La méthode des éléments finis : de la théorie à la pratique. Tome 1 : Concepts généraux

Patrick Ciarlet, Éric Lunéville

Les cours

P. Ciarlet & E. Lunéville

La méthode des éléments finis : de la théorie à la pratique

I. Concepts généraux

PARIS
LES PRESSES DE L'ENSTA
32, boulevard Victor, Paris 15^{e}
2009

Table des matières

Avant-propos

La simulation numérique est devenue un puissant moyen d'investigation qui tend à prendre une place de plus en plus importante, à côté de l'approche expérimentale classique, dans les sciences et techniques. Dans les domaines de la mécanique, de la physique ou de la chimie, l'utilisation de modèles fondés sur des systèmes d'équations aux dérivées partielles est naturelle (*élastodynamique, acoustique, électromagnétisme, aérodynamique, hydrodynamique, chimie ab-initio...*). Plus récemment, ce type de modélisation a fait son apparition dans d'autres domaines tels que la biologie et l'économie (finance en particulier). Hormis quelques cas particuliers, il n'est pas possible de résoudre analytiquement ces systèmes d'équations, et il est donc obligatoire d'avoir recours à des techniques d'approximation. Il en existe un certain nombre : méthodes semi-analytiques, méthodes probabilistes, méthodes des différences finies, méthodes spectrales et méthode des éléments finis pour ne citer que les principales. La méthode des éléments finis, introduite dans les années 1950, a connu depuis de nombreux développements, et est aujourd'hui présente dans de nombreux logiciels de simulation numérique. Tout ingénieur travaillant dans un environnement où la simulation numérique est un outil important, y sera confronté et il doit donc en connaître les principes fondamentaux, voire les derniers raffinements.

Comme toute technique d'approximation, la méthode des éléments finis doit être employée en respectant des règles, incluant notamment des contraintes d'utilisation ainsi que des facteurs de qualité de différentes natures. L'objectif de ce cours est de fournir les éléments nécessaires, allant des plus théoriques au plus concrets, à la maîtrise de ces contraintes et de ces facteurs. Dans cette optique, nous avons souhaité présenter les différents aspects de la méthode : cadre mathématique, analyse numérique et mise en œuvre efficiente en accordant autant d'importance à chacun de ces aspects et en fournissant les points clés ainsi que des éléments supplémentaires correspondant aux situations que l'on rencontre souvent en pratique. Cet ouvrage constitue le premier tome d'un cours dispensé à l'Ecole Nationale Supérieure de Techniques Avancées. Il est consacré à la présentation

des concepts fondamentaux de la méthode des éléments finis pour des problèmes stationnaires elliptiques. Dans un second tome sont présentés des approfondissements de la méthode des éléments finis, permettant en particulier, d'aborder les problèmes de valeurs propres, les problèmes transitoires ainsi que les problèmes mixtes.

Le premier chapitre est consacré à l'étude des problèmes *elliptiques* et, en particulier, à la *théorie variationnelle* des équations elliptiques (contexte de la méthode des éléments finis) qui est exposée dans un cadre fonctionnel rigoureux et requiert des connaissances élémentaires d'analyse. La plupart des outils opérationnels d'analyse fonctionnelle (analyse hilbertienne, théorie des distributions) sont rappelés, permettant à ceux qui ne les connaissent pas une lecture plus aisée. La méthode des éléments finis fait l'objet des chapitres 2 et 3. Nous en donnons, tout d'abord, une présentation concrète sur un exemple à la fois simple et significatif, et dans un second temps, un cadre formel qui permet de construire une grande variété d'éléments finis. Les principaux résultats d'*estimations d'erreurs* sont détaillés. Au chapitre 3, nous étudions les aspects pratiques et algorithmiques de cette méthode, liés à sa *mise en œuvre informatique*, et des *illustrations numériques* réalisées avec Matlab[1] sont fournies à titre d'exemple. Enfin, dans une annexe, nous présentons quelques considérations élémentaires – mais indispensables – sur la *résolution des systèmes linéaires* et, en particulier, les systèmes linéaires *creux* issus de l'approximation par éléments finis.

[1] Matlab est une marque déposée par The MathWorks, Inc.

1

Du côté de la théorie des équations elliptiques

1.1 Un peu de vocabulaire

De nombreux régimes établis de systèmes physiques (i.e. l'état atteint lorsque $t \to \infty$) se modélisent à l'aide d'équations aux dérivées partielles de nature elliptique[1]. Ainsi,

– la diffusion de la chaleur se modélise à l'aide de l'équation :

$$\boxed{\text{div } (k\nabla T) \;=\; 0}$$

où T désigne la température, k la conductivité thermique ;

– le mouvement élastique d'un corps homogène isotrope linéaire conduit au modèle :

$$\boxed{\sum_{j=1}^{n} \frac{\partial}{\partial x_j}\, \sigma_{ij}(\mathbf{u}) = 0,\ 1 \leq i \leq n}$$

où $\mathbf{u}$ désigne le vecteur déplacement dans $\mathbb{R}^n$ (habituellement $\mathbb{R}^3$) et

$$\sigma_{ij}(\mathbf{u}) = \lambda \left(\sum_{k=1}^{n} \varepsilon_{kk}(\mathbf{u}) \right) \delta_{ij} + 2\mu\varepsilon_{ij}(\mathbf{u})$$
$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)$$

sont respectivement les tenseurs des contraintes et des déformations, λ et μ les coefficients de Lamé, δ_{ij} le symbole de Kronecker ;

– l'équilibre électrostatique dans un corps conducteur est gouverné par l'équation :

$$\boxed{-\Delta\varphi = \frac{\rho}{\varepsilon}}$$

[1] On trouvera de nombreux exemples détaillés dans [13].

où φ désigne le potentiel électrostatique, ε la constante diélectrique du matériau, et ρ la densité de charge ;

– la flexion des poutres ou des plaques élastiques se modélise à l'aide d'un opérateur du quatrième ordre :

$$\boxed{-\Delta^2 u = 0}$$

u désignant le déplacement transverse de la plaque ou de la poutre et $\Delta^2 u = \Delta(\Delta u)$;

– la propagation des ondes acoustiques dans un milieu compressible homogène, en régime harmonique (dépendance en $\cos(\omega t)$ ou en $\sin(\omega t)$), conduit à l'équation de Helmholtz :

$$\boxed{\Delta p + \omega^2 p = 0}$$

où p désigne la pression acoustique.

Aucune de ces équations ne présente de courbe caractéristique. Par conséquent, les propriétés, les techniques d'étude et d'approximation leurs seront spécifiques et bien différentes de celles utilisées dans le cas hyperbolique (voir par exemple [15]). Afin d'appréhender ces techniques propres aux équations de nature elliptique, nous nous focaliserons sur l'exemple simple et académique que constitue l'équation de Laplace-Poisson :

$$\boxed{-\Delta u = f \ \text{ dans } \Omega} \tag{1.1}$$

où Ω est un ouvert borné[2] de $\mathbb{R}^n (n \geq 1)$, de frontière notée $\partial\Omega$, qu'on suppose "suffisamment régulière"[3], et f est une fonction donnée sur Ω.

Après quelques considérations sur les conditions aux limites que l'on peut adjoindre à l'équation de Laplace, nous introduisons la notion de solution classique (c'est-à-dire deux fois dérivable au sens classique). On peut facilement démontrer des résultats d'unicité de ces solutions, mais il est très difficile de prouver l'existence de telles solutions. C'est pourquoi, nous exposons en détail dans les sections 1.3 et 1.4, la théorie variationnelle des équations elliptiques qui permet d'accéder à des solutions dites faibles (fonctions dont les dérivées au sens des distributions sont de carré intégrable). Enfin, les propriétés fondamentales des solutions faibles (positivité, principe du maximum, régularité) sont indiquées à la section 1.5.

[2] Le cas des ouverts non bornés soulève des difficultés supplémentaires liées à des conditions de comportement des solutions à l'infini.

[3] Cette notion sera précisée par la suite, voir la définition 1.10.

• Conditions aux limites pour l'équation de Laplace

Il est clair que si v est une fonction harmonique régulière (i.e. $\Delta v = 0,\ v \in \mathcal{C}^2(\Omega)$) alors la fonction $u+v$ est encore solution de (1.1). C'est pourquoi il est nécessaire d'adjoindre des conditions aux limites sur $\partial\Omega$ à l'équation (1.1) pour pouvoir espérer recouvrer l'unicité de u.

Classiquement, on considère trois types de conditions aux limites :

$$u = g \ \text{ sur } \partial\Omega \qquad \text{condition de Dirichlet} \tag{1.2}$$

$$\frac{\partial u}{\partial n} = g \ \text{ sur } \partial\Omega \qquad \text{condition de Neumann} \tag{1.3}$$

$$\frac{\partial u}{\partial n} + \lambda u = g \ \text{ sur } \partial\Omega \qquad \text{condition de Fourier} \tag{1.4}$$

où $\mathbf{n}$ désigne le vecteur normal unitaire sortant à $\partial\Omega$, $\dfrac{\partial u}{\partial n} = \nabla u \cdot \mathbf{n}$, $\lambda \in \mathbb{R}$ et g est une fonction donnée sur $\partial\Omega$. Donnons à titre indicatif l'interprétation de ces

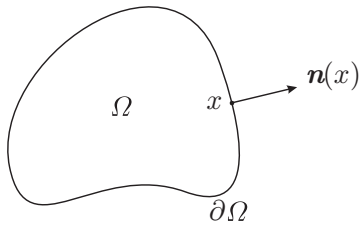


Fig. 1.1. Domaine Ω et choix de la normale

conditions en thermique.
– La condition de Dirichlet correspond à une température imposée sur la frontière, par exemple $T = T_{ext}$ (T_{ext} température extérieure) ;
– la condition de Neumann s'interprète comme une condition de flux de chaleur à travers $\partial\Omega$ ($g = 0$ paroi isolée, $g > 0$ apport d'énergie thermique à la paroi, $g < 0$ perte d'énergie thermique) ;
– la condition de Fourier traduit des pertes énergétiques proportionnelles à l'écart de température si $\lambda > 0$:

$$\frac{\partial(T - T_{ext})}{\partial n} = -\lambda(T - T_{ext}) \quad \text{(loi de Fourier).}$$

Lorsque l'on adjoint respectivement les conditions aux limites (1.2), (1.3) et (1.4) à l'équation (1.1) on obtient des problèmes aux limites que l'on désignera par la

suite respectivement par P_D (problème de Dirichlet), P_N (problème de Neumann) et P_F (problème de Fourier). On parlera de problème homogène lorsque la donnée g est nulle (pour les problèmes de Dirichlet ou de Neumann).

On rencontre également des problèmes où les conditions aux limites sont de nature différente sur des portions distinctes de la frontière. On parle alors de problème aux limites mixtes. Par exemple, le problème mixte de Dirichlet-Neumann prend la forme suivante :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g_1 & \text{sur } \Gamma_1 \\ \dfrac{\partial u}{\partial n} = g_2 & \text{sur } \Gamma_2 \end{cases} \tag{1.5}$$

On supposera toujours que $\overline{\Gamma_1 \cup \Gamma_2} = \partial\Omega$ et $\Gamma_1 \cap \Gamma_2 = \emptyset$.

• Autres conditions aux limites

Comme nous le verrons, dans le cas d'équations elliptiques du deuxième ordre (Δ par exemple) une condition aux limites sur $\partial\Omega$ suffit pour que le problème soit bien posé (unicité-existence) alors que pour les équations du quatrième ordre (Δ^2 par exemple) il faut imposer deux conditions aux limites sur $\partial\Omega$ pour que le problème soit bien posé, par exemple :

$$u = g_1 \text{ et } \frac{\partial u}{\partial n} = g_2 \qquad \text{sur } \partial\Omega$$

mais d'autres conditions sont également possibles ($\dfrac{\partial}{\partial n}(\Delta u) = g$, par exemple). Leur traitement est alors similaire au cas que nous allons étudier.

Signalons que l'on rencontre parfois des conditions aux limites "obliques" dans $\mathbb{R}^2$:

$$\alpha\frac{\partial u}{\partial n} + \beta\frac{\partial u}{\partial \tau} = g$$

où τ désigne un vecteur unitaire tangent à la frontière. L'étude de l'équation (1.1) munie de cette condition limite est délicate et nous ne l'aborderons pas ici (voir [18, 19]).

Enfin les conditions aux limites d'ordre supérieur à 1 pour l'équation (1.1) (par exemple $\partial_n^2 u = g$ sur $\partial\Omega$) conduisent à des problèmes mal posés (non existence) sauf cas très particulier.

Dans la section 1.2 nous présentons quelques résultats élémentaires liés aux solutions classiques. La théorie variationnelle des équations elliptiques fait l'objet des sections 1.3 et 1.4. Enfin, nous indiquons succinctement, dans la section 1.5, quelques propriétés importantes des solutions faibles.

1.2 Solutions classiques au problème de Laplace

Commençons par préciser la notion de solution classique (i.e. dérivable au sens classique) des problèmes de Dirichlet, Neumann ou Fourier. Pour que les conditions aux limites aient un sens, il faut imposer à la solution des propriétés de régularité sur la frontière $\partial\Omega$.

Définition 1.1. *On appelle solution classique du problème de Dirichlet (P_D) toute fonction $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ qui vérifie (P_D).*

Définition 1.2. *On appelle solution classique du problème de Neumann (P_N) (resp. Fourier (P_F)) toute fonction $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1\left(\overline{\Omega}\right)$ qui vérifie (P_N) (resp. (P_F)).*

Les démonstrations d'existence et d'unicité de solutions classiques s'appuient sur des principes du maximum qui sont liés aux propriétés des fonctions harmoniques.

1.2.1 Formules de Green

Les principales propriétés des fonctions harmoniques dérivent des formules de Green suivantes :

Proposition 1.1. (Formules de Green)
Soient $\mathbf{w} \in \left(\mathcal{C}^1(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)\right)^n$ et $v \in \mathcal{C}^1(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ on a :

$$\int_\Omega (\operatorname{div} \mathbf{w})\, v \, d\Omega = -\int_\Omega \mathbf{w} \cdot \nabla v \, d\Omega + \int_{\partial\Omega} \mathbf{w} \cdot \mathbf{n}\, v \, d\Gamma. \tag{1.6}$$

Soient $u, v \in \mathcal{C}^1(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ on a pour tout $i = 1$ à n :

$$\int_\Omega \frac{\partial u}{\partial x_i}\, v \, d\Omega = -\int_\Omega u \, \frac{\partial v}{\partial x_i}\, d\Omega + \int_{\partial\Omega} u\, v\, n_i \, d\Gamma \quad (n_i : i^{\text{ème}} \text{ composante de } n). \tag{1.7}$$

Soient $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1\left(\overline{\Omega}\right)$ et $v \in \mathcal{C}^1(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ on a :

$$\int_\Omega \Delta u\, v \, d\Omega = -\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\Gamma. \tag{1.8}$$

Soient $u, v \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1\left(\overline{\Omega}\right)$ on a :

$$\int_\Omega \Delta u\, v \, d\Omega = \int_\Omega u\, \Delta v \, d\Omega + \int_{\partial\Omega} \left(\frac{\partial u}{\partial n}\, v - \frac{\partial v}{\partial n}\, u \right) d\Gamma. \tag{1.9}$$

La formule (1.6) constitue la formule fondamentale de Stokes, dont la démonstration sort du cadre de ce cours (voir e.g. [1]). La formule (1.7) est une conséquence immédiate de (1.6) (prendre $\mathbf{w} = u\mathbf{e}_i$). La formule (1.8), dite première formule de Green, se déduit de (1.6) en prenant $\mathbf{w} = \nabla u$ et la formule (1.9), dite seconde formule de Green, découle trivialement de (1.8). Ces résultats sont valables moyennant des hypothèses de régularité de la frontière !

Considérons maintenant, une fonction harmonique $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ c'est-à-dire vérifiant :

$$\Delta u = 0 \quad \text{dans } \Omega, \tag{1.10}$$

on a alors les formules de moyenne suivantes :

Proposition 1.2. (Formules de la moyenne sphérique) *Pour toute boule B, de centre y et de rayon R, telle que $B \subset \Omega$ on a :*

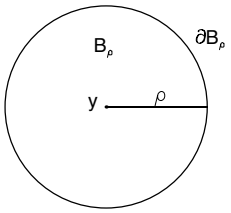
$$u(y) = \frac{1}{\omega_n R^{n-1}} \int_{\partial B} u d\Gamma \tag{1.11}$$

$$u(y) = \frac{n}{\omega_n R^n} \int_B u d\Omega \tag{1.12}$$

où ω_n désigne la mesure (i.e. la surface) de la sphère unité de $\mathbb{R}^n$.

Démonstration : soit $0 \leq \rho \leq R$, appliquons la formule de Green (1.8) dans la boule B_ρ de centre y et de rayon ρ avec $v = 1$:

$$\int_{\partial B_\rho} \frac{\partial u}{\partial n} d\Gamma = \int_{B_\rho} \Delta u \, d\Omega = 0$$



Posons $r = |x - y|$ et $\omega = \dfrac{x - y}{r}$. Pour tout x on a : $u(x) = u(y + r\omega)$ d'où on déduit :

$$\begin{aligned}
\int_{\partial B_\rho} \frac{\partial u}{\partial n} d\Gamma &= \int_{\partial B_\rho} \frac{\partial u}{\partial r}(y + \rho\omega)\, d\Gamma \\
&= \int_{|\omega|=1} \frac{\partial u}{\partial r}(y + \rho\omega) \rho^{n-1} d\omega \quad \text{(changement de variable } d\Gamma = \rho^{n-1} d\omega) \\
&= \rho^{n-1} \frac{\partial}{\partial \rho} \left(\int_{|\omega|=1} u(y + \rho\omega)\, d\omega \right) \qquad \left(\text{car } \frac{\partial}{\partial r} = \frac{\partial}{\partial \rho}\right) \\
0 &= \rho^{n-1} \frac{\partial}{\partial \rho} \left(\rho^{1-n} \int_{\partial B_\rho} u \, d\Gamma \right) \qquad \text{(changement de variable } d\omega = \rho^{1-n} d\Gamma)
\end{aligned}$$

d'où on tire que : $\rho^{1-n} \displaystyle\int_{\partial B_\rho} u \, d\Gamma = R^{1-n} \int_{\partial B_R} u \, d\Gamma.$

Or $\lim_{\rho\to 0} \rho^{1-n} \int_{\partial B_\rho} u\, d\Gamma = \omega_n u(y)$ car :

$$\begin{aligned}
\rho^{1-n} \int_{\partial B_\rho} u\, d\Gamma &= \rho^{1-n} \int_{\partial B_\rho} (u(y) + O(\rho))\, d\Gamma \qquad \text{(continuité uniforme de } u \text{ sur } B) \\
&= \rho^{1-n} \left(\rho^{n-1} u(y) \int_{|\omega|=1} d\omega + \rho^{n-1} \rho \int_{|\omega|=1} O(1)\, d\omega \right) \\
&= \omega_n u(y) + \rho \int_{|\omega|=1} O(1)\, d\omega \qquad (\longrightarrow 0 \text{ quand } \rho \longrightarrow 0).
\end{aligned}$$

d'où la formule de représentation (1.11).

Pour obtenir (1.12), on intègre la formule (1.11) pour $0 \leq \rho \leq R$:

$$\omega_n \rho^{n-1} u(y) = \int_{\partial B_\rho} u\, d\Gamma$$

ce qui donne :

$$\int_0^R \omega_n \rho^{n-1} u(y)\, d\rho = \int_{\rho=0}^R \int_{\partial B_\rho} u\, d\Gamma d\rho,$$

soit :

$$\omega_n u(y) \int_0^R \rho^{n-1} d\rho = \int_{B_R} u\, d\Omega,$$

et finalement

$$\omega_n u(y) \frac{R^n}{n} = \int_{B_R} u\, d\Omega.$$

■

1.2.2 Principe du Maximum

Une conséquence des formules de moyenne sphérique est le principe du maximum. Plus précisément on a :

Proposition 1.3. (Principe du maximum) *Soit $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ une fonction harmonique dans un ouvert borné connexe Ω. Si la fonction u n'est pas constante dans $\overline{\Omega}$ alors on a :*

$$\inf_{\partial\Omega} u < u(x) < \sup_{\partial\Omega} u \qquad \forall x \in \Omega. \tag{1.13}$$

En d'autres termes, le maximum ou le minimum d'une fonction harmonique non constante sur un ouvert borné Ω, est atteint sur la frontière $\partial\Omega$. En dimension 1, les seules fonctions harmoniques régulières sont de la forme $u(x) = ax + b$ et il est bien clair que le maximum et le minimum sont atteints aux extrémités d'un segment !

Démonstration : montrons que $u < \sup_{\partial\Omega} u$, la démonstration étant similaire pour l'autre inégalité. Supposons qu'il existe $y \in \Omega$ tel que $u(y) = \sup_{\overline{\Omega}} u = M$ et considérons l'ensemble :

$$\Omega_M = \{x \in \Omega \text{ tel que } u(x) = M\}.$$

$\Omega_M \neq \emptyset$ car $y \in \Omega_M$. Comme u est continu, Ω_M, qui est l'image réciproque de $\{M\}$ par u, est un fermé de Ω. Soit $z \in \Omega_M$, appliquons la formule de la moyenne (1.11) à la fonction $u - M$ (on a bien $\Delta(u - M) = 0$) dans la boule B_R centrée en z telle que $B_R \subset \Omega$:

$$0 = u(z) - M = \frac{n}{\omega_n R^n} \int_{B_R} (u - M) d\Gamma$$

Or $u - M \leq 0$ (définition de M), d'où on tire que $u = M$ dans B_R pour tout R tel que $B_R \subset \Omega$. Donc, Ω_M est un ouvert de Ω. Finalement, comme Ω est connexe, on a $\Omega_M = \Omega$. Ce qui montre que u est constant sur $\overline{\Omega}$, contredisant l'hypothèse. ∎

Le principe du maximum permet de démontrer simplement des résultats d'unicité pour le problème de Dirichlet et on énonce :

Théorème 1.1. (Unicité du problème de Dirichlet) *Le problème de Dirichlet (P_D) admet au plus une solution classique dans $\mathcal{C}^2(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$.*

Démonstration : soient (u_1, u_2) deux solutions classiques du problème (P_D). Posons $v = u_1 - u_2$. Alors $v \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$ et vérifie :

$$\begin{cases} \Delta v = 0 & \text{dans } \Omega \\ v = 0 & \text{sur } \partial\Omega \end{cases}$$

Comme $\inf_{\partial\Omega} v = \sup_{\partial\Omega} v = 0$, on déduit du principe du maximum (proposition 1.3) que $v \equiv 0$ sur $\overline{\Omega}$ et, par conséquent, que $u_1 \equiv u_2$ sur $\overline{\Omega}$. ∎

Le principe du maximum (1.13) ne permet pas de déduire l'unicité des solutions classiques des problèmes de Fourier, Neumann ni même des problèmes mixtes présentant une condition de Dirichlet. En effet, il n'y a aucun moyen simple d'estimer le minimum et le maximum de la solution sur la frontière $\partial\Omega$. Il existe cependant d'autres principes du maximum permettant d'aborder ces problèmes (voir [16]).

Signalons tout de même une évidence pour le problème de Neumann : la solution, si elle existe, n'est pas unique. En effet, si u est solution de (P_N) alors $\tilde{u} = u + \beta$, où β est un nombre réel quelconque, est encore une solution du problème de Neumann. En fait, on peut montrer que la solution est unique à une constante près.

La question de l'existence de solutions classiques aux problèmes de Dirichlet, Neumann et Fourier est une question difficile. En effet, la régularité du domaine intervient de façon cruciale et les démonstrations "directes" (méthode de Perron par exemple, voir [16]) sont assez complexes. Par ailleurs, nous allons voir que la théorie variationnelle permet d'obtenir facilement des résultats d'existence de solutions faibles, moyennant quoi, des résultats de régularité des solutions faibles assurent ensuite l'existence des solutions classiques.

1.3 Solutions faibles et formulations variationnelles

Considérons une solution classique $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1\left(\overline{\Omega}\right)$, solution du problème de Dirichlet homogène :

$$\left\{ \begin{array}{rcl} -\Delta u & = & f \quad \text{sur} \quad \Omega \quad \text{avec } f \in \mathcal{C}^0(\Omega) \\ u & = & 0 \quad \text{sur} \quad \partial\Omega \end{array} \right. \tag{1.14}$$

et v une fonction test appartenant à $\mathcal{C}^1(\Omega) \cap \mathcal{C}^0\left(\overline{\Omega}\right)$.

L'utilisation de la formule de Green (1.8) conduit à la relation :

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n} \, v \, d\Gamma \tag{1.15}$$

Par ailleurs, si on choisit v tel que $v = 0$ sur $\partial\Omega$, (1.15) se réduit à :

$$\int_\Omega \nabla u \cdot \nabla v d\Omega = \int_\Omega f v d\Omega. \tag{1.16}$$

La recherche d'une fonction u, nulle sur la frontière $\partial\Omega$, qui satisfait (1.16) pour toute fonction test v nulle sur $\partial\Omega$ s'appelle une formulation variationnelle du problème de Dirichlet (1.14).

Quels sont les avantages d'une telle formulation ?

– D'une part, on a abaissé le degré de dérivation sur la fonction u. Par conséquent, il peut exister des solutions u seulement dérivables vérifiant (1.16).

– En outre, on peut considérer des fonctions telles que toutes leurs dérivées au sens des distributions soient de carré intégrable, c'est-à-dire :

$$u \in L^2(\Omega) \;\; \text{telle que} \;\; \int_\Omega |\nabla u|^2 d\Omega < +\infty, \tag{1.17}$$

la formulation (1.16) ayant toujours un sens.

– Si la formulation variationnelle (1.16) admet de telles solutions, sont-elles également solutions classiques du problème écrit sous la forme d'E.D.P. (1.14) ? Rien ne le garantit. Formellement, ces solutions vérifient au sens des distributions

$$-\Delta u = f \quad \text{dans } \Omega.$$

C'est pourquoi on qualifie ces solutions de solutions faibles du problème de Dirichlet et la formulation variationnelle (1.16) de formulation faible du problème de Dirichlet.

– L'ensemble des fonctions qui satisfont (1.17) définit un espace de Hilbert, alors que l'ensemble des fonctions $\mathcal{C}^1$ définit seulement un espace de Banach (i.e. un espace vectoriel normé complet). La structure des espaces de Hilbert est beaucoup plus riche que celle des espaces de Banach (produit scalaire, théorème de projection, théorème de représentation de Riesz en particulier). Ceci va nous permettre de démontrer l'existence de solutions faibles de façon très générale.

– En fait cet espace de Hilbert s'interprète comme l'espace d'énergie des solutions. En effet, si on considère la fonctionnelle d'énergie, associée à la fonction u solution de (1.16) :

$$E(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 d\Omega - \int_\Omega f u d\Omega$$

un calcul élémentaire montre que pour tout v vérifiant (1.17) et tel que $v = 0$ sur $\partial\Omega$:

$$E(u+v) - E(u) = \int_\Omega \nabla u \cdot \nabla v d\Omega - \int_\Omega f v d\Omega + \frac{1}{2}\int_\Omega |\nabla v|^2 d\Omega,$$

montrant ainsi (cf. [9]) que la fonctionnelle E est différentiable en u, de différentielle :

$$dE(u).v = \int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_\Omega f \, v \, d\Omega.$$

En d'autres termes, les solutions de la formulation variationnelle sont les fonctions qui minimisent l'énergie E dans l'espace d'énergie. Cette interprétation physique rejoint le principe des travaux virtuels cher aux mécaniciens.

– Enfin, la formulation variationnelle s'interprète comme un système pseudo-matriciel. En effet, si $(w_i)_{i\geq 0}$ est une "base infinie" de l'espace dans lequel on cherche u, en prenant comme fonctions test les fonctions de base w_i, (1.16) se réécrit "formellement" :

$$\mathbb{A}U = F$$

avec

$$\mathbb{A}_{ij} = \int_\Omega \nabla w_j \cdot \nabla w_i d\Omega, \; u = \sum_j U_j w_j \text{ et } \; F_i = \int_\Omega f w_i d\Omega.$$

C'est sur cette idée qu'est fondée la méthode des éléments finis.

Nous allons préciser toutes ces notions en rappelant tout d'abord quelques éléments d'analyse fonctionnelle.

1.3.1 Espaces de Sobolev

Dans toute la suite, Ω désigne un ouvert borné de $\mathbb{R}^n$, de frontière $\partial\Omega$[4].

[4] La plupart des définitions et résultats énoncés par la suite demeurent valables si l'ouvert n'est pas borné.

Un élément $\alpha = (\alpha_1, \cdots, \alpha_n)$ de $\mathbb{N}^n$ est appelé un multi-indice, de longueur $|\alpha| = \sum_{j=1}^n \alpha_j$. La dérivée partielle d'ordre α est notée

$$\partial_\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Commençons par les espaces de Lebesgue de type L^p, pour $1 \leq p \leq \infty$. Nous mettons l'accent sur des fonctions à valeurs réelles, mais toutes les définitions qui suivent peuvent être étendues au cas de fonctions à valeurs complexes. Soit $d\Omega = dx_1 dx_2 \cdots dx_n$ la mesure de Lebesgue sur Ω, *considéré comme un sous-ensemble de* $\mathbb{R}^n$.

Définition 1.3. *L'espace $L^p(\Omega)$ a pour éléments les classes de fonctions f, définies presque partout, mesurables au sens de Lebesgue sur Ω, et telles que*

$$\begin{cases} \text{pour } 1 \leq p < \infty & \|f\|_{L^p(\Omega)} := \left\{ \int_\Omega |f|^p \, d\Omega \right\}^{1/p} < \infty \\ \text{pour } p = \infty & \|f\|_{L^\infty(\Omega)} := \sup_{x \in \Omega} ess|f(x)| < \infty \end{cases} . \tag{1.18}$$

Muni de la norme $\|\cdot\|_{L^p(\Omega)}$, $L^p(\Omega)$ est un espace de Banach.

Soit $p \in [1, \infty]$, alors $f_1 = f_2$ dans $L^p(\Omega)$ signifie que $f_1(x) = f_2(x)$ presque partout dans Ω. On peut ensuite définir les fonctions *localement* dans $L^p(\Omega)$, dans le sens suivant : si f appartient à $L^p(K)$ pour tout sous-ensemble mesurable et compact K de Ω, alors f est localement dans $L^p(\Omega)$, et on écrit $f \in L^p_{loc}(\Omega)$. Ainsi, $L^2(\Omega)$ est l'ensemble des fonctions (définies presque partout sur Ω), mesurables au sens de Lesbesgue, et de carré intégrable sur Ω. Muni du produit scalaire[5] :

$$(u, v)_{L^2(\Omega)} = \int_\Omega uv \, d\Omega$$

de norme associée :

$$\|u\|_{L^2(\Omega)} = \left(\int_\Omega u^2 d\Omega \right)^{1/2},$$

$L^2(\Omega)$ est un espace de Hilbert (préhilbertien complet pour cette norme). Notons encore une fois que l'égalité $v = w$ dans $L^2(\Omega)$ signifie que $v(x) = w(x)$ presque partout dans Ω.

[5] Comme remarqué plus haut, les mêmes résultats sont transposables, pour des fonctions à valeurs dans $\mathbb{C}$. On considère alors le produit scalaire hermitien $(u, v)_{L^2(\Omega)} = \int_\Omega u\bar{v} \, d\Omega$, de norme associée $\|u\|_{L^2(\Omega)} = \left(\int_\Omega |u|^2 d\Omega \right)^{1/2}$, pour lequel $L^2(\Omega)$ est un espace de Hilbert sur $\mathbb{C}$.

• Dérivation au sens des distributions

On note $\mathcal{D}(\Omega)$ l'ensemble des fonctions $\mathcal{C}^\infty$ sur Ω à support compact dans Ω et $\mathcal{D}'(\Omega)$ l'espace des formes linéaires continues sur $\mathcal{D}(\Omega)$ muni de sa topologie limite inductive, autrement dit l'espace des distributions sur Ω (voir par exemple [20, 6]). Pratiquement, on peut utiliser la *convergence au sens des limites* pour définir la topologie. Soit $(f_k)_k$ une suite d'éléments de $\mathcal{D}(\Omega)$, elle converge dans $\mathcal{D}(\Omega)$ vers f si, et seulement si,

(i) il existe un sous-ensemble compact K de Ω tel que $\mathrm{supp}(f_k) \subset K$, pour tout k ; $\mathrm{supp}(f_k - f) \subset K$, pour tout k ;
(ii) pour tout multi-indice α, $(\partial_\alpha f_k)_k$ converge uniformément dans K vers $\partial_\alpha f$.

