`, mais en en recalculant ses composantes chaque fois que cela s'avère nécessaire.

12.3 Assemblage

L'assemblage désigne l'ensemble des opérations réalisées dans un logiciel d'éléments finis afin d'évaluer les coefficients de la matrice et du membre de droite dans le système linéaire $AU = F$.

On reprend les notations de la section 9.3.1 ; on suppose notamment que la forme bilinéaire a_h intervenant dans le problème discret (9.23) s'écrit sous la forme

$$a_h(u_h, w_h) = \int_\Omega A_h(x, u_h, w_h)\,\mathrm{d}x, \tag{12.29}$$

où $A_h : \Omega \times V_h \times W_h \to \mathbb{R}$ est un opérateur qui dépend du problème modèle. Pour simplifier, on suppose que l'espace solution V_h et l'espace test discret W_h dans (9.23) sont les mêmes. Pour $i,j \in \{1,\ldots,N\}$ où $N = \dim(V_h) = \dim(W_h)$, on a donc

$$\mathcal{A}_{ij} = a_h(\varphi_j, \varphi_i) = \int_\Omega A_h(x, \varphi_j, \varphi_i)\,\mathrm{d}x, \tag{12.30}$$

où $\{\varphi_1,\ldots,\varphi_N\}$ sont les fonctions de forme dans V_h. Les coefficients $\mathcal{A}_{ij}$ s'évaluent par des quadratures selon la méthode décrite dans la section 9.1. L'algorithme 12.1 décrit l'assemblage de la matrice $\mathcal{A}$. On observera que la boucle principale s'effectue sur les mailles et non sur les nœuds. Le tableau de connectivité des fonctions de forme, `connect_forme`, permet de positionner la contribution de chaque maille au sein de la matrice $\mathcal{A}$.

Algorithme 12.1 Assemblage de la matrice $\mathcal{A}$

$\mathcal{A} = 0$
======Boucle sur les mailles
for $m \in \{1, \ldots, N_{\text{ma}}\}$ **do**
  ======Boucle sur les points de Gauß
  for $l \in \{1, \ldots, l_{\text{q}}\}$ **do**
    ======Boucles sur les fonctions de forme
    for $ni \in \{1, \ldots, n_{\text{f}}\}$ **do**; $i =$ `connect_forme`(ni, m)
      for $nj \in \{1, \ldots, n_{\text{f}}\}$ **do**; $j =$ `connect_forme`(nj, m)
        ======Accumuler
        $\mathcal{A}_{ij} = \mathcal{A}_{ij} +$ `poids`$(l, m) * A_b(\xi_{K_m,l},$ `theta`$(nj, l),$ `theta`$(ni, l))$
      end for
    end for
  end for
end for

Lorsque la matrice $\mathcal{A}$ est de très grande taille, le temps d'exécution de l'algorithme 12.1 peut être pénalisé par des accès mémoire trop fréquents aux coefficients de la matrice $\mathcal{A}$. On préfère alors introduire un tableau temporaire de taille $n_{\text{f}} \times n_{\text{f}}$ noté

$$\texttt{temp}(1\colon n_{\text{f}}, 1\colon n_{\text{f}}), \tag{12.31}$$

et procéder à l'accumulation des contributions

$$\texttt{poids}(l, m) * A_b(\xi_{lK_m}, \texttt{theta}(nj, l), \texttt{theta}(ni, l))$$

dans `temp`(ni, nj). La mise à jour de la matrice $\mathcal{A}$ se fait à l'extérieur de la boucle en l (sur les points de Gauß). Ces opérations sont décrites dans l'algorithme 12.2.

Afin de préciser quelques détails d'implémentation, on considère l'exemple d'un problème d'advection–diffusion–réaction pour lequel

$$A_b(x, \varphi_j, \varphi_i) = \nabla\varphi_i \cdot \sigma \cdot \nabla\varphi_j + \varphi_i(\beta \cdot \nabla\varphi_j) + \mu\varphi_i\varphi_j, \tag{12.32}$$

où σ, β et μ sont des fonctions données à valeurs dans $\mathbb{R}^{d,d}$, $\mathbb{R}^d$ et $\mathbb{R}$, respectivement. L'assemblage de la matrice $\mathcal{A}$ est présenté dans l'algorithme 12.3.

Algorithme 12.2 Assemblage de la matrice $\mathcal{A}$ (variante)

$\mathcal{A} = 0$
======Boucle sur les mailles
for $m \in \{1, \ldots, N_{\text{ma}}\}$ **do**
  `temp` $= 0$
  ======Boucle sur les points de Gauß
  for $l \in \{1, \ldots, l_{\text{q}}\}$ **do**
    ======Boucles sur les fonctions de forme
    for $ni \in \{1, \ldots, n_{\text{f}}\}$ **do**
      for $nj \in \{1, \ldots, n_{\text{f}}\}$ **do**
        ======Stocker provisoirement
        $\texttt{temp}(ni, nj) = \texttt{temp}(ni, nj)$
          $+ \texttt{poids}(l, m) * A_b(\xi_{K_m,l}, \texttt{theta}(nj, l), \texttt{theta}(ni, l))$
      end for
    end for
  end for
  ======Boucles sur les fonctions de forme
  for $ni \in \{1, \ldots, n_{\text{f}}\}$ **do**; $i = \texttt{connect\_forme}(ni, m)$
    for $nj \in \{1, \ldots, n_{\text{f}}\}$ **do**; $j = \texttt{connect\_forme}(nj, m)$
      ======Accumuler
      $\mathcal{A}_{ij} = \mathcal{A}_{ij} + \texttt{temp}(ni, nj)$
    end for
  end for
end for

On observera l'utilisation du tableau de connectivité géométrique et de celui de connectivité des fonctions de forme. À nouveau, on peut considérer la variante inspirée de l'algorithme 12.2 afin de réduire les accès mémoire.

L'assemblage du membre de droite procède de manière analogue. Par exemple, lorsque le membre de droite dans (9.23) s'écrit sous la forme

$$f_h(w_h) = \int_\Omega F_h(x, w_h)\, \mathrm{d}x, \tag{12.33}$$

où $F_h : \Omega \times V_h \to \mathbb{R}$ est un opérateur qui dépend des données du problème, l'assemblage du vecteur F peut se faire selon l'algorithme 12.4. Lorsque le

Algorithme 12.3 Assemblage de la matrice $\mathcal{A}$ pour un problème d'advection–diffusion–réaction

$\mathcal{A} = 0$
======Boucle sur les mailles
for $m \in \{1, \ldots, N_{\text{ma}}\}$ **do**
  ======Boucle sur les points de Gauß
  for $l \in \{1, \ldots, l_{\text{q}}\}$ **do**
    ======Évaluer les coordonnées cartésiennes du point de Gauß $\xi_{K_m,l}$
    for $k \in \{1, \ldots, d\}$ **do**
      $\texttt{xi\_l}(k) = \sum_{n=1}^{n_{\text{géo}}} \texttt{coord}(k, \texttt{connect\_géo}(n, m)) * \texttt{psi}(n, l)$
    end for
    ======Boucles sur les fonctions de forme
    for $ni \in \{1, \ldots, n_{\text{f}}\}$ **do**; $i = \texttt{connect\_forme}(ni, m)$
      for $nj \in \{1, \ldots, n_{\text{f}}\}$ **do**; $j = \texttt{connect\_forme}(nj, m)$
        ======Évaluer $\nabla\varphi_i \cdot \sigma \cdot \nabla\varphi_j$
        $x_1 = \sum_{k_1,k_2=1}^{d} \texttt{dphi\_dx}(k_1, ni, l, m) * \sigma_{k_1,k_2}(\texttt{xi\_l}) * \texttt{dphi\_dx}(k_2, nj, l, m)$
        ======Évaluer $\varphi_i(\beta \cdot \nabla\varphi_j)$
        $x_2 = \texttt{theta}(ni, l) \sum_{k_1=1}^{d} \beta_{k_1}(\texttt{xi\_l}) * \texttt{dphi\_dx}(k_1, nj, l, m)$
        ======Évaluer $\mu\varphi_i\varphi_j$
        $x_3 = \mu(\texttt{xi\_l}) * \texttt{theta}(ni, l) * \texttt{theta}(nj, l)$
        ======Accumuler
        $\mathcal{A}_{ij} = \mathcal{A}_{ij} + [x_1 + x_2 + x_3] * \texttt{poids}(l, m)$
      end for
    end for
  end for
end for

problème modèle fait intervenir des conditions de Neumann non-homogènes, le membre de droite dans (9.23) contient également une intégrale surfacique. Celle-ci est évaluée grâce à des quadratures surfaciques ; voir la section 9.1. Dans ces conditions, l'algorithme 12.4 est complété, après la boucle sur les éléments du maillage, par une boucle sur les faces de frontière où la condition de Neumann non-homogène est imposée. On utilise pour cela les tableaux `connect_g_front` et `i_cond_lim` définis dans la section 12.1.

Algorithme 12.4 Assemblage du membre de droite F

$F = 0$
======Boucle sur les mailles
for $m \in \{1, \ldots, N_{\text{ma}}\}$ **do**
  ======Boucle sur les points de Gauß
  for $l \in \{1, \ldots, l_{\text{q}}\}$ **do**
    ======Boucle sur les fonctions de forme
    for $ni \in \{1, \ldots, n_{\text{f}}\}$ **do**; $i =$ `connect_forme`(ni, m)
      ======Accumuler
      $F_i = F_i +$ `poids`$(l, m) * F_b(\xi_{K_m,l},$ `theta`$(ni, l))$
    end for
  end for
end for

12.4 Stockage

Les matrices issues de la méthode des éléments finis étant creuses, on réalise des économies substantielles de place mémoire en ne stockant que les coefficients non-nuls de ces matrices. Cette section présente deux formats de compression des matrices creuses, le format CSR (ou sa variante CSC) et le format Ellpack–Itpack.

12.4.1 Format CSR

Le format de compression CSR (de l'anglais, Compressed Sparse Row) est un des formats les plus couramment utilisés pour le stockage des matrices d'éléments finis. Le format CSC (de l'anglais, Compressed Sparse Column)

est analogue, les rôles joués par les lignes et les colonnes de la matrice étant simplement échangés.

On considère une matrice $\mathcal{A} \in \mathbb{R}^{N,N}$ (le format CSR peut également être utilisé pour compresser des matrices rectangulaires). Le format de compression CSR est basé sur l'utilisation de trois tableaux

$$\texttt{ia}(1\!:\!N\!+\!1), \qquad \texttt{ja}(1\!:\!nnz) \qquad \text{et} \qquad \texttt{aa}(1\!:\!nnz), \tag{12.34}$$

où nnz est le nombre d'éléments non-nuls dans la matrice $\mathcal{A}$. Ces tableaux sont définis de la manière suivante.

(i) Le tableau `ia`, à valeurs entières, stocke le nombre d'éléments non-nuls sur chaque ligne de $\mathcal{A}$. On pose $\texttt{ia}(1) = 1$ et pour $i \in \{1, \ldots, N\}$, $\texttt{ia}(i+1) - \texttt{ia}(i)$ est égal au nombre d'éléments non-nuls dans la i-ième ligne de $\mathcal{A}$. Clairement, on a $nnz = \texttt{ia}(N+1) - \texttt{ia}(1)$.

(ii) Le tableau `ja`, à valeurs entières, contient les numéros de colonne des éléments non-nuls de $\mathcal{A}$. Pour $i \in \{1, \ldots, N\}$, la liste

$$(\texttt{ja}(p))_{\texttt{ia}(i) \leqslant p \leqslant \texttt{ia}(i+1)-1}, \tag{12.35}$$

contient les numéros de colonne des éléments non-nuls de la i-ième ligne de $\mathcal{A}$. En général, on choisit d'ordonner le tableau `ja` de sorte que pour une ligne fixée, les numéros de colonne sont classés par ordre croissant.

(iii) Le tableau `aa` contient les valeurs des coefficients non-nuls de $\mathcal{A}$. Pour $i \in \{1, \ldots, N\}$, la liste

$$(\texttt{aa}(p))_{\texttt{ia}(i) \leqslant p \leqslant \texttt{ia}(i+1)-1}, \tag{12.36}$$

contient les valeurs des coefficients non-nuls de $\mathcal{A}$ situés sur la i-ième ligne. L'ordre des colonnes utilisé pour les tableaux `ja` et `aa` est le même si bien que pour $i \in \{1, \ldots, N\}$ et pour $p \in \{\texttt{ia}(i), \ldots, \texttt{ia}(i+1)-1\}$, on a

$$\texttt{aa}(p) = \mathcal{A}_{i,\texttt{ja}(p)}. \tag{12.37}$$

À titre d'exemple, on considère la matrice $\mathcal{A}$ d'ordre 5 telle que

$$\mathcal{A} = \begin{bmatrix} 1. & 0. & 0. & 0. & 2. \\ 3. & 4. & 0. & 5. & 0. \\ 6. & 0. & 7. & 8. & 9. \\ 0. & 0. & 10. & 11. & 0. \\ 0. & 0. & 0. & 0. & 12. \end{bmatrix}. \tag{12.38}$$

La compression au format CSR de cette matrice conduit aux trois tableaux suivants :

`ia` =	1	3	6	10	12	13						
`ja` =	1	5	1	2	4	1	3	4	5	3	4	5
`aa` =	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.