Définition 1.4. *Une forme linéaire et continue T définie sur $\mathcal{D}(\Omega)$ est appelée une distribution. L'espace des distributions est noté $\mathcal{D}'(\Omega)$.*
Soit $T \in \mathcal{D}'(\Omega)$ et $f \in \mathcal{D}(\Omega)$: l'action de T sur f est écrite à l'aide de crochets de dualité, c'est-à-dire

$$\langle T, f\rangle.$$

D'après la définition de la topologie de $\mathcal{D}(\Omega)$ au sens des limites, T est continue dès lors que

$$\forall (f_k)_k, f \in \mathcal{D}(\Omega) \text{ telles que } f_k \to f \text{ dans } \mathcal{D}(\Omega), \quad \langle T, f_k\rangle \to \langle T, f\rangle.$$

Quelques exemples sont proposés dans la suite, voir (1.19), (1.22). En ce qui concerne la convergence de suites d'éléments de $\mathcal{D}'(\Omega)$, nous utilisons la définition ci-dessous.

Définition 1.5. *Soit $(T_k)_k$ une suite d'éléments de $\mathcal{D}'(\Omega)$: elle converge dans $\mathcal{D}'(\Omega)$ vers T si, et seulement si, pour tout f dans $\mathcal{D}(\Omega)$, $\langle T_k, f\rangle \to \langle T, f\rangle$.*

A partir de là, on peut facilement prouver l'inclusion

$$L^1_{loc}(\Omega) \subset \mathcal{D}'(\Omega), \tag{1.19}$$

en *identifiant* tout élément f de $L^1_{loc}(\Omega)$ à une distribution, toujours notée f, selon

$$\langle f, g\rangle = \int_\Omega f\, g\, d\Omega, \ \forall g \in \mathcal{D}(\Omega). \tag{1.20}$$

Rappelons une propriété fort utile par la suite...

Proposition 1.4. *Soient f_1 et f_2 deux éléments de $L^1_{loc}(\Omega)$. La relation $\langle f_1, g\rangle = \langle f_2, g\rangle$ pour tout $g \in \mathcal{D}(\Omega)$ implique que $f_1(x) = f_2(x)$ presque partout dans Ω.*

Nous passons maintenant à la notion de dérivation au sens des distributions.

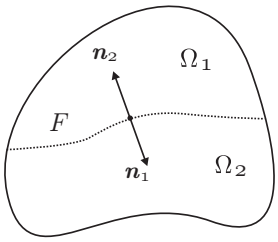
Définition 1.6. *Soit $T \in \mathcal{D}'(\Omega)$. Sa $j$ème dérivée partielle ($j = 1, \cdots, n$) est définie par*

$$\langle \frac{\partial T}{\partial x_j}, f \rangle = -\langle T, \frac{\partial f}{\partial x_j} \rangle, \quad \forall f \in \mathcal{D}(\Omega).$$

Bien sûr, toute dérivée d'une distribution est elle-même une distribution.

Proposition 1.5. *L'application $T \mapsto \partial_j T$ est linéaire et continue de $\mathcal{D}'(\Omega)$ dans $\mathcal{D}'(\Omega)$.*

Exemple de dérivation au sens des distributions : Soit Ω un ouvert borné partitionné en deux ouverts bornés Ω_1 et Ω_2 tels que $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ et $\overline{\Omega}_1 \cap \overline{\Omega}_2 = F$. Considérons une fonction $v \in \mathcal{C}^0\left(\overline{\Omega}\right)$ telle que $v_j = v_{|\Omega_j} \in \mathcal{C}^1\left(\overline{\Omega}_j\right)$ pour $j = 1, 2$.



Par définition de la dérivation au sens des distributions, on a $\forall i$:

$$\langle \frac{\partial v}{\partial x_i}, \varphi \rangle = -\langle v, \frac{\partial \varphi}{\partial x_i} \rangle = -\int_\Omega v \, \frac{\partial \varphi}{\partial x_i} d\Omega \quad \forall \varphi \in \mathcal{D}(\Omega)$$

soit :

$$\begin{aligned}
\langle \frac{\partial v}{\partial x_i}, \varphi \rangle &= -\int_{\Omega_1} v_1 \, \frac{\partial \varphi}{\partial x_i} \, d\Omega - \int_{\Omega_2} v_2 \, \frac{\partial \varphi}{\partial x_i} \, d\Omega \\
\begin{pmatrix} \text{formule} \\ \text{de Green} \end{pmatrix} &= \int_{\Omega_1} \frac{\partial v_1}{\partial x_i} \, \varphi \, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i} \, \varphi \, d\Omega - \int_F \left(v_{1|F} \left(n_1\right)_i + v_{2|F} \left(n_2\right)_i \right) \varphi \, d\Gamma \\
(\mathbf{n}_1 = -\mathbf{n}_2) &= \int_{\Omega_1} \frac{\partial v_1}{\partial x_i} \, \varphi \, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i} \, \varphi \, d\Omega - \int_F \left(v_{1|F} - v_{2|F} \right) \left(n_1\right)_i \varphi \, d\Gamma \\
(\text{continuité de } v) &= \int_{\Omega_1} \frac{\partial v_1}{\partial x_i} \, \varphi \, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i} \, \varphi \, d\Omega
\end{aligned}$$

ainsi $(\partial_i v)_{|\Omega_1} = \partial_i v_1$ et $(\partial_i v)_{|\Omega_2} = \partial_i v_2$, et $\nabla v \in \left(L^2(\Omega)\right)^n$ car $v_j \in \mathcal{C}^1(\overline{\Omega}_j)$ pour $j = 1, 2$.

Puisque $L^2(\Omega)$ est un sous-espace de $L^1_{loc}(\Omega)$ qui est lui-même inclus dans $\mathcal{D}'(\Omega)$ (par identification, cf. (1.20)), il est loisible de dériver ses éléments au sens des distributions. On introduit en conséquence l'espace de Sobolev :

$$\boxed{H^1(\Omega) = \left\{ v \in L^2(\Omega) \text{ tel que } \nabla v \in \left(L^2(\Omega)\right)^n \right\},}$$

où la dérivation $(\nabla \cdot)$ est comprise au sens des distributions.

Propriété 1.1 *Soit Ω un ouvert borné partitionné en deux ouverts bornés disjoints Ω_1 et Ω_2 ($\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$). Soit $v \in \mathcal{C}^0\left(\overline{\Omega}\right)$ telle que $v_j = v_{|\Omega_j} \in \mathcal{C}^1\left(\overline{\Omega}_j\right)$ pour $j = 1, 2$, alors $v \in H^1(\Omega)$.*

L'espace $H^1(\Omega)$ muni du produit scalaire :

$$(u,v)_{H^1(\Omega)} = \int_\Omega uv\,d\Omega + \int_\Omega \nabla u \cdot \nabla v\,d\Omega\ , \tag{1.21}$$

dont dérive la norme :

$$\|u\|_{H^1(\Omega)} = (u,u)^{1/2}_{H^1(\Omega)},$$

est un espace de Hilbert.

Proposition 1.6. *Soit $f \in L^2(\Omega)$. Alors f appartient à $H^1(\Omega)$ si, et seulement si, il existe $h_1, \cdots, h_n \in L^2(\Omega)$ tels que, pour $j = 1, \cdots, n$,*

$$\int_\Omega h_j g_j\,d\Omega = -\int_\Omega f\partial_j g_j\,d\Omega, \quad \forall g_j \in \mathcal{D}(\Omega).$$

Démonstration : Par définition, une fonction f appartient à $H^1(\Omega)$ si, et seulement si, $f \in L^2(\Omega)$ et $\partial_j f \in L^2(\Omega)$, $j = 1, \cdots, n$, où les dérivées partielles sont comprises au sens des distributions. D'après la définition 1.6, $\partial_1 f \in \mathcal{D}'(\Omega)$ et, pour tout g_1 dans $\mathcal{D}(\Omega)$, on a l'égalité

$$\langle \partial_1 f, g_1\rangle = -\langle f, \partial_1 g_1\rangle = -\int_\Omega f\frac{\partial g_1}{\partial x_1}\,d\Omega,$$

puisque $L^2(\Omega)$ peut-être identifié cf. (1.20) à un sous-espace de $\mathcal{D}'(\Omega)$. En utilisant à nouveau cette identification, on en déduit que $\partial_1 f$ appartient à $L^2(\Omega)$ *si, et seulement si*,

$$\exists h_1 \in L^2(\Omega), \quad \int_\Omega h_1 g_1\,d\Omega = -\int_\Omega f\frac{\partial g_1}{\partial x_1}\,d\Omega, \quad \forall g_1 \in \mathcal{D}(\Omega).$$

Ceci est vrai pour toutes les dérivées partielles ($j = 1, \cdots, n$). En d'autres termes, chaque composante du gradient ∇f est égale à un élément h_j de $L^2(\Omega)$, pour $j = 1, \cdots, n$. ∎

Poursuivons : disposant d'un multi-indice α, on peut définir récursivement, à l'aide de la définition 1.6, la dérivée partielle d'ordre α d'une distribution.

Définition 1.7. *Soit $T \in \mathcal{D}'(\Omega)$, sa dérivée partielle d'ordre α est définie par*

$$\langle \partial_\alpha T, f\rangle = (-1)^{|\alpha|}\langle T, \partial_\alpha f\rangle, \quad \forall f \in \mathcal{D}(\Omega).$$

Lorsque $\alpha = (0,0,0)$, il n'y a pas de dérivation !

La définition précédente nous permet de construire les espaces de Sobolev d'ordre m, $m \geq 2$. Pour $m \in \mathbb{N}$, on définit :

$$H^m(\Omega) = \left\{ v \in L^2(\Omega) \text{ tel que } \partial_\alpha v \in L^2(\Omega)\ ,\ \forall \alpha \in \mathbb{N}^n,\ |\alpha| \leq m \right\}.$$

C'est encore un espace de Hilbert, muni du produit scalaire :

$$(u,v)_{H^m(\Omega)} = \int_\Omega \left(\sum_{|\alpha| \leq m} \partial_\alpha u\,\partial_\alpha v \right) d\Omega.$$

On a notamment la suite d'inclusions[6] (pour $m \geq 2$) :

$$\mathcal{D}(\Omega) \subset H^m(\Omega) \subset H^1(\Omega) \subset L^2(\Omega). \tag{1.22}$$

On peut également définir des espaces de Sobolev d'ordre fractionnaire $s \in \mathbb{R}^+$, encore notés $H^s(\Omega)$, à l'aide de la théorie de l'interpolation (cf. [22]). En particulier, pour $s > s' \geq 0$ on a $H^s(\Omega) \subset H^{s'}(\Omega)$.

Par la suite, nous utiliserons essentiellement les espaces de Sobolev $H^1(\Omega)$ et $H^2(\Omega)$. Nous allons donc préciser quelques-unes de leurs propriétés[7].

• Frontière "suffisamment régulière"

Jusqu'à présent, les résultats sont valables dans tout ouvert (borné) Ω de $\mathbb{R}^n$, c'est-à-dire sans hypothèse sur sa frontière $\partial\Omega$. Pour pouvoir poursuivre, nous allons préciser la notion de frontière "suffisamment régulière", en deux temps.

Définition 1.8. *Soit Ω un ouvert borné de $\mathbb{R}^n$. Sa frontière $\partial\Omega$ est lipschitzienne si, et seulement si,*

- *en tout point x de $\partial\Omega$, il existe une application lipschitzienne (définie sur un hypercube de $\mathbb{R}^{n-1}$ à valeurs dans $\mathbb{R}$), dont le graphe représente localement $\partial\Omega$ dans un voisinage (ouvert) de x ;*
- *en tout point x de $\partial\Omega$, Ω est localement d'un seul côté de $\partial\Omega$.*

Dans cette définition, les deux aspects sont fondamentaux. Voici deux exemples d'applications lipschitziennes (on parle aussi de cartes locales).

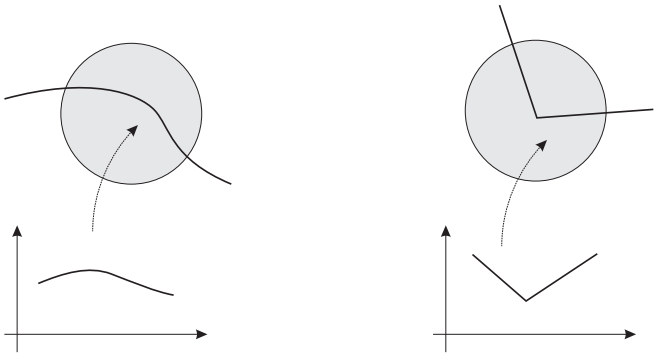


Fig. 1.2. Exemples de carte locale

[6] Ici, et dans toute la suite, la propriété $X \subset Y$ (pour 2 espaces vectoriels topologiques X et Y) signifie que l'injection de X dans Y est *continue*.

[7] Voir par ex. [25] pour des démonstrations "élémentaires". Pour un exposé beaucoup plus complet des propriétés des espaces de Sobolev, voir [2, 18].

Ci-dessous, nous donnons deux exemples d'ouverts à frontière non-lipschitzienne. Celui de gauche (un ouvert "fissuré"), parce qu'à l'extrémité de la "fissure", l'ouvert n'est pas localement d'un seul côté de sa frontière. Celui de droite (un cusp), parce qu'au point de rebroussement de sa frontière, l'application permettant de la représenter localement n'est pas lipschitzienne.

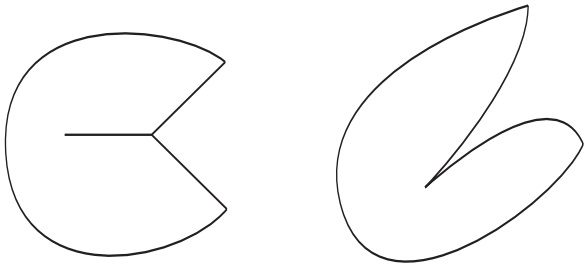


Fig. 1.3. Exemples de frontières non-lipschitziennes

Dans un ouvert de $\mathbb{R}^n$ à frontière lipschitzienne, on peut prouver des résultats de densité très utiles. Pour cela, définissons un ensemble de fonctions régulières *ad hoc*.

Définition 1.9. *On appelle $\mathcal{C}^\infty(\overline{\Omega})$ l'espace composé des restrictions à $\overline{\Omega}$ de fonctions de $\mathcal{C}^\infty(\mathbb{R}^n)$ à support compact dans $\mathbb{R}^n$.*

Proposition 1.7. *Soit $m \in \mathbb{N}$. Dans tout ouvert borné Ω à frontière lipschitzienne, $\mathcal{C}^\infty(\overline{\Omega})$ est dense dans $H^m(\Omega)$.*

Dans la suite, nous allons considérer des sous-classes d'ouverts à frontière lipschitzienne, que l'on rencontre souvent en pratique. Ces sous-classes sont regroupées sous le vocable ouvert à frontière "suffisamment régulière".

Définition 1.10. *Soit Ω un ouvert borné de $\mathbb{R}^n$ à frontière $\partial\Omega$ lipschitzienne. Sa frontière est dite "suffisamment régulière" si on se trouve dans un des cas suivants :*

- *pour $\Omega \subset \mathbb{R}^n$, la frontière de Ω est régulière, c'est-à-dire qu'en tout point x de $\partial\Omega$, l'application dont le graphe représente localement $\partial\Omega$ est de classe $\mathcal{C}^\infty$;*
- *pour $\Omega \subset \mathbb{R}^3$, la frontière de Ω est polyédrique (avec des faces planes), ou polyédrique curviligne (avec des faces courbes) ;*
- *pour $\Omega \subset \mathbb{R}^2$, la frontière de Ω est polygonale (avec des côtés droits), ou polygonale curviligne (avec des côtés courbes) ;*

Voici quelques représentants d'ouverts à frontière "suffisamment régulière".

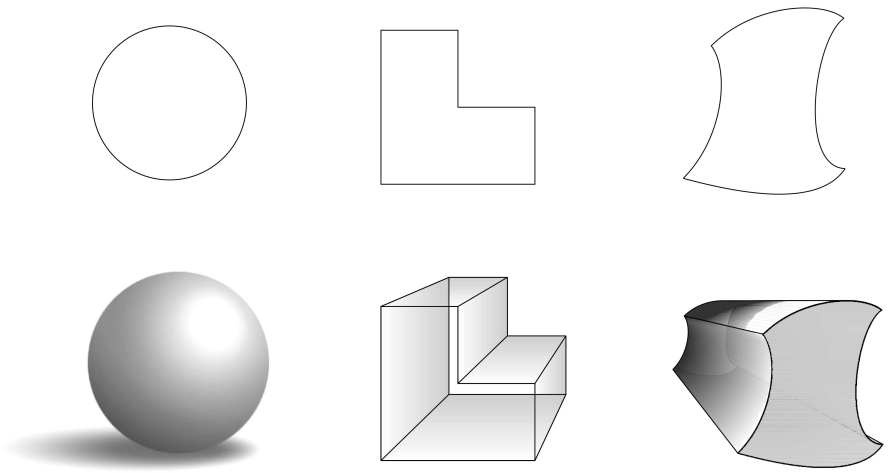


Fig. 1.4. Exemples de frontières "suffisamment" régulières

• Existence de trace

Soit $v \in H^1(\Omega)$, peut-on définir sa "valeur sur la frontière", c'est-à-dire sa trace sur $\partial\Omega$? Ce n'est pas une question naïve !

En effet, remarquons tout d'abord que les fonctions de $L^2(\Omega)$ n'ont pas nécessairement de trace (prendre par exemple la fonction $v(x) = x^{-1/4}$ qui appartient à $L^2(]0,1[)$). Par ailleurs, pour $n \geq 2$, les fonctions de $H^1(\Omega)$ ne sont généralement pas continues (prendre par exemple $n = 2$, $\Omega = B(0,R)$ et $v(r) = |\operatorname{Log} r|^k$ avec $k \in]0, \frac{1}{2}[$ pour s'en convaincre). La définition de la trace de telles fonctions n'est donc pas immédiate.

Toutefois, pour toute fonction $v \in \mathcal{C}^\infty\left(\overline{\Omega}\right)$ on peut définir sa valeur $\gamma_0 v$ sur la frontière $\partial\Omega$ qui vérifie :

$$\gamma_0 v \in L^2(\partial\Omega) = \left\{ w \text{ mesurable sur } \partial\Omega \text{ tel que } \quad \int_{\partial\Omega} w^2 d\Gamma < +\infty \right\}$$

où $d\Gamma$ désigne l'élément de surface porté par $\partial\Omega$. Muni du produit scalaire

$$(w,z)_{L^2(\partial\Omega)} = \int_{\partial\Omega} w\, z\, d\Gamma,$$

dont dérive la norme

$$\|w\|_{L^2(\partial\Omega)} = \left\{ \int_{\partial\Omega} w^2\, d\Gamma \right\}^{1/2},$$

$L^2(\partial\Omega)$ est un espace de Hilbert.

L'application γ_0 ainsi définie se prolonge aux fonctions de $H^1(\Omega)$ et plus précisément on énonce :

Théorème 1.2. (Existence de trace) *Soit Ω un ouvert borné de $\mathbb{R}^n$, à frontière "suffisamment régulière". Alors, l'application trace*

$$\gamma_0 : \begin{cases} \mathcal{C}^\infty(\overline{\Omega}) & \to & L^2(\partial\Omega) \\ v & \mapsto & \gamma_0 v = v_{|\partial\Omega} \end{cases}$$

se prolonge par continuité en une application linéaire continue, notée encore γ_0, de $H^1(\Omega)$ dans $L^2(\partial\Omega)$ et il existe une constante[8] *C indépendante de v telle que :*

$$\|\gamma_0 v\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^1(\Omega)} \tag{1.23}$$

L'application trace γ_0 n'est pas surjective de $H^1(\Omega)$ dans $L^2(\partial\Omega)$. En d'autres termes, il existe des fonctions de $L^2(\partial\Omega)$ qui ne sont pas la trace de fonctions de $H^1(\Omega)$. Par contre, l'opérateur de trace γ_0 est surjectif sur $H^{1/2}(\partial\Omega)$, espace de Sobolev à indice fractionnaire tel que $H^1(\partial\Omega) \subset H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega)$, dont l'étude sort du cadre de ce cours (voir [2, 18]). En particulier, on peut établir les propriétés suivantes sur

$$H^{1/2}(\partial\Omega) = \left\{ w \in L^2(\partial\Omega) \text{ tel que } \exists v \in H^1(\Omega),\ w = \gamma_0 v \right\}.$$

Proposition 1.8. *L'espace des traces $H^{1/2}(\partial\Omega)$ est un espace de Banach. Qui plus est, on peut munir $H^{1/2}(\partial\Omega)$ de la norme*

$$\|w\|_{H^{1/2}(\partial\Omega)} = \inf_{v \in H^1(\Omega) \ tq \ \gamma_0 v = w} \|v\|_{H^1(\Omega)}.$$

Enfin, $H^{1/2}(\partial\Omega)$ est dense dans $L^2(\partial\Omega)$.

On peut prouver que l'ensemble $\mathcal{D}(\Omega)$ est dense dans $L^2(\Omega)$ mais n'est pas dense dans $H^1(\Omega)$. C'est pourquoi, on introduit l'espace $H_0^1(\Omega)$, la fermeture de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$:

$$H_0^1(\Omega) = \overline{\mathcal{D}(\Omega)}^{\,H^1(\Omega)}$$

c'est-à-dire l'ensemble des limites des suites (pour la norme $H^1(\Omega)$) d'éléments de $\mathcal{D}(\Omega)$. C'est un espace de Hilbert comme sous-espace fermé d'un espace de Hilbert.

L'existence de trace permet d'identifier l'espace $H_0^1(\Omega)$:

[8] Ici et dans la suite, on notera habituellement les constantes strictement positives avec un C "générique" (pour alléger les notations). Ainsi, deux instances de C pourront faire référence à deux constantes différentes.

Théorème 1.3. (Caractérisation de $H_0^1(\Omega)$) *Si Ω est un ouvert borné de $\mathbb{R}^n$ à frontière "suffisamment régulière", alors :*

$$H_0^1(\Omega) = \left\{ v \in H^1(\Omega) \text{ tel que } \gamma_0 v = 0 \right\}$$

En utilisant le théorème 1.2 et la proposition 1.7, on peut généraliser la formule d'intégration par parties de Green (1.7).

Proposition 1.9. (Formule de Green) *Soient $u, v \in H^1(\Omega)$, on a :*

$$\int_\Omega \frac{\partial u}{\partial x_i}\, v \, d\Omega = -\int_\Omega u \, \frac{\partial v}{\partial x_i}\, d\Omega + \int_{\partial\Omega} u\, v\, n_i \, d\Gamma. \tag{1.24}$$

Démonstration : D'après la proposition 1.7, il existe deux suites $(u_k)_k$ et $(v_m)_m$ de $\mathcal{C}^\infty(\overline{\Omega})$ telles que

$$\lim_{k\to\infty} u_k = u \text{ et } \lim_{m\to\infty} v_m = v \text{ dans } H^1(\Omega).$$

On note que l'on peut appliquer la formule d'intégration par parties de Green (1.7) pour chaque couple (u_k, v_m).
Par ailleurs, $u_k \to u$ dans $L^2(\Omega)$ et $\partial_i u_k \to \partial_i u$ dans $L^2(\Omega)$. De même, $v_m \to v$ dans $L^2(\Omega)$ et $\partial_i v_m \to \partial_i v$ dans $L^2(\Omega)$. On en déduit que, lorsque k et m tendent vers l'infini :

$$\int_\Omega \frac{\partial u_k}{\partial x_i}\, v_m \, d\Omega \to \int_\Omega \frac{\partial u}{\partial x_i}\, v \, d\Omega \text{ et } \int_\Omega u_k \, \frac{\partial v_m}{\partial x_i}\, d\Omega \to \int_\Omega u \, \frac{\partial v}{\partial x_i}\, d\Omega.$$

Quant au terme frontière, on sait que, par continuité de l'application trace γ_0 (cf. théorème 1.2), $\lim_{k\to\infty} \gamma_0 u_k = \gamma_0 u$ et $\lim_{m\to\infty} \gamma_0 v_m = \gamma_0 v$ dans $L^2(\partial\Omega)$.
Enfin, comme la frontière $\partial\Omega$ est lipschitzienne, on a en particulier le résultat $n_i \in L^\infty(\partial\Omega)$ (voir [18]). On en conclut cette fois que, lorsque k et m tendent vers l'infini :

$$\int_{\partial\Omega} u_k\, v_m\, n_i \, d\Gamma \to \int_{\partial\Omega} u\, v\, n_i \, d\Gamma.$$

(Ci-dessus, on a omis d'écrire l'action de l'application trace γ_0 sur u_k, v_m, u, v).
Si maintenant on reprend la formule (1.7) pour chaque couple (u_k, v_m), on peut passer à la limite (en k et m) et retrouver le résultat (1.24). ∎

Le théorème de trace 1.2 se généralise aux dérivées normales sur la frontière :

Théorème 1.4. *L'application trace de la dérivée normale :*

$$\gamma_1 : \begin{cases} \mathcal{C}^\infty(\overline{\Omega}) & \to & L^2(\partial\Omega) \\ v & \mapsto & \gamma_1 v = \left(\dfrac{\partial v}{\partial n}\right)_{|\partial\Omega} \end{cases}$$

se prolonge par continuité en une application linéaire continue, notée encore γ_1, de $H^2(\Omega)$ dans $L^2(\partial\Omega)$ et on a l'inégalité :

$$\|\gamma_1 v\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^2(\Omega)} \tag{1.25}$$

Ce théorème permet de considérer les traces de dérivées normales de fonctions v de $H^2(\Omega)$ comme fonctions de $L^2(\partial\Omega)$, ce qui n'est pas possible si v appartient seulement à $H^1(\Omega)$. Les théorèmes de trace 1.2 et 1.4 permettent, par ailleurs, de généraliser par densité les formules de Green (1.6), (1.7), (1.8) et (1.9) aux espaces de Sobolev. On généralise ainsi la formule (1.8).

Proposition 1.10. (Formule de Green) *Soient $u \in H^2(\Omega)$ et $v \in H^1(\Omega)$, on a alors :*

$$\boxed{\int_\Omega \Delta u\, v\, d\Omega = -\int_\Omega \nabla u \cdot \nabla v\, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n}\, v\, d\Gamma} \tag{1.26}$$

Terminons ce survol des espaces de Sobolev en indiquant le lien qu'il existe entre ces espaces et les espaces $\mathcal{C}^k\left(\overline{\Omega}\right)$.

Proposition 1.11. (Régularité des espaces de Sobolev) *Soit Ω un ouvert borné de $\mathbb{R}^n$, à frontière "suffisamment régulière" et soit $m > \dfrac{n}{2} + k$ ($k \geq 0$), alors $H^m(\Omega)$ s'injecte continûment dans $\mathcal{C}^k(\overline{\Omega})$. Plus précisément, on a l'estimation :*

$$\|u\|_{\mathcal{C}^k(\overline{\Omega})} \left(= \sup_{x\in\overline{\Omega}} \sum_{|\alpha|\leq k} |\partial_\alpha u(x)| \right) \leq C\, \|u\|_{H^m(\Omega)}. \tag{1.27}$$

Nous aurons également besoin du résultat plus spécialisé ci-dessous. Soient[9]

$$\begin{cases} \Psi = \left\{ v \in H^1(\Omega) \text{ tel que } \Delta v \in L^2(\Omega) \right\}, \\ \Psi_D = \left\{ v \in \Psi \text{ tel que } \gamma_0 v = 0 \right\}, \quad \Psi_N = \left\{ v \in \Psi \text{ tel que } \gamma_1 v = 0 \right\}. \end{cases} \tag{1.28}$$

On munit Ψ (et donc Ψ_D et Ψ_N) de la norme du graphe, c'est-à-dire

$$\|u\|_\Psi = \left(\|u\|^2_{H^1(\Omega)} + \|\Delta u\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Proposition 1.12. (Régularité des espaces de solutions) *Soit Ω un ouvert borné de $\mathbb{R}^n$, $n = 1, 2, 3$, à frontière "suffisamment régulière", alors Ψ_D et Ψ_N s'injectent continûment dans $\mathcal{C}^0(\overline{\Omega})$.*

Par contre, le résultat est faux pour Ψ, qui ne s'injecte pas continûment dans $\mathcal{C}^0(\overline{\Omega})$. C'est l'adjonction d'une condition aux limites homogène qui permet de conclure.

On retiendra surtout de ce résultat général :
– en dimension 1, les fonctions de $H^1(\Omega)$ sont continues mais pas nécessairement dérivables.

[9] Le fait que l'on puisse définir la trace de la dérivée normale d'éléments de Ψ (pour construire Ψ_N) sera explicité en fin de §1.3.3.

– en dimension 2 ou 3, les fonctions de $H^1(\Omega)$ ne sont pas nécessairement continues, mais les fonctions de Ψ_D, Ψ_N et $H^2(\Omega)$ sont continues.

Bien que les fonctions appartenant à l'espace $H^1(\Omega)$ ne soient pas nécessairement continues si $n \geq 2$, on peut toutefois démontrer que des fonctions globalement H^1 présentent un saut nul sur une interface quelconque. En outre, si elles sont continues de chaque côté de l'interface elles sont en fait globalement continues. Ce dernier résultat constitue une forme de réciproque de l'exemple considéré au début de la sous-section. On énonce donc :

Propriété 1.2 *Soit Ω un ouvert borné partitionné en deux ouverts Ω_1 et Ω_2 tels que $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\overline{\Omega}_1 \cap \overline{\Omega}_2 = F$ et $v \in H^1(\Omega)$. On note $v_j = v_{|\Omega_j}$ pour $j = 1, 2$. On a alors :*

$$[v]_F \overset{def}{=} v_{1|F} - v_{2|F} = 0 \quad \text{p. p. sur } F.$$

En particulier, si $v_j \in \mathcal{C}^0(\overline{\Omega}_j)$ pour $j = 1, 2$ alors $v \in \mathcal{C}^0\left(\overline{\Omega}\right)$.

Démonstration : Comme $v \in H^1(\Omega)$, on a $\forall \varphi \in \mathcal{D}(\Omega)$:

$$\int_\Omega \frac{\partial v}{\partial x_i}\,\varphi\, d\Omega = \int_{\Omega_1} \frac{\partial v_1}{\partial x_i}\,\varphi\, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i}\,\varphi\, d\Omega \quad \text{où } v_j = v_{|\Omega_j}.$$

Par ailleurs, on a, car $\varphi \in \mathcal{D}(\Omega)$:

$$\begin{aligned}
\int_\Omega \frac{\partial v}{\partial x_i}\,\varphi\, d\Omega &= -\int_\Omega v\,\frac{\partial \varphi}{\partial x_i}\, d\Omega = -\int_{\Omega_1} v_1\,\frac{\partial \varphi}{\partial x_i}\, d\Omega - \int_{\Omega_2} v_2\,\frac{\partial \varphi}{\partial x_i}\, d\Omega \\
\begin{pmatrix}\text{formule}\\ \text{de Green}\end{pmatrix} &= \int_{\Omega_1} \frac{\partial v_1}{\partial x_i}\,\varphi\, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i}\,\varphi\, d\Omega - \int_F \left(v_{1|F}\,(n_1)_i + v_{2|F}\,(n_2)_i\right)\varphi_{|F} dF \\
(\text{car } \mathbf{n}_1 = -\mathbf{n}_2) &= \int_{\Omega_1} \frac{\partial v_1}{\partial x_i}\,\varphi\, d\Omega + \int_{\Omega_2} \frac{\partial v_2}{\partial x_i}\,\varphi\, d\Omega - \int_F \left(v_{1|F} - v_{2|F}\right)(n_1)_i\,\varphi_{|F} dF
\end{aligned}$$

Comme le vecteur normal est unitaire, il existe au moins une composante non nulle de ce vecteur, d'où on déduit par densité que $v_{1|F} = v_{2|F}$ dans $L^2(F)$ et par conséquent

$$v_{1|F} = v_{2|F} \quad \text{p. p. sur } F.$$

Lorsque $v_j \in \mathcal{C}^0\left(\overline{\Omega}_j\right)$, l'égalité précédente est en fait vraie partout, ce qui démontre que $v \in \mathcal{C}^0\left(\overline{\Omega}\right)$. ∎

Enfin, on rappelle un dernier résultat (cf. [8]), valable uniquement lorsque l'ouvert Ω est borné.

Théorème 1.5. ***(Rellich)*** *Soit Ω un ouvert borné de $\mathbb{R}^n$, à frontière "suffisamment régulière". Alors, de toute suite bornée de $H^1(\Omega)$, on peut extraire une sous-suite qui converge dans $L^2(\Omega)$.*

On dit alors que l'injection canonique de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte.

1.3.2 Théorèmes fondamentaux de l'analyse hilbertienne

Nous allons rappeler deux théorèmes fondamentaux de l'analyse hilbertienne qui nous serviront par la suite. Pour la démonstration de ces théorèmes nous renvoyons à [8].