Lors de l'assemblage de la matrice $\mathcal{A}$, on réalise plusieurs fois l'opération suivante : étant donné deux indices $i, j \in \{1, \ldots, N\}$, tels que $\mathcal{A}_{ij} \neq 0$, et une valeur `val`, rajouter `val` à $\mathcal{A}_{ij}$. Lorsque la matrice $\mathcal{A}$ est compressée selon le format CSR, cette opération peut s'implémenter selon l'algorithme 12.5. Enfin, dans le cadre de la résolution itérative du système linéaire $\mathcal{A}U = F$, on est amené à effectuer plusieurs fois des produits matrice–vecteur de la forme $\mathcal{A}X$ où X est un vecteur donné dans $\mathbb{R}^N$. Cette opération peut s'implémenter selon l'algorithme 12.6.

Algorithme 12.5 Mise à jour de $\mathcal{A}_{ij} \neq 0$ en format CSR

Input : $i, j \in \{1, \ldots, N\}$ et `val` $\in \mathbb{R}$
for $p \in \{$`ia`$(i), \ldots,$ `ia`$(i+1) - 1\}$ **do**
  if `ja`$(p) = j$ **then**
    `aa`$(p) =$ `aa`$(p) +$ `val`; Fin boucle en p
  end if
end for

Algorithme 12.6 Produit matrice–vecteur $Y = \mathcal{A}X$ en format CSR

Input : X
for $i \in \{1 \ldots, N\}$ **do**
  $Y_i = 0$
  for $p \in \{\mathtt{ia}(i), \ldots, \mathtt{ia}(i+1) - 1\}$ **do**
    $Y_i = Y_i + \mathtt{aa}(p) * X_{\mathtt{ja}(p)}$
  end for
end for

12.4.2 Format Ellpack–Itpack

Le format de compression CSR présente certains défauts, notamment pour une mise en œuvre sur des ordinateurs à architecture parallèle ou à accélération vectorielle. Les difficultés proviennent d'une part du fait que les lignes de la matrice compressée n'ont pas toutes la même longueur et d'autre part du fait que l'indice du premier coefficient non-nul de chaque ligne doit être recalculé chaque fois que de besoin.

Le format de compression Ellpack–Itpack permet de lever partiellement ces difficultés au prix d'une (légère) augmentation de la place mémoire nécessaire au stockage de la matrice compressée. Pour $i \in \{1, \ldots, N\}$, $N_{\max}(i)$ désigne le nombre d'éléments non-nuls dans la i-ième ligne de la matrice $\mathcal{A}$. On pose

$$N_{\max} = \max_{1 \leqslant i \leqslant N} N_{\max}(i). \tag{12.39}$$

Le format de compression Ellpack–Itpack est basé sur l'utilisation de deux tableaux

$$\mathtt{aa}(1\colon N, 1\colon N_{\max}) \qquad \text{et} \qquad \mathtt{ja}(1\colon N, 1\colon N_{\max}). \tag{12.40}$$

Ces tableaux sont définis de la manière suivante.

(i) Le tableau `aa` stocke les valeurs des coefficients non-nuls de $\mathcal{A}$. Pour $i \in \{1, \ldots, N\}$, la liste

$$(\mathtt{aa}(i,p))_{1 \leqslant p \leqslant N_{\max}(i)}, \tag{12.41}$$

contient les valeurs des coefficients non-nuls de $\mathcal{A}$ situés sur la i-ième ligne. Par convention, ces coefficients sont ordonnés de sorte

que les numéros de colonne sont classés par ordre croissant. Si $N_{\max}(i) < N_{\max}$, on pose conventionnellement $\texttt{aa}(i,p) = 0$ pour $p \in \{N_{\max}(i)+1, \ldots, N_{\max}\}$.

(ii) Le tableau `ja`, à valeurs entières, contient les numéros de colonne des éléments non-nuls de $\mathcal{A}$. Pour $i \in \{1, \ldots, N\}$, la liste

$$(\texttt{ja}(i,p))_{1 \leqslant p \leqslant N_{\max}(i)}, \tag{12.42}$$

contient les numéros de colonne des coefficients non-nuls de $\mathcal{A}$ situés sur la i-ième ligne. Si $N_{\max}(i) < N_{\max}$, on pose conventionnellement $\texttt{ja}(i,p) = \texttt{ja}(i, N_{\max}(i))$ pour $p \in \{N_{\max}(i)+1, \ldots, N_{\max}\}$.

À titre d'exemple, pour la matrice d'ordre 5 définie en (12.38), le format de compression Ellpack–Itpack conduit aux tableaux suivants :

$$\texttt{aa} = \begin{bmatrix} 1. & 2. & 0. & 0. \\ 3. & 4. & 5. & 0. \\ 6. & 7. & 8. & 9. \\ 10. & 11. & 0. & 0. \\ 12. & 0. & 0. & 0. \end{bmatrix} \quad \text{et} \quad \texttt{ja} = \begin{bmatrix} 1 & 5 & 5 & 5 \\ 1 & 2 & 4 & 4 \\ 1 & 3 & 4 & 5 \\ 3 & 4 & 4 & 4 \\ 5 & 5 & 5 & 5 \end{bmatrix}. \tag{12.43}$$

Enfin, l'algorithme 12.7 décrit la mise à jour d'un coefficient $\mathcal{A}_{ij} \neq 0$ en format Ellpack–Itpack et l'algorithme 12.8 décrit l'implémentation du produit matrice–vecteur dans ce même format. On observera que le fait d'avoir mis à zéro les coefficients $\texttt{aa}(i,p)$ pour $p \in \{N_{\max}(i)+1, \ldots, N_{\max}\}$ permet de ne pas effectuer de test dans la boucle en p.

Algorithme 12.7 Mise à jour de $\mathcal{A}_{ij} \neq 0$ en format Ellpack–Itpack

Input : $i, j \in \{1, \ldots, N\}$ et $\texttt{val} \in \mathbb{R}$
for $p \in \{1, \ldots, N_{\max}\}$ **do**
  if $\texttt{ja}(i,p) = j$ **then**
    $\texttt{aa}(i,p) = \texttt{aa}(i,p) + \texttt{val}$; Fin boucle en p
  end if
end for

Algorithme 12.8 Produit matrice–vecteur $Y = \mathcal{A}X$ en format Ellpack–Itpack

Input : X
for $i \in \{1, \ldots, N\}$ **do**
  $Y_i = 0$
  for $p \in \{1, \ldots, N_{\max}\}$ **do**
    $Y_i = Y_i + \mathtt{aa}(i,p) * X_{\mathtt{ja}(i,p)}$
  end for
end for

12.5 Mailleurs

Un mailleur est un logiciel permettant de mailler un domaine Ω de $\mathbb{R}^2$ ou de $\mathbb{R}^3$, c'est-à-dire de générer les différents tableaux décrits dans la section 12.1. Pour simplifier, on se place en deux dimensions d'espace ; pour une description plus générale, on renvoie à George et Borouchaki [**44**] ou à Frey et George [**41**].

Afin de décrire le principe général du fonctionnement d'un mailleur, on part des observations suivantes.

(i) Un domaine Ω est entièrement déterminé par sa frontière $\partial\Omega$, qui est une variété de dimension 1.

(ii) La frontière est décomposée en ses composantes connexes et chaque composante connexe est décomposée en *lacets élémentaires*. Chaque lacet élémentaire est une variété de dimension 1 qui est suffisamment régulière, en général de classe $\mathcal{C}^1$ au moins. Par exemple, les lacets élémentaires composant la frontière d'un carré sont ses quatre côtés. On pose $\partial\Omega = \bigcup_{e=1}^{E} \partial\Omega_e$ où $\partial\Omega_e$ désigne un des lacets élémentaires composant la frontière et E le nombre total de lacets élémentaires. Chaque lacet élémentaire est paramétré par une transformation $\gamma_e : [0,1] \to \partial\Omega_e$. Les sommets du lacet $\partial\Omega_e$ sont les deux points $\gamma_e(0)$ et $\gamma_e(1)$.

Un mailleur met en œuvre l'algorithme suivant.

(i) On construit un maillage de chaque lacet $\partial\Omega_e$, $e \in \{1, \ldots, E\}$, en maillant l'intervalle $[0, 1]$ et en transformant ce maillage par γ_e. On choisit donc un maillage $\bigcup_{i=0}^{I_e-1}[x_{e,i}, x_{e,i+1}]$ de l'intervalle $[0, 1]$ et on obtient un maillage de $\partial\Omega_e$ en considérant $\bigcup_{i=0}^{I_e-1}\gamma_e([x_{e,i}, x_{e,i+1}])$. La partition

$$\bigcup_{e=1}^{E}\bigcup_{i=0}^{I_e-1}\gamma_e([x_{e,i}, x_{e,i+1}]), \tag{12.44}$$

constitue un maillage de $\partial\Omega$ qui est la trace du maillage de Ω qui doit maintenant être généré.

(ii) On maille l'intérieur du domaine Ω en étendant le maillage de frontière (12.44). Plusieurs techniques peuvent être utilisées pour réaliser cette extension. Un exemple classique, permettant de générer des triangulations dites de *Delaunay*, est l'algorithme de Bowyer–Watson où le maillage de Ω est construit par induction en insérant des nouveaux sommets et en modifiant les arêtes ; voir, par exemple, [**38**, p. 346].

Disposer d'un bon mailleur est un aspect important dans une mise en œuvre robuste de la méthode des éléments finis. Les sites suivants :
`http://www-users.informatik.rwth-aachen.de/~roberts/software.html`
et
`http://www.andrew.cmu.edu/user/sowen/mesh.html`
contiennent une liste de mailleurs disponibles (gratuitement ou non) sur Internet et une bibliographie relativement exhaustive sur le sujet. Parmi les mailleurs développés en France, on peut citer les exemples suivants.

(i) Les logiciels **EMC2** (Éditeur de maillages et de contours en deux dimensions), **BAMG** (de l'anglais, Bidimensional Anisotropic Mesh Generator), **GHS3D** (mailleur tétraédrique commercialisé sous le nom de **TetMesh**) sont développés à l'Inria au sein du projet Gamma (P.L. George, H. Borouchaki, P. Frey, F. Hecht et collaborateurs). Le site Web du projet Gamma se trouve à l'adresse suivante :
`http://www-rocq1.inria.fr/gamma/eng.htm`

(ii) Le logiciel **Mefisto** est développé dans le laboratoire Jacques-Louis Lions de l'Université Paris VI par A. Perronnet, P. Joly, C. Doursat et collaborateurs. Il contient un mailleur bidimensionnel et tridimensionnel ainsi qu'une boîte à outils éléments finis pour résoudre des problèmes de thermique et d'élasticité. Ce logiciel est accessible à l'adresse suivante : `http://www.ann.jussieu.fr/~perronne/mefisto.gene.html`

12.6 Conditions aux limites de Dirichlet

On considère un problème modèle où on impose des conditions de Dirichlet non-homogènes de la forme $u|_{\partial\Omega} = g$ où g est une fonction fixée sur $\partial\Omega$ que l'on suppose suffisamment régulière. On considère un espace d'approximation V_h construit à partir d'un élément fini de Lagrange. On désigne par $\{\varphi_1, \ldots, \varphi_N\}$, où $N = \dim(V_h)$, les fonctions de forme dans V_h et par $\{a_1, \ldots, a_N\}$ les nœuds associés. On désigne par N_D le nombre de nœuds situés sur la frontière et, sans perte de généralité, on suppose que la numérotation des nœuds est telle que les nœuds situés sur la frontière sont $\{a_1, \ldots, a_{N_D}\}$. On pose

$$V_{h0} = \{v_h \in V_h \,;\, v_h|_{\partial\Omega} = 0\} = \mathrm{vect}\{\varphi_{N_D+1}, \ldots, \varphi_N\}, \tag{12.45}$$

et on considère le problème discret suivant :

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a(u_h, w_h) = f(w_h), \quad \forall w_h \in V_{h0}, \\ u_h|_{\partial\Omega} = \mathcal{I}_h^{\mathrm{Lag}\partial} g, \end{cases} \tag{12.46}$$

où $a \in \mathcal{L}(V_h \times V_h; \mathbb{R}), f \in \mathcal{L}(V_h; \mathbb{R})$ et où

$$\mathcal{I}_h^{\mathrm{Lag}\partial} g = \sum_{i=1}^{N_D} g(a_i)\varphi_i \tag{12.47}$$

est l'interpolé de Lagrange surfacique de la donnée de Dirichlet non-homogène. On retrouve, à un changement de notation près, le problème discret (5.32) étudié dans la section 5.1.4.

En posant $u_h = \sum_{i=1}^N U_i\varphi_i$, on déduit de (12.46) que $U_i = g(a_i)$ pour $i \in \{1, \ldots, N_D\}$, et $\sum_{j=1}^N a(\varphi_j, \varphi_i)U_j = f(\varphi_i)$ pour $i \in \{N_D + 1, \ldots, N\}$.

Pour un vecteur $X \in \mathbb{R}^N$, on introduit la décomposition bloc $X = (X^0, X^+)$ où $X^0 = (X_1, \ldots, X_{N_D})^T$ regroupe les composantes de X associées aux nœuds de Dirichlet et $X^+ = (X_{N_D+1}, \ldots, X_N)^T$ regroupe les autres composantes. On pose $N_0 = N_D$ et $N_+ = N - N_D$. On introduit le vecteur $F = (F^0, F^+) \in \mathbb{R}^N$ tel que

$$F_i^0 = g(a_i), \qquad \text{pour } i \in \{1, \ldots, N_0\}, \tag{12.48}$$

$$F_i^+ = f(\varphi_{N_0+i}), \qquad \text{pour } i \in \{1, \ldots, N_+\}, \tag{12.49}$$

et les matrices $\mathcal{A}^{+0} \in \mathbb{R}^{N_+,N_0}$ et $\mathcal{A}^{++} \in \mathbb{R}^{N_+,N_+}$ telles que

$$\mathcal{A}_{ij}^{+0} = a(\varphi_j, \varphi_{N_0+i}), \qquad \text{pour } i \in \{1, \ldots, N_+\} \text{ et } j \in \{1, \ldots, N_0\}, \tag{12.50}$$

$$\mathcal{A}_{ij}^{++} = a(\varphi_{N_0+j}, \varphi_{N_0+i}), \qquad \text{pour } i, j \in \{1, \ldots, N_+\}. \tag{12.51}$$

Le problème (12.46) conduit au système linéaire

$$\mathcal{A} \begin{bmatrix} U^0 \\ U^+ \end{bmatrix} = \begin{bmatrix} F^0 \\ F^+ \end{bmatrix} \qquad \text{avec} \qquad \mathcal{A} = \left[\begin{array}{c:c} \mathcal{I}_{N_0} & 0 \\ \hdashline \mathcal{A}^{+0} & \mathcal{A}^{++} \end{array}\right], \tag{12.52}$$

où $\mathcal{I}_{N_0}$ est la matrice identité d'ordre N_0.