On désigne par H un espace de Hilbert muni du produit scalaire $(\cdot,\cdot)_H$ et on note H' le dual de l'ensemble H, c'est-à-dire l'ensemble des formes linéaires continues sur H. On munit l'espace dual de la norme canonique :

$$\|\ell\|_{H'} = \sup_{v\in H, v\neq 0} \frac{|\ell(v)|}{\|v\|_H}$$

Deux propriétés essentielles des espaces de Hilbert sont d'une part, l'existence de projections sur tout convexe fermé non vide et d'autre part, la représentation d'une forme linéaire continue par le produit scalaire.

Théorème 1.6. (Projection) *Soit K un ensemble convexe, fermé et non vide de H. Alors pour tout $f \in H$ il existe un unique élément de K, noté $P_K f$, tel que :*

$$\|f - P_K f\|_H = \min_{v\in K} \|f - v\|_H \tag{1.29}$$

qui est caractérisé par :

$$(f - P_K f | v - P_K f)_H \leq 0 \qquad \forall v \in K. \tag{1.30}$$

En outre, l'application P_K est une contraction :

$$\|P_K f_1 - P_K f_2\|_H \leq \|f_1 - f_2\|_H \qquad \forall f_1, f_2 \in H. \tag{1.31}$$

Remarque 1.1. Lorsque K est un sous-espace vectoriel de H, la condition (1.30) devient $dJ(P_K f).v = 0$, $\forall v \in K$, si on a posé $J(u) = \|u - f\|_H^2$, rejoignant l'interprétation énergétique de la formulation variationnelle.

Théorème 1.7. (Représentation de Riesz-Fréchet) *Soit $\ell \in H'$, il existe un unique élément $f \in H$ tel que :*

$$\ell(v) \underset{notation}{=} \langle \ell, v\rangle_{H',H} = (f, v)_H \qquad \forall v \in H$$

et on a :

$$\|f\|_H = \|\ell\|_{H'}$$

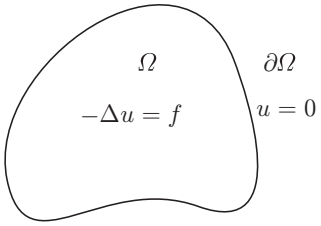
Remarque 1.2. Ce théorème repose sur la propriété de réflexivité des espaces de Hilbert (isomorphisme entre le bidual H''et H). En dimension finie, ce théorème est utilisé naturellement lorsque l'on identifie l'application dérivée avec le vecteur gradient !

1.3.3 Formulations variationnelles des problèmes de Dirichlet, Neumann et Fourier

Nous allons voir dans cette sous-section quels sont les espaces naturels qui apparaissent dans les formulations variationnelles des problèmes de Dirichlet, Neumann et Fourier.

• Problème de Dirichlet homogène

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \qquad (1.32)$$



On supposera par la suite que $f \in L^2(\Omega)$.

Si on suppose que $u \in H^2(\Omega)$ (ce qui implique $\Delta u \in L^2(\Omega)$!) alors, en vertu de la formule de Green (1.26), pour toute fonction test v qui appartient à $H^1(\Omega)$, on a :

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\Gamma$$

Si on choisit $v \in H_0^1(\Omega)$, compte tenu de la définition de $H_0^1(\Omega)$, on a :

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \qquad \forall v \in H_0^1(\Omega)$$

d'où la formulation variationnelle du problème de Dirichlet homogène (1.32) dans l'espace de Hilbert $H_0^1(\Omega)$:

$$\boxed{\begin{array}{l} \text{Trouver } u \in H_0^1(\Omega) \text{ telle que :} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \qquad \forall v \in H_0^1(\Omega) \end{array}} \qquad (1.33)$$

Remarque 1.3. Les fonctions test v et la solution u appartiennent au même espace, en l'occurence $H_0^1(\Omega)$. Par ailleurs, la condition aux limites de Dirichlet ($u = 0$) apparaît explicitement dans l'espace. On dit que la condition de Dirichlet est une condition essentielle qui s'oppose aux conditions aux limites dites naturelles qui n'apparaissent pas dans l'espace (voir la remarque 1.6).

On a établi que toute solution $u \in H^2(\Omega)$ de l'équation (1.32) vérifie la formulation variationnelle (1.33). En fait, la réciproque est également vérifiée. Choisissons dans (1.33) une fonction test $v \in \mathcal{D}(\Omega)$, ce qui est licite car $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, identifions $\partial_i u \in L^2(\Omega)$ à un élément de $\mathcal{D}'(\Omega)$, et dérivons au sens des distributions :

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, d\Omega$$
$$= \sum_{i=1}^n \langle \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \rangle = - \sum_{i=1}^n \langle \frac{\partial^2 u}{\partial x_i^2}, v \rangle = - \langle \Delta u, v \rangle.$$

Si on remplace également l'intégrale de droite dans (1.33) par le crochet de dualité, on trouve, au sens des distributions sur Ω :

$$-\Delta u = f \quad \text{dans} \quad \mathcal{D}'(\Omega).$$

Mais comme $f \in L^2(\Omega)$, et que $L^2(\Omega) \subset \mathcal{D}'(\Omega)$, on a donc montré que :

$$-\Delta u = f \quad \text{dans } L^2(\Omega),$$

c'est-à-dire au sens des fonctions (presque partout !).

On peut faire encore mieux ! En effet, lorsque $u \in H_0^1(\Omega)$ mais $u \notin H^2(\Omega)$, la formule de Green (1.26) n'est plus valide. Cependant, en travaillant au sens des distributions, on peut encore passer de (1.32) à (1.33). Cette fois, on utilise non seulement le fait que $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, mais surtout le résultat plus fort selon lequel $\mathcal{D}(\Omega)$ est par définition dense dans $H_0^1(\Omega)$. Soit donc $v \in H_0^1(\Omega)$ et $(v_k)_k$ une suite d'éléments de $\mathcal{D}(\Omega)$ qui converge vers v dans $H^1(\Omega)$. D'après (1.32), on a tout d'abord

$$\langle f, v_k \rangle = - \langle \Delta u, v_k \rangle = - \sum_{i=1}^n \langle \frac{\partial^2 u}{\partial x_i^2}, v_k \rangle = \sum_{i=1}^n \langle \frac{\partial u}{\partial x_i}, \frac{\partial v_k}{\partial x_i} \rangle.$$

Mais, d'une part, $f \in L^2(\Omega)$, et, d'autre part, $\partial_i u \in L^2(\Omega)$, pour $i = 1, n$. Ainsi, on peut remplacer les crochets de dualité initiaux et finaux par des intégrales.

$$\int_\Omega f v_k \, d\Omega = \sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v_k}{\partial x_i} \, d\Omega.$$

Enfin, comme $(v_k)_k$ converge vers v dans $H^1(\Omega)$, on sait que $(v_k)_k$ converge vers v dans $L^2(\Omega)$, mais aussi que $(\partial_i v_k)_k$ converge vers $\partial_i v$ dans $L^2(\Omega)$, pour $i = 1, n$. On peut donc passer à la limite à gauche et à droite, pour arriver à

$$\int_\Omega f v \, d\Omega = \sum_{i=1}^n \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, d\Omega \left(= \int_\Omega \nabla u \cdot \nabla v \, d\Omega \right).$$

Ceci correspond bien à la formulation variationnelle (1.33).

On a donc finalement établi le résultat suivant, qui permet de s'affranchir de l'hypothèse "artificielle" selon laquelle $u \in H^2(\Omega)$:

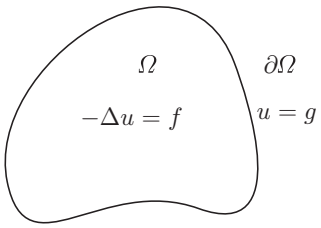
Proposition 1.13. *Soit $f \in L^2(\Omega)$. Alors $u \in H_0^1(\Omega)$ vérifie l'équation (1.32) si et seulement si u vérifie la formulation variationnelle (1.33).*

Par construction, on note que $u \in \Psi_D$.

Remarque 1.4. Dans l'énoncé de la proposition 1.13, on peut remplacer les espaces $L^2(\Omega)$ et $H_0^1(\Omega)$ respectivement par les espaces $\mathcal{C}^0(\Omega)$ et $\mathcal{C}^2(\Omega) \cap \mathcal{C}^1(\overline{\Omega})$. Ceci relie les solutions faibles aux solutions classiques, vues au §1.2.

• **Problème de Dirichlet non homogène**

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases} \qquad (1.34)$$



On suppose toujours que $f \in L^2(\Omega)$. Le choix de la régularité de la donnée g soulève une difficulté. En effet, rappelons que l'application trace γ_0 n'est pas surjective de $H^1(\Omega)$ dans $L^2(\partial\Omega)$. C'est pourquoi, nous allons faire l'hypothèse suivante sur g :

$$\begin{cases} g \in H^{1/2}(\partial\Omega) \text{ ou, ce qui est équivalent,} \\ \text{il existe } \tilde{g} \in H^1(\Omega) \text{ tel que } \gamma_0 \tilde{g} = g. \end{cases} \qquad (1.35)$$

La fonction $\tilde{g}$ s'appelle un relèvement de la fonction g, et de plus $\|\tilde{g}\|_{H^1(\Omega)}$ permet de mesurer g. En effet, d'après la proposition 1.8, on a $\|g\|_{H^{1/2}(\partial\Omega)} \leq \|\tilde{g}\|_{H^1(\Omega)}$.

Nous sommes maintenant en mesure d'établir la formulation variationnelle du problème de Dirichlet non homogène.

Supposons que la solution u de (1.34) appartienne à $H^2(\Omega)$ et multiplions par une fonction test $v \in H_0^1(\Omega)$ l'équation sur Ω. En utilisant la formule de Green (1.26), on obtient :

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} \, v \, d\Gamma = \int_\Omega f \, v \, d\Omega$$

Compte tenu du fait que $v \in H_0^1(\Omega)$, on aboutit à la formulation variationnelle :

$$\left\| \begin{array}{l} \text{trouver } u \in H^1(\Omega) \text{ tel que } u = g \text{ sur } \partial\Omega \text{ et} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f \, v d\Omega \qquad \forall v \in H_0^1(\Omega) \end{array} \right. \qquad (1.36)$$

Cette formulation n'est pas satisfaisante car l'espace des fonctions test $(H_0^1(\Omega))$ n'est pas le même que l'espace des solutions $(H^1(\Omega))$· C'est pourquoi, on effectue le changement d'inconnue :

$$\tilde{u} = u - \tilde{g}$$

qui conduit à la "bonne" formulation variationnelle :

$$\boxed{\begin{array}{l}\text{Trouver } \tilde{u} \in H_0^1(\Omega) \text{ tel que} \\ \displaystyle\int_\Omega \nabla\tilde{u}\cdot\nabla v d\Omega = \int_\Omega f v d\Omega - \int_\Omega \nabla\tilde{g}\cdot\nabla v d\Omega \qquad \forall v \in H_0^1(\Omega)\end{array}} \tag{1.37}$$

u étant alors donné par $u = \tilde{u} + \tilde{g}$.

Remarque 1.5. On s'est en fait ramené au cas homogène !

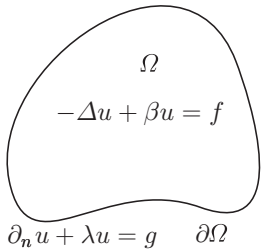
On aurait également pu raisonner au sens des distributions, lorsque $u \notin H^2(\Omega)$, pour passer de (1.34) à la formulation variationnelle (1.37). Réciproquement, à partir de (1.37) et de la relation $\tilde{u} = u - \tilde{g}$, on revient à la relation $-\Delta u = f$, comme dans le cas homogène. Et comme $\tilde{u} = 0$ sur $\partial\Omega$, on retrouve bien la seconde relation $u = g$ sur $\partial\Omega$. Bref, on peut passer de (1.37) à (1.34). On a donc cette fois le résultat :

Proposition 1.14. *Soient $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$ et $\tilde{g}$ un relèvement de g. Alors $u \in H^1(\Omega)$ vérifie l'équation (1.34) si et seulement si $\tilde{u} = u - \tilde{g}$ vérifie la formulation variationnelle (1.37).*

• Problèmes de Neumann et Fourier

Nous allons traiter simultanément le cas de Neumann et celui de Fourier en introduisant le problème général :

$$\left\{\begin{array}{ll} -\Delta u + \beta u = f & \text{dans } \Omega \\ \dfrac{\partial u}{\partial n} + \lambda u = g & \text{sur } \partial\Omega \end{array}\right. \tag{1.38}$$



où λ, β sont des constantes réelles[10], $f \in L^2(\Omega)$ et $g \in L^2(\partial\Omega)$.

Supposons encore une fois que $u \in H^2(\Omega)$. Multiplions la première équation de (1.38) par une fonction test $v \in H^1(\Omega)$ et intégrons sur Ω. En vertu de (1.26) on obtient :

$$\int_\Omega \nabla u\cdot\nabla v\, d\Omega + \beta\int_\Omega uv\, d\Omega = \int_\Omega fv\, d\Omega + \int_{\partial\Omega}\frac{\partial u}{\partial n}\, v\, d\Gamma$$

qui compte-tenu de la condition aux limites, conduit à la formulation variationnelle du problème (1.38) dans $H^1(\Omega)$:

[10] Le cas $\lambda = 0$ correspond au cas de la condition de Neumann.

$$\boxed{\begin{array}{l}\text{Trouver } u \in H^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \beta \int_\Omega uv d\Omega + \lambda \int_{\partial\Omega} uv \, d\Gamma \\ \qquad\qquad = \displaystyle\int_\Omega fv \, d\Omega + \int_{\partial\Omega} gv \, d\Gamma, \ \forall v \in H^1(\Omega)\end{array}} \tag{1.39}$$

Remarque 1.6. La principale différence avec le cas du problème de Dirichlet réside dans le fait que la condition aux limites apparaît explicitement dans la formulation variationnelle (1.39) et non plus dans l'espace. On parle de condition aux limites naturelle. Lorsque $g = 0$ et $\lambda = 0$, elle disparaît même complètement!

On a encore une propriété d'équivalence entre la formulation variationnelle et la formulation forte.

Proposition 1.15. *Soit $u \in H^2(\Omega)$. Alors u vérifie les équations (1.38) si et seulement si u vérifie la formulation variationnelle (1.39).*

Démonstration : La démonstration est ici un peu plus instructive que les précédentes, car il faut faire réapparaître la condition aux limites. Soit $u \in H^2(\Omega)$ vérifiant (1.39). Prenons $v \in \mathcal{D}(\Omega) \subset H^1(\Omega)$, on a alors

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \beta \int_\Omega uv \, d\Omega = \int_\Omega fv \, d\Omega \qquad \text{car } v = 0 \text{ sur } \partial\Omega.$$

D'après la formule de Green (1.26), il vient :

$$\int_\Omega (-\Delta u + \beta u)v \, d\Omega + \underbrace{\int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\Gamma}_{= 0 \text{ car } v \in \mathcal{D}(\Omega)} = \int_\Omega f \, v \, d\Omega$$

d'où on tire que :

$$\langle -\Delta u + \beta u, v\rangle = \langle f, v\rangle \quad \forall v \in \mathcal{D}(\Omega)$$

soit : $-\Delta u + \beta u = f$ au sens des distributions et même au sens des fonctions de $L^2(\Omega)$ car $u \in H^2(\Omega)$ et $f \in L^2(\Omega)$. On a alors en particulier que :

$$\int_\Omega (-\Delta u + \beta u)v \, d\Omega = \int_\Omega f \, v \, d\Omega, \quad \forall v \in H^1(\Omega).$$

Prenons maintenant $v \in H^1(\Omega)$ dans (1.39) et appliquons à nouveau la formule de Green :

$$\int_\Omega (-\Delta u + \beta u)v \, d\Omega + \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\Gamma + \lambda \int_{\partial\Omega} uv \, d\Gamma = \int_\Omega fv \, d\Omega + \int_{\partial\Omega} gv \, d\Gamma,$$

compte tenu de ce que nous venons d'établir, on déduit que :

$$\int_{\partial\Omega} \left(\frac{\partial u}{\partial n} + \lambda u - g \right) v \, d\Gamma = 0 \qquad \forall v \in H^1(\Omega)$$

En rappelant que l'ensemble des traces sur $\partial\Omega$ des fonctions de $H^1(\Omega)$ est dense dans $L^2(\partial\Omega)$ (cf. proposition 1.8), on déduit de l'égalité précédente que :

$$\int_{\partial\Omega} \left(\frac{\partial u}{\partial n} + \lambda u - g \right) \omega d\Gamma = 0 \qquad \forall \omega \in L^2(\partial\Omega)$$

ce qui prouve que $\dfrac{\partial u}{\partial n} + \lambda u = g$ presque partout sur $\partial\Omega$. ∎

Comme pour le problème de Dirichlet homogène, on aimerait s'affranchir de l'hypothèse "artificielle" $u \in H^2(\Omega)$. Malheureusement, l'idée de raisonner par densité de $\mathcal{D}(\Omega)$ (et donc au sens des distributions) échoue, puisque $\mathcal{D}(\Omega)$ *n'est pas dense dans* $H^1(\Omega)$. La façon correcte de procéder est de "généraliser" directement la formule de Green (1.26) en une formule d'intégration par parties valable pour des couples (u, v) de fonctions-test moins réguliers, pour ce qui concerne u (on conserve l'hypothèse $v \in H^1(\Omega)$).
Intuitivement, si on reprend (1.26), on remarque que pour $v \in H^1(\Omega)$, on a :

- $\nabla v \in L^2(\Omega)^n$: pour que le deuxième terme, $\displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega$, ait un sens, il suffit que $\nabla u \in L^2(\Omega)^n$. Ainsi, la fonction-test u doit toujours appartenir à $H^1(\Omega)$.
- $\gamma_0 v \in H^{1/2}(\partial\Omega)$: pour que le troisième terme $\displaystyle\int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\Gamma$ ait un sens, il suffit cette fois que la trace de la dérivée normale de u, $\gamma_1 u = \dfrac{\partial u}{\partial n}_{|\partial\Omega}$, définisse une forme linéaire et continue sur $H^{1/2}(\partial\Omega)$. En d'autres termes, il suffit que $\gamma_1 u$ appartienne à l'espace dual de $H^{1/2}(\partial\Omega)$ (voir les rappels du §1.3.2). On remplace dans ce cas ce deuxième terme par

  $$\langle \gamma_1 u, \gamma_0 v \rangle_{(H^{1/2}(\partial\Omega))', H^{1/2}(\partial\Omega)},$$

  où $\langle \cdot, \cdot \rangle_{(H^{1/2}(\partial\Omega))', H^{1/2}(\partial\Omega)}$ représente l'action d'une forme linéaire et continue sur $H^{1/2}(\partial\Omega)$. D'après [17], notons que l'on peut identifier le dual de $H^{1/2}(\partial\Omega)$ à l'espace de Sobolev défini sur $\partial\Omega$ d'indice $-1/2$. On écrit donc

  $$H^{-1/2}(\partial\Omega) = \Big(H^{1/2}(\partial\Omega)\Big)',$$

  avec les inclusions $H^{1/2}(\partial\Omega) \subset L^2(\partial\Omega) \subset H^{-1/2}(\partial\Omega)$.
  La question est alors : à quelle(s) condition(s) sur $u \in H^1(\Omega)$ peut-on garantir que $\gamma_1 u$ appartienne à $H^{-1/2}(\partial\Omega)$. La réponse est donnée dans [17] : cette propriété est vraie, dès lors que $u \in H^1(\Omega)$, et que $\Delta u \in L^2(\Omega)$, où la dérivation est comprise au sens des distributions. C'est-à-dire, si $u \in \Psi$, avec Ψ défini en (1.28). Qui plus est, on peut démontrer que l'application trace de la dérivée normale γ_1 est linéaire et continue de Ψ dans $H^{-1/2}(\partial\Omega)$.
- Lorsque $u \in \Psi$, on constate que le premier terme $-\displaystyle\int_\Omega \Delta u \, v \, d\Omega$ a automatiquement un sens !

En conclusion, nous aboutissons à la généralisation de (1.26) ci-dessous.

Proposition 1.16. (Formule de Green) *Soient $u \in \Psi$ et $v \in H^1(\Omega)$, on a :*

$$\boxed{\int_\Omega \Delta u \, v \, d\Omega = -\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \langle \frac{\partial u}{\partial n}_{|\partial\Omega}, v \rangle_{H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)}} \qquad (1.40)$$

Quelques commentaires additionnels, tirés de [17].

Les résultats qui précèdent sont des cas particuliers de l'étude ci-dessous. Soit

$$H(\mathrm{div};\Omega) = \left\{ \mathbf{v} \in L^2(\Omega)^n \text{ tel que } \mathrm{div}\,\mathbf{v} \in L^2(\Omega) \right\},$$

où la divergence est comprise au sens des distributions. On munit $H(\mathrm{div};\Omega)$ de la norme

$$\|\mathbf{u}\|_{H(\mathrm{div};\Omega)} = \left(\|\mathbf{u}\|^2_{L^2(\Omega)^n} + \|\,\mathrm{div}\,\mathbf{u}\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Proposition 1.17. (Trace normale) *Soit Ω un ouvert borné de $\mathbb{R}^n$, à frontière "suffisamment régulière". Alors :*

- *$\mathcal{C}^\infty(\overline{\Omega})^n$ est dense dans $H(\mathrm{div};\Omega)$;*
- *l'application trace normale*

$$\gamma_n : \left\{ \begin{array}{ccc} \mathcal{C}^\infty(\overline{\Omega})^n & \to & H^{-1/2}(\partial\Omega) \\ \mathbf{v} & \mapsto & \gamma_n \mathbf{v} = (\mathbf{v}\cdot\mathbf{n})_{|\partial\Omega} \end{array} \right.$$

se prolonge par continuité en une application linéaire continue, encore notée γ_n, de $H(\mathrm{div};\Omega)$ dans $H^{-1/2}(\partial\Omega)$ et on a l'inégalité

$$\|\gamma_n \mathbf{v}\|_{H^{-1/2}(\partial\Omega)} \leq \|\mathbf{v}\|_{H(\mathrm{div};\Omega)}.$$

- *Soient $\mathbf{u} \in H(\mathrm{div};\Omega)$ et $v \in H^1(\Omega)$, on a la formule d'intégration par parties de type Stokes :*

$$\int_\Omega (\mathrm{div}\,\mathbf{u})\, v\, d\Omega = -\int_\Omega \mathbf{u}\cdot\nabla v\, d\Omega + \langle \gamma_n \mathbf{u}, v\rangle_{H^{-1/2}(\partial\Omega),H^{1/2}(\partial\Omega)}. \tag{1.41}$$

L'étude qui précède et qui a notamment amené à la proposition 1.16 est un cas particulier de la proposition 1.17, puisque, si $u \in \Psi$, alors $\nabla u \in H(\mathrm{div};\Omega)$...

On en déduit pour finir une version plus générale de la proposition 1.15 pour résoudre le problème de Neumann et Fourier, lorsque $f \in L^2(\Omega)$ et $g \in H^{-1/2}(\partial\Omega)$. La formulation variationnelle devient alors

$$\boxed{\begin{array}{l} \text{Trouver } u \in H^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u\cdot\nabla v\, d\Omega + \beta\int_\Omega uv d\Omega + \lambda\int_{\partial\Omega} uv\, d\Gamma \\ \qquad\qquad = \displaystyle\int_\Omega fv\, d\Omega + \langle g, v\rangle_{H^{-1/2}(\partial\Omega),H^{1/2}(\partial\Omega)},\ \forall v \in H^1(\Omega) \end{array}} \tag{1.42}$$

Proposition 1.18. *Soient $f \in L^2(\Omega)$ et $g \in H^{-1/2}(\partial\Omega)$. Alors $u \in H^1(\Omega)$ vérifie les équations (1.38) si et seulement si u vérifie la formulation variationnelle (1.42).*

Démonstration : laissée en exercice. ∎

Par construction, on note que la solution du problème de Neumann homogène (c'est-à-dire avec le paramètre $\lambda = 0$ et la donnée $g = 0$) est telle que $u \in \Psi_N$.

1.3.4 D'autres exemples de formulations variationnelles

L'analyse que nous venons de faire s'étend à des situations plus générales : opérateurs à coefficients variables, équations d'ordre supérieur. Par ailleurs, on peut donner des formulations encore plus faibles.

• Problème à coefficient variable

Considérons une fonction $k(x)$, éventuellement discontinue, et le problème à coefficient variable :

$$\begin{cases} -\operatorname{div}(k(x)\nabla u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.43}$$

En multipliant par une fonction test $v \in H_0^1(\Omega)$ et en intégrant par parties, on aboutit à la formulation variationnelle suivante :

$$\left\| \begin{array}{l} \text{Trouver } u \in H_0^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega k(x)\nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega, \qquad \forall v \in H_0^1(\Omega). \end{array} \right. \tag{1.44}$$

Cette formulation variationnelle a bien un sens même si k est une fonction discontinue. Il suffit que la fonction k appartienne à $L^\infty(\Omega)$.

Remarque 1.7. En suivant la même idée, on pourrait considérer le cas d'un coefficient k dépendant de la solution u, conduisant alors à des problèmes non-linéaires plus difficiles à étudier car les résultats dépendent fortement de la nature de la fonction qui à u associe $k(u)$.

• Problèmes elliptiques d'ordre 2

L'équation de Laplace n'est qu'un cas particulier d'équations elliptiques. En effet, considérons l'opérateur aux dérivées partielles P défini par :

$$Pu = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu$$

qui lorsque :

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j > 0 \quad \forall \mathbf{X} \in \mathbb{R}^n, \quad \mathbf{X} \neq 0$$

est dit uniformément fortement elliptique (le cas du laplacien correspond à $a_{ij} = \delta_{ij}$).

Le problème de Dirichlet homogène :

$$\begin{cases} Pu = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.45}$$

admet pour formulation variationnelle :

Trouver $u \in H_0^1(\Omega)$ tel que :

$$\sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, d\Omega + \sum_{i=1}^{n} \int_\Omega b_i \frac{\partial u}{\partial x_i} \, v \, d\Omega + \int_\Omega c \, u \, v \, d\Omega = \int_\Omega f \, v \, d\Omega, \ \forall v \in H_0^1(\Omega) \tag{1.46}$$

les coefficients a_{ij}, b_i et c pouvant dépendre de x.

• Problèmes d'ordre supérieur

A titre d'exemple considérons le bilaplacien muni de conditions aux limites de Dirichlet et tangentielle :

$$\begin{cases} \Delta^2 u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \Delta u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.47}$$

Multiplions par une fonction test $v \in H^2(\Omega)$ telle que $v = 0$ sur $\partial\Omega$. Formellement, appliquons deux fois la formule de Green (1.26), on obtient :

$$\int_\Omega \Delta^2 u v \, d\Omega = \int_\Omega \Delta u \Delta v \, d\Omega + \int_{\partial\Omega} \left(\frac{\partial}{\partial n} (\Delta u) \, v - \Delta u \, \frac{\partial v}{\partial n} \right) d\Gamma.$$

Compte tenu des conditions aux limites, il vient :

$$\int_\Omega \Delta u \Delta v \, d\Omega = \int_\Omega f v \, d\Omega \quad \forall v \in H^2(\Omega), \ v = 0 \text{ sur } \partial\Omega \tag{1.48}$$

qui constitue une formulation variationnelle de (1.47) dans l'espace :

$$H^2(\Omega) \cap H_0^1(\Omega) = \left\{ v \in H^2(\Omega), v = 0 \text{ sur } \partial\Omega \right\}.$$

La condition $u = 0$ sur $\partial\Omega$ apparaît comme une condition essentielle du problème et la condition $\Delta u = 0$ sur $\partial\Omega$ comme une condition naturelle. Notons que l'ordre de dérivation a diminué de deux dans la formulation variationnelle.

• Données moins régulières

Dans les exemples précédents, on a toujours considéré $f \in L^2(\Omega)$. En fait on peut se contenter de données moins régulières. Par exemple pour le problème de Dirichlet, choisissons $f \in H^{-1}(\Omega)$, le dual de $H_0^1(\Omega)$, inclus dans l'espace des distributions $\mathcal{D}'(\Omega)$. On introduit alors la formulation faible suivante du problème de Dirichlet homogène :

$$\left\| \begin{array}{l} \text{Trouver } u \in H_0^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \langle f, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \qquad \forall v \in H_0^1(\Omega) \end{array} \right. \tag{1.49}$$

dont la solution vérifie :

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{au sens des distributions sur } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{array} \right.$$

$\langle \cdot, \cdot \rangle_{H^{-1}(\Omega), H_0^1(\Omega)}$ désigne la dualité entre $H_0^1(\Omega)$ et son dual $H^{-1}(\Omega)$.

Il est bien clair que l'on ne peut pas considérer des données aussi peu régulières que l'on veut (distribution quelconque) sous peine de perdre l'existence et l'unicité.

L'exemple le plus "simple" est le suivant. Si, pour le problème de Neumann et Fourier, on choisit $f \in H^{-1}(\Omega)$, on perd l'existence !? En effet, la formulation variationnelle (1.42) est posée pour v quelconque dans $H^1(\Omega)$. Or, f se trouve dans le dual de $H_0^1(\Omega)$: on ne peut donc pas définir *a priori* l'action de f sur un élément v de $H^1(\Omega) \setminus H_0^1(\Omega)$... Cet exemple illustre en particulier la différence entre condition aux limites essentielle (pour le problème de Dirichlet) et condition aux limites naturelle (pour le problème de Neumann et Fourier).

1.4 Existence et unicité des solutions faibles

Toutes les formulations variationnelles que nous avons présentées à la section précédente s'inscrivent dans le cadre abstrait suivant :

$$\boxed{\begin{array}{l} \text{Trouver } \; u \in H \text{ tel que :} \\ a(u, v) = \ell(v) \qquad \forall v \in H \end{array}} \tag{1.50}$$

où
– H est un espace de Hilbert, muni du produit scalaire $(\cdot, \cdot)_H$,
– $a(\cdot, \cdot)$ est une forme bilinéaire continue sur H,
– $\ell(\cdot)$ est une forme linéaire continue sur H.
Ainsi, on a pour les problèmes de Dirichlet (1.32) et Neumann-Fourier (1.38) :

	Dirichlet	Neumann − Fourier
H	$H_0^1(\Omega)$	$H^1(\Omega)$
$a(u,v)$	$\displaystyle\int_\Omega \nabla u \cdot \nabla v d\Omega$	$\displaystyle\int_\Omega (\nabla u \cdot \nabla v + \beta uv) d\Omega + \int_{\partial\Omega} \lambda uv d\Gamma$
$\ell(v)$	$\displaystyle\int_\Omega fv d\Omega$	$\displaystyle\int_\Omega fv d\Omega + \int_{\partial\Omega} gv d\Gamma$

On choisit ici des données $f \in L^2(\Omega)$ et $g \in L^2(\partial\Omega)$.

Avant d'exposer la théorie générale de l'existence et de l'unicité du problème abstrait (1.50), nous allons mettre en évidence à l'aide d'exemples, les principes sous-jacents à cette théorie.

1.4.1 Le problème de Neumann

Plaçons-nous dans le cas $\beta = 1$ et λ=0. On a alors :

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + uv) d\Omega = (u,v)_{H^1(\Omega)} \quad \text{(produit scalaire sur } H^1(\Omega))$$

et

$$\ell(v) = \int_\Omega f v d\Omega + \int_{\partial\Omega} g v d\Omega.$$

L'application

$$\ell \; : \; \left\{ \begin{array}{ccc} H^1(\Omega) & \to & \mathbb{R} \\ v & \mapsto & \ell(v) \end{array} \right.$$

est clairement linéaire et continue sur $H^1(\Omega)$ car :

$$\left| \int_\Omega f v \, d\Omega \right| \underset{(Cauchy-Schwarz)}{\leq} \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} \tag{1.51}$$

et

$$\begin{array}{l} \left| \int_{\partial\Omega} g v \, d\Gamma \right| \underset{(Cauchy-Schwarz)}{\leq} \|g\|_{L^2(\partial\Omega)} \|\gamma_0 v\|_{L^2(\partial\Omega)} \\ \qquad \underset{(continuité \; de \; la \; trace \; (1.23))}{\leq} C \, \|g\|_{L^2(\partial\Omega)} \|v\|_{H^1(\Omega)} \end{array} \tag{1.52}$$

L'application du théorème de Riesz fournit alors l'existence et l'unicité d'une solution $u \in H^1(\Omega)$ à la formulation variationnelle (1.50) du problème de Neumann et l'estimation de continuité suivante :

$$\|u\|_{H^1(\Omega)} = \|\ell\|_{\left(H^1(\Omega)\right)'} = \sup_{v \neq 0} \frac{|\ell(v)|}{\|v\|_{H^1(\Omega)}}$$

Or on a, en vertu des estimations (1.51) et (1.52) :

$$\sup_{v \neq 0} \frac{|\ell(v)|}{\|v\|_{H^1(\Omega)}} \leq \max(1, C) \, \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right)$$

On énonce donc le résultat :

Théorème 1.8. (Unicité et existence pour le problème de Neumann) *Soit $f \in L^2(\Omega)$ et $g \in L^2(\partial\Omega)$ alors il existe une unique solution faible $u \in H^1(\Omega)$ au problème de Neumann :*

$$\left\{ \begin{array}{ll} -\Delta u + u = f & \text{dans } \Omega \\ \dfrac{\partial u}{\partial n} = g & \text{sur } \partial\Omega \end{array} \right.$$

De plus, la solution u dépend continûment des données f et g, c'est-à-dire qu'il existe une constante C indépendante de u, f et g telle que :

$$\|u\|_{H^1(\Omega)} \leq C \left(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right).$$

Remarque 1.8. Cette notion de dépendance continue est également appelée stabilité par rapport aux données.

1.4.2 Le problème de Dirichlet - L'inégalité de Poincaré

Dans le cas du problème de Dirichlet homogène, la forme bilinéaire

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

ne définit pas, a priori, une norme équivalente à la norme induite par $H^1(\Omega)$ sur $H_0^1(\Omega)$:

$$\|v\|^2_{H^1(\Omega)} = \int_\Omega v^2 d\Omega + \int_\Omega |\nabla v|^2 d\Omega.$$

On a seulement l'inégalité :

$$a(v,v) = \int_\Omega |\nabla v|^2 d\Omega \leq \|v\|^2_{H^1(\Omega)} \qquad \forall v \in H^1(\Omega). \tag{1.53}$$

Par conséquent, le théorème de représentation de Riesz ne s'applique pas avec le produit scalaire usuel de $H^1(\Omega)$.

Cependant, on peut démontrer une inégalité réciproque, satisfaite seulement pour les fonctions appartenant à $H_0^1(\Omega)$! C'est l'inégalité de Poincaré.

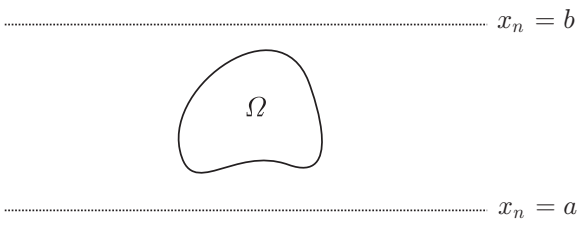
Proposition 1.19. (Inégalité de Poincaré) *Soit Ω un ouvert borné de $\mathbb{R}^n$, il existe une constante C_p, dépendant seulement de Ω telle que :*

$$\int_\Omega v^2 d\Omega \leq C_p \int_\Omega |\nabla v|^2 d\Omega \qquad \forall v \in H_0^1(\Omega) \tag{1.54}$$

Remarque 1.9. Cette inégalité est trivialement fausse dans $H^1(\Omega)$ (prendre $v = 1$)!

Démonstration : Par densité de $\mathcal{D}(\Omega)$ dans $H_0^1(\Omega)$, on se ramène aux fonctions v qui sont $\mathcal{C}^\infty$ dans Ω et à support compact dans Ω. On note $\tilde{v}$ le prolongement de v par 0 à l'extérieur de Ω ($\tilde{v}$ est défini sur $\mathbb{R}^n$).

Comme Ω est supposé borné dans $\mathbb{R}^n$, il est nécessairement contenu dans une bande de l'espace $\{(x_1, \cdots, x_n) \in \mathbb{R}^n$ tels que $a \leq x_n \leq b\}$: on pose $x = (x', x_n)$ avec $x' = (x_1, x_2, \cdots, x_{n-1})$.



v étant régulier, on peut écrire :

$$\tilde{v}\left(x', x_n\right) = \int_a^{x_n} \frac{\partial \tilde{v}}{\partial x_n}\left(x', t\right) dt$$

qui, en vertu de Cauchy-Schwarz, conduit à l'estimation :

$$\left|\tilde{v}\left(x', x_n\right)\right|^2 \leq (x_n - a) \int_a^{x_n} \left|\frac{\partial \tilde{v}}{\partial x_n}\left(x', t\right)\right|^2 dt \leq (x_n - a) \int_{-\infty}^{+\infty} \left|\frac{\partial \tilde{v}}{\partial x_n}\left(x', t\right)\right|^2 dt.$$

A l'aide du théorème de Fubini, ceci permet de montrer :

$$\int_{\mathbb{R}^{n-1}} \left|\tilde{v}\left(x', x_n\right)\right|^2 dx' \leq (x_n - a) \int_{\mathbb{R}^n} \left|\frac{\partial \tilde{v}}{\partial x_n}(x', t)\right|^2 dx' dt.$$

En intégrant, entre a et b suivant x_n, il vient finalement :

$$\int_{\mathbb{R}^n} \tilde{v}^2 dx \leq \int_a^b (x_n - a)\, dx_n \int_{\mathbb{R}^n} \left|\frac{\partial \tilde{v}}{\partial x_n}\right|^2 dx = \frac{(b-a)^2}{2} \int_{\mathbb{R}^n} \left|\frac{\partial \tilde{v}}{\partial x_n}\right|^2 dx$$

qui fournit l'inégalité souhaitée, car $\tilde{v} = 0$ en dehors de Ω :

$$\|v\|^2_{L^2(\Omega)} \leq \frac{1}{2}(b-a)^2 \int_\Omega |\nabla v|^2 d\Omega \quad \forall v \in \mathcal{D}(\Omega).$$

Par densité, cette dernière inégalité reste vraie pour tout $v \in H_0^1(\Omega)$. ∎

Remarque 1.10. D'après la démonstration fournie, il suffit que Ω soit un ouvert borné dans une direction pour que l'inégalité de Poincaré soit vérifiée. En effet, après un éventuel changement de base dans $\mathbb{R}^n$, on peut toujours supposer que l'on se trouve dans le cas où $\Omega \subset \{(x_1, \cdots, x_n) \in \mathbb{R}^n \text{ tels que } a \leq x_n \leq b\}$.

De l'inégalité de Poincaré (1.54), on tire immédiatement que :

$$\int_\Omega |\nabla v|^2 d\Omega \geq \frac{1}{1 + C_p} \|v\|^2_{H^1(\Omega)}, \qquad \forall v \in H_0^1(\Omega). \tag{1.55}$$

Ceci prouve, compte tenu de (1.53), que

$$a(v, v) = \int_\Omega |\nabla v|^2 d\Omega$$

définit une norme équivalente à la norme $H^1(\Omega)$, sur l'espace $H_0^1(\Omega)$.
Usuellement, on introduit la semi-norme de $H^1(\Omega)$:

$$|v|_1 = \left(\int_\Omega |\nabla v|^2 d\Omega\right)^{1/2}. \tag{1.56}$$

C'est donc une norme sur $H_0^1(\Omega)$, associée au produit scalaire :

$$(u, v)_{H_0^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, d\Omega. \tag{1.57}$$

Ce résultat nous permet maintenant d'appliquer le théorème de Riesz en choisissant le produit scalaire (1.57) sur $H_0^1(\Omega)$. On obtient ainsi un théorème d'existence et d'unicité pour le problème de Dirichlet homogène, car $\ell(v) = \int_\Omega f v \, d\Omega$ demeure une forme linéaire continue sur $H_0^1(\Omega)$, compte tenu de l'estimation :

$$\left| \int_\Omega f \, v \, d\Omega \right| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \underset{(Poincaré)}{\leq} C_p^{1/2} \, \|f\|_{L^2(\Omega)} |v|_1.$$

On énonce donc le résultat :

Théorème 1.9. (Existence et unicité pour le problème de Dirichlet) *Soit $f \in L^2(\Omega)$ alors il existe une unique solution faible $u \in H_0^1(\Omega)$ au problème de Dirichlet homogène :*

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

qui dépend continûment de la donnée f : il existe $C > 0$ indépendante de u et f telle que

$$\|u\|_{H^1(\Omega)} \leq (1 + C_p)^{1/2} |u|_1 \leq C \|f\|_{L^2(\Omega)}.$$

Lorsque Ω est borné (dans toutes les directions), il existe une démonstration par l'absurde de l'inégalité de Poincaré s'appuyant sur l'injection compacte de $H_0^1(\Omega)$ dans $L^2(\Omega)$ (voir le théorème 1.5). Dans le cas où Ω est de plus connexe, cette démonstration permet, par ailleurs, d'établir des généralisations très utiles de l'inégalité de Poincaré :

- Inégalité de Poincaré-Wirtinger

$$\int_\Omega v^2 d\Omega \leq C \left(\int_\Omega |\nabla v|^2 d\Omega + \frac{1}{\text{mes}\Omega} \left(\int_\Omega v \, d\Omega \right)^2 \right) \quad \forall v \in H^1(\Omega). \tag{1.58}$$

- Inégalité de Poincaré-Friedrichs ($\Gamma \subset \partial\Omega, \ \text{mes}\Gamma > 0$)

$$\|v\|_{H^1(\Omega)}^2 \leq C \left(\int_\Omega |\nabla v|^2 d\Omega + \left(\int_\Gamma v \, d\Omega \right)^2 \right) \quad \forall v \in H^1(\Omega). \tag{1.59}$$

Indiquons l'idée de la démonstration de l'inégalité de Poincaré-Friedrichs.

On raisonne par l'absurde. Soit (y_k) une suite d'éléments de $H^1(\Omega)$ telle que :

$$\|y_k\|_{H^1(\Omega)} = 1 \quad \text{et} \quad |y_k|_+ = \int_\Omega |\nabla y_k|^2 \, d\Omega + \left(\int_\Gamma y_k d\Gamma \right)^2 \leq \frac{1}{k}, \quad \forall k \geq 1.$$

(y_k) est une suite bornée de $H^1(\Omega)$, qui s'injecte de façon compacte dans $L^2(\Omega)$. Par conséquent, il existe une sous-suite $(y_{k'})$ qui converge vers $y \in L^2(\Omega)$ ($\|y_{k'} - y\|_{L^2(\Omega)} \longrightarrow 0$)). Vérifions que $\nabla y = 0$ au sens des distributions. Soit $\varphi \in \mathcal{D}(\Omega)$:

$$\begin{aligned}\langle \partial_i y, \varphi\rangle &= -\langle y, \partial_i\varphi\rangle = -\int_\Omega y\partial_i\varphi\, d\Omega = -\int_\Omega \lim_{k'\to\infty} y_{k'}\partial_i\varphi\, d\Omega \\ &= -\lim_{k'\to\infty}\int_\Omega y_{k'}\partial_i\varphi\, d\Omega = -\lim_{k'\to\infty}\langle y_{k'}, \partial_i\varphi\rangle = \lim_{k'\to\infty}\langle \partial_i y_{k'}, \varphi\rangle \\ &= \lim_{k'\to\infty}\int_\Omega \partial_i y_{k'}\varphi\, d\Omega = 0.\end{aligned}$$

Ainsi, $y \in H^1(\Omega)$ et $\nabla y = 0$; on a donc $\|y_{k'} - y\|_{H^1(\Omega)} \longrightarrow 0$. En passant à la limite, on a, par continuité de la trace γ_0, $\displaystyle\int_\Gamma y\, d\Gamma = \lim_{k'\to\infty}\int_\Gamma y_{k'}\, d\Gamma = 0$. Par conséquent, d'une part, y est égale à une constante sur Ω (car Ω est connexe) et, d'autre part, l'intégrale de cette constante sur Γ est nulle ! Ainsi, $y = 0$ ce qui contredit l'hypothèse $\|y\|_{H^1(\Omega)} = \lim_{k'\to\infty}\|y_{k'}\|_{H^1(\Omega)} = 1$.

• Le problème de Dirichlet non homogène

L'existence et l'unicité d'une solution au problème de Dirichlet non homogène découlent de l'analyse précédente.

En effet, rappelons que sous l'hypothèse $g \in H^{1/2}(\partial\Omega)$ (voir (1.35)), on se ramène à une formulation variationnelle en $\tilde{u} = u - \tilde{g}$ posée dans $H^1_0(\Omega)$ avec

$$a(\tilde{u}, v) = \int_\Omega \nabla\tilde{u}\cdot\nabla v\, d\Omega \quad \text{et} \quad \ell(v) = \int_\Omega fv\, d\Omega - \int_\Omega \nabla\tilde{g}\cdot\nabla v\, d\Omega$$

Il suffit donc de vérifier que $\ell(v)$ est une forme linéaire continue sur $H^1_0(\Omega)$ pour la norme (1.56). Or on a, en utilisant les inégalités de Cauchy-Schwarz et de Poincaré :

$$\begin{aligned}|\ell(v)| &\le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\nabla\tilde{g}\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \\ &\le C_p\|f\|_{L^2(\Omega)}|v|_1 + \|\nabla\tilde{g}\|_{L^2(\Omega)}|v|_1\end{aligned}$$

Ceci prouve qu'il existe une unique solution $\tilde{u} \in H^1_0(\Omega)$ à la formulation variationnelle (1.37). Mais ça n'est pas fini...

Théorème 1.10. (Existence et unicité pour le problème de Dirichlet non homogène) *Soient $f \in L^2(\Omega)$ et $g \in H^{1/2}(\partial\Omega)$. Alors il existe une unique solution faible $u \in H^1(\Omega)$ au problème de Dirichlet non homogène :*

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases}$$

qui dépend continûment des données f et g : il existe $C > 0$ indépendante de u, f et g telle que

$$\|u\|_{H^1(\Omega)} \le C(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}).$$

Démonstration : On a déjà établi l'existence et l'unicité de la solution $\tilde{u} \in H^1_0(\Omega)$ de la formulation variationnelle (1.37). Par ailleurs, si on choisit $v = \tilde{u}$ dans (1.37), on trouve :

$$\begin{aligned}|\tilde{u}|_1^2 &\le \|f\|_{L^2(\Omega)}\|\tilde{u}\|_{L^2(\Omega)} + |\tilde{g}|_1\,|\tilde{u}|_1 \\ &\le \left(C_p^{1/2}\|f\|_{L^2(\Omega)} + |\tilde{g}|_1\right)|\tilde{u}|_1, \quad \text{d'où} \\ |\tilde{u}|_1 &\le C_p^{1/2}\|f\|_{L^2(\Omega)} + |\tilde{g}|_1. \end{aligned} \tag{1.60}$$

Existence de u. Comme $\tilde{u}$ est solution de (1.37), $u = \tilde{u} + \tilde{g}$ appartient à $H^1(\Omega)$ car $\tilde{u} \in H_0^1(\Omega)$ et $\tilde{g} \in H^1(\Omega)$ par hypothèse, et par conséquent u est solution du problème de Dirichlet non homogène.
Unicité de u. Soient u_1 et u_2 deux solutions du problème de Dirichlet non homogène. Posons $w = u_1 - u_2$. Il est clair que w vérifie le problème de Dirichlet homogène :

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

dont l'unique solution est $w = 0$ en vertu du théorème 1.9. Ce qui prouve que u est unique.
Dépendance continue de u par rapport aux données. D'après la proposition 1.8, on peut choisir un relèvement $\tilde{g}_0$ de g tel que $\|\tilde{g}_0\|_{H^1(\Omega)} \leq 2\|g\|_{H^{1/2}(\partial\Omega)}$. Sinon, ceci contredirait le fait que $\|g\|_{H^{1/2}(\partial\Omega)} = \inf_{v \in H^1(\Omega) \ tq \ \gamma_0 v = g} \|v\|_{H^1(\Omega)}$! Pour ce relèvement, on note $\tilde{u}_0$ la solution de la formulation variationnelle (1.37). On a :

$$\begin{aligned} \|u\|_{H^1(\Omega)} = \|\tilde{u}_0 + \tilde{g}_0\|_{H^1(\Omega)} &\leq \|\tilde{u}_0\|_{H^1(\Omega)} + \|\tilde{g}_0\|_{H^1(\Omega)} \\ &\leq |\tilde{u}_0|_1 + \|\tilde{u}_0\|_{L^2(\Omega)} + \|\tilde{g}_0\|_{H^1(\Omega)} \\ \text{(Poincaré)} \qquad &\leq (1 + C_p^{1/2})|\tilde{u}_0|_1 + \|\tilde{g}_0\|_{H^1(\Omega)} \end{aligned}$$

qui compte tenu de l'estimation (1.60) et du choix de $\tilde{g}_0$, permet d'établir le résultat de continuité. En effet, on écrit :

$$\begin{aligned} \|u\|_{H^1(\Omega)} &\leq (1 + C_p^{1/2})(C_p^{1/2}\|f\|_{L^2(\Omega)} + |\tilde{g}_0|_1) + \|\tilde{g}_0\|_{H^1(\Omega)} \\ &\leq (1 + C_p^{1/2})C_p^{1/2}\|f\|_{L^2(\Omega)} + 2(2 + C_p^{1/2})\|g\|_{H^{1/2}(\partial\Omega)} \\ &\leq \max((1 + C_p^{1/2})C_p^{1/2}, 2(2 + C_p^{1/2}))\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)}\right) \end{aligned}$$

c'est-à-dire une dépendance continue par rapport à g et f. ∎

1.4.3 Le théorème de Lax-Milgram

Les démonstrations d'existence et d'unicité que nous venons de voir sur des cas particuliers mettent en évidence deux points essentiels :
– le fait que la forme bilinéaire $a(u, v)$ définisse une norme équivalente sur l'espace de Hilbert H.
– l'utilisation du théorème de représentation de Riesz.
Le premier point découle d'une propriété très forte dite de coercivité de la forme bilinéaire $a(\cdot, \cdot)$.

Définition 1.11. (Coercivité) *On dit qu'une forme bilinéaire est coercive sur H si et seulement si il existe une constante $\alpha > 0$ telle que :*

$$a(v, v) \geq \alpha\|v\|_H^2 \qquad \forall v \in H. \tag{1.61}$$

Remarque 1.11. En dimension finie, cette propriété exprime le caractère définipositif d'une application bilinéaire.

Cette notion nous permet de formuler un résultat général d'existence et d'unicité pour le problème variationnel abstrait (1.50), qui, de surcroît, englobe le cas des formes bilinéaires non nécessairement symétriques :

Théorème 1.11. (Lax-Milgram) *Soit H un espace de Hilbert, $a(\cdot,\cdot)$ une forme bilinéaire continue et coercive sur H, $\ell(\cdot)$ une forme linéaire continue sur H. Alors il existe une unique solution $u \in H$ au problème*

$$a(u,v) = \ell(v) \qquad \forall v \in H.$$

En outre, la solution u dépend continûment de la forme linéaire ℓ :

$$\|u\|_H \leq C \|\ell\|_{H'}.$$

Démonstration : D'une part, d'après le théorème de représentation de Riesz, il existe un unique élément $L \in H$ tel que :

$$\ell(v) = (L,v)_H, \quad \forall v \in H.$$

D'autre part, l'application $v \mapsto a(u,v)$ définit une forme linéaire continue sur H. Toujours grâce au théorème de représentation, il existe un unique élément de H, noté Au, tel que :

$$(Au,v)_H = a(u,v), \quad \forall v \in H,$$

avec l'estimation :

$$\|Au\|_H = \sup_{v\neq 0} \frac{|a(u,v)|}{\|v\|_H} \leq C_a \|u\|_H \quad (C_a \text{ module de continuité de } a(\cdot,\cdot)). \tag{1.62}$$

Ceci montre que l'application

$$A : \left\{ \begin{array}{ccc} H & \longrightarrow & H \\ u & \mapsto & Au \end{array} \right.$$

est continue sur H. Elle vérifie de plus, compte tenu de la coercivité :

$$(Au,u)_H \geq \alpha \|u\|_H^2. \tag{1.63}$$

Le problème $a(u,v) = \ell(v) \ \ \forall v \in H$, est alors équivalent au problème :

$$(Au - L, v)_H = 0 \quad \forall v \in H,$$

et donc à l'équation fonctionnelle :

$$Au = L \quad \text{dans } H. \tag{1.64}$$

On définit l'opérateur S sur H de la façon suivante :

$$Sv = -\rho Av + \rho L + v \qquad \text{avec } \ \rho > 0.$$

On a :

$$\|Sv_1 - Sv_2\|_H^2 = \|-\rho A(v_1 - v_2) + v_1 - v_2\|_H^2$$

soit en développant et en utilisant les estimations (1.62) et (1.63) :

$$\begin{aligned} \|Sv_1 - Sv_2\|_H^2 &= \|v_1 - v_2\|_H^2 + \rho^2 \|Av_1 - Av_2\|_H^2 - 2\rho (Av_1 - Av_2, v_1 - v_2)_H \\ &\leq (1 - 2\rho\alpha + \rho^2 C_a^2) \|v_1 - v_2\|_H^2 \end{aligned}$$

En prenant $0 < \rho < \dfrac{2\alpha}{C_a^2}$, on obtient :

$$\|Sv_1 - Sv_2\|_H \leq k_\rho \|v_1 - v_2\|_H \quad \text{avec } k_\rho < 1, \tag{1.65}$$

qui prouve que S est une application contractante de H.
En utilisant le théorème du point fixe de Banach on en déduit que S admet un unique point fixe $u = Su$ dans H. D'après la définition de S, u vérifie (1.64). Enfin, on a, en vertu de (1.63) :

$$\alpha\|u\|_H^2 \leq (Au,u)_H = (L,u)_H \leq \|L\|_H\|u\|_H = \|\ell\|_{H'}\|u\|_H,$$

qui établit que la solution u dépend continûment de ℓ : $\|u\|_H \leq \alpha^{-1}\|l\|_{H'}$. ∎

Remarque 1.12. Lorsque la forme bilinéaire $a(\cdot,\cdot)$ est de plus symétrique, la démonstration devient immédiate car $a(\cdot,\cdot)$ définit une norme équivalente sur l'espace de Hilbert H et le théorème de Riesz répond à la question (nous avons déjà utilisé cette propriété précédemment). Par ailleurs, si on introduit la fonctionnelle :

$$J(u) = \frac{1}{2}a(u,u) - l(u),$$

on a (cf. [9]) :

$$dJ(u).v = (Au - L, v), \text{ ou } \nabla J(u) = Au - L,$$

montrant ainsi que la solution u, lorsque la forme bilinéaire est symétrique, minimise la fonctionnelle J. Dans le contexte de l'optimisation quadratique, la coercivité s'interprète comme une condition de stricte convexité de la fonctionnelle J.

• Exemple d'application du théorème de Lax-Milgram

On suppose ici que l'ouvert Ω est (borné) et connexe. Considérons le problème de Fourier ($\beta = 0$ et $\lambda > 0$) :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \qquad \text{avec } f \in L^2(\Omega) \\ \dfrac{\partial u}{\partial n} + \lambda u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.66}$$

dont une formulation variationnelle dans $H^1(\Omega)$ est :

$$\left\| \begin{array}{l} \text{Trouver } u \in H^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega + \lambda \int_{\partial\Omega} uv \, d\Gamma = \int_\Omega fv \, d\Omega \qquad \forall v \in H^1(\Omega). \end{array} \right. \tag{1.67}$$

Appliquons le théorème de Lax-Milgram avec $H = H^1(\Omega)$,

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, d\Omega + \lambda \int_{\partial\Omega} uv \, d\Gamma \qquad \text{et} \quad \ell(v) = \int_\Omega fv \, d\Omega$$

Comme $f \in L^2(\Omega)$, $\ell(\cdot)$ est une forme linéaire continue sur $H^1(\Omega)$ (voir (1.51)). $a(\cdot,\cdot)$ est clairement une forme bilinéaire sur $H^1(\Omega)$. Elle est continue sur $H^1(\Omega)$ car :

$$\left| \int_\Omega \nabla u \cdot \nabla v \, d\Omega \right| \leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}$$

et

$$\left|\int_{\partial\Omega} uv\, d\Gamma\right| \leq \|u\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)} \leq C^2\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \quad \text{(théorème 1.2)}.$$

Par ailleurs, comme on a :

$$\left(\int_{\partial\Omega} v\, d\Gamma\right)^2 \leq \text{mes}(\partial\Omega)\int_{\Omega} v^2 d\Gamma \quad (\text{d'après Cauchy} - \text{Schwarz})$$

on déduit de l'inégalité de Poincaré-Friedrichs (1.59) que $\forall v \in H^1(\Omega)$:

$$\int_{\partial\Omega} v^2 d\Gamma + \int_{\Omega} |\nabla v|^2 d\Omega \geq C\int_{\Omega} \left(v^2 + |\nabla v|^2\right) d\Omega$$

conduisant à l'estimation :

$$a(v,v) \geq \min(1,\lambda)\left(\int_{\Omega} |\nabla v|^2 d\Omega + \int_{\partial\Omega} v^2 d\Gamma\right) \geq C\ \min(1,\lambda)\|v\|^2_{H^1(\Omega)}$$

qui prouve que $a(\cdot,\cdot)$ est coercive sur $H^1(\Omega)$. La formulation variationnelle (1.67) admet donc une unique solution $u \in H^1(\Omega)$.

Lorsque $\lambda = 0$, le problème de Fourier dégénère en un problème de Neumann dont la formulation variationnelle découle de (1.67) :

$$\left\| \begin{array}{l} \text{Trouver } u \in H^1(\Omega) \text{ tel que :} \\ \displaystyle\int_{\Omega} \nabla u \cdot \nabla v\, d\Omega = \int_{\Omega} fv\, d\Omega \qquad \forall v \in H^1(\Omega). \end{array} \right.$$

On constate, en choisissant $v = 1$ dans cette formulation, que f doit vérifier la condition, dite de compatibilité :

$$\int_{\Omega} f\, d\Omega = 0.$$

C'est une condition nécessaire sur f pour que le problème soit bien posé.
En outre, si cette condition est satisfaite il est clair qu'il n'y a pas unicité de la solution car toutes les fonctions constantes u_{cste} vérifient :

$$\int_{\Omega} \nabla u_{cste} \cdot \nabla v\, d\Omega = 0 \quad \forall v \in H^1(\Omega).$$

A l'aide de l'inégalité de Poincaré-Wirtinger (1.58), on peut montrer que si la condition de compatibilité est vérifiée, alors le problème de Neumann précédent admet une solution dans $H^1(\Omega)$, unique à une constante additive près. Pour cela, on se place dans le sous-espace (de $H^1(\Omega)$) égal à $\{v \in H^1(\Omega)$ tel que $\displaystyle\int_{\Omega} v\, d\Omega = 0\}$. Dans ce sous-espace, $|\cdot|_1$ est bien une norme équivalente à $\|\cdot\|_{H^1(\Omega)}$, d'après (1.58).

Remarque 1.13. On a toujours supposé que $a(\cdot,\cdot)$ est à valeurs dans $\mathbb{R}$. Il existe des situations où $a(\cdot,\cdot)$ est à valeurs complexes. La coercivité de $a(\cdot,\cdot)$ prend alors la forme :

$$|a(v,v)| \geq \alpha \|v\|_H^2 \qquad (|\cdot| \text{ désignant le module})\cdot$$

On peut encore énoncer le théorème de Lax-Milgram, sa démonstration étant cependant différente de celle proposée auparavant (on prouve directement l'injectivité et la surjectivité de l'opérateur A associé à la forme bilinéaire $a(\cdot,\cdot)$, voir par exemple [21]).

Une généralisation du théorème de Lax-Milgram qu'il est utile de connaître est le théorème de Stampacchia qui fournit un résultat d'existence et d'unicité pour l'inéquation variationnelle [11] :

$$\left\| \begin{array}{l} \text{Trouver } u \in K, \text{ tel que :} \\ a(u, v-u) \geq \ell(v-u) \quad \forall v \in K \end{array} \right. \tag{1.68}$$

où K est un convexe fermé non-vide de l'espace de Hilbert H.

Sous les hypothèses du théorème 1.11, on démontre qu'il existe une unique solution $u \in K$ au problème (1.68).

En effet, à l'aide du théorème de représentation de Riesz on se ramène au problème :

$$(Au, v-u)_H \geq (L, v-u)_H \quad \forall v \in K$$

faisant apparaître naturellement l'opérateur P_K de projection sur K (voir théorème 1.6). On introduit alors l'opérateur :

$$Sv = P_K(\rho L - \rho A v + v), \quad \rho > 0$$

pour lequel on démontre, en vertu de (1.31), que pour ρ petit, il existe $0 < k_\rho < 1$ tel que :

$$\|Sv_1 - Sv_2\|_H \leq \|P_K(v_1 - v_2 - \rho\,(Av_1 - Av_2))\|_H \leq k_\rho \|v_1 - v_2\|_H.$$

On conclut par un argument de point fixe.

1.5 Quelques propriétés des solutions faibles

Terminons ce chapitre, en indiquant brièvement, quelques exemples de résultats portant sur les solutions faibles : principe de positivité, principe du maximum et régularité des solutions en fonction de celles des données, qui sont à mettre en regard des propriétés des solutions classiques présentées à la section 1.2.

[11] On rencontre ce type d'inéquations pour des problèmes de contact.

1.5.1 Principe du Maximum

Nous allons indiquer, sur le cas du problème de Dirichlet, la démarche qui permet d'obtenir un principe de positivité sur la solution faible et par conséquent un principe du maximum.

Considérons le problème de Dirichlet suivant :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases} \tag{1.69}$$

avec $f \in L^2(\Omega)$ et $g \in H^{1/2}(\partial\Omega)$.

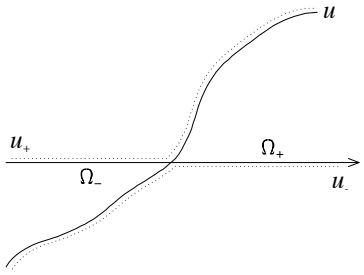
On a le principe de positivité suivant :

Proposition 1.20. (Positivité) *Soit $f \geq 0$ presque partout sur Ω et $g \geq 0$ presque partout sur $\partial\Omega$ (resp. $f \leq 0$ et $g \leq 0$). Alors la solution $u \in H^1(\Omega)$ du problème de Dirichlet (1.69) est positive (resp. négative) presque partout sur Ω.*

Démonstration : Nous traitons le cas $f \geq 0$, $g \geq 0$. L'autre cas est similaire. Remarquons que $u \in H^1(\Omega)$ est tel que :

$$u = g \text{ sur } \partial\Omega, \quad \int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \quad \forall v \in H_0^1(\Omega). \tag{1.70}$$

On pose $u_- = \text{Inf}(u, 0)$, $u_+ = \text{Sup}(u, 0)$ et $\Omega_- = \text{Support}(u \leq 0)$, $\Omega_+ = \text{Support}(u \geq 0)$.



On peut montrer (cf. [16]), et on l'admet ici, que si $v \in H^1(\Omega)$ (resp. $H_0^1(\Omega)$) alors v_+ et v_- appartiennent à $H^1(\Omega)$ (resp. $H_0^1(\Omega)$) et que :

$$\nabla v_-(x) = \begin{cases} \nabla v(x) & \text{si } x \in \Omega_- \\ 0 & \text{si } x \notin \Omega_- \end{cases} \quad \text{et} \quad \nabla v_+(x) = \begin{cases} \nabla v(x) & \text{si } x \in \Omega_+ \\ 0 & \text{si } x \notin \Omega_+ \end{cases} \tag{1.71}$$

Choisissons, $v = u_-$ dans (1.70). Ceci est licite puisqu'on a $u_- \in H_0^1(\Omega)$:

$$(u_-)_{|\partial\Omega} = \inf(g, 0) = 0 \quad \text{si } g \geq 0.$$

Il en découle, sachant que $f \geq 0$ et $u_- \leq 0$:

$$0 \geq \int_{\Omega_-} f u_- d\Omega = \int_{\Omega_-} \nabla u \cdot \nabla u_- d\Omega = \int_{\Omega_-} \nabla(u_+ + u_-) \cdot \nabla u_- d\Omega = \int_{\Omega_-} |\nabla u_-|^2 d\Omega.$$

(Ci-dessus, on a utilisé les propriétés $u = u_+ + u_-$, ainsi que $\nabla u_+ \perp \nabla u_-$ presque partout).
Ce qui prouve que $\nabla u_- = 0$ et donc $u_- = 0$ car $u_- \in H_0^1(\Omega)$. La démonstration est similaire si $f \leq 0$ et $g \leq 0$ (choisir $v = u_+$). ∎

On déduit de cette proposition le principe du maximum suivant :

Théorème 1.12. (Principe du maximum) *Soient $f \in L^2(\Omega)$ et $g \in H^{1/2}(\partial\Omega)$. Alors l'unique solution $u \in H^1(\Omega)$ du problème de Dirichlet (1.69) vérifie le principe du maximum suivant :*

$$\begin{array}{lll} f \geq 0 & \Longrightarrow & u \geq \inf\limits_{\partial\Omega} g \qquad \text{p. p. sur } \Omega \\ f \leq 0 & \Longrightarrow & u \leq \sup\limits_{\partial\Omega} g \qquad \text{p. p. sur } \Omega \\ f = 0 & \Longrightarrow & \inf\limits_{\partial\Omega} g \leq u \leq \sup\limits_{\partial\Omega} g \qquad \text{p. p. sur } \Omega \end{array}$$

Démonstration : Cas $f \geq 0$. Notons $K = \inf\limits_{\partial\Omega} g$. Si $K = -\infty$, il est clair que $u \geq -\infty$ presque partout ! Si K est finie, elle vérifie $-\Delta K = 0$ dans Ω. Posons $v = K - u$. Par construction, v vérifie les équations :

$$\begin{cases} -\Delta v = -f & (\leq 0) \qquad \text{dans } \Omega \\ v = K - g & (\leq 0) \qquad \text{sur } \partial\Omega \end{cases}$$

qui prouve, en vertu de la proposition précédente, que $v \leq 0$, soit $u \geq K$. Les autres cas se traitent de la même façon. ∎

Remarque 1.14. Pour le problème de Dirichlet modifié :

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases}$$

on a un résultat encore plus précis :

$$\min(\inf_{\Omega} f, \inf_{\partial\Omega} g) \ \leq \ u \ \leq \ \max(\sup_{\Omega} f, \sup_{\partial\Omega} g) \qquad \text{p. p. sur } \Omega,$$

dont nous laissons la démonstration à titre d'exercice.