Une première approche consiste à éliminer le vecteur U^0 dans (12.52), ce qui conduit au système linéaire

$$\mathcal{A}^{++} U^+ = F^+ - \mathcal{A}^{+0} F^0. \tag{12.53}$$

Ce système linéaire présente deux avantages par rapport à (12.52) : il est de taille plus petite et si la forme bilinéaire a est symétrique, la matrice $\mathcal{A}^{++}$ est symétrique. L'inconvénient est qu'il faut assembler séparément les matrices $\mathcal{A}^{+0}$ et $\mathcal{A}^{++}$.

Une deuxième approche consiste à assembler d'abord la matrice $\mathcal{A}^* \in \mathbb{R}^{N,N}$ de coefficients

$$\mathcal{A}_{ij}^* = a(\varphi_j, \varphi_i) \qquad \text{pour } i, j \in \{1, \ldots, N\}, \tag{12.54}$$

sans tenir compte des conditions aux limites de Dirichlet (on assemble ainsi la matrice associée au problème de Neumann). Clairement, les N_0 premières lignes de la matrice $\mathcal{A}^*$ diffèrent de celles de la matrice $\mathcal{A}$ définie en (12.52).

Puis, on corrige ces lignes en annulant tous les coefficients sauf les coefficients diagonaux auxquels on affecte la valeur 1. On obtient ainsi la matrice $\mathcal{A}$. Soit $R = (R^0, R^+) \in \mathbb{R}^N$; on désigne par $\mathbb{K}_k$ *l'espace de Krylov*

$$\mathbb{K}_k = \text{vect}\{R, \mathcal{A}R, \ldots, \mathcal{A}^{k-1}R\}. \tag{12.55}$$

Une observation importante est que si la forme bilinéaire a est symétrique, la restriction de la matrice $\mathcal{A}$ à $\mathbb{K}_k$ est symétrique pourvu que $R^0 = 0$. En d'autres termes, même si la matrice $\mathcal{A}$ n'est pas symétrique, on peut appliquer la méthode du gradient conjugué afin d'approcher la solution de (12.52) pourvu que la forme bilinéaire a soit symétrique et coercive et que la méthode ait été initialisée avec un vecteur satisfaisant les conditions aux limites de Dirichlet.

Une troisième approche consiste à imposer les conditions aux limites de Dirichlet non-homogènes par pénalisation. On assemble la matrice $\mathcal{A}^* \in \mathbb{R}^{N,N}$ définie en (12.54) ainsi que le vecteur $F^* \in \mathbb{R}^N$ tel que $F_i^* = f(\varphi_i)$ pour $i \in \{1, \ldots, N\}$. Puis, on choisit un paramètre ϵ suffisamment petit et pour $i \in \{1, \ldots, N_0\}$, on rajoute ϵ^{-1} au coefficient $\mathcal{A}_{ii}^*$ et $\epsilon^{-1} g(a_i)$ au coefficient F_i^*. Pour l'analyse de cette méthode de pénalisation et ses variantes, on pourra consulter, entre autres, [7].

ANNEXE A • BASES MATHÉMATIQUES DE LA MÉTHODE DES ÉLÉMENTS FINIS

Cette annexe contient quelques rappels sur les bases mathématiques de la méthode des éléments finis. Pour approfondir les notions présentées ci-dessous, on pourra consulter, entre autres, Adams [**3**], Aubin [**8**], Bartle [**13**], Brezis [**21**], Rudin [**65**], Yosida [**78**] ou Zeidler [**79**].

A.1 Espaces de Banach

A.1.1 Espaces vectoriels normés

Soit V un espace vectoriel réel (tout ce qui suit s'adapte facilement aux espaces vectoriels complexes).

- **Norme.** Une norme sur V est une application de V dans $\mathbb{R}_+$, généralement notée $\|\cdot\|_V$, satisfaisant les trois propriétés suivantes :

  (i) (homogénéité de degré 1) $\forall c \in \mathbb{R}, \forall v \in V, \|cv\|_V = |c|\,\|v\|_V$;

  (ii) (inégalité triangulaire) $\forall (v, w) \in V \times V, \|v + w\|_V \leqslant \|v\|_V + \|w\|_V$;

  (iii) $(\|v\|_V = 0) \Longrightarrow (v = 0)$.

  Un *espace vectoriel normé* est un espace vectoriel muni d'une norme. Une application de V dans $\mathbb{R}_+$ qui ne satisfait que les propriétés (i) et (ii) est appelée une *semi-norme*.

- **Normes équivalentes.** On dit que deux normes $\|\cdot\|_{V,1}$ et $\|\cdot\|_{V,2}$ sont *équivalentes* sur V s'il existe deux constantes strictement positives c_1 et c_2 telles que

$$\forall v \in V, \qquad c_1\|v\|_{V,2} \leqslant \|v\|_{V,1} \leqslant c_2\|v\|_{V,2}.$$

  Si l'espace vectoriel V est de dimension finie, toutes ses normes sont équivalentes. Ce résultat est faux en dimension infinie.

- **Produit scalaire.** Un produit scalaire sur V est une application de $V \times V$ dans $\mathbb{R}$, généralement notée $(\cdot,\cdot)_V$, telle que :

  (i) (bilinéarité) $\forall v \in V$, les applications $V \ni w \longmapsto (v,w)_V \in \mathbb{R}$ et $V \ni w \longmapsto (w,v)_V \in \mathbb{R}$ sont linéaires ;

  (ii) (symétrie) $\forall (v,w) \in V \times V$, $(v,w)_V = (w,v)_V$;

  (iii) (positivité) $\forall v \in V$, $(v,v)_V \geqslant 0$;

  (iv) $((v,v)_V = 0) \Longrightarrow (v = 0)$.

  Un *espace euclidien* est un espace vectoriel équipé d'un produit scalaire. Un produit scalaire induit une norme sur V en posant pour tout $v \in V$,

$$\|v\|_V = (v,v)_V^{\frac{1}{2}}. \tag{A.1}$$

  On dispose alors de *l'inégalité de Cauchy–Schwarz* : pour tout $(v,w) \in V \times V$,

$$(v,w)_V \leqslant \|v\|_V \|w\|_V. \tag{A.2}$$

- **Intérieur, adhérence et densité.** Soit V un espace vectoriel normé. On dit qu'un ensemble $O \subset V$ est *ouvert* (dans V) si pour tout $x \in O$, il existe $r > 0$ tel que la boule de centre x et de rayon r est incluse dans O, c'est-à-dire

$$\{y \in V\,;\, \|x-y\|_V < r\} \subset O. \tag{A.3}$$

  On dit qu'un ensemble $F \subset V$ est *fermé* (dans V) si son complémentaire dans V est ouvert. Soit E un ensemble inclus dans V.

  - On appelle *intérieur de* E la réunion de tous les ouverts inclus dans E ; cet ensemble est noté $\overset{\circ}{E}$.

- On appelle *adhérence de* E dans V l'intersection de tous les fermés de V contenant E ; cet ensemble est noté $\overline{E}^V$ ou, plus simplement, $\overline{E}$ si aucune ambiguïté n'est possible.
- On dit que l'ensemble E est *dense dans* V si $\overline{E}^V = V$.

A.1.2 Espaces de Banach

- **Suites de Cauchy.** Une suite de Cauchy dans un espace vectoriel V équipé d'une norme $\|\cdot\|_V$ est une suite $(v_n)_{n\geqslant 0}$ d'éléments de V satisfaisant la propriété suivante :

$$\forall \epsilon > 0, \quad \exists N \geqslant 0 \quad \text{tel que} \quad \forall n \geqslant N,\ \forall p \geqslant 0, \quad \|v_{n+p} - v_n\|_V \leqslant \epsilon. \tag{A.4}$$

  Un espace vectoriel normé est dit **complet** si toute suite de Cauchy converge dans V. Si les normes $\|\cdot\|_{V,1}$ et $\|\cdot\|_{V,2}$ sont équivalentes, V équipé de la norme $\|\cdot\|_{V,1}$ est complet si et seulement si V équipé de la norme $\|\cdot\|_{V,2}$ est complet. Par ailleurs, un résultat classique est qu'un espace vectoriel de dimension finie est complet.

- Un **espace de Banach** est un espace vectoriel normé complet.

- **Applications linéaires continues.** Soient V et W deux espaces vectoriels normés et soit $A : V \to W$ une application linéaire de V dans W. On dit que A est continue si[1]

$$\sup_{v\in V} \frac{\|Av\|_W}{\|v\|_V} < +\infty. \tag{A.5}$$

  On note $\mathcal{L}(V; W)$ l'espace vectoriel des applications linéaires continues de V dans W. En posant pour $A \in \mathcal{L}(V; W)$,

$$\|A\|_{\mathcal{L}(V;W)} = \sup_{v\in V} \frac{\|Av\|_W}{\|v\|_V}, \tag{A.6}$$

1. On adopte la convention de notation suivante : dans les expressions du type $\sup_{v\in V}(\cdot)$, il est entendu que le supremum est pris pour $v \in V \setminus \{0\}$. La même convention est adoptée pour les expressions du type $\inf_{v\in V}(\cdot)$.

on munit $\mathcal{L}(V;W)$ d'une norme. De plus, si W est un espace de Banach, on montre que $\mathcal{L}(V;W)$ est un espace de Banach pour la norme (A.6), et ce indépendamment de la complétude de V.

- **Dualité.** Soit V un espace vectoriel normé. L'espace vectoriel $\mathcal{L}(V;\mathbb{R})$ est noté V' et est appelé *l'espace dual* de V. Un élément $A \in V'$ s'appelle une *forme linéaire* sur V. L'action d'une forme linéaire $A \in V'$ sur un vecteur $v \in V$ est également notée $\langle A, v\rangle_{V',V}$. En posant pour $A \in V'$,

$$\|A\|_{V'} = \sup_{v\in V} \frac{\langle A, v\rangle_{V',V}}{\|v\|_V}, \tag{A.7}$$

on munit V' d'une norme. L'espace V' est toujours un espace de Banach pour cette norme puisque $\mathbb{R}$, de dimension finie, est complet.

- **Espaces de Banach réflexifs.** Soit V un espace de Banach. On note V'' *l'espace bidual* de V, c'est-à-dire l'espace dual de V'. On montre que l'application linéaire $J_V : V \to V''$ définie pour tout $v \in V$ et pour tout $v' \in V'$ par

$$\langle J_V v, v'\rangle_{V'',V'} = \langle v', v\rangle_{V',V}, \tag{A.8}$$

est une isométrie. Par conséquent, cette application est injective. Par contre, elle n'est pas nécessairement surjective. Lorsque l'application J_V est surjective, les espaces V et V'' sont isomorphes ; on dit que l'espace de Banach V est réflexif.

- **Opérateur transposé.** Soient V et W deux espaces vectoriels normés et soit $A \in \mathcal{L}(V, W)$. L'opérateur transposé de A est une application linéaire de W' dans V' qui est notée A^T. Cette application est définie de la manière suivante : pour tout $w' \in W'$ et pour tout $v \in V$, on a

$$\langle A^T w', v\rangle_{V',V} = \langle w', Av\rangle_{W',W}. \tag{A.9}$$

Lorsque V est un espace de Banach réflexif, on dit que l'opérateur $A \in \mathcal{L}(V;V')$ est *auto-adjoint* si $A^T = A$ modulo l'identification de V et de V''.

- **Espaces polaires.** Soit V un espace vectoriel normé. Pour un sous-espace vectoriel $Y \subset V$, on pose

$$Y^{\perp} = \{v' \in V' \,;\, \forall y \in Y,\ \langle v', y\rangle_{V',V} = 0\}. \tag{A.10}$$

De même, pour un sous-espace vectoriel $Z \subset V'$, on pose

$$Z^{\perp} = \{v \in V \,;\, \forall z \in Z,\ \langle z, v\rangle_{V',V} = 0\}. \tag{A.11}$$

Les espaces $Y^{\perp} \subset V'$ et $Z^{\perp} \subset V$ sont appelés les espaces polaires de Y et de Z, respectivement.

- **Formes bilinéaires continues.** Soient V et W deux espaces vectoriels normés. On dit qu'une application $a : V \times W \to \mathbb{R}$ est une forme bilinéaire continue sur $V \times W$ si :

(i) $\forall v \in V$, $a(v, \cdot)$ est une application linéaire de W dans $\mathbb{R}$;

(ii) $\forall w \in W$, $a(\cdot, w)$ est une application linéaire de V dans $\mathbb{R}$;

(iii) $\sup_{(v,w)\in V\times W} \frac{a(v,w)}{\|v\|_V \|w\|_W} < +\infty$.