1.5.2 Régularité des solutions

Les formulations variationnelles ne fournissent que des solutions faibles (appartenant à $H^1(\Omega)$ pour le problème de Laplace). Une question naturelle est de se demander si les solutions faibles trouvées ne sont pas plus régulières.
Cette question est très technique, car il convient d'étudier attentivement ce qui se "passe à la frontière du domaine", la régularité de la frontière et la forme de l'ouvert (frontière $\mathcal{C}^\infty$, ouvert polygonal ou polyédrique, convexe ou pas ...) intervenant de façon cruciale.
Nous indiquons, pour commencer, deux résultats de régularité lorsque la frontière est supposée **très régulière** (voir [24, 8, 18] pour les démonstrations de ces résultats et des résultats plus généraux). Puis, nous poursuivons en examinant la situation, lorsque l'ouvert borné Ω est un polygone de $\mathbb{R}^2$, ou un polyèdre (lipschitzien) de $\mathbb{R}^3$. Nous reprenons ici des résultats dus à Grisvard [18, 19]. Et, pour finir, nous énonçons un résultat de régularité "locale", valable à l'intérieur de l'ouvert.

• **Problème de Dirichlet** (frontière $\mathcal{C}^\infty$)

Soient $f \in H^k(\Omega)$ et $g \in L^2(\partial\Omega)$ telle qu'il existe $\tilde{g} \in H^\ell(\Omega)$ avec $\ell \geq 1$ et $\tilde{g} = g$ sur $\partial\Omega$, alors la solution $u \in H^1(\Omega)$ du problème de Dirichlet :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases}$$

appartient à $H^m(\Omega)$ avec $m = \text{Min}(k+2, \ell)$.

Notons, qu'en vertu des injections des espaces de Sobolev dans les espaces $\mathcal{C}^p(\overline{\Omega})$ (voir proposition 1.11), il faut que $k > \dfrac{n}{2}$ et $\ell > \dfrac{n}{2}+2$ pour que $H^m(\Omega) \subset \mathcal{C}^2\left(\overline{\Omega}\right)$. Ce qui fournit des conditions suffisantes pour que u soit une solution classique du problème de Dirichlet.

Retenons, par exemple que :

$$\begin{cases} f \in L^2(\Omega) \\ g = 0 \end{cases} \quad \Longrightarrow \quad u \in H^2(\Omega)$$

et plus généralement :

$$\begin{cases} f \in H^k(\Omega) \\ g = 0 \end{cases} \quad \Longrightarrow \quad u \in H^{k+2}(\Omega).$$

• **Problème de Neumann** (frontière $\mathcal{C}^\infty$)

Soient $f \in H^k(\Omega)$ et $g \in L^2(\partial\Omega)$ telle qu'il existe $\tilde{g} \in H^\ell(\Omega)$ avec $\ell \geq 2$ et $\dfrac{\partial\tilde{g}}{\partial n} = g$ sur $\partial\Omega$, alors la solution $u \in H^1(\Omega)$ du problème de Neumann :

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega \\ \dfrac{\partial u}{\partial n} = g & \text{sur } \partial\Omega \end{cases}$$

appartient à $H^m(\Omega)$ avec $m = \text{Min}(k+2, \ell)$.

C'est une solution classique si $k > \dfrac{n}{2}$ et $\ell > \dfrac{n}{2} + 2$·

Attention, ces résultats ne se prolongent pas naturellement aux problèmes mixtes Dirichlet-Neumann. La régularité des solutions de ces problèmes est liée à la nature du raccord des conditions aux limites (voir [18, 19]).

• **Problème de Dirichlet homogène** (Ω polygone ou polyèdre)

On considère ici le problème : trouver $u \in H_0^1(\Omega)$ telle que

$$-\Delta u = f \text{ dans } \Omega,$$

avec une donnée f appartenant à $L^2(\Omega)$.

– *Si Ω est convexe*, alors on a automatiquement $u \in H^2(\Omega)$.

– *Si Ω n'est pas convexe*, alors l'appartenance de u à $H^2(\Omega)$ n'est pas garantie.

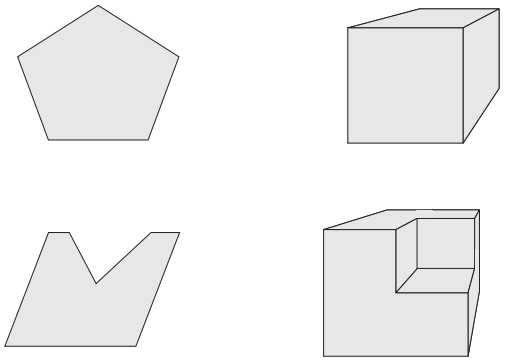


Fig. 1.5. Domaines convexe et non convexe en 2D et 3D

Plus précisément, si l'on en revient aux espaces de Sobolev d'ordre fractionnaire (déjà évoqués auparavant), on peut établir qu'il existe $\sigma_D \in]1/2, 1[$, qui dépend *uniquement* de Ω, tel que

(i) $\forall f \in L^2(\Omega)$, $u \in \bigcap_{s<\sigma_D} H^{1+s}(\Omega)$;

(ii) $\exists f \in L^2(\Omega)$ telle que $u \notin H^{1+\sigma_D}(\Omega)$.

Enfin, en dehors d'un voisinage ω des coins rentrants de $\partial\Omega$ en 2D, ou des coins et arêtes rentrants en 3D, on a le résultat

(iii) $u_{|\Omega\setminus\overline{\omega}} \in H^2(\Omega \setminus \overline{\omega})$.

De façon "plus imagée", pour une donnée f dans $L^2(\Omega)$, l'exposant limite de régularité de la solution du problème de Dirichlet homogène est 2 si Ω est convexe, et $1 + \sigma_D$ si Ω n'est pas convexe, avec $1 + \sigma_D < 2$. Qui plus est, dans le second cas, la "perte de régularité" de la solution est *localisée* autour des coins (et arêtes) rentrants de la frontière.

• **Problème de Neumann homogène** (Ω polygone ou polyèdre)

On considère maintenant le problème : trouver $u \in H^1(\Omega)$ telle que

$$-\Delta u + u = f \text{ dans } \Omega, \qquad \frac{\partial u}{\partial n} = 0 \text{ sur } \partial\Omega,$$

avec une donnée f appartenant à $L^2(\Omega)$.

Les resultats sont semblables à ceux observés pour le problème de Dirichlet homogène.

– *Si Ω est convexe*, alors $u \in H^2(\Omega)$.

– *Si Ω n'est pas convexe*, alors il existe $\sigma_N \in]1/2, 1[$, qui dépend *uniquement* de Ω, tel que

(i) $\forall f \in L^2(\Omega)$, $u \in \bigcap_{s<\sigma_N} H^{1+s}(\Omega)$;

(ii) $\exists f \in L^2(\Omega)$ telle que $u \notin H^{1+\sigma_N}(\Omega)$.

Enfin, en dehors d'un voisinage ω' des coins (et arêtes) rentrants de $\partial\Omega$, on a le résultat :

(iii) $u_{|\Omega\setminus\overline{\omega}'} \in H^2(\Omega \setminus \overline{\omega}')$.

Lorsque Ω est un polygone de $\mathbb{R}^2$, on peut établir que $\sigma_D = \sigma_N$. Par contre, lorsque Ω est un polyèdre (lipschitzien) de $\mathbb{R}^3$, on peut avoir $\sigma_D \neq \sigma_N$.
Enfin, le même type de résultats peut être obtenu dans des polygones ou des polyèdres curvilignes. Dans ce cas, la convexité ou la non-convexité du domaine est à prendre en compte uniquement au voisinage de coins (et arêtes) de la frontière.

• Régularité locale de la solution d'un problème de Laplace

Soi $u \in H^1(\Omega)$ solution du problème

$$\begin{cases} -\Delta u = f \text{ dans } \Omega \\ \text{condition aux limites sur } \partial\Omega \end{cases}$$

Alors, à l'aide d'une fonction de troncature et d'une partition de l'unité [6], on peut établir le résultat ci-dessous.

Soit $B(C, R)$ une boule de centre C et de rayon R, telle que $B(C, R) \subset \Omega$. Si $f \in H^k(B(C, R))$, alors $u \in H^{k+2}(B(C, R'))$ pour tout $R' < R$.

En d'autres termes, pour le problème de Laplace (et indépendamment des conditions aux limites), si la donnée f est "localement" H^k–régulière, la solution u est "localement" H^{k+2}–régulière.

2

Introduction à la méthode des éléments finis

Il existe principalement deux techniques de discrétisation des problèmes elliptiques : la méthode des différences finies s'appuyant sur la formulation forte du problème (donc sur les solutions classiques) et la méthode des éléments finis basée sur la formulation variationnelle (donc sur les solutions faibles).
Les principes fondamentaux présidant à la discrétisation par différences finies des problèmes elliptiques étant similaires à ceux rencontrés lors de la discrétisation des équations hyperboliques (cf. [15]), nous ne développons ici que la théorie de l'approximation par éléments finis.
La méthode des éléments finis consiste à remplacer l'espace des fonctions "tests" (de dimension infinie !) de la formulation variationnelle, par un espace de fonctions tests approché et de dimension finie. L'espace approché est généralement inclus dans l'espace initial. On parle alors d'approximation interne ou de Galerkin et, sous certaines conditions, il converge vers l'espace initial (dans un sens à préciser).
Nous consacrons la première section à des considérations générales sur l'approximation de Galerkin. Nous exposons ensuite le principe de construction des éléments finis et les propriétés des espaces d'approximation ainsi obtenus.

2.1 Approximation de Galerkin

Rappelons que les problèmes elliptiques que nous avons étudiés au chapitre précédent (principalement Dirichlet, Neumann, Fourier) s'inscrivent dans le cadre hilbertien abstrait suivant :

$$\boxed{\begin{array}{l}\text{Trouver } u \in V \text{ tel que :} \\ a(u,v) = \ell(v) \quad \forall v \in V\end{array}} \tag{2.1}$$

où
– V est un espace de Hilbert, muni de la norme $\|\cdot\|_V$,
– $a(\cdot,\cdot)$ est une forme bilinéaire continue et coercive sur V, c'est-à-dire qu'il existe deux constantes M et $\alpha > 0$ telles que :

$$|a(v,w)| \leq M\|v\|_V\|w\|_V \quad \forall (v,w) \in V, \tag{2.2}$$
$$a(v,v) \geq \alpha\|v\|_V^2 \quad \forall v \in V. \tag{2.3}$$

– ℓ est une forme linéaire continue sur V.
D'après le théorème de Lax-Milgram (théorème 1.11), sous les hypothèses précédentes, le problème abstrait (2.1) admet une unique solution notée, par la suite, $u \in V$. L'objet des sous-sections qui suivent est l'étude de l'approximation du problème (2.1).

2.1.1 Approximation interne

Considérons V_h un sous-espace de V, de dimension finie $n = n(h)$ où h est un paramètre positif destiné à tendre vers 0, avec $\lim_{h\to 0} n(h) = \infty$ (on suppose implicitement ici, et dans la suite, que V est de dimension infinie). V_h représente une approximation de dimension finie de l'espace V.

On introduit le problème variationnel approché suivant :

$$\boxed{\begin{array}{l}\text{Trouver } u_h \in V_h \text{ tel que :} \\ a(u_h, v_h) = \ell(v_h), \quad \forall v_h \in V_h\end{array}} \tag{2.4}$$

Une telle approximation est qualifiée d'approximation interne ou de Galerkin.

Sous les hypothèses (2.2) et (2.3), le problème approché (2.4) admet une unique solution $u_h \in V_h$, car on peut toujours appliquer le théorème de Lax-Milgram (V_h est encore un espace de Hilbert, en tant que sous-espace fermé de V).

Dans la suite, on note $(\cdot|\cdot)$ le produit scalaire dans $\mathbb{R}^n$.

Proposition 2.1. (existence et unicité de la solution approchée) *Sous les hypothèses (2.2) et (2.3), le problème variationnel approché (2.4) admet une unique solution $u_h \in V_h \subset V$.*

Nous allons donner une autre démonstration, très instructive, de ce résultat. Mentionnons tout d'abord une interprétation immédiate du problème approché (2.4).

Lemme 2.1. *Soit $(\varphi_1, \cdots, \varphi_n)$ une base de l'espace vectoriel V_h. Alors le problème approché (2.4) est équivalent au système linéaire posé dans $\mathbb{R}^n$:*

$$\mathbb{A}\overrightarrow{\lambda} = \overrightarrow{L} \tag{2.5}$$

avec $\mathbb{A}_{ij} = a(\varphi_j, \varphi_i)$, $\left(\overrightarrow{\lambda}\right)_i = \lambda_i$ *et* $\left(\overrightarrow{L}\right)_i = \ell(\varphi_i) \quad 1 \leq i, j \leq n$

et l'on a

$$u_h = \sum_{j=1}^{n} \lambda_j \varphi_j. \tag{2.6}$$

Démonstration : on choisit $v_h = \varphi_i$ dans la formulation variationnelle (2.4), ce qui nous conduit aux équations :

$$\sum_{j=1}^{n} \lambda_j a(\varphi_j, \varphi_i) = \ell(\varphi_i) \qquad \forall i = 1, n,$$

qui constitue le système linéaire (2.5), écrit ligne par ligne.
La réciproque est immédiate par linéarité de a. ∎

Démonstration de la proposition 2.1 : La matrice $\mathbb{A}$ est inversible. En effet, soit $\overrightarrow{\lambda}$ tel que $\mathbb{A}\overrightarrow{\lambda} = 0$. En effectuant le produit scalaire avec $\overrightarrow{\lambda}$, on trouve :

$$0 = (\mathbb{A}\overrightarrow{\lambda}\,|\,\overrightarrow{\lambda}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \lambda_j a(\varphi_j, \varphi_i) \right) \lambda_i = a\left(\sum_{j=1}^{n} \lambda_j \varphi_j, \sum_{i=1}^{n} \lambda_i \varphi_i \right).$$

D'après l'hypothèse de coercivité (2.3), que l'on peut utiliser car $\varphi_i \in V_h \subset V$ (approximation interne), on en déduit :

$$\alpha \| \sum_{i=1}^{n} \lambda_i \varphi_i \|_V^2 \leq 0,$$

c'est-à-dire :

$$\sum_{i=1}^{n} \lambda_i \varphi_i = 0,$$

et, comme $(\varphi_i)_i$ est une base de V_h, on conclut que $\lambda_i = 0,\ \forall i = 1, n$, soit $\overrightarrow{\lambda} = 0$. ∎

Remarque 2.1. En fait, en reprenant la démonstration précédente, on constate aisément que la coercivité de la forme bilinéaire $a(\cdot, \cdot)$ implique que la matrice $\mathbb{A}$ est définie-positive (donc inversible) : pour tout $\overrightarrow{\lambda} \in \mathbb{R}^n \setminus \{0\}$, on a $(\mathbb{A}\overrightarrow{\lambda}\,|\,\overrightarrow{\lambda}) > 0$.

Remarque 2.2. Comme le montre (2.5), le problème approché est équivalent à la résolution d'un système linéaire. C'est évidemment cette forme que l'on exploite numériquement.

Remarque 2.3. On notera l'importance capitale de l'hypothèse d'approximation interne $V_h \subset V$. Lorsque $V_h \not\subset V$, l'approximation est qualifiée d'externe. Ce cas est plus complexe à étudier car la forme bilinéaire $a(\cdot, \cdot)$ peut ne plus être définie sur V_h, n'être plus coercive sur V_h. Il est alors nécessaire d'introduire une forme bilinéaire approchée $a_h(\cdot, \cdot)$ qui complique singulièrement l'étude (voir par exemple [21]).

Nous examinons maintenant sous quelles hypothèses l'approximation interne converge.

• Convergence de l'approximation de Galerkin

Afin d'obtenir un résultat de convergence de la solution approchée u_h du problème (2.4) vers la solution u du problème continu (2.1), nous allons estimer directement l'écart :

$$e_h = \|u - u_h\|_V.$$

Notons au passage que cette approche diffère d'une approche classique par différences finies s'appuyant sur les notions de stabilité et de consistance [15].

Commençons par démontrer le lemme fondamental de Céa :

Lemme 2.2. *Il existe une constante $C > 0$, indépendante de V_h, telle que :*

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V \tag{2.7}$$

Démonstration : Soit $v_h \in V_h$ quelconque. En notant que $(v_h - u_h) \in V_h$, on a, par différence entre (2.1) et (2.4) :

$$a\left(u - u_h, v_h - u_h\right) = 0.$$

En introduisant u, ceci donne :

$$a\left(u - u_h, u - u_h\right) = a\left(u - u_h, u - v_h\right), \quad \forall v_h \in V_h.$$

D'après (2.3) (coercivité de $a(\cdot,\cdot)$) on a la première majoration :

$$\alpha \|u - u_h\|_V^2 \leq a\left(u - u_h, u - u_h\right).$$

D'après (2.2) (continuité de $a(\cdot,\cdot)$) on a la seconde majoration :

$$a\left(u - u_h, u - v_h\right) \leq M \|u - u_h\|_V \|u - v_h\|_V.$$

Ceci montre que :

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \|u - v_h\|_V, \quad \forall v_h \in V_h. \qquad \blacksquare$$

Remarque 2.4. Dans le cas où la forme bilinéaire $a(\cdot,\cdot)$ est symétrique, on obtient une meilleure constante : $C = (M/\alpha)^{1/2}$, en développant $a(u - v_h, u - v_h)$ en fonction de $a\left(u - u_h, u - u_h\right)$.

L'inégalité (2.7) montre que l'erreur e_h est proportionnelle[1] à l'erreur d'approximation entre les espaces V_h et V, cette erreur caractérisant la façon dont ces espaces sont "proches". Donnons une condition suffisante sur l'espace V_h pour que cette erreur tende vers 0 lorsque $h \to 0$ et par conséquent, que l'approximation converge.

Théorème 2.1. (convergence de l'approximation interne) *On suppose qu'il existe un sous-espace W dense de V et, pour chaque h, une application r_h de W dans V_h tels que :*

$$\lim_{h \to 0} \|v - r_h v\|_V = 0, \quad \forall v \in W. \tag{2.8}$$

Alors, on a :

$$\lim_{h \to 0} \|u - u_h\|_V = 0. \tag{2.9}$$

[1] Bien sûr, on sait que $\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V$!

Démonstration : Soit $\varepsilon > 0$, comme W est supposé dense dans V, il existe $w \in W$ tel que :

$$\|u - w\|_V \leq \frac{\varepsilon}{2C} \quad (C \text{ constante du lemme de Céa})$$

D'après (2.8), il existe $h(\varepsilon) > 0$ tel que :

$$\|w - r_h w\|_V \leq \frac{\varepsilon}{2C} \quad \forall h \leq h(\varepsilon).$$

D'après le lemme de Céa, (2.7) donne pour $h \leq h(\varepsilon)$:

$$\|u - u_h\|_V \leq C\,(\|u - w\|_V + \|w - r_h w\|_V) \leq \varepsilon$$

ce qui prouve (2.9). ∎

Remarque 2.5. L'espace V est souvent un espace de Sobolev de type $H^k(\Omega)$ ($k \geq 1$), dans lequel un espace du type $\mathcal{C}^m(\overline{\Omega})$ (avec $m \geq 0$ *ad hoc*) est dense (cf. [16]). Les espaces de fonctions dérivables au sens classique $\mathcal{C}^m(\overline{\Omega})$ étant plus faciles à manipuler lors de l'estimation fondamentale $\|w - r_h w\|_V$, ils jouent le rôle de W. Néanmoins, dans certains cas, on peut se passer de l'intermédiaire W, en réalisant des estimations directes (voir §2.3.1).

La relation (2.9) prouve la convergence de l'approximation interne en norme V. On précise la vitesse de convergence à l'aide de la définition suivante.

Définition 2.1. (vitesse de convergence) *On dit que l'approximation interne est convergente à l'ordre k s'il existe une constante $C > 0$, indépendante de h telle que :*

$$\|u - u_h\|_V \leq C\,h^k.$$

2.1.2 Un exemple d'approximation : les bases hilbertiennes

Un cas particulier d'approximation interne provient de l'utilisation d'une base hilbertienne dont nous rappelons la définition ci-après. On appelle base hilbertienne de l'espace de Hilbert V, toute suite $(w_j)_{j\geq 1}$ d'éléments de V telle que :
– $\forall i, j,\ (w_i, w_j)_V = \delta_{ij}$;
– l'espace vectoriel engendré par les combinaisons linéaires *finies* des (w_j) est dense dans V.
On pose $V_h = Vect\,(w_1, \cdots, w_m)$ avec $h = 1/m$, et u_h l'unique solution du problème approché (2.4). Cette approximation porte le nom d'approximation de Ritz-Galerkin.

En vertu de l'application du théorème de convergence 2.1, on démontre :

Théorème 2.2. (approximation hilbertienne) *Pour tout $m \geq 1$, le problème approché (2.4) admet une unique solution u_h qui converge vers la solution u du problème continu (2.1).*

Démonstration : L'existence et l'unicité de u_h découle de la proposition 2.1. Pour montrer la convergence, on applique le théorème 2.1 avec $W = V$ et $r_h = \pi_m$: la projection orthogonale de V sur V_h. Il est facile de montrer que la deuxième propriété de base hilbertienne implique que :

$\lim_{h\to 0} \|v - \pi_m v\|_V = 0 \quad \forall v \in V.$ ∎

• *Exemple d'application*

Dans l'ouvert $\Omega =]0,1[\times]0,1[$, les fonctions propres $(w_{m,p})_{m,p\geq 1}$ du problème de Dirichlet homogène :

$$w_{m,p}(x,y) = \sin(m\pi x)\sin(p\pi y)$$

constituent une base hilbertienne de $H_0^1(\Omega)$. En prenant $V_h = Vect\,(w_{m,p})_{m,p=1,N}$ avec $h = 1/N$, on obtient alors un espace d'approximation de dimension finie de l'espace $H_0^1(\Omega)$ qui permet de résoudre le problème de Dirichlet approché :

$$\text{Trouver } u_h \in V_h \text{ tel que } \int_\Omega k(x,y)\nabla u_h \cdot \nabla v_h d\Omega = \int_\Omega f v_h d\Omega \qquad \forall v_h \in V_h.$$

On peut évidemment utiliser d'autres bases hilbertiennes. Signalons, en particulier, les bases de polynômes orthogonaux (e.g. polynômes de Legendre) qui conduisent aux méthodes spectrales. Ces méthodes sont très précises mais aboutissent à la résolution d'un système linéaire "plein", inconvénient que ne présente pas la méthode des éléments finis. Pour une introduction aux méthodes spectrales on pourra consulter [5, 4].

2.2 La méthode des éléments finis

Dans un premier temps, nous allons dégager à travers un exemple simple les grands principes de la méthode des éléments finis. Ensuite nous présentons le cadre général et formel qui permet de construire une grande variété d'approximations par éléments finis. Enfin, nous abordons brièvement quelques aspects de la convergence des méthodes d'éléments finis.

2.2.1 Principes de la méthode des éléments finis

Rappelons que le problème de Neumann admet pour formulation variationnelle dans $H^1(\Omega)$:

$$\text{Trouver } u \in H^1(\Omega) \text{ tel que}$$
$$\int_\Omega (\nabla u \cdot \nabla v + uv)\, d\Omega = \int_\Omega f v d\Omega + \int_{\partial\Omega} g v d\Gamma \quad \forall v \in H^1(\Omega). \qquad (2.10)$$

Nous allons nous placer dans le cas où Ω est un ouvert borné de $\mathbb{R}^2$, que nous supposerons, en outre, polygonal[2].

[2] Ceci constitue une simplification ayant essentiellement des répercussions au niveau de l'analyse de l'erreur ainsi que de la mise en œuvre.

Pour fabriquer une approximation de Galerkin de ce problème, il nous faut construire un espace d'approximation $V_h \subset H^1(\Omega)$ de dimension finie.

Une première idée naturelle consiste à construire un espace d'approximation à partir de fonctions affines par morceaux. Afin que ces fonctions appartiennent à $H^1(\Omega)$, il est suffisant que ces fonctions soient continues sur $\overline{\Omega}$ (voir la propriété 1.1).
En vue de construire un tel espace, donnons-nous un maillage de l'ouvert Ω, par exemple une partition de L triangles (en dimension 2) :

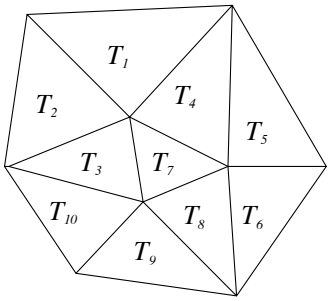


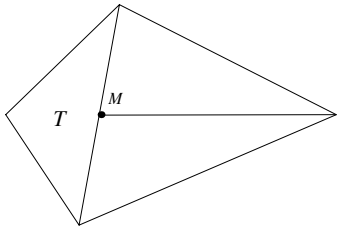
Fig. 2.1. Triangulation

Nous allons supposer que cette partition vérifie des propriétés assez naturelles :

$$\begin{cases} - \text{ Tout triangle } T_\ell \text{ est d'intérieur non vide (i.e. } \overset{\circ}{T}_\ell \neq \emptyset) \\ - \ \overset{\circ}{T}_\ell \cap \overset{\circ}{T}_{\ell'} = \emptyset \text{ si } \ell \neq \ell' \\ - \ \bigcup_\ell T_\ell = \overline{\Omega} \\ - \text{ toute arête d'un triangle est soit une arête d'un autre triangle} \\ \text{soit une arête portée par la frontière } \partial\Omega. \end{cases} \tag{2.11}$$

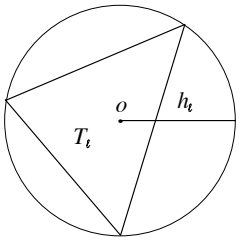
La dernière propriété exclut, en particulier, la situation suivante :

Le sommet M
n'est pas un sommet
du triangle T



On note $(M_j)_{j=1,N}$ l'ensemble de tous les sommets du maillage. Ces sommets M_j sont encore appelés nœuds du maillage.

On note $h = \max_\ell h_\ell$ où
h_ℓ est le rayon
du cercle circonscrit
au triangle T_ℓ



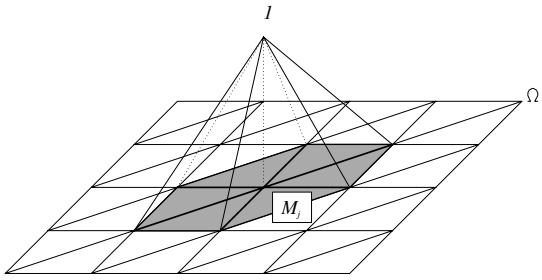
On introduit l'espace suivant, indexé par h :

$$\boxed{V_h = \left\{ v_h \in \mathcal{C}^0\left(\overline{\Omega}\right),\ v_{h|T_\ell} \text{ est affine}, \forall \ell = 1, L \right\}} \tag{2.12}$$

Proposition 2.2. *L'espace V_h est un espace vectoriel de dimension N dont une base est donnée par les fonctions continues et affines par morceaux $(w_j)_{j=1,N}$ qui satisfont aux relations :*

$$w_j(M_i) = \delta_{ij} \qquad \forall i, j = 1, N. \tag{2.13}$$

Fonction de base w_j :



• Propriétés des fonctions de base et preuve de la proposition 2.2

– w_j est continue sur $\overline{\Omega}$ donc $w_j \in V_h$.
– $(w_j)_{j=1,N}$ forme une famille libre de V_h car :

$$\sum_j \alpha_j w_j = 0 \text{ sur } \overline{\Omega} \ \Rightarrow \sum_j \alpha_j w_j\left(M_i\right) = 0,\ \forall i = 1, N \Rightarrow \alpha_i = 0,\ \forall i = 1, N.$$

– $(w_j)_{j=1,N}$ est une famille génératrice de V_h car :

$$\text{si } v_h \in V_h \text{ alors } v_h(M) = \sum_j v_h(M_j) w_j(M),\ \forall M \in \overline{\Omega}. \tag{2.14}$$

Pourquoi ? Considérons d'une part $v_{h|T_\ell}$, et d'autre part $\left(\sum_j v_h(M_j)w_j\right)_{|T_\ell}$. Ces deux fonctions sont affines et, si les trois sommets de T_ℓ sont $\{M_{i_1}, M_{i_2}, M_{i_3}\}$, on a, pour $k = 1, 2, 3$:

$$\left(\sum_j v_h(M_j)w_j\right)_{|T_\ell}(M_{i_k}) = \sum_j v_h(M_j)w_j(M_{i_k}) = v_h(M_{i_k}).$$

Ces deux fonctions coïncident en valeur aux trois sommets : elles sont donc égales. Comme c'est valable pour tout triangle T_ℓ, on a bien (2.14), d'après la propriété $\overline{\Omega} = \cup_\ell T_\ell$.
Ceci achève la démonstration de la proposition 2.2. ∎

Remarque 2.6. Revenons sur les contraintes imposées sur la partition (cf. (2.11)). Si un triangle T est tel que l'une de ses arêtes est incluse dans l'union de deux (ou plus) arêtes d'autres triangles, alors la trace – affine par rapport à l'abscisse curviligne – de $v_{h|T}$ sur cette arête est définie par trois (ou plus) valeurs : il y aura un problème dans la définition de cette trace (voir le théorème 2.3 *infra* pour une version plus détaillée de ce raisonnement).

Si T_ℓ et T_m sont deux triangles différents tels que $\overset{\circ}{T}_\ell \cap \overset{\circ}{T}_m \neq \emptyset$, on remarque que la donnée de $v_{h|\overset{\circ}{T}_\ell \cap \overset{\circ}{T}_m}$ définit complétement v_h sur $T_\ell \cup T_m$ (par prolongement). Dans ce cas, on va donc disposer de six valeurs pour déterminer une fonction affine en 2 dimensions, ce qui pose là encore un problème.

Enfin, si un triangle T est tel que $\overset{\circ}{T}_\ell = \emptyset$, alors T_ℓ est un segment. Dans ce cas, puisqu'on sait que $\overline{\Omega} = \cup_\ell T_\ell$, il existe ℓ tel que $T \subset T_\ell$ (on se souvient que l'intersection $T \cap T_\ell$ est soit vide, soit égale à une arête entière), et on peut tout simplement supprimer T de la partition.

L'égalité (2.14) montre que les composantes d'une fonction v_h dans la base $(w_j)_{j=1,N}$ coïncident avec ses valeurs nodales.
– Enfin, notons que w_j possède une propriété remarquable :

$$\text{support } w_j = \bigcup_{\ell \ t.q. \ M_j \in T_\ell} T_\ell \tag{2.15}$$

Par conséquent, le support de w_j est restreint (voir figure ci-dessus)!

On considère maintenant le problème variationnel approché :

$$\int_\Omega (\nabla u_h \cdot \nabla v_h + u_h v_h) d\Omega = \int_\Omega f v_h d\Omega + \int_{\partial\Omega} g v_h d\Gamma \quad \forall v_h \in V_h \tag{2.16}$$

Quelles sont les propriétés intéressantes qui découlent du choix de l'espace d'approximation V_h défini par (2.12) ?