On note $\mathcal{L}(V \times W; \mathbb{R})$ l'espace vectoriel des formes bilinéaires continues sur $V \times W$. En posant pour $a \in \mathcal{L}(V \times W; \mathbb{R})$,

$$\|a\|_{V,W} = \sup_{(v,w)\in V\times W} \frac{a(v,w)}{\|v\|_V \|w\|_W}, \tag{A.12}$$

on définit une norme sur $\mathcal{L}(V \times W; \mathbb{R})$ et on munit cet espace d'une structure d'espace de Banach.

Lorsque $V = W$, on dit que la forme bilinéaire $a \in \mathcal{L}(V \times V; \mathbb{R})$ est *symétrique* si pour tout $(v, w) \in V \times V$, $a(v, w) = a(w, v)$. On dit que la forme bilinéaire a est *positive* si pour tout $v \in V$, $a(v, v) \geqslant 0$. Un exemple simple de forme bilinéaire continue, symétrique et positive sur $V \times V$, où V est un espace euclidien, est le produit scalaire $(\cdot, \cdot)_V$.

A.1.3 Espaces de Hilbert

- Un **espace de Hilbert** est un espace euclidien qui est complet pour la norme induite par le produit scalaire.

- **Théorème de représentation de Riesz.** Soit V un espace de Hilbert. Pour tout $v' \in V'$, il existe un et un seul $u \in V$ tel que pour tout $v \in V$,

$$\langle v', v\rangle_{V',V} = (u, v)_V. \tag{A.13}$$

L'application $V' \ni v' \longmapsto u \in V$ est un isomorphisme isométrique. Une conséquence de ce résultat est que les espaces de Hilbert sont réflexifs.

A.1.4 Opérateurs bijectifs dans les espaces de Banach

Dans cette section, V et W désignent deux espaces de Banach. Pour $A \in \mathcal{L}(V; W)$, on note $\mathrm{Ker}(A)$ le noyau de A et $\mathrm{Im}(A)$ son image. L'objectif de cette section est de caractériser les opérateurs A qui sont bijectifs, c'est-à-dire tels que $\mathrm{Ker}(A) = \{0\}$ (A est injectif) et $\mathrm{Im}(A) = W$ (A est surjectif). En général, l'injectivité d'un opérateur est plus simple à montrer que sa surjectivité. En effet, montrer l'injectivité de A revient à prouver que pour tout $v \in V$, $Av = 0$ implique $v = 0$. Par ailleurs, montrer la surjectivité de A revient à prouver que pour tout $w \in W$, on peut trouver un antécédent $v_w \in V$ tel que $Av_w = w$; or, il n'est pas toujours évident d'exhiber un tel antécédent. L'objectif des résultats ci-dessous est de caractériser les opérateurs bijectifs à partir de propriétés relativement simples à vérifier.

- **Liens entre noyau et image.** Pour tout $A \in \mathcal{L}(V; W)$, on a :

(i) $\mathrm{Ker}(A) = (\mathrm{Im}(A^T))^\perp$;

(ii) $\mathrm{Ker}(A^T) = (\mathrm{Im}(A))^\perp$;

(iii) $\overline{\mathrm{Im}(A)} = (\mathrm{Ker}(A^T))^\perp$;

(iv) $\overline{\mathrm{Im}(A^T)} \subset (\mathrm{Ker}(A))^\perp$.

- Les deux théorèmes suivants, dus à Banach, jouent un rôle fondamental dans la caractérisation des opérateurs bijectifs.

Théorème A.1 (Image fermée). *Soit $A \in \mathcal{L}(V; W)$. Les propositions suivantes sont équivalentes :*

(i) $\mathrm{Im}(A)$ *est fermé ;*

(ii) $\mathrm{Im}(A^T)$ *est fermé ;*

(iii) $\mathrm{Im}(A) = (\mathrm{Ker}(A^T))^\perp$;

(iv) $\mathrm{Im}(A^T) = (\mathrm{Ker}(A))^\perp$.

Théorème A.2 (Application ouverte). *Si $A \in \mathcal{L}(V;W)$ est surjectif et si O est ouvert dans V, alors $A(O)$ est ouvert dans W.*

Le théorème de l'application ouverte a pour conséquence le lemme suivant.

Lemme A.3. *Soit $A \in \mathcal{L}(V;W)$. Les propositions suivantes sont équivalentes :*

(i) $\mathrm{Im}(A)$ *est fermé ;*

(ii) *il existe* $\alpha > 0$ *tel que pour tout* $w \in \mathrm{Im}(A)$, *il existe* $v_w \in V$ *tel que* $Av_w = w$ *et* $\alpha\|v_w\|_V \leqslant \|w\|_W$.

Une conséquence importante de ce lemme est que si $A \in \mathcal{L}(V;W)$ est bijectif, alors son inverse est nécessairement continu.

- **Caractérisation des opérateurs surjectifs.** Les deux résultats suivants sont une conséquence du théorème de l'image fermée et du théorème de l'application ouverte.

Lemme A.4. *Soit $A \in \mathcal{L}(V;W)$. Les propositions suivantes sont équivalentes :*

(i) $A : V \to W$ *est surjectif;*

(ii) $A^T : W' \to V'$ *est injectif et* $\mathrm{Im}(A^T)$ *est fermé dans* V' ;

(iii) *il existe* $\alpha > 0$ *tel que*

$$\forall w' \in W', \qquad \|A^T w'\|_{V'} \geqslant \alpha \|w'\|_{W'}. \tag{A.14}$$

Lemme A.5. *Soit $A \in \mathcal{L}(V;W)$. Les propositions suivantes sont équivalentes :*

(i) $A^T : W' \to V'$ *est surjectif;*

(ii) $A : V \to W$ *est injectif et* $\mathrm{Im}(A)$ *est fermé dans* W ;

(iii) *Il existe* $\alpha > 0$ *tel que*

$$\forall v \in V, \qquad \|Av\|_W \geqslant \alpha \|v\|_V. \tag{A.15}$$

- **Caractérisation des opérateurs bijectifs.** Une conséquence importante des lemmes A.4 et A.5 est le théorème suivant.

Théorème A.6. *Soit* $A \in \mathcal{L}(V; W)$. *Les propositions suivantes sont équivalentes :*

(i) $A : V \to W$ *est bijectif;*

(ii) $A^T : W' \to V'$ *est bijectif;*

(iii) A^T *est injectif,* A *est injectif et* $\mathrm{Im}(A)$ *est fermé dans* W *;*

(iv) A^T *est injectif et il existe* $\alpha > 0$ *tel que*

$$\forall v \in V, \qquad \|Av\|_W \geqslant \alpha \|v\|_V. \tag{A.16}$$

- **Le théorème BNB dans les espaces de Banach.** Le cadre naturel du théorème BNB, qui a été énoncé dans la section 2.1.2 dans le cadre simplifié des espaces de Hilbert, est celui des espaces de Banach. Plus précisément, soit V un espace de Banach et soit W un espace de Banach réflexif. Soit $a \in \mathcal{L}(V \times W; \mathbb{R})$ et $f \in W'$. Le théorème BNB s'énonce de la manière suivante : le problème

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u, w) = \langle f, w \rangle_{W',W}, \quad \forall w \in W, \end{cases} \tag{A.17}$$

admet une et une seule solution si et seulement si les deux conditions (BNB1) et (BNB2) sont satisfaites, c'est-à-dire si et seulement si

$$\exists \alpha > 0, \qquad \inf_{v \in V} \sup_{w \in W} \frac{a(v, w)}{\|v\|_V \|w\|_W} \geqslant \alpha, \tag{A.18}$$

$$\forall w \in W, \quad (\forall v \in V,\ a(v, w) = 0) \implies (w = 0). \tag{A.19}$$

Afin d'interpréter ces conditions à l'aide d'opérateurs bijectifs dans les espaces de Banach, on associe à la forme bilinéaire a l'opérateur

$A \in \mathcal{L}(V; W')$ tel que pour tout $v \in V$, Av est la forme linéaire continue sur W définie par

$$\forall w \in W, \qquad \langle Av, w\rangle_{W',W} = a(v, w). \tag{A.20}$$

Le problème (A.17) revient à chercher $u \in V$ tel que $Au = f$ dans W'. Puisque l'espace W est réflexif, l'opérateur transposé $A^T : W'' \equiv W \to V'$ peut être défini de la manière suivante :

$$\forall (v, w) \in V \times W, \qquad \langle A^T w, v\rangle_{V',V} = \langle Av, w\rangle_{W',W} = a(v, w). \tag{A.21}$$

Grâce aux lemmes ci-dessus, les conditions (BNB1) et (BNB2) s'interpètent comme suit :

(BNB1) $\iff$ (Ker$(A) = \{0\}$ et Im(A) fermé dans W) $\iff$ (A^T surjectif),
(BNB2) $\iff$ (Ker$(A^T) = \{0\}$) $\iff$ (A^T injectif).

Du théorème A.6 on déduit que les conditions (BNB1) et (BNB2) sont équivalentes à la bijectivité de A, donc au fait que le problème (A.17) admet une et une seule solution.

- **Opérateurs coercifs.** On dit qu'un opérateur $A \in \mathcal{L}(V; V')$ est *monotone* si

$$\forall v \in V, \qquad \langle Av, v\rangle_{V',V} \geqslant 0. \tag{A.22}$$

On dit qu'un opérateur $A \in \mathcal{L}(V; V')$ est *coercif* s'il existe une constante $\alpha > 0$ telle que

$$\forall v \in V, \qquad \langle Av, v\rangle_{V',V} \geqslant \alpha \|v\|_V^2. \tag{A.23}$$

La notion d'opérateur coercif n'est pertinente que *dans un cadre hilbertien*. En effet, en posant pour $(v, w) \in V \times V$,

$$(v, w)_V = \frac{1}{2}(\langle Av, w\rangle_{V',V} + \langle Aw, v\rangle_{V',V}),$$

on vérifie qu'on équipe V d'un produit scalaire dont la norme induite est équivalente à la norme $\|\cdot\|_V$; par conséquent, l'existence d'un opérateur coercif dans $\mathcal{L}(V; V')$ implique que V est un espace de Hilbert.

On vérifie facilement qu'une *condition suffisante* pour qu'un opérateur $A \in \mathcal{L}(V; V')$ soit bijectif est qu'il soit coercif. Par ailleurs, lorsque l'opérateur $A \in \mathcal{L}(V; V')$ est monotone et auto-adjoint, la coercivité est équivalente à la bijectivité.

A.2 Espaces de fonctions régulières

- **Espaces $\mathcal{C}^k$.** Soit un entier $k \geqslant 0$. Soit Ω un ouvert borné de $\mathbb{R}^d$. On désigne par $\overline{\Omega}$ l'adhérence de Ω dans $\mathbb{R}^d$. Soit E un sous-ensemble de $\overline{\Omega}$. Les espaces $\mathcal{C}^k(E)$ sont définis de la manière suivante. Pour $k = 0$, $\mathcal{C}^0(E)$ désigne l'espace vectoriel des fonctions continues sur E à valeurs dans $\mathbb{R}$. Pour $k \geqslant 1$, $\mathcal{C}^k(E)$ désigne l'espace des fonctions k-fois continûment différentiables sur E à valeurs dans $\mathbb{R}$.
  On note $(x_1, \ldots, x_d)$ les coordonnées cartésiennes dans $\mathbb{R}^d$. Pour $i \in \{1, \ldots, d\}$ et pour $f \in \mathcal{C}^k(E)$, $\partial_i f$ désigne la dérivée partielle de f par rapport à x_i. Pour un entier $\alpha_i \in \{0, \ldots, k\}$, $\partial_i^{\alpha_i} f$ désigne la dérivée partielle de f d'ordre α_i par rapport à x_i avec la convention que $\partial_i^0 f = f$. En posant pour $k \geqslant 0$ et pour $f \in \mathcal{C}^k(E)$,

$$\|f\|_{\mathcal{C}^k(E)} = \max_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| \leqslant k}} \left(\sup_{x \in E} |\partial^\alpha f(x)| \right), \tag{A.24}$$

  on munit $\mathcal{C}^k(E)$ d'une norme. Dans la définition ci-dessus, α est un *multi-indice* de $\mathbb{N}^d$, c'est-à-dire $\alpha = (\alpha_1, \ldots, \alpha_d)^T$ avec $\alpha_i \in \mathbb{N}$ pour tout $i \in \{1, \ldots, d\}$, $|\alpha| = \sum_{i=1}^d \alpha_i$ et $\partial^\alpha f(x) = (\partial_1^{\alpha_1} \ldots \partial_d^{\alpha_d} f)(x)$. On introduit également la semi-norme suivante :

$$|f|_{\mathcal{C}^k(E)} = \max_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| = k}} \left(\sup_{x \in E} |\partial^\alpha f(x)| \right). \tag{A.25}$$

  Il s'agit d'une semi-norme et non d'une norme car $|f|_{\mathcal{C}^k(E)} = 0$ si f est dans $\mathbb{P}_{k-1}$.

  Proposition A.7. *Équipé de la norme définie en* (A.24), *$\mathcal{C}^k(E)$ est un espace de Banach.*

- **Espaces de Hölder.** Soit un réel $\omega \in]0, 1]$. On désigne par $\mathcal{C}^{0,\omega}(E)$ le sous-espace vectoriel de $\mathcal{C}^0(E)$ constitué des fonctions telles que

$$\sup_{(x,y) \in E \times E} \frac{|f(x) - f(y)|}{\|x - y\|_{\mathbb{R}^d}^\omega} < +\infty. \tag{A.26}$$

Pour un entier $k \geqslant 1$, $\mathcal{C}^{k,\omega}(E)$ désigne le sous-espace vectoriel de $\mathcal{C}^k(E)$ constitué des fonctions telles que pour tout multi-indice γ tel que $|\gamma| = k$, la fonction $\partial^\gamma f$ est dans $\mathcal{C}^{0,\omega}(E)$. Les espaces $\mathcal{C}^{k,\omega}(E)$ sont appelés des *espaces de Hölder*. Les éléments de $\mathcal{C}^{k,\omega}(E)$ sont appelés les *fonctions höldériennes* sur E d'ordre k et d'exposant ω. Lorsque $k = 0$ et $\omega = 1$, les éléments de $\mathcal{C}^{0,1}(E)$ sont appelés les *fonctions lipschitziennes* sur E.