Remarquons pour commencer que, d'après le lemme 2.1, le problème discrétisé (2.16) est équivalent au système linéaire, posé dans $\mathbb{R}^N$,

$$\mathbb{A}\overrightarrow{U} = \overrightarrow{B} \tag{2.17}$$

avec :

$$\begin{cases} \mathbb{A}_{ij} = \displaystyle\int_\Omega \nabla w_j \cdot \nabla w_i \, d\Omega + \int_\Omega w_j w_i \, d\Omega \\ B_i = \displaystyle\int_\Omega f w_i \, d\Omega + \int_{\partial\Omega} g w_i \, d\Omega \\ U_i = \ i^{ème} \text{ composante de la solution } u_h \text{ dans la base } (w_j)\,. \end{cases} \tag{2.18}$$

Par construction, $\mathbb{A}$ est symétrique.

• **Norme des éléments de l'espace d'approximation** V_h

Soit $\overrightarrow{V}$ le vecteur de $\mathbb{R}^N$ des composantes de v_h par rapport à la base (w_j). On note que :

$$\begin{aligned} (u_h, v_h)_{H^1(\Omega)} &= (\sum_i U_i w_i, \sum_j V_j w_j)_{H^1(\Omega)} = \sum_{i,j} (w_i, w_j)_{H^1(\Omega)} U_i V_j \\ &= \sum_{i,j} \mathbb{A}_{ji} U_i V_j = \sum_j \left(\sum_i \mathbb{A}_{ji} U_i \right) V_j \\ &= \sum_j \left(\mathbb{A}\overrightarrow{U} \right)_j V_j = (\mathbb{A}\overrightarrow{U} \,|\, \overrightarrow{V}). \end{aligned}$$

On a également,

$$\|v_h\|^2_{H^1(\Omega)} = (\mathbb{A}\overrightarrow{V} \,|\, \overrightarrow{V}).$$

On en déduit que la matrice $\mathbb{A}$ de $\mathbb{R}^{N\times N}$ est (symétrique) définie-positive.
Si on introduit les matrices $\mathbb{K}$ et $\mathbb{M}$ de $\mathbb{R}^{N\times N}$, respectivement définies par

$$\mathbb{K}_{ij} = \int_\Omega \nabla w_j \cdot \nabla w_i d\Omega, \qquad \mathbb{M}_{ij} = \int_\Omega w_j w_i d\Omega,$$

on en déduit cette fois que

$$\|v_h\|^2_{L^2(\Omega)} = (\mathbb{M}\overrightarrow{V} \,|\, \overrightarrow{V}) \text{ et } |v_h|^2_1 = (\mathbb{K}\overrightarrow{V} \,|\, \overrightarrow{V}).$$

La matrice $\mathbb{K}$ est appelée matrice de rigidité : par construction, elle est (symétrique) positive, et on a de plus la propriété $(\mathbb{K}\overrightarrow{V} \,|\, \overrightarrow{V}) = 0 \Longleftrightarrow V_j = cste,\ j = 1, N$. La matrice $\mathbb{M}$ est appeleée matrice de masse : par construction, elle est (symétrique) définie-positive et on a, bien sûr, l'égalité $\mathbb{A} = \mathbb{K} + \mathbb{M}$. Cette terminologie est

issue de la mécanique et plus précisément des équations de l'élasticité linéaire d'un corps, la rigidité faisant référence à l'opérateur aux dérivées partielles de l'élasticité et la masse, bien évidemment, à la masse du corps. Notons au passage que si $\overrightarrow{1}$ désigne le vecteur de composantes 1, on a les relations suivantes :

$$\mathbb{K}\overrightarrow{1} = \overrightarrow{0} \text{ et } (\mathbb{M}\overrightarrow{1}|\overrightarrow{1}) = \text{mes}(\Omega).$$

• Propriété de matrice creuse

La propriété de support (2.15) est essentielle. En effet, d'après cette propriété, on vérifie aisément que :

$$\mathbb{A}_{ij} = 0 \quad \text{si } M_i \text{ et } M_j \text{ n'appartiennent pas au même triangle.} \tag{2.19}$$

Ceci montre que la matrice $\mathbb{A}$ présente un très grand nombre de zéros : on parle de matrice creuse (voir le chapitre 3, §3.2). Par voie de conséquence, d'une part le stockage du système linéaire occupe peu de place en mémoire, et d'autre part on peut appliquer des méthodes spécifiques de résolution (voir l'annexe et, plus généralement, [23]).

• Interpolation de Lagrange

Les inconnues du système linéaire s'appellent des degrés de liberté de l'approximation. On constate d'après (2.14) que $U_i = u_h(M_i)$, pour $i = 1, N$. La résolution du système linéaire fournit directement l'approximation de la solution u aux nœuds M_i du maillage. On parle dans ce cas d'approximation par éléments finis de type Lagrange, chacun des degrés de liberté du système étant supporté par un nœud du maillage.

• Caractère local de la méthode des éléments finis

Plaçons nous dans un triangle T_ℓ de sommets $(M_{\ell_1}, M_{\ell_2}, M_{\ell_3})$. Les seules fonctions de base globales $(w_j)_{j=1,N}$ dont la restriction à T_ℓ n'est pas nulle sont w_{ℓ_1}, w_{ℓ_2} et w_{ℓ_3}.
On définit les fonctions de base locales dans le triangle T_ℓ :

$$\tau_{\ell_i} = w_{\ell_i|T_\ell} \qquad 1 \leq i \leq 3. \tag{2.20}$$

τ_{ℓ_i} est un polynôme de degré 1 en (x, y), puisque $w_{\ell_i|T_\ell}$ est affine par définition ; ainsi, τ_{ℓ_i} appartient à $P^1[T_\ell]$, l'espace des polynômes définis sur T_ℓ de degré 1 en (x, y). Il est facile de vérifier que :

$$\tau_{\ell_i}(M) = \lambda_{\ell_i}(M) \qquad 1 \leq i \leq 3 \tag{2.21}$$

où $(\lambda_{\ell_i})_{1\leq i\leq 3}$ désignent les coordonnées barycentriques [3] de M par rapport aux points $(M_{\ell_i})_{1\leq i\leq 3}$. Par ailleurs, $(\tau_{\ell_i})_{1\leq i\leq 3}$ forme une base de $P^1[T_\ell]$.

En effet $M \mapsto \lambda_{\ell_i}(M)$ appartient à $P^1[T_\ell]$ et par définition :

$$\lambda_{\ell_i}\left(M_{\ell_j}\right) = \delta_{ij} \quad 1 \leq i \leq 3.$$

Comme τ_{ℓ_i} vérifie également (d'après (2.13) et (2.20)) :

$$\tau_{\ell_i}\left(M_{\ell_j}\right) = \delta_{ij} \quad 1 \leq i \leq 3,$$

on en déduit la relation (2.21) car $P^1[T_\ell]$ est de dimension 3.

Si on raisonne localement, on peut se ramener à une configuration de référence, sur un triangle fixe. En effet, les fonctions de base locales (τ_{ℓ_i}) se déduisent des fonctions de base $(\widehat{\tau}_i)$, dites de référence et définies sur le triangle unité $\widehat{T}$ de sommets $\widehat{M}_1(0,0)$, $\widehat{M}_2(1,0)$, et $\widehat{M}_3(0,1)$.
On introduit la transformation affine F_ℓ :

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_1\hat{x} + b_1\hat{y} + c_1 \\ a_2\hat{x} + b_2\hat{y} + c_2 \end{pmatrix},$$

qui transforme[4] le triangle $\widehat{T}$ en le triangle T_ℓ, avec les correspondances :

$$F_\ell\left(\widehat{M}_i\right) = M_{\ell_i} \qquad 1 \leq i \leq 3. \tag{2.22}$$

On a, puisque F_ℓ est affine :

$$M = F_\ell\left(\widehat{M}\right) = \sum_{i=1}^{3} \widehat{\lambda}_i\left(\widehat{M}\right) M_{\ell_i} \tag{2.23}$$

où $\widehat{\lambda}_i\left(\widehat{M}\right)$ sont les coordonnées barycentriques de $\widehat{M}$ par rapport aux points $\left(\widehat{M}_i\right)_{1\leq i\leq 3}$:

$$\begin{cases} \widehat{\lambda}_1\left(\widehat{x},\widehat{y}\right) = 1 - \widehat{x} - \widehat{y} \\ \widehat{\lambda}_2\left(\widehat{x},\widehat{y}\right) = \widehat{x} \\ \widehat{\lambda}_3\left(\widehat{x},\widehat{y}\right) = \widehat{y} \end{cases}. \tag{2.24}$$

[3] On rappelle que les coordonnées barycentriques $(\lambda_i)_{i=1,2,3}$ associées à trois points non alignés $(M_i)_{i=1,2,3}$ (de coordonnées (x_i, y_i)) permettent de répérer tout point M de $\mathbb{R}^2$. Si on note (x, y) les coordonnées de M, alors $\lambda_1(x, y)$, $\lambda_2(x, y)$, $\lambda_3(x, y)$ sont définies de façon unique par les trois équations linéaires :

$$\textstyle\sum_{i=1,3} \lambda_i(x,y)x_i = x,\ \sum_{i=1,3} \lambda_i(x,y)y_i = y,\ \sum_{i=1,3} \lambda_i(x,y) = 1.$$

[4] Par abus, si on confond le point $\widehat{M}$ avec le vecteur $\overrightarrow{O\widehat{M}}$ la transformation F_ℓ s'écrit encore $F_\ell(\widehat{M}) = \mathbb{B}_\ell\widehat{M} + \overrightarrow{C}_\ell$ avec $\mathbb{B}_\ell \in \mathbb{R}^{2\times 2}$ et $\overrightarrow{C}_\ell \in \mathbb{R}^2$.

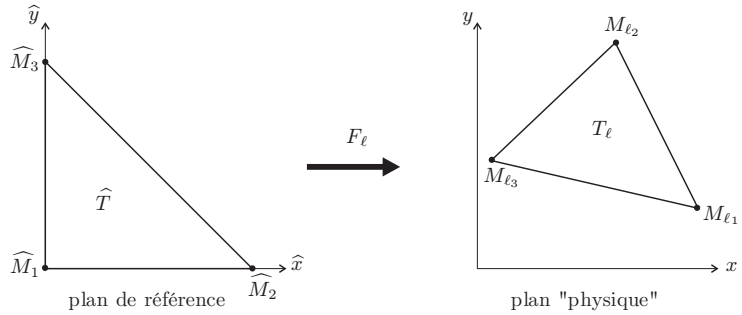


Fig. 2.2. Transformation géométrique

On pose par la suite $\widehat{\tau}_i = \widehat{\lambda}_i$, $1 \leq i \leq 3$, car les fonctions de base locales sur le triangle de référence sont les coordonnées barycentriques dans ce triangle.

Il est facile de voir que, dès lors que les points $(M_{\ell_i})_{1\leq i\leq 3}$ ne sont pas alignés – T_{ℓ_i} n'est pas dégénéré – alors F_ℓ est inversible (d'inverse affine) et on a d'après (2.23)-(2.24) :

$$\begin{cases} \tau_{\ell_i} = \widehat{\tau}_i \circ F_\ell^{-1} \\ \widehat{\tau}_i = \tau_{\ell_i} \circ F_\ell \end{cases} \tag{2.25}$$

Par conséquent, on peut construire les fonctions de base locales τ_{ℓ_i}, puis les fonctions de base globales w_j, à l'aide des seules fonctions de base de référence données par (2.24). Ceci fournit un procédé de construction général d'une base de l'espace d'approximation, qui est par ailleurs simple à mettre en œuvre algorithmiquement et informatiquement.

Ce processus de construction permet en outre d'effectuer rapidement et simplement le calcul de la matrice du système linéaire (2.17). En effet, dans cette matrice, intervient le terme :

$$\mathbb{K}_{ij} = \int_\Omega \nabla w_i \cdot \nabla w_j d\Omega.$$

D'après la propriété de support de (w_j) ((2.15) et (2.20)) on a :

$$\mathbb{K}_{ij} = \sum_{T_\ell \text{ tq } M_i, M_j \in T_\ell} \int_{T_\ell} \nabla \tau_{\ell_i} \cdot \nabla \tau_{\ell_j} d\Omega. \tag{2.26}$$

Remarque 2.7. Bien évidemment, lorsque $i \neq j$ il existe au plus deux triangles ayant M_i et M_j pour sommets (les triangles contenant l'arête $[M_i, M_j]$), alors que lorsque $i = j$ il peut en exister beaucoup (tous les triangles de sommet M_i) !

On est donc ramené au calcul des <u>matrices dites élémentaires</u> :

$$\mathbb{K}_{ij}^\ell = \int_{T_\ell} \nabla \tau_{\ell_i} \cdot \nabla \tau_{\ell_j} d\Omega. \tag{2.27}$$

Notons $dF_\ell(\widehat{M})$ la différentielle de F_ℓ en $\widehat{M}$. Soient maintenant $[dF_\ell(\widehat{M})]$ et $J_{F_\ell}(\widehat{M})$ respectivement la matrice jacobienne de F_ℓ en $\widehat{M}$ (i.e. la représentation matricielle de $dF_\ell(\widehat{M})$), et son jacobien en $\widehat{M}$, égal à $det([dF_\ell(\widehat{M})])$. On a la formule suivante (cf. [6]) pour le changement de variable $M = F_\ell(\widehat{M})$:

$$\int_{T_\ell} v(M)\, d\Omega = \int_{\widehat{T}} v \circ F_\ell(\widehat{M})\, |J_{F_\ell}(\widehat{M})|\, d\widehat{\Omega}.$$

Dans le cas qui nous intéresse, rappelons qu'on tire de l'identité $\widehat{\tau}_i(\widehat{M}) = \tau_{\ell_i} \circ F_\ell(\widehat{M})$ les relations $d\widehat{\tau}_i(\widehat{M}) = d\tau_{\ell_i}(M) \circ dF_\ell(\widehat{M})$ et $\widehat{\nabla}\widehat{\tau}_i(\widehat{M}) = [dF_\ell(\widehat{M})]^t\, \nabla\tau_{\ell_i}(M)$ (cf. [9]). On aboutit alors à :

$$\mathbb{K}^\ell_{ij} = \int_{\widehat{T}} (\mathbb{C}^\ell(\widehat{M})\nabla\widehat{\tau}_i(\widehat{M})) \cdot \widehat{\nabla}\widehat{\tau}_j(\widehat{M}) |J_{F_\ell}(\widehat{M})| d\widehat{\Omega}, \tag{2.28}$$

où la matrice (symétrique) $\mathbb{C}^\ell(\widehat{M})$ est égale à :

$$\mathbb{C}^\ell(\widehat{M}) = [dF_\ell(\widehat{M})]^{-1}([dF_\ell(\widehat{M})]^{-1})^t.$$

Ici, F_ℓ est affine[5], et on a :

$$|J_{F_\ell}(\widehat{M})| = \left| \overrightarrow{M_{\ell_1} M_{\ell_3}} \times \overrightarrow{M_{\ell_1} M_{\ell_2}} \right| = 2\,\mathrm{Mes}(T_\ell),$$

et d'autre part $\mathbb{C}^\ell(\widehat{M})$ est une matrice indépendante de $\widehat{M}$. Ce qui permet d'en conclure que :

$$\begin{aligned}\mathbb{K}^\ell_{ij} \quad &= 2\,\mathrm{Mes}\,(T_\ell) \left\{ \mathbb{C}^\ell_{11} \int_{\widehat{T}} \partial_{\widehat{x}}\widehat{\tau}_i \partial_{\widehat{x}}\widehat{\tau}_j d\widehat{x}d\widehat{y} + \mathbb{C}^\ell_{22} \int_{\widehat{T}} \partial_{\widehat{y}}\widehat{\tau}_i \partial_{\widehat{y}}\widehat{\tau}_j d\widehat{x}d\widehat{y} \right. \\ &\left. + \mathbb{C}^\ell_{12} \left(\int_{\widehat{T}} \partial_{\widehat{x}}\widehat{\tau}_i \partial_{\widehat{y}}\widehat{\tau}_j d\widehat{x}d\widehat{y} + \int_{\widehat{T}} \partial_{\widehat{y}}\widehat{\tau}_i \partial_{\widehat{x}}\widehat{\tau}_j d\widehat{x}d\widehat{y} \right) \right\},\end{aligned}$$

montrant ainsi que le calcul de $\mathbb{K}^\ell_{ij}$ et, par conséquent, celui de $\mathbb{K}_{ij}$, ne requièrent que la connaissance des fonctions de base de référence.

On traite de la même façon le terme :

$$\mathbb{M}_{ij} = \int_\Omega w_i w_j d\Omega.$$

On obtient :

$$\mathbb{M}_{ij} = \sum_{T_\ell \text{ tq } M_i, M_j \in T_\ell} \mathbb{M}^\ell_{ij}$$

[5] Noter que $[dF_\ell(\widehat{M})] = \mathbb{B}_\ell$, $J_{F_\ell}(\widehat{M}) = det(\mathbb{B}_\ell)$ et $\widehat{\nabla}\widehat{\tau}_i(\widehat{M}) = \mathbb{B}_\ell^t\, \nabla\tau_{\ell_i}(M)$.

avec cette fois :

$$\mathbb{M}_{ij}^{\ell} = \int_{T_\ell} \tau_{\ell_i} \tau_{\ell_j} d\Omega = \int_{\widehat{T}} \widehat{\tau}_i(\widehat{M}) \widehat{\tau}_j(\widehat{M}) |J_{F_\ell}(\widehat{M})| d\widehat{\Omega}, \tag{2.29}$$

soit dans notre cas

$$\mathbb{M}_{ij}^{\ell} = 2\,\mathrm{Mes}(T_\ell) \int_{\widehat{T}} \widehat{\tau}_i \widehat{\tau}_j d\widehat{x} d\widehat{y}.$$

Remarque 2.8. Dans la pratique, on n'utilise pas la formule sommatoire (2.26) mais un algorithme différent, car il est coûteux (d'un point de vue informatique) de détecter les triangles qui possèdent à la fois M_i et M_j pour sommets. Nous y reviendrons au chapitre 3 (§3.1).

L'espace d'approximation dont nous venons de décrire la construction et les principales propriétés constitue l'approximation par éléments finis P^1-Lagrange. Cet exemple met en évidence le processus général de construction d'une méthode d'éléments finis :

– définition des fonctions de base locales sur le triangle de référence ;
– construction des fonctions de base globales (par transport et assemblage des fonctions de base locales), engendrant un espace d'approximation interne.

2.2.2 Processus général de construction des éléments finis

Commençons par une première généralisation de l'approximation P^1 par morceaux.

• Eléments finis de Lagrange d'ordre k

Définition 2.2. *On appelle élément fini de Lagrange d'ordre k un triplet (K, Σ, P) où K est un fermé borné non vide de $\mathbb{R}^n$, Σ un ensemble de n_K points $\left(M_i^K\right)_{i=1,n_K}$ appartenant à K et P un espace vectoriel de polynômes contenant $\mathbb{P}^k(K)$ (espace des polynômes de degré au plus k sur K) tels que :*

$$\forall(\alpha_1, \alpha_2, \cdots, \alpha_{n_K}) \in \mathbb{R}^{n_K},\ \exists! p \in P \text{ tel que } p(M_i^K) = \alpha_i,\ \forall i = 1, n_K. \tag{2.30}$$

Un point de Σ s'appelle un degré de liberté de Lagrange et lorsque la propriété (2.30) est vérifiée on dit que P est Σ-unisolvant. Cette propriété exprime le fait que l'application :

$$\begin{aligned} S\ :\ P &\longrightarrow \mathbb{R}^{n_K} \\ p\ &\mapsto\ (p(M_1^K), \cdots, p(M_{n_K}^K)) \end{aligned}$$

est bijective ou en d'autres termes, que toute fonction de l'espace d'approximation est déterminée de façon unique par les valeurs qu'elle prend aux degrés de liberté. Une façon simple de prouver l'unisolvance consiste donc à vérifier que $\dim P = n_K$ et à prouver que l'application S est injective ou surjective.

Remarque 2.9. Dans la définition 2.2, K peut prendre n'importe quelle forme : triangle, quadrangle, etc. en 2D ; tétraèdre, pentaèdre, hexaèdre, cube, etc. en 3D. Par ailleurs, les points $(M_i^K)_i$ ne sont pas nécessairement des sommets ! Enfin, l'ordre de l'élément fini qui apparaît dans la définition est lié à la précision de l'aproximation, notion qui sera abordée ultérieurement (voir le §2.3.1).

En vertu de la propriété d'unisolvance (2.30), il existe n_K fonctions de base locales telles que :

$$\tau_i^K(M_j^K) = \delta_{ij} \quad \forall i,j = 1, n_K. \tag{2.31}$$

Ainsi, $\left(\tau_i^K\right)_{i=1,n_K}$ forme une famille libre et engendre le sous-espace vectoriel P.

• Exemples d'éléments finis

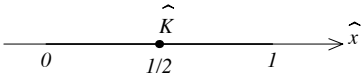
Indiquons, à titre d'exemple, quelques éléments finis classiquement utilisés.

Dimension 1

On note $\widehat{S} = [0,1]$ le segment unité.

– Elément fini de Lagrange d'ordre 0

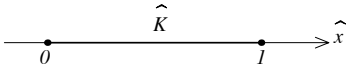
$$\left|\begin{array}{l} \widehat{K} = \widehat{S} \\ \widehat{\Sigma} = \left\{\frac{1}{2}\right\} \\ P^0 = \left\{p\left(\widehat{x}\right) = a,\ a \in \mathbb{R}\right\} \end{array}\right.$$



La propriété d'unisolvance (2.30) est trivialement vérifiée et on a pour fonction de base : $\widehat{\tau}_1\left(\widehat{x}\right) = 1$. Notons que l'on peut choisir pour degré de liberté n'importe quel point du segment $\widehat{S}$.

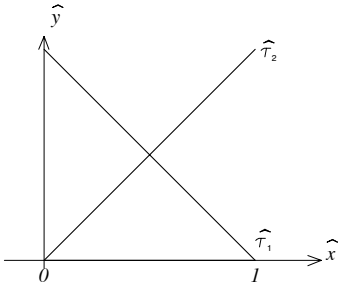
– Elément fini de Lagrange d'ordre 1

$$\left|\begin{array}{l} \widehat{K} = \widehat{S} \\ \widehat{\Sigma} = \{0,1\} \\ P^1 = \left\{p\left(\widehat{x}\right) = a\widehat{x} + b,\ (a,b) \in \mathbb{R}^2\right\} \end{array}\right.$$



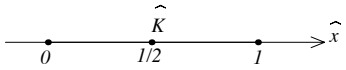
La propriété (2.30) est encore vérifiée. On trouve :

$$\widehat{\tau}_1\left(\widehat{x}\right) = 1 - \widehat{x}$$
$$\widehat{\tau}_2\left(\widehat{x}\right) = \widehat{x}$$



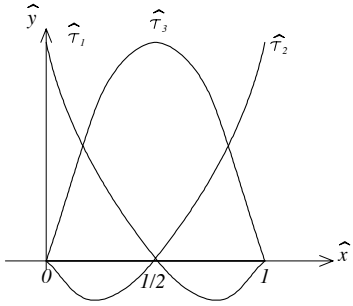
– Elément fini de Lagrange d'ordre 2

$$\left|\begin{array}{l}\widehat{K}=\widehat{S}\\ \widehat{\Sigma}=\left\{0,1,\frac{1}{2}\right\}\\ P^2=\left\{p\left(\widehat{x}\right)=a\widehat{x}^2+b\widehat{x}+c,\ (a,b,c)\in\mathbb{R}^3\right\}\end{array}\right.$$



La propriété (2.30) est vérifiée et les fonctions de base sont cette fois données par :

$$\begin{array}{l}\widehat{\tau}_1\left(\widehat{x}\right)=\left(1-\widehat{x}\right)\left(1-2\widehat{x}\right)\\ \widehat{\tau}_2\left(\widehat{x}\right)=\widehat{x}\left(2\widehat{x}-1\right)\\ \widehat{\tau}_3\left(\widehat{x}\right)=4\widehat{x}\left(1-\widehat{x}\right)\end{array}$$

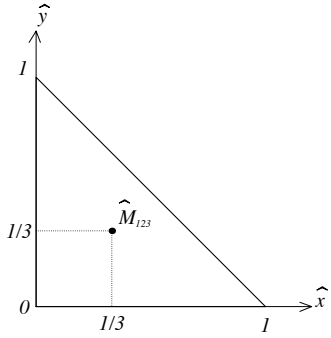


Dimension 2

En s'appuyant sur le triangle de référence $\widehat{T}$, on construit classiquement des approximations P^0, P^1 et P^2. On note $\widehat{M}_1(0,0), \widehat{M}_2(1,0)$ et $\widehat{M}_3(0,1)$ les sommets de $\widehat{T}$.

– Elément fini P^0

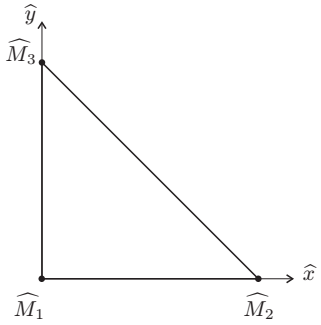
$$\left|\begin{array}{l}\widehat{K}=\widehat{T}\\ \widehat{\Sigma}=\left\{\widehat{M}_{123}=\left(\frac{1}{3},\frac{1}{3}\right)\right\}\\ \widehat{P}=P^0\left(\widehat{T}\right)=\left\{p\left(\widehat{x},\widehat{y}\right)=a\right\}\end{array}\right.$$



Il est clair que $\widehat{P}$ est $\widehat{\Sigma}$-unisolvant et que $\widehat{\tau}_{123}\left(\widehat{M}\right)=1$. Ce choix conduit à des espaces d'approximation V_h contenant des fonctions qui, quoique régulières par morceaux, ne sont pas globalement continues. Par conséquent, d'après la contraposée de la propriété 1.2, on n'aura pas l'inclusion $V_h \subset H^1(\Omega)$ (nécessaire dans le cadre de l'approximation interne). Par contre, ces éléments sont adaptés pour l'approximation de l'espace $L^2(\Omega)$.

– Elément fini P^1

$$\left|\begin{array}{l}\widehat{K}=\widehat{T}\\ \widehat{\Sigma}=\left\{\widehat{M}_1,\widehat{M}_2,\widehat{M}_3\right\}\\ \widehat{P}=P^1\left(\widehat{T}\right)=\left\{p\left(\widehat{x},\widehat{y}\right)=a\widehat{x}+b\widehat{y}+c\right\}\end{array}\right.$$

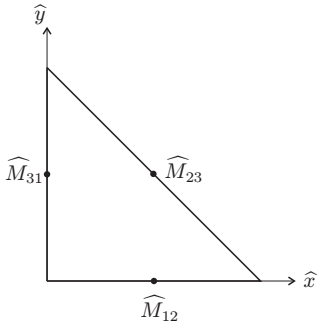


Dès lors que les trois sommets du triangle ne sont pas alignés, la propriété d'unisolvance (2.30) est vérifiée, et on a :

$$\widehat{\tau}_1\left(\widehat{x},\widehat{y}\right)=1-\widehat{x}-\widehat{y}\qquad\widehat{\tau}_2\left(\widehat{x},\widehat{y}\right)=\widehat{x}\qquad\widehat{\tau}_3\left(\widehat{x},\widehat{y}\right)=\widehat{y}$$

Remarquons que l'on aurait très bien pu choisir pour $\widehat{\Sigma}$:

$$\widehat{\Sigma}=\left\{\widehat{M}_{12}\left(\frac{1}{2},0\right),\widehat{M}_{23}\left(\frac{1}{2},\frac{1}{2}\right),\widehat{M}_{31}\left(0,\frac{1}{2}\right)\right\}$$



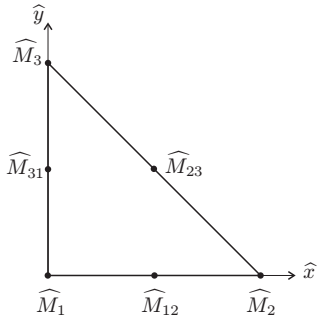
qui conduit également à un élément fini de type P^1 avec :

$$\widehat{\tau}_{12}\left(\widehat{x},\widehat{y}\right)=1-2\widehat{y}\qquad\widehat{\tau}_{23}\left(\widehat{x},\widehat{y}\right)=2(\widehat{x}+\widehat{y})-1\qquad\widehat{\tau}_{31}\left(\widehat{x},\widehat{y}\right)=1-2\widehat{x}$$

Dans le cadre des approximations internes, cet élément est beaucoup moins intéressant que le précédent car il ne permet pas de construire, après assemblage des fonctions de base locales, des fonctions de H^1 (la continuité aux interfaces ne sera pas garantie) !

– Elément fini P^2

$$\left|\begin{array}{l}\widehat{K}=\widehat{T}\\ \widehat{\Sigma}=\left\{\widehat{M}_1,\widehat{M}_2,\widehat{M}_3,\widehat{M}_{12},\widehat{M}_{23},\widehat{M}_{31}\right\}\\ \widehat{P}=P^2\left(\widehat{T}\right)\end{array}\right.$$



On a $P^2\left(\widehat{T}\right) = \left\{p\left(\widehat{x},\widehat{y}\right) = a\widehat{x}^2 + b\widehat{y}^2 + c\widehat{x}\widehat{y} + d\widehat{x} + e\widehat{y} + f\right\}$.
Lorsque $\dim\widehat{P} = \mathrm{card}\widehat{\Sigma}$, pour montrer que $\widehat{P}$ est $\widehat{\Sigma}$-unisolvant, il suffit d'exhiber les fonctions de base locales $\widehat{\tau}_i$.

Une technique rapide de calcul des fonctions de base $\widehat{\tau}_i$ consiste à écrire les équations des droites sur lesquelles sont situés les points où doit s'annuler la fonction de base que l'on cherche à calculer. Par exemple, pour le calcul de $\widehat{\tau}_1$, on a $\widehat{M}_3, \widehat{M}_2$ et $\widehat{M}_{23}$ qui appartiennent à la droite d'équation : $1-\widehat{x}-\widehat{y} = 0$ et $\widehat{M}_{31}, \widehat{M}_{12}$ qui appartiennent à la droite d'équation : $2\widehat{x} + 2\widehat{y} - 1 = 0$ d'où on choisit :

$$\widehat{\tau}_1\left(\widehat{x},\widehat{y}\right) = \alpha\left(1-\widehat{x}-\widehat{y}\right)\left(2\widehat{x}+2\widehat{y}-1\right)$$

En écrivant que $\widehat{\tau}_1(0,0) = 1$ on obtient la constante $\alpha = -1$.

En procédant de même pour les autres fonctions de base, on obtient successivement :

$$\begin{array}{ll}
\widehat{\tau}_1\left(\widehat{x},\widehat{y}\right) = \left(1-\widehat{x}-\widehat{y}\right)\left(1-2\widehat{x}-2\widehat{y}\right) & \widehat{\tau}_{12}\left(\widehat{x},\widehat{y}\right) = 4\widehat{x}\left(1-\widehat{x}-\widehat{y}\right) \\
\widehat{\tau}_2\left(\widehat{x},\widehat{y}\right) = \widehat{x}\left(2\widehat{x}-1\right) & \widehat{\tau}_{23}\left(\widehat{x},\widehat{y}\right) = 4\widehat{x}\widehat{y} \\
\widehat{\tau}_3\left(\widehat{x},\widehat{y}\right) = \widehat{y}\left(2\widehat{y}-1\right) & \widehat{\tau}_{31}\left(\widehat{x},\widehat{y}\right) = 4\widehat{y}\left(1-\widehat{x}-\widehat{y}\right)
\end{array}$$

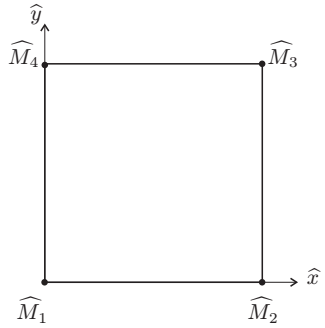
Attention, dans ce procédé, il faut s'assurer que les fonctions que l'on a ainsi construites appartiennent bien à l'espace $\widehat{P}$; ce qui est le cas si on a utilisé le nombre minimum de droites (ici 2) !