Proposition A.8. *Soit un entier $k \geqslant 0$ et soit un réel $\omega \in]0, 1]$. Équipé de la norme*

$$\|f\|_{\mathcal{C}^{k,\omega}(E)} = \|f\|_{\mathcal{C}^k(E)} + \max_{\substack{\gamma \in \mathbb{N}^d \\ |\gamma| = k}} \left(\sup_{(x,y) \in E \times E} \frac{|\partial^\gamma f(x) - \partial^\gamma f(y)|}{\|x - y\|^\omega_{\mathbb{R}^d}} \right), \tag{A.27}$$

$\mathcal{C}^{k,\omega}(E)$ est un espace de Banach.

A.3 Intégration et espaces de Lebesgue

- **Ensembles et fonctions mesurables.** On montre qu'il existe :

(i) un ensemble de parties de $\mathbb{R}^d$ noté $\mathbb{T}_d$ et appelé la tribu borélienne de $\mathbb{R}^d$ (l'ensemble $\mathbb{T}_d$ est suffisamment riche pour contenir tous les ouverts et fermés de $\mathbb{R}^d$, il est stable par passage au complémentaire et par réunion dénombrable) ;

(ii) une application $\lambda : \mathbb{T}_d \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ appelée *mesure de Lebesgue* et telle que
   - $\lambda(\varnothing) = 0$;
   - si $(T_n)_{n \in \mathbb{N}}$ est une suite d'éléments de $\mathbb{T}_d$ deux à deux disjoints, alors
   $$\lambda\left(\bigcup_{n \in \mathbb{N}} T_n \right) = \sum_{n \in \mathbb{N}} \lambda(T_n); \tag{A.28}$$
   - pour un pavé $P = \prod_{i=1}^d (a_i, b_i)$ (où la notation (a_i, b_i) désigne l'un quelconque des intervalles $[a_i, b_i]$, $[a_i, b_i[$, $]a_i, b_i]$ ou $]a_i, b_i[$), on a $\lambda(P) = \prod_{i=1}^d (b_i - a_i)$ avec la convention que ce produit vaut 0 si l'un des facteurs $(b_i - a_i)$ est nul, même si d'autres facteurs valent $+\infty$.

Un élément $T \in \mathbb{T}_d$ est appelé un ensemble mesurable (pour la mesure de Lebesgue) et le réel $\lambda(T)$ (pouvant éventuellement valoir $+\infty$) est appelé sa mesure. Dans cet aide-mémoire, on utilise la notation plus explicite

$$\text{mes}(T) = \lambda(T). \tag{A.29}$$

Soit Ω un ouvert de $\mathbb{R}^d$. On dit qu'une fonction $f : \Omega \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ est *mesurable* si pour tout $\alpha \in \mathbb{R}$, l'ensemble réciproque $f^{-1}(]\alpha, +\infty])$ est mesurable. On dit qu'une fonction e est « en escalier » si elle peut s'écrire sous la forme

$$e = \sum_{i=1}^{N} \alpha_i \chi_{A_i}, \tag{A.30}$$

où, pour tout $i \in \{1, \ldots, N\}$, $\alpha_i \in \mathbb{R}$, $A_i \in \mathbb{T}_d$ et χ_{A_i} désigne la *fonction indicatrice* de l'ensemble mesurable A_i ($\chi_{A_i}(x) = 1$ si $x \in A_i$ et $\chi_{A_i}(x) = 0$ sinon). On vérifie facilement qu'une fonction en escalier est mesurable. Par ailleurs, on montre qu'une fonction mesurable peut toujours s'écrire comme limite simple d'une suite de fonctions en escalier.

- **Intégrale de Lebesgue.** Soit e la fonction en escalier définie en (A.30). On suppose que e est positive. On définit l'intégrale (de Lebesgue) de e sur $\mathbb{R}^d$ comme le réel

$$\int_{\mathbb{R}^d} e \, d\lambda = \sum_{i=1}^{N} \alpha_i \lambda(A_i) \in \overline{\mathbb{R}}_+. \tag{A.31}$$

Pour une fonction mesurable et *positive*, on pose

$$\int_{\mathbb{R}^d} f \, d\lambda = \sup_{\substack{e \text{ en escalier} \\ 0 \leqslant e \leqslant f}} \left(\int_{\mathbb{R}^d} e \, d\lambda \right) \in \overline{\mathbb{R}}_+, \tag{A.32}$$

et on dit que f est *intégrable* sur $\mathbb{R}^d$ si $\int_{\mathbb{R}^d} f \, d\lambda < +\infty$. Pour une fonction mesurable f, on pose $f = f^+ - f^-$ avec $f^+ = \max(f, 0)$ et $f^- = \max(-f, 0)$. On dit que f est intégrable sur $\mathbb{R}^d$ si f^+ et f^- le sont ; dans ces conditions, on pose

$$\int_{\mathbb{R}^d} f \, d\lambda = \int_{\mathbb{R}^d} f^+ \, d\lambda - \int_{\mathbb{R}^d} f^- \, d\lambda. \tag{A.33}$$

Enfin, on dit qu'une fonction mesurable f est intégrable sur un ouvert Ω de $\mathbb{R}^d$ si la fonction $\chi_\Omega f$ est intégrable sur $\mathbb{R}^d$ où χ_Ω désigne la fonction indicatrice de Ω. On note $\mathcal{L}^1(\Omega)$ l'espace des fonctions intégrables sur Ω.

Dans cet aide-mémoire, on omet, sauf dans quelques situations, le symbole $d\lambda$ dans les intégrales, étant entendu que celles-ci sont *toujours* considérées pour la mesure de Lebesgue. De plus, l'ensemble des fonctions mesurables étant suffisamment riche pour contenir toutes les fonctions habituellement rencontrées en sciences de l'ingénieur, on suppose implicitement que toutes les fonctions considérées dans cet aide-mémoire sont mesurables.

- **Propriétés vraies presque partout et classes de fonctions.** On dit qu'une propriété est vraie presque partout sur un ouvert Ω si elle est satisfaite partout sauf sur un sous-ensemble de Ω qui est de mesure nulle. Par exemple, on dit que deux fonctions f et g sont égales presque partout sur Ω si $\lambda(\{x \in \Omega\,;\, f(x) \neq g(x)\}) = 0$. Si $f = g$ presque partout sur Ω et si f est intégrable sur Ω, alors g est intégrable sur Ω et on a $\int_\Omega g = \int_\Omega f$. La relation $(f \mathcal{R} g) \Leftrightarrow (f = g$ presque partout sur $\Omega)$ étant une relation d'équivalence sur $\mathcal{L}^1(\Omega)$, on introduit l'espace quotient $L^1(\Omega) = \mathcal{L}^1(\Omega)/\mathcal{R}$. Un élément $f \in L^1(\Omega)$ est une classe de fonctions ; dans la pratique, l'objet f peut être vu comme une fonction qui est définie à un ensemble de mesure nulle près, c'est-à-dire qu'il est invariant si on change la valeur de $f(x)$ pour $x \in A$ et A est un ensemble de mesure nulle. Pour simplifier, un élément de $L^1(\Omega)$ est appelé une fonction. Par la suite, on considère également l'espace $\mathcal{L}^1_{\text{loc}}(\Omega)$ composé des fonctions intégrables sur tout compact de Ω et l'espace quotient $L^1_{\text{loc}}(\Omega) = \mathcal{L}^1_{\text{loc}}(\Omega)/\mathcal{R}$.

- **Espaces de Lebesgue.** Soit un réel $p \in [1, +\infty]$. On pose

$$L^p(\Omega) = \{f \text{ mesurable}\,;\, \|f\|_{L^p(\Omega)} < +\infty\}, \tag{A.34}$$

avec pour $p \in [1, +\infty[$,

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f|^p\right)^{\frac{1}{p}}, \tag{A.35}$$

et

$$\begin{aligned}\|f\|_{L^\infty(\Omega)} &= \sup_{x\in\Omega} \operatorname{ess} |f(x)| \\ &= \inf\{M \in \mathbb{R}_+\,;\, |f(x)| \leqslant M \text{ presque partout dans } \Omega\}.\end{aligned} \tag{A.36}$$

En particulier, on a

$$\|f\|_{L^1(\Omega)} = \int_\Omega |f|. \tag{A.37}$$

Théorème A.9 (Fischer–Riesz). *Soit $p \in [1, +\infty]$. Équipé de la norme $\|\cdot\|_{L^p(\Omega)}$, $L^p(\Omega)$ est un espace de Banach.*

Théorème A.10. *Soit un réel $p \in [1, +\infty[$. Alors, l'espace dual de $L^p(\Omega)$ peut être identifié avec $L^{p'}(\Omega)$ où $p' \in]1, +\infty]$ désigne le réel conjugué de p tel que $\frac{1}{p} + \frac{1}{p'} = 1$ (avec la convention que $p' = +\infty$ si $p = 1$).*

Pour $p = 2$, on obtient l'espace $L^2(\Omega)$ qui joue un rôle important dans l'analyse de la méthode des éléments finis. Cet espace peut être muni d'une structure hilbertienne grâce au produit scalaire

$$(f, g)_{L^2(\Omega)} = \int_\Omega fg. \tag{A.38}$$

Dans cet aide-mémoire, on préfère utiliser les notations $\|\cdot\|_{0,\Omega}$ pour $\|\cdot\|_{L^2(\Omega)}$ et $(\cdot,\cdot)_{0,\Omega}$ pour $(\cdot,\cdot)_{L^2(\Omega)}$, car ces notations sont cohérentes avec celles employées pour les espaces de Sobolev. En particulier, on a

$$\|f\|_{0,\Omega} = \left(\int_\Omega |f|^2\right)^{\frac{1}{2}}. \tag{A.39}$$

A.4 Distributions et espaces de Sobolev

- **Fonctions indéfiniment dérivables à support compact.** Soit Ω un ouvert de $\mathbb{R}^d$. Soit φ une fonction définie sur Ω à valeurs dans $\mathbb{R}$. Le *support* de φ dans Ω est défini comme l'ensemble

$$\text{support}(\varphi) = \overline{\{x \in \Omega;\ \varphi(x) \neq 0\}}^{\Omega}. \tag{A.40}$$

On désigne par $\mathcal{D}(\Omega)$ l'espace vectoriel constitué des fonctions de $\mathcal{C}^\infty(\Omega)$ dont le support dans Ω est compact. L'espace $\mathcal{D}(\Omega)$ est suffisamment riche pour satisfaire les deux résultats suivants.

Théorème A.11. *Soit un réel* $p \in [1, +\infty[$. *Alors,* $\mathcal{D}(\Omega)$ *est dense dans* $L^p(\Omega)$.

Théorème A.12. *Soit* $f \in L^1_{\text{loc}}(\Omega)$ *telle que* $\int_\Omega f\varphi = 0$ *pour tout* $\varphi \in \mathcal{D}(\Omega)$. *Alors,* $f = 0$.

- **Distributions.** On dit qu'une application linéaire

$$u : \mathcal{D}(\Omega) \ni \varphi \longmapsto \langle u, \varphi\rangle_{\mathcal{D}',\mathcal{D}} \in \mathbb{R}, \tag{A.41}$$

est une distribution sur Ω si la propriété suivante est satisfaite : pour tout compact K inclus dans Ω, il existe un entier n et une constante c tels que

$$\forall \varphi \in \mathcal{D}(\Omega),\ \text{supp}(\varphi) \subset K, \quad \langle u, \varphi\rangle_{\mathcal{D}',\mathcal{D}} \leqslant c \max_{\substack{\alpha\in\mathbb{N}^d \\ |\alpha|\leqslant n}} \left(\sup_{x\in K} |\partial^\alpha \varphi(x)| \right). \tag{A.42}$$

La notion de distribution sur Ω est une généralisation du concept de fonction sur Ω. En particulier, toute fonction $f \in L^1_{\text{loc}}(\Omega)$ peut être vue comme une distribution en posant pour tout $\varphi \in \mathcal{D}(\Omega)$,

$$\langle f, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \int_\Omega f\varphi. \tag{A.43}$$

Par contre, il existe des distributions sur Ω qui ne peuvent pas être représentées par des fonctions, par exemple la *masse de Dirac* δ_a définie pour tout $\varphi \in \mathcal{D}(\Omega)$ par $\langle \delta_a, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \varphi(a)$ où $a \in \Omega$.

- **Dérivation au sens des distributions.** Un des points d'orgue de la théorie des distributions est le fait que toute distribution est dérivable au sens suivant : soit $u \in \mathcal{D}'(\Omega)$ et soit $i \in \{1, \ldots, d\}$; alors, en posant

$$\partial_i u : \mathcal{D}(\Omega) \ni \varphi \longmapsto \langle \partial_i u, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = -\langle u, \partial_i \varphi\rangle_{\mathcal{D}',\mathcal{D}}, \tag{A.44}$$

on définit une distribution sur Ω. Plus généralement, si α est un multi-indice, en posant

$$\partial^\alpha u : \mathcal{D}(\Omega) \ni \varphi \longmapsto \langle \partial^\alpha u, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = (-1)^{|\alpha|} \langle u, \partial^\alpha \varphi\rangle_{\mathcal{D}',\mathcal{D}}, \tag{A.45}$$

on définit une distribution sur Ω. On pose conventionnellement $\nabla u = (\partial_1 u, \ldots, \partial_d u)^T$ et $\partial^0 u = u$.