Pour des raisons géométriques, il est parfois plus facile de mailler un domaine de calcul en quadrangles. On introduit, par conséquent, des éléments finis quadrangulaires. Présentons-les sur le carré unité $\widehat{C}$ de sommets $\widehat{M}_1(0,0)$, $\widehat{M}_2(1,0)$, $\widehat{M}_3(1,1)$ et $\widehat{M}_4(0,1)$. Les espaces de polynômes $\widehat{P}$ associés ne coïncident plus avec les espaces $P^k(\widehat{C})$ mais avec des espaces de polynômes de degré au plus k en chaque variable, notés $Q^k(\widehat{C})$:

$$Q^k\left(\widehat{C}\right) = \left\{ p(\widehat{x},\widehat{y}) = \sum_{0\le i,j\le k} \alpha_{ij}\widehat{x}^i\widehat{y}^j,\ \alpha_{ij}\in\mathbb{R} \right\}.$$

– Elément fini Q^1

$$\left|\begin{array}{l}
\widehat{K} = \widehat{C} \\
\widehat{\Sigma} = \left\{\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \widehat{M}_4\right\} \\
\widehat{P} = Q^1\left(\widehat{C}\right) = \left\{p\left(\widehat{x},\widehat{y}\right) = a\widehat{x}\widehat{y} + b\widehat{x} + c\widehat{y} + d\right\}
\end{array}\right.$$

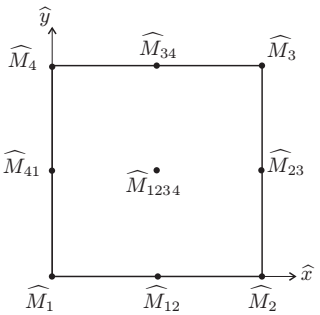


dont les fonctions de base sont données par :

$$\widehat{\tau}_1(\widehat{x},\widehat{y}) = (1-\widehat{x})(1-\widehat{y}) \qquad \widehat{\tau}_2(\widehat{x},\widehat{y}) = \widehat{x}(1-\widehat{y})$$
$$\widehat{\tau}_3(\widehat{x},\widehat{y}) = \widehat{x}\widehat{y} \qquad \widehat{\tau}_4(\widehat{x},\widehat{y}) = \widehat{y}(1-\widehat{x})$$

– Elément fini Q^2

$$\left|\begin{array}{l} \widehat{K} = \widehat{C} \\ \widehat{\Sigma} = \left\{ \widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \widehat{M}_4, \widehat{M}_{12}, \right. \\ \qquad \left. \widehat{M}_{23}, \widehat{M}_{34}, \widehat{M}_{41}, \widehat{M}_{1234} \right\} \\ \widehat{P} = Q^2\left(\widehat{C}\right) = \left\{ p(\widehat{x},\widehat{y}) = \displaystyle\sum_{0\le i,j\le 2} \alpha_{ij}\widehat{x}^i\widehat{y}^j \right\} \end{array}\right.$$



Donnons à titre indicatif la première fonction de base :

$$\widehat{\tau}_1(\widehat{x},\widehat{y}) = (1-\widehat{x})(1-2\widehat{x})(1-\widehat{y})(1-2\widehat{y}).$$

Remarque 2.10. Pour des raisons pratiques (réduction de la taille des matrices) on peut choisir de "supprimer" le nœud central $\widehat{M}_{1234}$, en déterminant un hyperplan de Q^2 contenant P^2 et sur lequel la propriété d'unisolvance sera vérifiée avec l'ensemble des degrés de liberté $\left\{\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \widehat{M}_4, \widehat{M}_{12}, \widehat{M}_{23}, \widehat{M}_{34}, \widehat{M}_{41}\right\}$. Cette technique conduit à un nouvel élément fini, dit Q^2-Serendip, qui reste un élément fini de Lagrange d'ordre 2.

Pour passer du carré unité $\widehat{C}$ à un quadrilatère convexe Q_ℓ quelconque (le passage du plan de référence au plan "physique"), on introduit une transformation F_ℓ, comme pour les triangles. Cette fois-ci, F_ℓ n'est pas systématiquement affine[6], mais de la forme $F_\ell \in \widehat{Q}_1 \times \widehat{Q}_1$, c'est-à-dire :

$$F_\ell\begin{pmatrix}\widehat{x}\\ \widehat{y}\end{pmatrix} = \begin{pmatrix} a_1^\ell\widehat{x}\widehat{y} + b_1^\ell\widehat{x} + c_1^\ell\widehat{y} + d_1^\ell \\ a_2^\ell\widehat{x}\widehat{y} + b_2^\ell\widehat{x} + c_2^\ell\widehat{y} + d_2^\ell \end{pmatrix}.$$

En effet, on a *a priori* besoin de 2×4 paramètres pour déterminer la transformation F_ℓ, puisque le quadrilatère Q_ℓ est défini par les coordonnées de ses 4 sommets. On peut vérifier (cf. [17]) qu'il existe $F_\ell \in \widehat{Q}_1 \times \widehat{Q}_1$ inversible unique telle que :

$$F_\ell\left(\widehat{M}_i\right) = M_{\ell_i} \qquad 1 \le i \le 4. \tag{2.32}$$

(A rapprocher de (2.22) pour les triangles).
Qui plus est, les côtés de $\widehat{C}$, $[\widehat{M}_i, \widehat{M}_{i+1}]$ pour $1 \le i \le 4$ (avec la convention

[6] On peut vérifier par le calcul que F_ℓ est affine si, et seulement si, Q_ℓ est un parallélogramme.

$\widehat{M}_5 = \widehat{M}_1$), sont transportés sur les côtés correspondants de Q_ℓ, $[M_{\ell_i}, M_{\ell_{i+1}}]$ pour $1 \leq i \leq 4$, et enfin la *restriction* de F_ℓ à chacun de ces côtés est *affine*.
Les fonctions de base sur Q_ℓ sont définies par "transport" des polynômes de $Q^k(\widehat{C})$, *via* F_ℓ. Plus précisément, elles sont définies par :

$$Q^k(Q_\ell) = \{p_\ell = \widehat{p} \circ F_\ell^{-1},\ \widehat{p} \in Q^k(\widehat{C})\}.$$

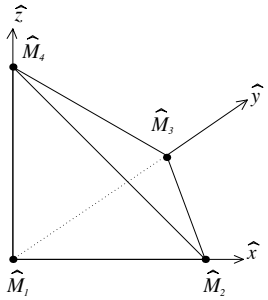
Dans le cas général[7], $Q^k(Q_\ell)$ *n'est pas un espace de polynômes*, mais on a l'inclusion $Q^k(Q_\ell) \subset \mathcal{C}^\infty(Q_\ell)$. Il est donc plus simple de représenter ses éléments en passant par F_ℓ (et F_ℓ^{-1}), en se ramenant à la référence sur le carré $\widehat{C}$.

Dimension 3

Indiquons en dimension 3 les éléments finis P^1 et Q^1 construits sur un principe identique à la dimension 2.

– Elément fini P^1
On désigne par $\widehat{T}'$ le tétraèdre de référence de sommets $\widehat{M}_1(0,0,0)$, $\widehat{M}_2(1,0,0)$, $\widehat{M}_3(0,1,0)$ et $\widehat{M}_4(0,0,1)$.

$$\left|\begin{array}{l} \widehat{K} = \widehat{T}' \\ \widehat{\Sigma} = \left\{\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \widehat{M}_4\right\} \\ \widehat{P} = P^1\left(\widehat{T}'\right) = \{p\,(\widehat{x},\widehat{y},\widehat{z}) = a\widehat{x} + b\widehat{y} + c\widehat{z} + d\} \end{array}\right.$$

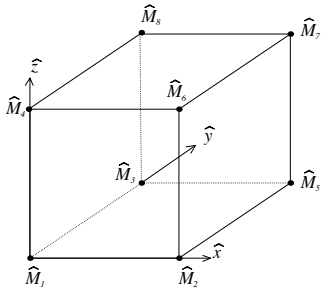


Les fonctions de base associées à cet élément sont :

$$\widehat{\tau}_1\,(\widehat{x},\widehat{y},\widehat{z}) = 1 - \widehat{x} - \widehat{y} - \widehat{z},\ \widehat{\tau}_2\,(\widehat{x},\widehat{y},\widehat{z}) = \widehat{x},\ \widehat{\tau}_3\,(\widehat{x},\widehat{y},\widehat{z}) = \widehat{y},\ \widehat{\tau}_4\,(\widehat{x},\widehat{y},\widehat{z}) = \widehat{z}.$$

– Elément fini Q^1
On note $\widehat{C}$ le cube unité de sommets $\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \widehat{M}_4$ et $\widehat{M}_5(1,1,0)$, $\widehat{M}_6(1,0,1)$, $\widehat{M}_7(1,1,1)$, $\widehat{M}_8(0,1,1)$.

$$\left|\begin{array}{l} \widehat{K} = \widehat{C} \\ \widehat{\Sigma} = \left\{\widehat{M}_i, i = 1,8\right\} \\ \widehat{P} = Q^1\left(\widehat{C}\right) = \left\{p(\widehat{x},\widehat{y},\widehat{z}) = \displaystyle\sum_{0\leq i,j,l\leq 1} \alpha_{ijk}\widehat{x}^i\widehat{y}^j\widehat{z}^l\right\} \end{array}\right.$$



[7] L'espace $Q^k(Q_\ell)$ est un espace de polynômes en (x, y) uniquement dans le cas particulier où Q_ℓ est un parallélogramme. En effet, dans le cas général, F_ℓ^{-1} est une *fraction rationnelle*.

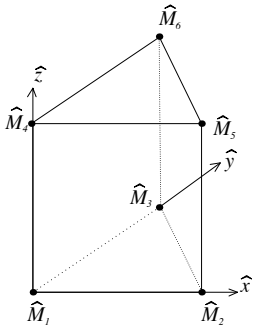
On a par exemple :

$$\widehat{\tau}_1(\widehat{x},\widehat{y},\widehat{z}) = (1-\widehat{y})(1-\widehat{x})(1-\widehat{z}) \quad \text{et} \quad \widehat{\tau}_2(\widehat{x},\widehat{y},\widehat{z}) = \widehat{x}(1-\widehat{y})(1-\widehat{z})$$

et nous laissons le soin au lecteur de déterminer les autres fonctions de base.

Il n'est pas difficile de construire les éléments finis P^2 sur $\widehat{T}'$ et Q^2 sur $\widehat{C}$, comportant respectivement 10 et 27 degrés de liberté. Signalons que l'on rencontre très souvent les éléments finis d'ordre 1 et 2 sur le prisme :

$$\left|\begin{array}{l}\widehat{K} = \widehat{T}\otimes\widehat{S}\\ \widehat{\Sigma} = \left\{\widehat{M}_i, i=1,6\right\}\\ \widehat{P} = \{p(\widehat{x},\widehat{y},\widehat{z}) = a\widehat{x}\widehat{z} + b\widehat{y}\widehat{z} + c\widehat{x} + d\widehat{y} + e\widehat{z} + f\}\end{array}\right.$$



Dimension n

On peut avoir à résoudre numériquement des problèmes en dimension $n \geq 4$. Le principe de construction des éléments finis P^1 et Q^1 est le même que précédemment.

– Elément fini P^1

Soit $\widehat{T}^{(n)}$ le simplexe de référence, de sommets $\widehat{M}_1(0,0,0,\cdots,0)$, $\widehat{M}_2(1,0,0,\cdots,0)$, $\widehat{M}_3(0,1,0,\cdots,0)$, ..., $\widehat{M}_{n+1}(0,0,0,\cdots,1)$.

On introduit

$$\left|\begin{array}{l}\widehat{K} = \widehat{T}^{(n)}\\ \widehat{\Sigma} = \left\{\widehat{M}_i,\ i=1,n+1\right\}\\ \widehat{P} = P^1\left(\widehat{T}^{(n)}\right) = \left\{p(\widehat{x}_1,\cdots,\widehat{x}_n) = \sum_{k=1,n} a_k\widehat{x}_k + a_{n+1}\right\}\end{array}\right.$$

Les fonctions de base sont

$$\widehat{\tau}_1(\widehat{x}_1,\cdots,\widehat{x}_n) = 1 - \sum_{k=1,n}\widehat{x}_k, \quad \widehat{\tau}_i(\widehat{x}_1,\cdots,\widehat{x}_n) = \widehat{x}_{i-1},\ 2\leq i\leq n+1.$$

– Elément fini Q^1

Soit $\widehat{C}^{(n)}$ l'hypercube unité, de sommets $(\widehat{M}_i)_{1\leq i\leq 2^n}$. Les coordonnées de $\widehat{M}_i$ correspondent à la valeur de $i-1$, exprimée en base 2. Par exemple : $\widehat{M}_1(0,\cdots,0)$, $\widehat{M}_2(0,\cdots,0,1)$, $\widehat{M}_3(0,\cdots,0,1,0)$, $\widehat{M}_4(0,\cdots,0,1,1)$, ..., $\widehat{M}_{2^n}(1,\cdots,1)$.
On introduit

$$\left|\begin{array}{l}\widehat{K}=\widehat{C}^{(n)}\\ \widehat{\Sigma}=\left\{\widehat{M}_i,\ i=1,2^n\right\}\\ \widehat{P}=Q^1\left(\widehat{C}^{(n)}\right)=\left\{p\left(\widehat{x}_1,\cdots,\widehat{x}_n\right)=\sum_{0\leq i_1,i_2,\cdots,i_n\leq 1}a_{i_1,i_2,\cdots,i_n}\widehat{x}_1^{i_1}\widehat{x}_2^{i_2}\cdots\widehat{x}_n^{i_n}\right\}\end{array}\right.$$

On aura pour fonctions de base

$$\widehat{\tau}_1(\widehat{x}_1,\cdots,\widehat{x}_n)=\prod_{k=1,n}(1-\widehat{x}_k),\quad \widehat{\tau}_2(\widehat{x}_1,\cdots,\widehat{x}_n)=\widehat{x}_n\prod_{k=1,n-1}(1-\widehat{x}_k),\quad \cdots,$$

$$\widehat{\tau}_{2^n}(\widehat{x}_1,\cdots,\widehat{x}_n)=\prod_{k=1,n}\widehat{x}_k.$$

Les éléments finis que nous venons de présenter se généralisent à tout ordre k. Les espaces P que l'on engendre ainsi sont égaux à P^k, Q^k ou des espaces de polynômes "intermédiaires".

A l'aide de bijections définies de $\mathbb{R}^n$ sur $\mathbb{R}^n$, on "transporte" à partir de là les éléments finis sur n'importe quel domaine. On introduit la définition suivante :

Définition 2.3. *Les éléments finis de Lagrange* $\left(\widehat{K},\widehat{\Sigma},\widehat{P}\right)$ *et* (K,Σ,P) *sont dits équivalents s'il existe une bijection* F *de* $\widehat{K}$ *sur* K *telle que :*

$$\Sigma=F\left(\widehat{\Sigma}\right) \tag{2.33}$$

$$P=\left\{p:K\to\mathbb{R}\ \textit{tel que}\ p\circ F\in\widehat{P}\right\} \tag{2.34}$$

Lorsque F *est une application affine, les éléments finis* $\left(\widehat{K},\widehat{\Sigma},\widehat{P}\right)$ *et* (K,Σ,P) *sont dits affinement équivalents.*

Remarque 2.11. Dans la pratique, on construit la transformation géométrique F pour se ramener à l'élément fini de référence dont les fonctions de base sont intrinsèques (cf. les exemples précédents).

• Construction de l'espace d'approximation

Nous allons maintenant décrire le procédé de construction d'espaces d'approximation interne V_h de $H^1(\Omega)$ à l'aide d'éléments finis de Lagrange. On supposera dans toute la suite que Ω est un ouvert borné non vide de $\mathbb{R}^n$ $(n=1,2,3)$, dont la frontière est polyédrique si $n=3$ et polygonale si $n=2$.

– *Maillage*

On considère un maillage $\mathcal{T}_h$ de $\overline{\Omega}$, c'est-à-dire une partition finie du domaine de calcul en polyèdres (segments si $n = 1$, polygones si $n = 2$) :

$$\overline{\Omega} = \bigcup_{K_\ell \in \mathcal{T}_h} K_\ell \tag{2.35}$$

telle que :

$$\left|\begin{array}{l} - \text{ tout polyèdre } K_\ell \text{ est d'intérieur non vide } (\overset{\circ}{K_\ell} \neq \emptyset) \\ - \ \overset{\circ}{K_\ell} \cap \overset{\circ}{K_m} = \emptyset \qquad \forall \ell \neq m \\ - \text{ toute face d'un polyèdre } K_\ell \in \mathcal{T}_h \text{ est :} \\ \quad \bullet \text{ soit une face d'un autre polyèdre } K_m \in \mathcal{T}_h \\ \quad \bullet \text{ soit une partie de la frontière } \partial\Omega. \end{array}\right. \tag{2.36}$$

Le paramètre h caractérise la finesse du maillage :

$$h = \max_{K_\ell \in \mathcal{T}_h} h_{K_\ell} \tag{2.37}$$

avec h_{K_ℓ} le rayon de la plus petite sphère contenant le polyèdre K_ℓ.

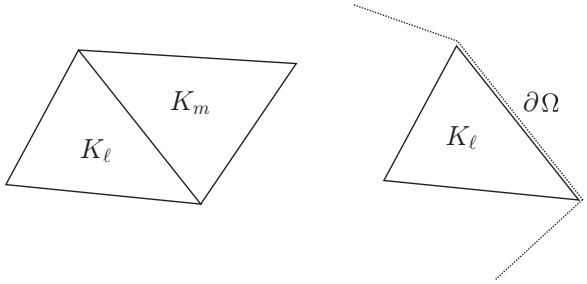


Fig. 2.3. Polyèdres adjacents et polyèdre frontière

Remarque 2.12. Il n'est pas nécessaire que tous les polyèdres K_ℓ soient du même type. Dans $\mathbb{R}^2$, on peut ainsi mélanger des triangles et des quadrangles dès lors que les propriétés (2.35) et (2.36) sont satisfaites.

On suppose qu'à chaque polyèdre $K_\ell \in \mathcal{T}_h$ est associé un élément fini de Lagrange d'ordre k : $E_\ell = (K_\ell, \Sigma_\ell, P_\ell)$. Dans un souci de simplicité, on suppose que k est identique pour tous les polyèdres. On note n_ℓ le nombre de degrés de liberté de l'élément fini E_ℓ ($n_\ell = \text{card}\Sigma_\ell$) et on désigne par $\left(\tau_i^\ell\right)_{i=1,n_\ell}$ les fonctions de base locales associées.

Nous allons voir maintenant sous quelles hypothèses supplémentaires on peut construire l'espace d'approximation :

$$V_h = \left\{ v_h \in \mathcal{C}^0\left(\overline{\Omega}\right) \text{ tel que } v_{h|K_\ell} \in P_\ell, \ \forall K_\ell \in \mathcal{T}_h \right\} \tag{2.38}$$

– *Construction des fonctions de base globales*

On note L le nombre d'éléments finis E_ℓ associés au maillage $\mathcal{T}_h$ et on définit l'ensemble de tous les nœuds du maillage :

$$\Sigma = \bigcup_{\ell=1,L} \Sigma_\ell$$

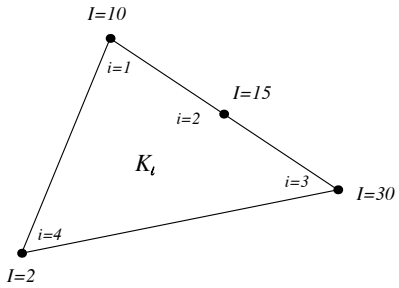
que l'on munit de la numérotation globale : $I = 1, N$ avec $N = \text{card}\Sigma$.

Pour *chaque* élément fini E_ℓ, on définit l'application ℓ_g :

$$\left| \begin{array}{ll} \{1, 2, \cdots, n_\ell\} & \to \{1, 2, \cdots, N\} \\ i & \mapsto I = \ell_g(\ell, i) \end{array} \right. \tag{2.39}$$

où $\ell_g(\ell, i)$ est le numéro global du $i^{ème}$ degré de liberté local du $\ell^{ème}$ élément fini.

$\ell_g(\ell, 1) = 10$
$\ell_g(\ell, 2) = 15$
$\ell_g(\ell, 3) = 30$
$\ell_g(\ell, 4) = 2$



On considère la fonction w_I associée au nœud M_I du maillage, construite en réunissant toutes les fonctions de base locales τ_i^ℓ – définies sur K_ℓ – attachées au nœud M_I, c'est-à-dire celles vérifiant $\tau_i^\ell(M_I) = 1$. La fonction ℓ_g nous permet alors d'exprimer la fonction de base globale w_I sur K_ℓ :

$$w_I = \begin{cases} \tau_i^\ell & \text{si } \exists i \text{ tel que } \ell_g(\ell, i) = I \\ 0 & \text{sinon} \end{cases} \tag{2.40}$$

Par construction, on a les propriétés immédiates suivantes :

Proposition 2.3. *La fonction w_I est un polynôme par morceaux vérifiant :*

$$w_I(M_J) = \delta_{I,J} \qquad \forall I, J = 1, N. \tag{2.41}$$

et le support de la fonction w_I est la réunion de tous les polyèdres K_ℓ ayant le nœud M_I pour degré de liberté, en d'autres termes :

$$\text{support}(w_I) = \bigcup_{\ell \ t.q. \ \exists i \ t.q. \ \ell_g(\ell,i)=I} K_\ell. \tag{2.42}$$

La famille de fonctions $(w_I)_{I=1,N}$ ainsi construite, forme une famille libre d'après la propriété (2.41) et constitue donc une base de l'espace $V_h = Vect\,(w_I)_{I=1,N}$. C'est pourquoi on introduit la définition :

Définition 2.4. (fonction de base globale) *On appelle fonction de base globale attachée au nœud M_I la fonction w_I définie par (2.40).*

Remarque 2.13. Rappelons que la propriété de support est fondamentale car elle conduit à des matrices creuses.

Ce processus de construction porte le nom d'assemblage. Nous allons l'illustrer plus concrètement sur deux exemples et mettre par ailleurs en évidence le caractère continu ou non des fonctions de base globales.

• **Exemples d'assemblage en dimension 2**

– Soient deux éléments finis adjacents $(K_1, \Sigma_1, P_1), (K_2, \Sigma_2, P_2)$ avec :

$P_1 = P^1\,(K_1)$, $P_2 = Q^1\,(K_2)$, $\Sigma_1 = \{M_4, M_5, M_3\}$ et $\Sigma_2 = \{M_2, M_1, M_4, M_3\}$.

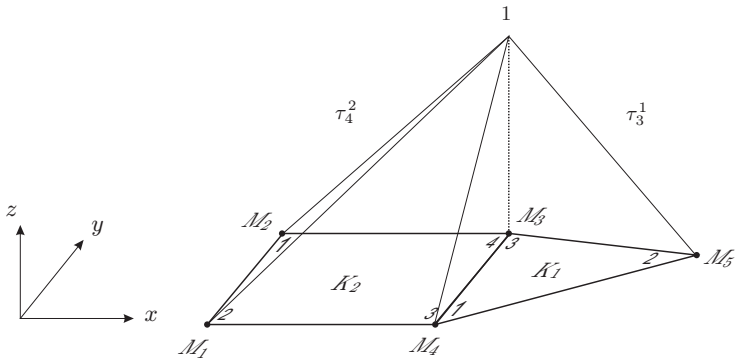


Fig. 2.4. Assemblage $P^1 - Q^1$

$\tau_3^1 \in P^1\,(K_1)$ et $\tau_4^2 \in Q^1\,(K_2)$ (mais n'appartient pas à $P^1\,(K_2)$).

On a, en vertu de (2.40) :

$$\begin{cases} w_3 = \tau_3^1 & \text{sur } K_1 \\ w_3 = \tau_4^2 & \text{sur } K_2 \end{cases}$$

Or les restrictions de τ_4^2 et τ_3^1 sur le segment $[M_3, M_4]$ sont des polynômes de degré 1 dans la variable s (paramétrage du segment $[M_3, M_4]$) qui prennent les valeurs 1 et 0 respectivement aux points M_3 et M_4.
En effet, pour $\tau_{3\,|[M_3,M_4]}^1$ c'est immédiat. Comme τ_3^1 est un élément de $P^1(K_1)$: $\tau_3^1(x,y) = ax + by + c$, d'où $\tau_{3\,|[M_3,M_4]}^1(s) = ax(s) + by(s) + c$. Par ailleurs, $x(s)$ et $y(s)$ sont toutes deux affines par rapport à s : $x(s) = \alpha s + \beta$, $y(s) = \gamma s + \delta$. Il

vient alors $\tau_{3\,|[M_3,M_4]}^1(s) = (a\alpha + b\gamma)s + (a\beta + b\delta + c)$.
Pour $\tau_{4\,|[M_3,M_4]}^2$, on raisonne par retour au carré de référence $\widehat{C}$ (rappel : $K_2 = F_2(\widehat{C})$). On a par définition $\tau_4^2 = \widehat{\tau}_4 \circ F_2^{-1}$, et on utilise la propriété selon laquelle F_2^{-1} est affine sur $[M_3, M_4]$, sachant que $\widehat{\tau}_4$ est également affine sur $[\widehat{M_2}, \widehat{M_3}]$.
Les restrictions de τ_4^2 et τ_3^1 sur le segment $[M_3, M_4]$ sont donc identiques ce qui prouve que w_3 est continue à l'interface des deux polygones K_1, K_2 et par conséquent continue sur $K_1 \cup K_2$.

– Considérons maintenant le cas où $P_1 = P^2(K_1)$ et $P_2 = Q^1(K_2)$ avec :

$$\Sigma_1 = \{M_3, M_4, M_5, M_6, M_7, M_8\} \quad \text{et} \quad \Sigma_2 = \{M_2, M_1, M_4, M_3\}.$$

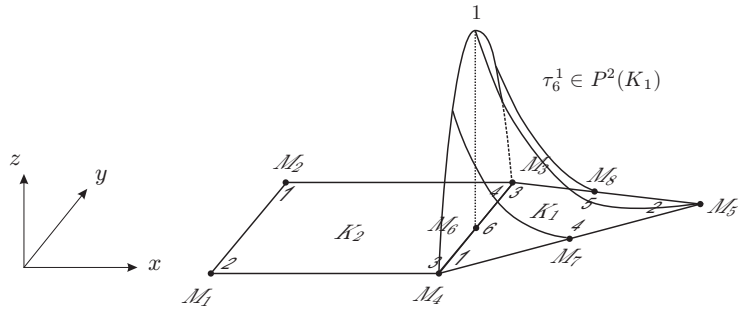


Fig. 2.5. Assemblage $P^2 - Q^1$ (non conforme H^1)

Comme le nœud M_6 n'est pas un degré de liberté de K_2 on a :

$$\begin{cases} w_6 = 0 & \text{sur } K_2 \\ w_6 = \tau_6^1 & \text{sur } K_1 \end{cases}$$

ce qui montre que w_6 n'est pas continue sur $K_1 \cup K_2$.

Ce dernier exemple montre que tout degré de liberté situé sur une face F d'un polygone K doit être un degré de liberté de tout polygone adjacent à K par cette face F, pour que la fonction de base globale attachée à ce degré de liberté soit continue sur F (c'est une condition *nécessaire*, mais ce n'est pas une condition *suffisante* !).

Par ailleurs, si un élément fini (K, Σ, P) possède un degré de liberté M_I qui n'appartient à aucune des faces de K (en d'autres termes, un degré de liberté situé à l'intérieur de K), par construction on aura :

$$\begin{cases} w_I = \tau_i & \text{sur } K \\ w_I = 0 & \text{sinon} \end{cases}$$

Par conséquent, pour que la fonction w_I soit continue à la traversée des faces de K, il est nécessaire et suffisant que :

$$w_{I|F} = 0 \quad \forall F, \text{ face de } K. \tag{2.43}$$

Ces observations nous conduisent au résultat de continuité suivant, assurant l'existence de l'espace d'approximation V_h défini par (2.38).

Proposition 2.4. *Soit $\mathcal{T}_h$ un maillage de Ω, vérifiant (2.36). Alors toute fonction de base globale w_I, définie par (2.40) est continue sur $\overline{\Omega}$ si :*

- *tout degré de liberté local appartenant à une face F d'un polyèdre K est degré de liberté de tout polyèdre adjacent à K par la face F,* (2.44)
- *la restriction (F, Σ_F, P_F) de tout élément fini (K, Σ, P) sur une de ses faces F :*
  $$\Sigma_F = \{M \in \Sigma \text{ tel que } M \in F\} \text{ et } P_F = \{p_{|F},\ p \in P\}$$
  est un élément fini de Lagrange, (2.45)
- *si deux éléments $(K_\ell, \Sigma_\ell, P_\ell)$, (K_m, Σ_m, P_m) sont adjacents par la face F alors $P_{\ell|F} = P_{m|F}$.* (2.46)

La condition (2.45) implique (2.43). En effet, soit M_I un degré de liberté qui n'appartient à aucune face de K et τ_I sa fonction de base associée. τ_I vérifie par définition :

$$\tau_I(M) = 0 \quad \forall M \in \Sigma_F$$

et comme $\tau_{I|F} \in P_F$ d'après (2.45) on a donc: $\tau_{I|F} \equiv 0$ sur F car (F, Σ_F, P_F) est un élément fini qui vérifie donc (2.30).

Nous sommes maintenant en mesure d'énoncer le théorème de construction suivant :

Théorème 2.3. *Soit $\mathcal{T}_h$ un maillage de $\overline{\Omega}$ vérifiant (2.35) et (2.36), sur lequel sont définis les éléments finis $(K_\ell, \Sigma_\ell, P_\ell)_{\ell=1,L}$ satisfaisant aux hypothèses (2.44), (2.45) et (2.46). Alors l'espace $V_h = \{v_h \in \mathcal{C}^0(\overline{\Omega}) \text{ t. q. } v_{h|K_\ell} \in P_\ell,\ \forall K_\ell \in \mathcal{T}_h\}$ admet pour base $(w_I)_{I=1,N}$, où les w_I sont définies par (2.40), avec $N = \text{card}\Sigma$, et on a :*

$$v_h(M) = \sum_{I=1,N} v_h(M_I) w_I(M) \qquad \forall v_h \in V_h. \tag{2.47}$$

En outre, les inclusions suivantes sont vérifiées :

$$V_h \subset H^1(\Omega) \quad et \quad V_h \subset \mathcal{C}^0(\overline{\Omega}).$$

Démonstration : D'après la proposition 2.4, les fonctions de base globales w_I sont continues. Par construction, cf. (2.40), elles appartiennent donc à V_h. Qui plus est, d'après (2.41), elles forment une famille libre de V_h et engendrent clairement V_h. La représentation (2.47) découle de (2.41) car :

$$v_h(M_I) = \sum_{J=1,N} \alpha_J w_J(M_I) = \sum_{J=1,N} \alpha_J \delta_{IJ} = \alpha_I.$$

Enfin, l'inclusion $V_h \subset \mathcal{C}^0(\overline{\Omega})$ résulte de la définition de V_h, et l'inclusion $V_h \subset H^1(\Omega)$ découle de la propriété 1.1. ∎

Les éléments finis de Lagrange permettent également de construire des approximations internes de $H_0^1(\Omega)$.

On note $\mathcal{I}$ le sous-ensemble des indices de $\{1, 2, \cdots, N\}$ tels que $M_I \notin \partial\Omega$ et on pose $q = \text{card}\mathcal{I}$.

On a alors le corollaire suivant :

Corollaire 2.1. *Sous les mêmes hypothèses que celles du théorème 2.3, l'espace :*

$$V_h^0 = \left\{ v_h = \sum_{I \in \mathcal{I}} \alpha_I w_I, \quad (\alpha_I) \in \mathbb{R}^q \right\}$$

est un espace d'approximation interne de $H_0^1(\Omega)$.

Démonstration : V_h^0 est un sous-espace de V_h. D'après le théorème 2.3, on a donc $V_h^0 \subset H^1(\Omega)$. Il reste à vérifier que $w_{I|\partial\Omega} = 0$, $\forall I \in \mathcal{I}$.

Soit $I \in \mathcal{I}$ et soit M un point de la frontière $\partial\Omega$, que l'on suppose situé sur la face F d'un polyèdre K. Par définition de $\mathcal{I}$, M_I n'appartient pas à $\partial\Omega$. Si $M_I \notin K$, d'après la propriété de support (2.42) $w_I = 0$ sur F. Si $M_I \in K$, on a $w_I(M_s) = 0$ pour tout degré de liberté M_s appartenant à F. Mais d'après (2.45), (F, Σ_F, P_F) définit un élément fini, par conséquent $w_I(M) = 0$. ∎

Remarque 2.14. Les fonctions de base w_I ne sont jamais dérivables sur $\overline{\Omega}$, par conséquent les éléments finis de Lagrange ne permettent pas de construire des espaces d'approximation interne de $H^2(\Omega)$. En effet, les fonctions de $H^2(\Omega)$ qui sont régulières par morceaux sont nécessairement globalement dans $\mathcal{C}^1(\overline{\Omega})$ (démonstration analogue à celle de la propriété 1.2).