La notion de dérivation au sens des distributions est une extension de la dérivation usuelle[1]. Si $u \in \mathcal{C}^1(\Omega)$, on vérifie facilement que sa dérivée classique et sa dérivée au sens des distributions coïncident. Dans cet aide-mémoire, il est implicitement entendu que *toutes les dérivées sont prises au sens des distributions*.

- **Espaces de Sobolev.** Soit un entier $s \geqslant 0$. L'espace de Sobolev $H^s(\Omega)$ est défini comme suit :

$$H^s(\Omega) = \{v \in \mathcal{D}'(\Omega) \,;\, \partial^\alpha v \in L^2(\Omega),\ |\alpha| \leqslant s\}. \tag{A.46}$$

Pour $s = 0$, on a $H^0(\Omega) = L^2(\Omega)$. On montre que pour tout $s \geqslant 0$, $H^s(\Omega)$ est un espace de Hilbert lorsqu'il est équipé du produit scalaire

$$(v, w)_{s,\Omega} = \sum_{|\alpha| \leqslant s} \int_\Omega \partial^\alpha v \, \partial^\alpha w. \tag{A.47}$$

La norme induite est notée

$$\|v\|_{s,\Omega} = \left(\sum_{|\alpha| \leqslant s} \|\partial^\alpha v\|_{0,\Omega}^2 \right)^{\frac{1}{2}}. \tag{A.48}$$

On considère également la semi-norme

$$|v|_{s,\Omega} = \left(\sum_{|\alpha| = s} \|\partial^\alpha v\|_{0,\Omega}^2 \right)^{\frac{1}{2}}. \tag{A.49}$$

La norme $\|\cdot\|_{s,\Omega}$ est équivalente à la norme $\sum_{|\alpha| \leqslant s} \|\partial^\alpha(\cdot)\|_{0,\Omega}$; dans cet aide-mémoire, on utilise indifféremment l'une ou l'autre de ces normes.

1. C'est pourquoi on utilise la même notation pour la dérivée au sens des distributions et la dérivée au sens usuel.

On introduit également l'espace

$$H_0^s(\Omega) = \overline{\mathcal{D}(\Omega)}^{H^s(\Omega)}. \tag{A.50}$$

Lorsque l'ouvert Ω est borné, cet espace est strictement inclus dans $H^s(\Omega)$.

Pour $s = 1$, on obtient l'espace de Hilbert

$$H^1(\Omega) = \{v \in L^2(\Omega)\,;\, \forall i \in \{1,\dots,d\},\ \partial_i v \in L^2(\Omega)\}, \tag{A.51}$$

avec le produit scalaire

$$(v,w)_{1,\Omega} = \int_\Omega vw + \int_\Omega \nabla v \cdot \nabla w = \int_\Omega vw + \sum_{i=1}^{d} \int_\Omega \partial_i v\, \partial_i w, \tag{A.52}$$

et la norme induite

$$\|v\|_{1,\Omega} = \left(\int_\Omega v^2 + \int_\Omega (\nabla v)^2\right)^{\frac{1}{2}} = \left(\|v\|_{0,\Omega}^2 + |v|_{1,\Omega}^2\right)^{\frac{1}{2}}. \tag{A.53}$$

La norme $\|\cdot\|_{1,\Omega}$ est équivalente à la norme $\|\cdot\|_{0,\Omega} + |\cdot|_{1,\Omega}$; dans cet aide-mémoire, on utilise indifféremment l'une ou l'autre de ces normes. Enfin, par définition, on a

$$H_0^1(\Omega) = \overline{\mathcal{D}(\Omega)}^{H^1(\Omega)}. \tag{A.54}$$

Une propriété importante des espaces de Sobolev est que

$$H^s(\Omega) \subset L^\infty(\Omega) \cap \mathcal{C}^{0,\alpha}(\overline{\Omega}) \qquad \text{si} \qquad s > \tfrac{d}{2}, \tag{A.55}$$

où $\alpha = 1 - \frac{d}{2s}$. La propriété (A.55) implique les fonctions de $H^s(\Omega)$ sont continues[1] dès que $s > \frac{d}{2}$. Ainsi, les fonctions de $H^1(\Omega)$ sont continues en dimension 1 et les fonctions de $H^2(\Omega)$ sont continues en dimension 2 et 3.

- **Théorie des traces.** La frontière $\partial\Omega$ étant un ensemble de mesure nulle, on ne peut pas donner un sens à $v|_{\partial\Omega}$ lorsque $v \in L^2(\Omega)$. Par contre, on peut

1. Plus rigoureusement, pour $v \in H^s(\Omega)$, il existe une fonction continue dans la classe de fonctions définies presque partout qui est représentée par v.

considérer l'application $\gamma_0 : \mathcal{C}^0(\overline{\Omega}) \ni v \longmapsto v|_{\partial\Omega} \in \mathcal{C}^0(\partial\Omega)$. On pose

$$H^{\frac{1}{2}}(\partial\Omega) = \left\{ v \in L^2(\partial\Omega) \, ; \, \frac{v(x) - v(y)}{\|x - y\|^{\frac{d+1}{2}}} \in L^2(\partial\Omega \times \partial\Omega) \right\}. \qquad \text{(A.56)}$$

On a le résultat remarquable suivant : si Ω est un ouvert borné de frontière lipschitzienne, on peut étendre l'application γ_0 à

$$\gamma_0 : H^1(\Omega) \ni v \longmapsto v|_{\partial\Omega} \in H^{\frac{1}{2}}(\partial\Omega), \qquad \text{(A.57)}$$

de sorte que :

(i) l'application γ_0 est surjective ;

(ii) le noyau de γ_0 est l'espace $H_0^1(\Omega)$.

De plus, du théorème de l'application ouverte on déduit qu'il existe une constante c telle que pour tout $g \in H^{\frac{1}{2}}(\partial\Omega)$, il existe $v_g \in H^1(\Omega)$ tel que $v_g|_{\partial\Omega} = g$ et $\|v_g\|_{1,\Omega} \leqslant c \, \|g\|_{\frac{1}{2},\partial\Omega}$.

On peut également étendre la notion de dérivée normale si on suppose que la frontière de Ω est de classe $\mathcal{C}^2$. Dans ces conditions, on montre que l'application $\gamma_1 : \mathcal{C}^1(\overline{\Omega}) \ni v \longmapsto (v|_{\partial\Omega}, n \cdot \nabla v|_{\partial\Omega}) \in \mathcal{C}^1(\partial\Omega) \times \mathcal{C}^0(\partial\Omega)$, où n est la normale extérieure à Ω, s'étend à $H^2(\Omega)$, que l'application étendue est surjective sur $H^{\frac{3}{2}}(\partial\Omega) \times H^{\frac{1}{2}}(\partial\Omega)$ et que son noyau est l'espace $H_0^2(\Omega)$. On a posé $H^{\frac{3}{2}}(\partial\Omega) = \{v \in H^1(\partial\Omega) \, ; \, \forall \alpha \text{ tel que } |\alpha| = 1, \, \partial^\alpha v \in H^{\frac{1}{2}}(\partial\Omega)\}$.

- **Inégalité de Poincaré et variantes.** Soit Ω un ouvert borné de $\mathbb{R}^d$. Il existe une constante ℓ_Ω telle que

$$\forall v \in H_0^1(\Omega), \qquad \|v\|_{0,\Omega} \leqslant \ell_\Omega \|\nabla v\|_{0,\Omega}. \qquad \text{(A.58)}$$

Cette inégalité est utilisée dans la section 5.1.1 pour prouver le caractère bien posé du problème de Dirichlet.

Une version un peu plus générale de l'inégalité de Poincaré consiste à considérer une forme linéaire continue sur $H^1(\Omega)$, que l'on désigne par f et dont le noyau ne contient pas les fonctions constantes sur Ω ; en d'autres termes, $f(\chi_\Omega) \neq 0$ où χ_Ω est la fonction indicatrice de Ω. On suppose que l'ouvert Ω est borné, connexe et de frontière lipschitzienne. Alors, il existe une

constante c_Ω telle que

$$c_\Omega \|v\|_{1,\Omega} \leqslant \|\nabla v\|_{0,\Omega} + |f(v)|. \tag{A.59}$$

De plus, si $f(\chi_\Omega) = 1$, en posant $Z = \{v \in H^1(\Omega)\,; f(v) = 0\}$, il existe une constante c^*_Ω telle que

$$\forall v \in Z, \qquad c^*_\Omega \|v\|_{1,\Omega} \leqslant \|\nabla v\|_{0,\Omega}. \tag{A.60}$$

Cette inégalité est connue sous le nom *d'inégalité de Poincaré–Friedrichs*. Le cas particulier où $f(v) = \frac{1}{\mathrm{mes}(\Omega)} \int_\Omega v$ conduit à *l'inégalité de Poincaré–Wirtinger*; voir le lemme 5.8.

NOMENCLATURE

Convention générale

Dans les estimations d'erreur, on désigne par c une constante générique dont la valeur numérique peut changer à chaque occurrence.

Vecteurs, matrices et espaces vectoriels

$(X_1, \ldots, X_m)$ ou $(X_i)_{1 \leqslant i \leqslant m}$	composantes du vecteur X dans une base de $\mathbb{R}^m$
$(X, Y)_m$	produit scalaire euclidien dans $\mathbb{R}^m$: $(X, Y)_m = \sum_{i=1}^m X_i Y_i$
$\|X\|_m$	norme euclidienne sur $\mathbb{R}^m$ (induite par le produit scalaire euclidien) : $\|X\|_m = (\sum_{i=1}^m X_i^2)^{\frac{1}{2}}$
$\mathbb{R}^{m,n}$	espace vectoriel des matrices à m lignes, n colonnes et à coefficients réels ; un coefficient générique d'une matrice $\mathcal{Z} \in \mathbb{R}^{m,n}$ est noté $\mathcal{Z}_{ij}$ pour $i \in \{1, \ldots, m\}$ et $j \in \{1, \ldots, n\}$
$\mathrm{Ker}(\mathcal{Z})$	noyau de la matrice $\mathcal{Z}$
$\mathrm{Im}(\mathcal{Z})$	image de la matrice $\mathcal{Z}$
$\mathcal{Z}^T$	matrice transposée de $\mathcal{Z}$: pour $\mathcal{Z} \in \mathbb{R}^{m,n}$, $\mathcal{Z}^T \in \mathbb{R}^{n,m}$ avec $\mathcal{Z}_{ij}^T = \mathcal{Z}_{ji}$ pour $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, m\}$
$\mathrm{tridiag}(a, b, c)$	matrice tridiagonale de terme générique $\delta_{i,i-1} a + \delta_{ii} b + \delta_{i,i+1} c$
$\mathcal{Z}X$	produit matrice–vecteur : pour $\mathcal{Z} \in \mathbb{R}^{m,n}$, $X \in \mathbb{R}^n$ et $i \in \{1, \ldots, m\}$, $(\mathcal{Z}X)_i = \sum_{j=1}^n \mathcal{Z}_{ij} X_j$

$\mathcal{A}$	matrice de rigidité ; voir (2.14) page 34
$\mathcal{M}$	matrice de masse ; voir (10.16) page 231
$\mathcal{I}_m$	matrice identité dans $\mathbb{R}^{m,m}$
$\text{vect}\{v_1, \ldots, v_m\}$	espace vectoriel engendré par la famille $\{v_1, \ldots, v_m\}$
$\dim(V)$	dimension de l'espace vectoriel V

Dans l'espace physique $\mathbb{R}^d$ (avec $d = 1$, 2 ou 3), on utilise également les notations suivantes.

$\sigma\cdot\xi$	produit matrice–vecteur : pour $\sigma \in \mathbb{R}^{d,d}$, $\xi \in \mathbb{R}^d$ et $i \in \{1, \ldots, d\}$, $(\sigma\xi)_i = \sum_{j=1}^d \sigma_{ij}\xi_j$
$\zeta\cdot\sigma\cdot\xi$	pour $\sigma \in \mathbb{R}^{d,d}$, $\zeta \in \mathbb{R}^d$ et $\xi \in \mathbb{R}^d$, $\zeta\cdot\sigma\cdot\xi = \sum_{i,j=1}^d \zeta_i\sigma_{ij}\xi_j$
$\sigma{:}\tau$	double contraction matricielle : pour $\sigma \in \mathbb{R}^{d,d}$ et $\tau \in \mathbb{R}^{d,d}$, $\sigma{:}\tau = \sum_{i,j=1}^d \sigma_{ij}\tau_{ij}$
$\zeta\otimes\xi$	produit tensoriel entre vecteurs : pour $\zeta \in \mathbb{R}^d$ et $\xi \in \mathbb{R}^d$, $\zeta\otimes\xi \in \mathbb{R}^{d,d}$ avec $(\zeta\otimes\xi)_{ij} = \zeta_i\xi_j$ pour $i,j \in \{1, \ldots, d\}$
$\zeta\times\xi$	produit vectoriel dans $\mathbb{R}^3$ de composantes $(\zeta_2\xi_3 - \zeta_3\xi_2, \zeta_3\xi_1 - \zeta_1\xi_3, \zeta_1\xi_2 - \zeta_2\xi_1)$
$\text{tr}(\sigma)$	trace de la matrice $\sigma \in \mathbb{R}^{d,d}$: $\text{tr}(\sigma) = \sum_{i=1}^d \sigma_{ii}$

Formes linéaires et bilinéaires

$\mathcal{L}(V;\mathbb{R})$ ou V'	espace vectoriel des formes linéaires continues sur un espace vectoriel normé V
$\|f\|_{V'}$	norme de f dans V' ; voir (A.7) page 313
$\mathcal{L}(V \times W;\mathbb{R})$	espace vectoriel des formes bilinéaires continues sur $V \times W$ où V et W sont des espaces vectoriels normés
$\|a\|_{V,W}$	norme de a dans $\mathcal{L}(V \times W;\mathbb{R})$; voir (A.12) page 314

Opérateurs différentiels

$(x_1, \ldots, x_d)$	coordonnées cartésiennes du point courant de $\mathbb{R}^d$

$\partial_i u$	dérivée (au sens des distributions) de u par rapport à x_i ; voir page 324
$\partial_i^m u$	dérivée (au sens des distributions) d'ordre m de u par rapport à x_i ; $m \in \mathbb{N}$
$\partial_{ij} u$	dérivée (au sens des distributions) de u par rapport à x_i et x_j
α	multi-indice : $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$ avec $\alpha_i \in \mathbb{N}$ pour tout $i \in \{1, \ldots, d\}$
$\lvert\alpha\rvert$	longueur du multi-indice α : $\lvert\alpha\rvert = \alpha_1 + \ldots + \alpha_d$
$\partial^\alpha u$	$\partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d} u$
∇u	gradient de u : $\nabla u = (\partial_1 u, \ldots, \partial_d u)^T \in \mathbb{R}^d$ si u est à valeurs dans $\mathbb{R}$ ou $\nabla u = (\partial_j u_i)_{1 \leqslant i \leqslant m,\, 1 \leqslant j \leqslant d} \in \mathbb{R}^{m,d}$ si u est à valeurs dans $\mathbb{R}^m$
$\nabla \cdot u$	divergence de u : $\nabla \cdot u = \sum_{i=1}^d \partial_i u_i$ où u est à valeurs dans $\mathbb{R}^d$
$\nabla \times u$	rotationnel de u : $\nabla \times u = (\partial_2 u_3 - \partial_3 u_2, \partial_3 u_1 - \partial_1 u_3, \partial_1 u_2 - \partial_2 u_1)^T$ où u est à valeurs dans $\mathbb{R}^3$
Δu	Laplacien de u : $\Delta u = \sum_{i=1}^d \partial_{ii} u$
∇_h	opérateur de gradient discret ; voir (4.21) page 81
$\nabla_h \cdot$	opérateur de divergence discrète ; voir (4.22) page 81

Fonctions et espaces fonctionnels

$u\lvert_E$	restriction de la fonction u à l'ensemble E
$\mathcal{C}^0(E)$	espace vectoriel des fonctions continues sur E
$\mathcal{C}^k(E)$	espace vectoriel des fonctions k-fois continûment différentiables sur E
$\mathcal{C}^{0,1}(E)$	espace vectoriel des fonctions lipschitziennes sur E
$\mathcal{C}^{k,\omega}(E)$	espace vectoriel des fonctions höldériennes sur E d'ordre k et d'exposant ω
$L^1(\Omega)$	espace vectoriel des fonctions intégrables (au sens de Lebesgue) sur Ω

$\|f\|_{L^1(\Omega)}$	norme de f dans $L^1(\Omega)$; voir (A.37) page 323
$L^2(\Omega)$	espace vectoriel des fonctions dont le carré est intégrable (au sens de Lebesgue) sur Ω
$\|f\|_{0,\Omega}$	norme de f dans $L^2(\Omega)$; voir (A.39) page 323
$L^\infty(\Omega)$	espace vectoriel des fonctions bornées presque partout sur Ω
$\|f\|_{L^\infty(\Omega)}$	norme de f dans $L^\infty(\Omega)$; voir (A.36) page 322
$\mathcal{D}(\Omega)$	espace vectoriel des fonctions de classe $\mathcal{C}^\infty$ sur Ω et dont le support dans Ω est compact
$\mathcal{D}'(\Omega)$	espace vectoriel des distributions sur Ω ; voir page 324
$H^s(\Omega)$	espace de Sobolev ; voir (A.46) page 325
$H^1(\Omega)$	espace de Sobolev ; voir (A.51) page 326
$H_0^1(\Omega)$	espace de Sobolev ; voir (A.54) page 326
$H^{\frac{1}{2}}(\partial\Omega)$	espace de Sobolev ; voir (A.56) page 327
$H(\mathrm{div};\Omega)$	$\{v \in [L^2(\Omega)]^d \; ; \; \nabla\cdot v \in L^2(\Omega)\}$
$H(\mathrm{rot};\Omega)$	$\{v \in [L^2(\Omega)]^3 \; ; \; \nabla\times v \in [L^2(\Omega)]^3\}$

Espaces polynômiaux

$\mathbb{P}_k$	espace vectoriel des polynômes en les variables $(x_1, \ldots, x_d)$, à coefficients réels et de degré global inférieur ou égal à k ; voir (3.9) page 50 ainsi que la définition 1.4 en dimension 1
$\mathbb{Q}_k$	espace vectoriel des polynômes en les variables $(x_1, \ldots, x_d)$, à coefficients réels et de degré inférieur ou égal à k en chaque variable ; voir (3.12) page 52
$\mathbb{RT}_0$	espace vectoriel des polynômes de Raviart–Thomas de plus bas degré ; voir (4.34) page 86
$\mathbb{N}_0$	espace vectoriel des polynômes de Nédélec de plus bas degré ; voir (4.49) et (4.50) page 90

Éléments finis et espaces d'approximation

$\{K, P, \Sigma\}$	élément fini ; voir la définition 4.1 page 75
n_{f}	nombre de fonctions de forme et de degrés de liberté de l'élément fini
$\{\theta_1, \ldots, \theta_{n_{\mathrm{f}}}\}$	fonctions de forme de l'élément fini
$\{\sigma_1, \ldots, \sigma_{n_{\mathrm{f}}}\}$	degrés de liberté de l'élément fini
$\{a_1, \ldots, a_{n_{\mathrm{f}}}\}$	nœuds d'un élément fini de Lagrange
$\mathcal{I}_K^{\mathrm{Lag}}$	opérateur d'interpolation local associé à un élément fini de Lagrange ; voir (3.4) page 46
$\mathcal{I}_K$	opérateur d'interpolation local associé à un élément fini $\{K, P, \Sigma\}$; voir (4.5) page 78
$V(K)$	domaine de l'opérateur d'interpolation local $\mathcal{I}_K$
$\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$	élément fini de référence pour la génération d'éléments finis
$\{\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_{\mathrm{f}}}\}$	fonctions de forme de l'élément fini de référence
$\{\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{n_{\mathrm{f}}}\}$	degrés de liberté de l'élément fini de référence
$\mathcal{I}_{\widehat{K}}$	opérateur d'interpolation local associé à l'élément fini de référence ; voir (4.7) page 78
$V(\widehat{K})$	domaine de l'opérateur d'interpolation local $\mathcal{I}_{\widehat{K}}$
$P^k_{\mathrm{c},h}$ et $P^k_{\mathrm{c},h,0}$	espaces d'éléments finis de Lagrange $\mathbb{P}_k$; voir (3.44) et (3.46) page 66
$Q^k_{\mathrm{c},h}$ et $Q^k_{\mathrm{c},h,0}$	espaces d'éléments finis de Lagrange $\mathbb{Q}_k$; voir (3.45) et (3.47) page 66
$P^k_{\mathrm{td},h}$	espace d'éléments finis totalement discontinu ; voir (4.20) page 81
$P^1_{\mathrm{pt},h}$ et $P^1_{\mathrm{pt},h,0}$	espaces d'éléments finis de Crouzeix–Raviart ; voir (4.31) page 84 et (5.15) page 108
D_h	espace d'éléments finis de Raviart–Thomas ; voir (4.42) page 88
R_h	espace d'éléments finis de Nédélec ; voir (4.56) page 93

Maillages et quadratures

$\mathcal{T}_h$	maillage ; voir (3.19) page 55
Ω_h	domaine recouvert par le maillage ; voir (3.22) page 56
$\{\mathcal{T}_h\}_{h>0}$	famille de maillages
$h_K = \text{diam}(K)$	diamètre de la maille $K \in \mathcal{T}_h$
h	diamètre maximum des mailles de $\mathcal{T}_h$: $h = \max_{K\in\mathcal{T}_h} h_K$
$\{\widehat{K}, \widehat{P}_{\text{géo}}, \widehat{\Sigma}_{\text{géo}}\}$	élément fini géométrique servant à générer le maillage ; voir la définition 3.15 page 57
T_K	transformation envoyant l'élément géométrique de référence $\widehat{K}$ dans une maille K
$n_{\text{géo}}$	nombre de fonctions de forme de l'élément fini géométrique
$N_{\text{géo}}$	nombre de nœuds géométriques dans le maillage
N_{ma}	nombre de mailles
N_{ar}, N_{ar}^{i}, N_{ar}^{∂}	nombre d'arêtes, d'arêtes internes et d'arêtes de frontière
N_{fa}, N_{fa}^{i}, N_{fa}^{∂}	nombre de faces, de faces internes et de faces de frontière
N_{so}, N_{so}^{i}, N_{so}^{∂}	nombre de sommets, de sommets intérieurs et de sommets de frontière
$\mathcal{E}_h$, $\mathcal{E}_h^{\text{i}}$, $\mathcal{E}_h^{\partial}$	ensemble des arêtes, des arêtes intérieures et des arêtes de frontière
$\mathcal{F}_h$, $\mathcal{F}_h^{\text{i}}$, $\mathcal{F}_h^{\partial}$	ensemble des faces, des faces intérieures et des faces de frontière
l_{q}	nombre de points de Gauß de la quadrature ; voir la définition 9.1 page 213
k_{q}	ordre de la quadrature
$\{\xi_{K,1}, \ldots, \xi_{K,l_{\text{q}}}\}$	points de Gauß pour la quadrature sur la maille K ; voir (9.6) page 215
$\{\omega_{K,1}, \ldots, \omega_{K,l_{\text{q}}}\}$	poids de la quadrature sur la maille K ; voir (9.5) page 215

Programmation des éléments finis

`connect_forme`	voir p. 294
`connect_g_front`	voir p. 291

`connect_géo`	voir p. 289
`coord`	voir p. 289
`dphi_dx`	voir p. 295
`dpsi_dx`	voir p. 292
`dtheta_dx`	voir p. 294
`i_cond_lim`	voir p. 291
`i_dom`	voir p. 292
`inv_jac`	voir p. 295
`poids`	voir p. 293
`psi`	voir p. 292
`theta`	voir p. 294
`vois`	voir p. 290

Autres symboles

$\mathrm{card}(E)$	cardinal de l'ensemble E
δ_{ij}	symbole de Kronecker : $\delta_{ij} = 1$ si $i = j$ et $\delta_{ij} = 0$ si $i \neq j$
$\mathcal{L}(E;F)$	espace vectoriel des opérateurs continus de E dans F où E et F sont des espaces vectoriels normés
$\mathrm{mes}(E)$	mesure (de Lebesgue) de l'ensemble $E \subset \mathbb{R}^d$; voir (A.29) page 321
$\overline{\Omega}$	adhérence de l'ensemble Ω ; voir page 312

BIBLIOGRAPHIE

[1] ABRAMOWITZ (M.) et STEGUN (I.), *Handbook of Mathematical Functions*. Dover, New York, NY, 9e édition, 1972.

[2] ACHCHAB (B.), AGOUZAL (A.), BARANGER (J.) et MAITRE (J.-F.), « Estimateur d'erreur a posteriori hiérarchique. Application aux éléments finis mixtes », *Numer. Math.*, vol. 80, p. 159–179, 1998.

[3] ADAMS (R.), *Sobolev Spaces*, vol. 65 (coll. *Pure and Applied Mathematics*). Academic Press, New York, NY, 1975.

[4] AINSWORTH (M.) et ODEN (J.), *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, New York, NY, 2000.

[5] AMROUCHE (C.) et GIRAULT (V.), « On the existence and regularity of the solution of Stokes problem in arbitrary dimension », *Proc. Japan Acad.*, vol. 67, p. 171–175, 1991.

[6] ARNOLD (D.), BREZZI (F.) et FORTIN (M.), « A stable finite element for the Stokes equations », *Calcolo*, vol. 21, p. 337–344, 1984.

[7] AUBIN (J.-P.), « Approximation des problèmes aux limites non homogènes pour des opérateurs non linéaires », *J. Math. Anal. Appl.*, vol. 30, p. 510–521, 1970.

[8] AUBIN (J.-P.), *Applied Functional Analysis*. Pure and Applied Mathematics. Wiley-Interscience, New York, NY, 2e édition, 2000.

[9] BABUŠKA (I.) et AZIZ (A.), « Survey lectures on the mathematical foundations of the finite element method », dans AZIZ (A.), éditeur, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*. Academic Press, New York, NY, 1972.

[10] BABUŠKA (I.) et RHEINBOLT (W.), « Error estimates for adaptive finite element method computations », *SIAM J. Numer. Anal.*, vol. 15, p. 736–754, 1978.

[11] BANK (R.) et SMITH (R.), « A posteriori error estimates based on hierarchical bases », *SIAM J. Numer. Anal.*, vol. 30, n° 4, p. 921–935, 1993.

[12] BANK (R.) et WEISER (A.), « Some a posteriori error estimators for elliptic partial differential equations », *Math. Comp.*, vol. 44, p. 283–301, 1985.

[13] BARTLE (R.), *A Modern Theory of Integration*, vol. 32 (coll. *Graduate Studies in Mathematics*). American Mathematical Society, Providence, RI, 2001.

[14] BECKER (R.) et RANNACHER (R.), « An optimal control approach to a posteriori error estimation in finite element methods », *Acta Numerica*, vol. 10, p. 1–102, 2001.

[15] BERCOVIER (M.) et PIRONNEAU (O.), « Error estimates for finite element solution of the Stokes problem in the primitive variables », *Numer. Math.*, vol. 33, p. 211–224, 1979.

[16] BOSSAVIT (A.), *Electromagnétisme en vue de la modélisation*, vol. 14 (coll. *Mathématiques et Applications*). Springer-Verlag, Paris, France, 1993. Voir également *Computational Electromagnetism, Variational Formulations, Complementary, Edge Elements*, Academic Press, New York, NY, 1998.

[17] BRAACK (M.) et ERN (A.), « A posteriori control of modeling errors and discretization errors », *Multiscale Model. Simul.*, vol. 1, n° 2, p. 221–238, 2003.

[18] BRAESS (D.), *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, Royaume-Uni, 2e édition, 1997.

[19] BRANDT (A.), « Multi-level adaptive solutions to boundary value problems », *Math. Comp.*, vol. 31, p. 333–390, 1977.

[20] BRENNER (S.) et SCOTT (R.), *The Mathematical Theory of Finite Element Methods*, vol. 15 (coll. *Texts in Applied Mathematics*). Springer, New York, NY, 1994.

[21] BREZIS (H.), *Analyse fonctionnelle. Théorie et applications*. Mathématiques appliquées pour la maîtrise. Masson, Paris, France, 1983.

[22] BREZZI (F.) et FORTIN (M.), *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York, NY, 1991.

[23] BRIGGS (W.), *A Multigrid Tutorial*. SIAM, Philadelphia, PA, 1987.

[24] BROOKS (A.) et HUGHES (T.), « Streamline Upwind/Petrov–Galerkin formulations for convective dominated flows with particular emphasis on the incompressible Navier–Stokes equations », *Comput. Methods Appl. Mech. Engrg.*, vol. 32, p. 199–259, 1982.

[25] BURDEN (R.) et FAIRES (J.), *Numerical Analysis*. PWS Publishing Company, Boston, MA, 5e édition, 1993.

[26] BURMAN (E.) et ERN (A.), « Stabilized Galerkin approximation of convection–diffusion–reaction equations : discrete maximum principle and convergence », *Math. Comp.*, 2005.

[27] CIARLET (P.), *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, Pays-Bas, 1978.

[28] CIARLET (P.), « Basic error estimates for elliptic problems », dans CIARLET (P.) et LIONS (J.-L.), éditeurs, *Finite Element Methods*, vol. II (coll. *Handbook of Numerical Analysis*), chap. 2. North-Holland, Amsterdam, Pays-Bas, 1991.

[29] CLÉMENT (P.), « Approximation by finite element functions using local regularization », *RAIRO, Anal. Num.*, vol. 9, p. 77–84, 1975.

[30] CODINA (R.), « Comparison of some finite element methods for solving the diffusion–convection–reaction equations », *Comput. Methods Appl. Mech. Engrg.*, vol. 156, p. 185–210, 1998.

[31] CROISILLE (J.-P.), « Finite volume box schemes and mixed methods », *ESAIM Math. Model. Numer. Anal.*, vol. 34, n° 2, p. 1087–1106, 2000.

[32] CROUZEIX (M.) et RAVIART (P.-A.), « Conforming and nonconforming finite element methods for solving the stationnary Stokes equations », *RAIRO, Anal. Num.*, vol. 3, p. 33–75, 1973.

[33] DAUTRAY (R.) et LIONS (J.-L.), *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 4. Integral equations and numerical methods*. Springer-Verlag, Berlin, Allemagne, 1990.

[34] DAVIS (P.) et RABINOWITZ (P.), *Methods of Numerical Integration*. Academic Press, New York, NY, 2e édition, 1984.

[35] EL ALAOUI (L.) et ERN (A.), « Residual and hierarchical a posteriori error estimates for nonconforming mixed finite element methods », *ESAIM Math. Model. Numer. Anal.*, vol. 38, n° 6, p. 903–929, 2004.

[36] ERN (A.) et GUERMOND (J.-L.), *Éléments finis : théorie, applications, mise en œuvre*, vol. 36 (coll. *SMAI Mathématiques & Applications*). Springer-Verlag, Paris, France, 2002.

[37] ERN (A.) et GUERMOND (J.-L.), « Evaluation of the condition number in linear systems arising in finite element approximations », 2004. Rapport de recherche 2004–265, Cermics, Ecole nationale des ponts et chaussées, Champs-sur-Marne, France.

[38] ERN (A.) et GUERMOND (J.-L.), *Theory and Practice of Finite Elements*, vol. 159 (coll. *Applied Mathematical Sciences*). Springer-Verlag, New York, NY, 2004.

[39] FORMAGGIA (L.) et PEROTTO (S.), « New anisotropic a priori error estimates », *Numer. Math.*, vol. 89, p. 641–667, 2001.

[40] FRANCA (L.) et FREY (S.), « Stabilized finite element methods. II. The incompressible Navier–Stokes equations », *Comput. Methods Appl. Mech. Engrg.*, vol. 99, p. 209–233, 1992.

[41] FREY (P.) et GEORGE (P.-L.), *Maillages*. Hermès, Paris, France, 1999.

[42] GALDI (G.), *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol. I*, vol. 38 (coll. *Springer Tracts in Natural Philosophy*). Springer-Verlag, New York, NY, 1994.

[43] GEORGE (A.) et LIU (J.), *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1981. Prentice-Hall Series in Computational Mathematics.

[44] GEORGE (P.-L.) et BOROUCHAKI (H.), *Delaunay Triangulation and Meshing. Application to Finite Elements*. Editions Hermès, Paris, France, 1998.

[45] GILBARG (D.) et TRUDINGER (N.), *Elliptic Partial Differential Equations of Second Order*, vol. 4 (coll. *Classics in Mathematics*). Springer-Verlag, Berlin, Allemagne, 2001.

[46] GIRAULT (V.) et RAVIART (P.-A.), *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, vol. 5 (coll. *Springer Series in Computational Mathematics*). Springer-Verlag, Berlin, Allemagne, 1986.

[47] GOLUB (G.) et VAN LOAN (C.), *Matrix Computations*. John Hopkins University Press, Baltimore, MD, 2e édition, 1989.

[48] GRISVARD (P.), *Singularities in Boundary Value Problems*. Masson, Paris, France, 1992.

[49] GUERMOND (J.-L.), « Stabilization of Galerkin approximations of transport equations by subgrid modeling », *ESAIM Math. Model. Numer. Anal.*, vol. 33, no 6, p. 1293–1316, 1999.

[50] GUERMOND (J.-L.), « Subgrid stabilization of Galerkin approximations of linear monotone operators », *IMA J. Numer. Anal.*, vol. 21, p. 165–197, 2001.

[51] HACKBUSCH (W.), *Multigrid Methods and Applications*, vol. 4 (coll. *Springer Series in Computational Mathematics*). Springer-Verlag, Berlin, Allemagne, 1985.

[52] JOHNSON (C.), *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, Cambridge, Royaume-Uni, 1987.

[53] JOHNSON (C.), NÄVERT (U.) et PITKÄRANTA (J.), « Finite element methods for linear hyperbolic equations », *Comput. Methods Appl. Mech. Engrg.*, vol. 45, p. 285–312, 1984.

[54] KARNIADAKIS (G.) et SPENCER (J.), *Spectral/hp Element Methods for CFD*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, NY, 1999.

[55] KNABNER (P.) et ANGERMANN (L.), *Numerical Methods for Elliptic and Parabolic Partial Differential Equations*, vol. 44 (coll. *Texts in Applied Mathematics*). Springer-Verlag, New York, NY, 2003.

[56] LASCAUX (P.) et THEODOR (R.), *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, vol. I et II. Masson, Paris, France, 2e édition, 1993.

[57] MCCORMICK (S.), *Multigrid Methods*, vol. 3 (coll. *SIAM Frontiers Series*). SIAM, Philadelphia, PA, 1987.

[58] NEČAS (J.), « Sur une méthode pour résoudre les équations aux dérivées partielles de type elliptique, voisine de la variationnelle », *Ann. Scuola Norm. Sup. Pisa*, vol. 16, p. 305–326, 1962.

[59] NÉDÉLEC (J.-C.), « A new family of mixed finite elements in $\mathbb{R}^3$ », *Numer. Math.*, vol. 50, p. 57–81, 1986.

[60] ORTEGA (J.), *Matrix Theory, a Second Course*. Plenum, New York, NY, 1987.

[61] QUARTERONI (A.) et VALLI (A.), *Numerical Approximation of Partial Differential Equations*, vol. 23 (coll. *Springer Series in Computational Mathematics*). Springer-Verlag, Berlin, Allemagne, 2e édition, 1997.

[62] RAVIART (P.-A.) et THOMAS (J.-M.), « A mixed finite element method for second order elliptic problems », dans GALLIGANI (I.) et MEGENES (E.), éditeurs, *Mathematical Aspects of the Finite Element Method*, vol. 606 (coll. *Lecture Notes in Mathematics*). Springer–Verlag, New York, NY, 1977.

[63] RAVIART (P.-A.) et THOMAS (J.-M.), *Introduction à l'analyse numérique des équations aux dérivées partielles*. Masson, Paris, France, 1983.

[64] RODRÍGUEZ (R.), « Some remarks on the Zienkiewicz–Zhu estimator », *Numer. Methods Partial Differential Equations*, vol. 10, n° 5, p. 625–635, 1994.

[65] RUDIN (W.), *Analyse réelle et complexe*. Masson, Paris, France, 4e édition, 1987.

[66] SAAD (Y.), *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, MA, 1996.

[67] SAAD (Y.) et SCHULTZ (M.), « GMRES : a generalized minimal residual algorithm for solving nonsymmetric linear systems », *SIAM J. Sci. Statist. Comput.*, vol. 7, p. 856–869, 1986.

[68] SCHIEWECK (F.), « A posteriori error estimates with post-processing for nonconforming finite elements », *ESAIM Math. Model. Numer. Anal.*, vol. 36, n° 3, p. 489–503, 2002.

[69] SCOTT (R.) et ZHANG (S.), « Finite element interpolation of nonsmooth functions satisfying boundary conditions », *Math. Comp.*, vol. 54, n° 190, p. 483–493, 1990.

[70] SIEBERT (K.), « An a posteriori error estimator for anisotropic refinement », *Numer. Math.*, vol. 73, p. 373–398, 1996.

[71] STROUD (A.), *Approximate Calculation of Multiple Integrals*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[72] TEMAM (R.), *Navier–Stokes Equations*, vol. 2 (coll. *Studies in Mathematics and its Applications*). North-Holland, Amsterdam, Pays-Bas, 1977.

[73] TOBISKA (L.) et VERFÜRTH (R.), « Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations », *SIAM J. Numer. Anal.*, vol. 33, nº 1, p. 107–127, 1996.

[74] VAN DER VORST (H.), « Bi-CGStab : a more stably converging variant of CG-S for the solution of nonsymmetric linear systems », *SIAM J. Sci. Statist. Comput.*, vol. 13, p. 631–644, 1992.

[75] VERFÜRTH (R.), « Error estimates for a mixed finite element approximation of the Stokes equation », *RAIRO, Anal. Num.*, vol. 18, p. 175–182, 1984.

[76] VERFÜRTH (R.), *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley, Chichester, Royaume-Uni, 1996.

[77] XU (J.) et ZIKATANOV (L.), « A monotone finite element scheme for convection–diffusion equations », *Math. Comp.*, vol. 68, nº 228, p. 1429–1446, 1999.

[78] YOSIDA (K.), *Functional Analysis*. Classics in Mathematics. Springer-Verlag, Berlin, Allemagne, 1995. Retirage de la 6e édition (1980).

[79] ZEIDLER (E.), *Applied Functional Analysis*, vol. 108 (coll. *Applied Mathematical Sciences*). Springer-Verlag, New York, NY, 1995.

[80] ZIENKIEWICZ (O.) et ZHU (J.), « A simple error estimator and adaptive procedure for practical engineering analysis », *Int. J. Numer. Methods Engrg.*, vol. 24, p. 337–357, 1987.

INDEX

D

E

F

Q

R

S

AIDE-MÉMOIRE DE L'INGÉNIEUR

Alexandre Ern

ÉLÉMENTS FINIS

Cet aide-mémoire présente les fondements théoriques de la méthode des éléments finis, des applications aux sciences de l'ingénieur et les bases de sa mise en œuvre numérique, en abordant successivement :

- les principales notions dans le cadre introductif des éléments finis unidimensionnels ;
- les fondements théoriques de la méthode, notamment la méthode de Galerkin et les propriétés interpolantes des éléments finis usuellement rencontrés dans les simulations numériques ;
- diverses applications de la méthode ;
- les techniques d'estimation d'erreur *a posteriori* ;
- la mise en œuvre numérique de la méthode et sa programmation.

Cet ouvrage constitue un outil de travail incontournable pour les ingénieurs en bureaux d'études et pour les élèves-ingénieurs et étudiants de niveau master dans le domaine.

ALEXANDRE ERN

est professeur à l'École nationale des ponts et chaussées. Il y préside le département de première année et est responsable des enseignements de calcul scientifique et éléments finis. Il anime l'équipe « Mécanique des fluides » au Cermics. Il est l'auteur de trois ouvrages et d'une quarantaine de publications en mathématiques appliquées et en modélisation numérique.

L'USINE NOUVELLE

ISBN 2 10 007303 6

www.dunod.com

DUNOD