2.2.3 Extension des éléments finis

Les éléments finis de Lagrange ne permettent pas de traiter toutes les situations. En particulier, ils ne conduisent pas à des approximations internes de $H^2(\Omega)$ utiles pour la résolution des problèmes elliptiques du quatrième ordre (cf. §1.3.4). C'est pourquoi on généralise la notion de degrés de liberté et d'éléments finis de la façon suivante :

Soit K un domaine fermé, d'intérieur non vide, de $\mathbb{R}^n$, $n = 1, 2, 3$ (non nécessairement un polyèdre).

– On appelle degré de liberté local sur K toute distribution[8] sur K.
– On appelle alors élément fini sur K, tout triplet (K, Σ, P) avec :
$\Sigma = \{\varphi_i, i = 1, n_K,\ \varphi_i \text{ degré de liberté local sur } K\}$
P espace vectoriel de fonctions sur K tel que :

$$\forall(\alpha_1, \alpha_2, \cdots, \alpha_{n_K}) \in \mathbb{R}^{n_K},\ \exists! p \in P \text{ tel que } \varphi_i(p) = \alpha_i,\ \forall i = 1, n_K. \qquad (2.48)$$

Il s'agit bien d'une généralisation des éléments finis de Lagrange. En effet, un degré de liberté de Lagrange M_i est représenté de façon équivalente par la distribution :

$$\varphi_i = \delta_{M_i} \qquad \text{(masse de Dirac au point } M_i)\,;$$

l'égalité $\varphi_i(p) = \alpha_i$ devenant $p\,(M_i) = \alpha_i$, la propriété (2.48) coïncide ainsi avec la propriété (2.30) de la définition des éléments finis de Lagrange.

Par ailleurs, la propriété (2.48) assure l'existence et l'unicité des fonctions de base locales $(\tau_i)_{i=1,n_K}$ vérifiant :

$$\varphi_j(\tau_i) = \delta_{ij} \qquad \forall i, j = 1, n_K. \qquad (2.49)$$

Deux exemples intéressants qu'autorise cette définition plus générale, sont les éléments finis d'Hermite et ceux de type "moment".

• Eléments finis d'Hermite

Ces éléments finis font intervenir la valeur de la dérivée des fonctions en quelques points d'un domaine K. En d'autres termes, on choisit :

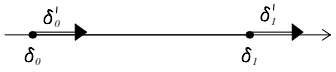
$$\Sigma \subset \left\{\delta'_{M_i}, \delta_{M_i},\ i = 1, n_K\right\}.$$

En règle générale, on conserve les degrés de liberté de Lagrange aux mêmes points qu'auparavant.

Dimension 1

Sur le segment $\widehat{S} = [0, 1]$, on choisit :

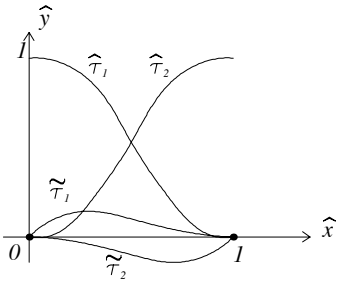
$$\left|\begin{array}{l} \widehat{K} = \widehat{S} \\ \widehat{\Sigma} = \{\delta_0, \delta_1, \delta'_0, \delta'_1\} \\ \widehat{P} = P^3([0, 1]) \end{array}\right.$$



[8] Attention ! le domaine K est fermé. Dans ce cas, les distributions agissent sur des éléments de $\mathcal{C}^\infty(K)$, qui en particulier peuvent être non nuls sur la frontière ∂K.

Un polynôme de degré 3 étant entièrement déterminé par les valeurs et les dérivées qu'il prend aux points 0 et 1, $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$ définit donc bien un élément fini au sens de la définition (2.48). C'est l'élément fini P^3 – Hermite. Il admet pour fonctions de base :

$$\begin{aligned}
\widehat{\tau}_1(\widehat{x}) &= (\widehat{x}-1)^2(2\widehat{x}+1)\\
\widehat{\tau}_2(\widehat{x}) &= \widehat{x}^2(3-2\widehat{x})\\
\tilde{\tau}_1(\widehat{x}) &= \widehat{x}(\widehat{x}-1)^2\\
\tilde{\tau}_2(\widehat{x}) &= \widehat{x}^2(\widehat{x}-1)
\end{aligned}$$

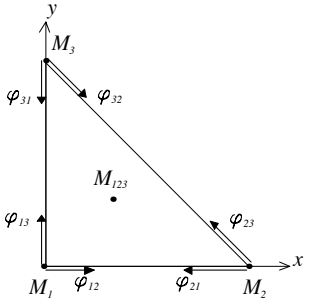


Dimension 2

On peut de même définir l'élément fini P^3 – Hermite en 2D sur le triangle $\widehat{T}$, avec :

$\widehat{K} = \widehat{T}$;
$\widehat{\Sigma} = \{\delta_{M_1}, \delta_{M_2}, \delta_{M_3}, \delta_{M_{123}}, \varphi_{12}, \varphi_{21}, \varphi_{23}, \varphi_{32}, \varphi_{13}, \varphi_{31}\}$;
$\widehat{P} = P^3\left(\widehat{K}\right)$.

(ci-dessus, φ_{ij} désigne la distribution $\varphi_{ij} : \ f \mapsto \nabla f(M_i) \cdot \overrightarrow{M_i M_j}$).



La propriété (2.48) est facile à vérifier, ce qui montre que $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$ définit bien un élément fini.

• Eléments finis de type "moment"

Un degré de liberté de type moment est représenté par la distribution :

$$\varphi = (-1)^{|\alpha|}\partial_\alpha 1_K \qquad (1_K \text{ fonction indicatrice de } K)$$

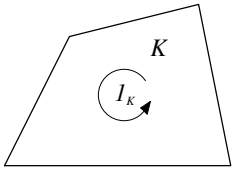
où α est un multi-indice, éventuellement nul.

On a pour toute fonction $v \in \mathcal{C}^\infty(K)$:

$$\langle \varphi, v \rangle = \langle 1_K, \partial_\alpha v(x) \rangle = \int_K \partial_\alpha v(x) d\Omega.$$

En dimension n, on peut ainsi considérer l'élément fini d'ordre 0 associé au moment d'ordre 0. Soit K un domaine fermé, d'intérieur non vide, on choisit par exemple :

$$\Sigma = \{1_K\} \quad \text{et} \quad P = P^0(K)$$



Un tel degré de liberté n'est attaché à aucun nœud de K, il représente la valeur moyenne sur K de la solution !

• Construction de l'espace d'approximation

Le principe de construction est similaire à celui utilisé pour les éléments finis de Lagrange. On se donne un maillage $\mathcal{T}_h$ vérifiant les hypothèses (2.35) et (2.36). Sur chaque domaine K_ℓ on se donne un élément fini $(K_\ell, \Sigma_\ell, P_\ell)$ de fonctions de base $\left(\tau_i^\ell\right)_{i=1,n_\ell}$ (n_ℓ désigne toujours cardΣ_ℓ).

On impose de surcroît les conditions suivantes pour $m \geq 0$:
– les fonctions de base locales τ_i^ℓ sont de classe $\mathcal{C}^{m+1}$ sur K_ℓ,

– dès que le support d'un degré de liberté φ_i^ℓ n'est pas inclus dans une face F (de la frontière de K_ℓ), alors :

$$\partial_\alpha \tau_{i|F}^\ell = 0 \quad \forall |\alpha| \leq m$$

– soit F une face commune à deux domaines K_ℓ et K_m, alors tout degré de liberté de K_ℓ dont le support est inclus dans F est également un degré de liberté de K_m.

– notons $D^\alpha P_\ell$ l'ensemble des dérivées d'ordre α des fonctions de P_ℓ, alors on a :

$$D^\alpha P_{\ell|F} = D^\alpha P_{m|F}, \quad \forall |\alpha| \leq m$$

Sous ces hypothèses, on peut construire les fonctions de base globales w_I, en décidant de considérer comme un (unique) degré de liberté global tout degré de liberté local partagé entre plusieurs éléments finis, ce qui conduit en utilisant l'application ℓ_g définie par (2.39) à la définition :

$$\begin{cases} w_I = \tau_i^\ell & \text{sur } K_\ell \text{ si } \exists i \text{ tel que } \ell_g(\ell, i) = I \\ w_I = 0 & \text{sinon} \end{cases} \tag{2.50}$$

On introduit alors l'espace d'approximation :

$$V_h = \underset{I=1,N}{Vect} \ (w_I)$$

où $N = \text{card}\Sigma$ et $\Sigma = \bigcup_{\ell=1,L} \Sigma_\ell$.

On peut alors montrer sous les hypothèses précédentes que les fonctions de base globales w_I appartiennent à $\mathcal{C}^m(\overline{\Omega})$ et, comme elles sont régulières par morceaux, on en déduit que V_h est un espace d'approximation interne de $H^{m+1}(\Omega)$ (voir [21] pour une généralisation de la propriété 1.1).

Il est facile par exemple de montrer que l'élément fini P^3-*Hermite* en *dimension 1* conduit à une approximation interne dans $H^2(\Omega)$. Par contre, en *dimension 2*, on peut vérifier que l'élément fini P^3-*Hermite* n'est pas globalement dans $\mathcal{C}^1(\overline{\Omega})$ à cause du degré de liberté central $\delta_{M_{123}}$.

2.3 Analyse numérique de la méthode des éléments finis

2.3.1 Convergence de la méthode des éléments finis

Nous allons examiner dans cette sous-section quelques aspects de l'analyse de la convergence des éléments finis. Nous nous intéresserons seulement au cas des éléments finis de Lagrange. Les preuves de convergence s'appuient, d'une part, sur le lemme de Céa (cf. lemme 2.2), les éléments finis étant un cas particulier d'approximation de Galerkin, et d'autre part, sur des estimations très techniques de l'erreur d'interpolation dont nous ne présenterons que les grandes lignes.

Nous nous plaçons dans le cas d'un problème elliptique sur $H^1(\Omega)$ de la forme :

$$\begin{cases} \text{Trouver } u \in H^1(\Omega) \text{ tel que} \\ a(u, v) = \ell(v) \quad \forall v \in H^1(\Omega) \end{cases} \tag{2.51}$$

Ici, $a(\cdot, \cdot)$ et $\ell(\cdot)$ vérifient les hypothèses du théorème de Lax-Milgram (cf. théorème 1.11). Le problème de Neumann (énoncé au §1.4.1) s'inscrit dans ce cadre.

Sur l'ouvert borné Ω, que l'on suppose toujours polyédrique, on se donne le maillage $\mathcal{T}_h$: $(K_\ell)_{\ell=1,L}$ qui satisfait les hypothèses (2.35) et (2.36). A chaque polyèdre K_ℓ est associé l'élément fini de Lagrange d'ordre k : $(K_\ell, \Sigma_\ell, P_\ell)$, de fonctions de base locales $\left(\tau_i^\ell\right)_{i=1,n_\ell}$.

On suppose que les éléments finis vérifient les conditions de compatibilité (2.44), (2.45) et (2.46), de sorte que les fonctions de base globales $(w_I)_{I=1,N}$ engendrent un espace vectoriel V_h de dimension N inclus dans $\mathcal{C}^0\left(\overline{\Omega}\right)$ et $H^1(\Omega)$.

On introduit le problème discret :

$$\begin{cases} \text{Trouver } u_h \in V_h \text{ tel que} \\ a\left(u_h, v_h\right) = \ell\left(v_h\right) \quad \forall v_h \in V_h \end{cases} \tag{2.52}$$

qui constitue une approximation de Galerkin du problème (2.51).

Rappelons que d'après le lemme de Céa (lemme 2.2) on a :

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}. \tag{2.53}$$

Il faut donc estimer $\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}$.

Pour ce faire, on introduit l'opérateur d'interpolation :

$$\begin{array}{rccl} \pi_h : & \mathcal{C}^0(\overline{\Omega}) & \longrightarrow & V_h \\ & v & \mapsto & \pi_h v = \displaystyle\sum_{I=1,N} v(M_I) w_I \end{array} \tag{2.54}$$

où M_I désigne le nœud attaché à la fonction de base globale w_I. Autrement dit, $\pi_h v$ est l'unique fonction de V_h qui prend les mêmes valeurs que v aux nœuds $(M_I)_{I=1,N}$.

Il est clair, d'après (2.53), que pour tout $w \in \mathcal{C}^0(\overline{\Omega})$:

$$\boxed{\|u - u_h\|_{H^1(\Omega)} \leq C \|u - \pi_h w\|_{H^1(\Omega)}} \tag{2.55}$$

Quelle fonction w choisir ?
Compte-tenu du fait que l'opérateur d'interpolation π_h est défini sur $\mathcal{C}^0\left(\overline{\Omega}\right)$, on supposera dans les calculs que la fontion u est suffisamment régulière. Typiquement, elle appartient à Ψ_D, Ψ_N (voir (1.28)), ou $H^2(\Omega)$, et d'après les propositions 1.11 ou 1.12, on peut alors choisir $w = u$ dans (2.55).

Remarque 2.15. Il est fondamental de noter que, dans les cas qui nous intéressent, il n'est pas nécessaire d'estimer la quantité $\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}$ pour $u \in H^1(\Omega)$ *quelconque*, mais pour u appartenant à un sous-espace de $H^1(\Omega)$, et même de $H^1(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$! Dans ce cas, il n'est plus nécessaire de faire appel à la théorie "générale" de convergence, voir le théorème 2.1, qui requiert l'utilisation d'un sous-espace W dense dans $H^1(\Omega)$. On choisit de raisonner "directement" (voir la suite...).

Nous allons maintenant estimer $\|u-\pi_h u\|_{H^1(\Omega)}$, que l'on appelle erreur d'interpolation.

Remarque 2.16. Bien entendu, il n'y a aucune raison pour que l'on ait $u_h = \pi_h u$!

Cette erreur se décompose suivant chaque polyèdre de $\mathcal{T}_h$:

$$\|u - \pi_h u\|_{H^1(\Omega)} = \left(\sum_{\ell=1,L} \|u - \pi_h u\|^2_{H^1(K_\ell)} \right)^{1/2}$$

On est donc amené à estimer une erreur d'interpolation locale sur chaque polyèdre K_ℓ.

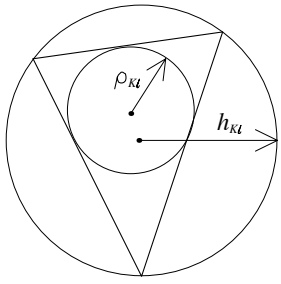
• Erreur d'interpolation locale

En vertu de la définition des fonctions de base globales (w_I) on a :

$$\pi_h u_{|K_\ell} = \sum_{i=1,n_\ell} u\left(M_i^\ell\right) \tau_i^\ell \underset{def}{=} \pi_{K_\ell} u$$

où M_i^ℓ désigne le degré de liberté local attaché à la fonction de base locale τ_i^ℓ. Supposons que l'élément fini $(K_\ell, \Sigma_\ell, P_\ell)$ soit affinement équivalent à l'élément fini de référence $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$ (voir définition 2.3) dont la taille est indépendante de h, et notons :

h_{K_ℓ} le rayon de la plus petite boule circonscrite à K_ℓ et ρ_{K_ℓ} le rayon de la plus grande boule inscrite dans K_ℓ (pour des quadrilatères, ρ_{K_ℓ} est le minimum des quatre rayons des plus grandes boules inscrites dans les triangles ayant pour sommets trois des quatre sommets du quadrilatère). Par définition, on a la relation $h_{K_\ell} \geq \rho_{K_\ell}$.



Introduisons, pour $m \geq 0$, $|v|_{m,K_\ell}$, la semi-norme de $H^m(K_\ell)$ définie par :

$$|v|^2_{m,K_\ell} = \sum_{\alpha \in \mathbb{N}^n, |\alpha|=m} \int_{K_\ell} \left(\partial_\alpha v(x)\right)^2 d\Omega.$$

On a le résultat fondamental suivant :

Proposition 2.5. (erreur d'interpolation locale) *Soit $(K_\ell, \Sigma_\ell, P_\ell)$ un élément fini de Lagrange d'ordre k, i.e. :*

$$P^k(K_\ell) \subset P_\ell \subset H^{k+1}(K_\ell),$$

et affinement équivalent à l'élément fini $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$. Alors il existe une constante C qui ne dépend que de $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$, telle que pour tout entier $0 \leq m \leq k+1$, on ait, $\forall v \in H^{k+1}(K_\ell)$, l'estimation suivante :

$$|v - \pi_{K_\ell} v|_{m,K_\ell} \leq C \, \frac{(h_{K_\ell})^{k+1}}{(\rho_{K_\ell})^m} \, |v|_{k+1,K_\ell} \tag{2.56}$$

Pour établir ce résultat, nous allons procéder en plusieurs étapes. Bien sûr, comme on peut se ramener à un domaine de référence, la première idée est d'effectuer un changement de variable pour passer de K_ℓ à $\hat{K}$. Ensuite, on réalise une estimation directe sur le domaine de référence $\hat{K}$, qui sera par définition indépendante de K_ℓ, puis on conclut en revenant à K_ℓ.
Tout d'abord, nous établissons des estimations entre semi-normes, évaluées sur K_ℓ et sur $\hat{K}$ respectivement. Comme il est indiqué, on suppose que $(K_\ell, \Sigma_\ell, P_\ell)$ est affinement équivalent à $\left(\hat{K}, \hat{\Sigma}, \hat{P}\right)$. En d'autres termes, la transformation F_ℓ : $\hat{K} \to K_\ell$ est affine, et on rappelle que l'on peut écrire :

$$M = F_\ell(\hat{M}) = \mathbb{B}_\ell \hat{M} + \overrightarrow{C}_\ell, \quad \mathbb{B}_\ell \in \mathbb{R}^{n\times n} \text{ inversible}, \ \overrightarrow{C}_\ell \in \mathbb{R}^n.$$

Dans ce cas, on vérifie que $[dF_\ell(\hat{M})] = \mathbb{B}_\ell$, $J_{F_\ell}(\hat{M}) = det(\mathbb{B}_\ell)$, et enfin $\hat{M} = \mathbb{B}_\ell^{-1} M - \mathbb{B}_\ell^{-1} \overrightarrow{C}_\ell$. Notons qu'on peut facilement borner la norme de $\mathbb{B}_\ell$ et celle de son inverse.

Lemme 2.3. *On a les majorations :*

$$\|\mathbb{B}_\ell\| \leq \frac{h_{K_\ell}}{\hat{\rho}}, \ \|\mathbb{B}_\ell^{-1}\| \leq \frac{\hat{h}}{\rho_{K_\ell}}. \tag{2.57}$$

Démonstration : Classiquement, on a

$$\|\mathbb{B}_\ell\| = \frac{1}{R} \sup_{\hat{y}, \ |\hat{y}|=R} |\mathbb{B}_\ell \hat{y}|,$$

pour $R > 0$ quelconque. Or $\hat{\rho}$ étant par définition égal au rayon de la plus grande boule inscrite dans $\hat{K}$, pour tout vecteur $\hat{y}$ de norme $2\hat{\rho}$, il existe deux points $\hat{M}_1$ et $\hat{M}_2$ de $\hat{K}$ tels que $\hat{y} = \overrightarrow{\hat{M}_1 \hat{M}_2}$. Les points $M_1 = F_\ell(\hat{M}_1)$ et $M_2 = F_\ell(\hat{M}_2)$ appartiennent bien sûr à K_ℓ et, par définition de h_{K_ℓ}, on a $|\overrightarrow{M_1 M_2}| \leq 2h_{K_\ell}$. Si on applique la formule précédente pour calculer $\|\mathbb{B}_\ell\|$, on en déduit le résultat de gauche de (2.57).
Pour le résultat de droite, on procède de façon similaire, en passant cette fois de K_ℓ à $\hat{K}$. ∎

Pour être complet, on note que, pour évaluer $|det(\mathbb{B}_\ell)|$, on peut appliquer la formule $|det(\mathbb{B}_\ell)| = \text{mes}(K_\ell)/\text{mes}(\hat{K})$.

Maintenant, on associe à une fonction $\hat{w}$ définie sur $\hat{K}$, la fonction w définie sur K par $\hat{w}(\hat{M}) = w(M)$ (et réciproquement). On peut écrire $w = \hat{w} \circ F_\ell^{-1}$, ou $\hat{w} = w \circ F_\ell$. Les transformations F_ℓ et F_ℓ^{-1} étant de classe $\mathcal{C}^\infty$ respectivement sur $\overline{\hat{K}}$ et sur $\overline{K_\ell}$, on a en particulier les résultats $w \in H^p(K_\ell) \Longleftrightarrow \hat{w} \in H^p(\hat{K})$, pour tout p.

Lemme 2.4. *Pour tout $q \geq 0$, on a les estimations suivantes :*

$$|v|_{q,K_\ell} \leq C_q \, (\rho_{K_\ell})^{-q} \, |det(\mathbb{B}_\ell)|^{1/2} \, |v \circ F_\ell|_{q,\hat{K}}, \quad \forall v \in H^q(K_\ell) \tag{2.58}$$

$$|\hat{v}|_{q,\hat{K}} \leq C'_q \, (h_{K_\ell})^q \, |det(\mathbb{B}_\ell)|^{-1/2} \, |\hat{v} \circ F_\ell^{-1}|_{q,K_\ell} \quad \forall \hat{v} \in H^q(\hat{K}). \tag{2.59}$$

Ci-dessus, les constantes C_q et C'_q ne dépendent que de q (et de la dimension n).

Démonstration : Commençons par établir la majoration (2.58) dans les cas $q = 0$ et $q = 1$. On rappelle la formule de changement de variable

$$\int_{K_\ell} w \, d\Omega = |det(\mathbb{B}_\ell)| \int_{\hat{K}} \hat{w} \, d\hat{\Omega}.$$

Pour $q = 0$, il suffit de choisir $w = v^2$ (et $C_0 = 1$).
Pour $q = 1$, on remplace cette fois w par $|\nabla v|^2$. On se souvient que $\nabla v(M) = (\mathbb{B}_\ell^{-1})^t \, \hat{\nabla}\hat{v}(\hat{M})$. La formule de changement de variable donne cette fois

$$\begin{aligned} |v|^2_{1,K_\ell} = \int_{K_\ell} |\nabla v|^2 \, d\Omega &= |det(\mathbb{B}_\ell)| \int_{\hat{K}} |(\mathbb{B}_\ell^{-1})^t \hat{\nabla}\hat{v}|^2 \, d\hat{\Omega} \\ &\leq |det(\mathbb{B}_\ell)| \, \|\mathbb{B}_\ell^{-1}\|^2 \int_{\hat{K}} |\hat{\nabla}\hat{v}|^2 \, d\hat{\Omega} \\ &\leq \hat{h}^2 \, |det(\mathbb{B}_\ell)| \, (\rho_{K_\ell})^{-2} \, |\hat{v}|^2_{1,\hat{K}}. \end{aligned}$$

On obtient (2.58) pour $q = 1$, avec $C_1 = \hat{h}$.
Pour $q \geq 2$, on raisonne dérivée partielle par dérivée partielle. Détaillons la méthode pour $q = 2$, qui se généralise sans peine au cas $q > 2$. On note que

$$|\frac{\partial^2 v}{\partial x_i \partial x_j}(M)| = |d^2 v(M).(\mathbf{e}_i, \mathbf{e}_j)| \leq \sup_{\mathbf{x}_1, \mathbf{x}_2, \ |\mathbf{x}_1| = |\mathbf{x}_2| = 1} |d^2 v(M).(\mathbf{x}_1, \mathbf{x}_2)| =: \|d^2 v(M)\|.$$

Par ailleurs, pour tout couple de vecteurs $(\mathbf{x}_1, \mathbf{x}_2)$, on a la relation de différentiation des fonctions composées suivante (cf. [9]),

$$d^2 v(M).(\mathbf{x}_1, \mathbf{x}_2) = d^2 \hat{v}(\hat{M}).(\mathbb{B}_\ell^{-1}\mathbf{x}_1, \mathbb{B}_\ell^{-1}\mathbf{x}_2),$$

de sorte que

$$\|d^2 v(M)\| \leq \|d^2 \hat{v}(\hat{M})\| \, \|\mathbb{B}_\ell^{-1}\|^2.$$

Pour chaque dérivée partielle, on a donc obtenu une majoration par la même quantité. Or, le nombre de dérivées partielles d'ordre 2 est fini : soit D_2 leur nombre (qui dépend de la dimension n). On trouve alors

$$|v|^2_{2,K_\ell} = \sum_{\alpha\in\mathbb{N}^2,|\alpha|=2}\int_{K_\ell}|\partial_\alpha v(M)|^2\,d\Omega \le D_2\,\|\mathbb{B}_\ell^{-1}\|^4\int_{K_\ell}\|d^2\hat v(F_\ell^{-1}(M))\|^2\,d\Omega$$
$$\le D_2\,\hat h^4\,|det(\mathbb{B}_\ell)|\,(\rho_{K_\ell})^{-4}\int_{\hat K}\|d^2\hat v(\hat M)\|^2\,d\hat\Omega.$$

Pour pouvoir conclure, il suffit de remarquer que l'on a une inégalité de type

$$\|d^2\hat v(\hat M)\| \le D_2'\max_{\alpha\in\mathbb{N}^2,|\alpha|=2}|\partial_\alpha\hat v(\hat M)|,$$

avec une constante $D_2'>0$ (qui dépend de n), ce qui conduit à la majoration

$$|v|^2_{2,K_\ell} \le D_2(D_2')^2\,\hat h^4\,|det(\mathbb{B}_\ell)|\,(\rho_{K_\ell})^{-4}\,|v|^2_{2,\hat K},$$

c'est-à-dire (2.58) pour $q=2$, avec $C_2 = D_2^{1/2}D_2'\,\hat h^2$. On procède de même pour $q>2$.

Pour établir les majorations (2.59), il suffit de raisonner en effectuant le changement de variable inverse. ∎

Le dernier résultat, que nous admettons, permet de comparer la semi-norme d'une fonction de H^{r+1}, à la norme complète de la même fonction, corrigée d'un polynôme de degré au plus r. Nous l'appliquons dans le domaine de référence $\hat K$.

Proposition 2.6. *Soit ω un ouvert borné de $\mathbb{R}^n$, de frontière suffisamment régulière, et soit $r\in\mathbb{N}$. Alors, il existe $C_r>0$ qui ne dépend que de ω et de r, telle que*

$$\inf_{p\in P^r(\omega)}\|w+p\|_{H^{r+1}(\omega)} \le C_r|w|_{r+1,\omega},\ \forall w\in H^{r+1}(\omega). \tag{2.60}$$

Nous pouvons maintenant passer à la

Démonstration de la proposition 2.5 : Comme la transformation F_ℓ est affine, on a la propriété $\hat\pi\hat v(\hat M) = \pi_{K_\ell}v(M)$, avec $\hat\pi$ l'opérateur d'interpolation sur le domaine de référence $\hat K$. D'après (2.58) appliquée à $q=m$, on en déduit imédiatement que

$$|v-\pi_{K_\ell}v|_{m,K_\ell} \le C\,(\rho_{K_\ell})^{-m}\,|det(\mathbb{B}_\ell)|^{1/2}\,|\hat v-\hat\pi\hat v|_{m,\hat K}.$$

Puisque $m\le k+1$, on a $|\hat v-\hat\pi\hat v|_{m,\hat K}\le\|\hat v-\hat\pi\hat v\|_{H^{k+1}(\hat K)}$. Ensuite, comme on considère ici un élément fini de Lagrange d'ordre k, on a $\hat\pi\hat p=\hat p$, pour tout $\hat p\in P^k(\hat K)$:

$$\|\hat v-\hat\pi\hat v\|_{H^{k+1}(\hat K)} = \inf_{\hat p\in P^k(\hat K)}\|(\hat v-\hat p)-\hat\pi(\hat v-\hat p)\|_{H^{k+1}(\hat K)}.$$

Et, qui plus est,

$$\|(\hat v-\hat p)-\hat\pi(\hat v-\hat p)\|_{H^{k+1}(\hat K)} = \|(Id-\hat\pi)(\hat v+\hat p)\|_{H^{k+1}(\hat K)} \le C_{k+1}\,\|\hat v+\hat p\|_{H^{k+1}(\hat K)},$$

avec C_{k+1} le module de continuité de l'opérateur $(Id-\hat\pi)$ de $H^{k+1}(\hat K)$ dans lui-même. On peut donc utiliser la proposition 2.6 avec $r=k$, pour trouver (on a inclus dans C toutes les constantes précédentes)

$$|v-\pi_{K_\ell}v|_{m,K_\ell} \le C\,(\rho_{K_\ell})^{-m}\,|det(\mathbb{B}_\ell)|^{1/2}|\hat v|_{k+1,\hat K}.$$

La conclusion de la démonstration suit par l'application de la majoration (2.59), pour $q=k+1$:

$$|v-\pi_{K_\ell}v|_{m,K_\ell} \le C\,(h_{K_\ell})^{k+1}(\rho_{K_\ell})^{-m}\,|v|_{k+1,K_\ell}.$$

∎

Le cas minimal d'application de cette proposition correspond à la situation $v \in H^2(K_\ell)$ et l'utilisation d'éléments finis d'ordre 1 (P^1 ou Q^1-Lagrange), qui constitue une situation courante. On a alors les estimations suivantes ($m = 0$ ou 1 et $k = 1$ dans (2.56)) :

$$\boxed{\begin{array}{c} \|v - \pi_{K_\ell} v\|_{L^2(K_\ell)} \leq C h_{K_\ell}^2 |v|_{2,K_\ell} \\ \\ |v - \pi_{K_\ell} v|_{1,K_\ell} \leq C \dfrac{h_{K_\ell}^2}{\rho_{K_\ell}} |v|_{2,K_\ell} \end{array}} \tag{2.61}$$

On observe dans la seconde estimation de (2.61) la présence du coefficient ρ_{K_ℓ} au dénominateur, coefficient destiné à tendre vers zéro. Pour obtenir une estimation utilisable en pratique, on se place donc dans la situation où :

$$\exists \sigma > 0,\ \forall h,\ \forall K_\ell \in \mathcal{T}_h,\ \frac{h_{K_\ell}}{\rho_{K_\ell}} \leq \sigma. \tag{2.62}$$

On dit que la famille de maillages $(\mathcal{T}_h)_h$ est régulière.
Cette seconde estimation devient alors :

$$\boxed{|v - \pi_{K_\ell} v|_{1,K_\ell} \leq C\, \sigma\, h_{K_\ell} |v|_{2,K_\ell}} \tag{2.63}$$

Remarque 2.17. L'estimation en norme $L^2(\Omega)$ est bien meilleure que celle en semi-norme $H^1(\Omega)$. Cela résulte du fait l'on contrôle assez mal les dérivées d'une fonction à l'aide de ses valeurs nodales. En outre, l'estimation (2.56) montre que l'ordre de l'erreur est limité, d'une part, par la régularité de la fonction v et d'autre part, par l'ordre des éléments finis. Par exemple, il est inutile d'utiliser des éléments finis d'ordre supérieur à 1 lorsque la solution cherchée est seulement dans $H^2(\Omega)$.

On a observé, à la fin du chapitre précédent, que la solution u d'un problème de Laplace pouvait ne pas appartenir à $H^2(\Omega)$, notamment lorsque l'ouvert Ω est un polyèdre ou un polygone non-convexe. C'est pourquoi, on pourra utiliser avec profit le résultat plus fin ci-dessous.

Proposition 2.7. (erreur d'interpolation locale) *Soit $(K_\ell, \Sigma_\ell, P_\ell)$ un élément fini de Lagrange d'ordre k vérifiant (2.62), et affinement équivalent à l'élément fini $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$. Soit $s \in]1/2, 1[$. Alors il existe une constante C qui ne dépend que de $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$ et de s, telle que, $\forall v \in H^{s+1}(K_\ell)$, on a les estimations suivantes :*

$$\boxed{\begin{array}{c} \|v - \pi_{K_\ell} v\|_{L^2(K_\ell)} \leq C h_{K_\ell}^{1+s} |v|_{s+1,K_\ell} \\ \\ |v - \pi_{K_\ell} v|_{1,K_\ell} \leq C\, \sigma\, h_{K_\ell}^{s} |v|_{s+1,K_\ell} \end{array}} \tag{2.64}$$

Nous ne donnons pas le sens exact de la semi-norme H^{s+1}. Il convient simplement de savoir que l'on a la propriété générale $\sum_{\ell=1,L} |u|^2_{s+1,K_\ell} \leq C'|u|^2_{s+1,\Omega}$.

Lorsque $\Omega \subset \mathbb{R}^2$, la condition (2.62) est équivalente, dans le cas des triangles, à la condition angulaire suivante :

$$\exists \theta_0 > 0,\ \forall h,\ \forall K_\ell \in \mathcal{T}_h,\ \theta_{K_\ell} \geq \theta_0 \quad (\theta_{K_\ell} \text{ plus petit angle de } K_\ell).$$

Elle traduit un non-aplatissement des triangles. Lorsque $\Omega \subset \mathbb{R}^3$, la condition

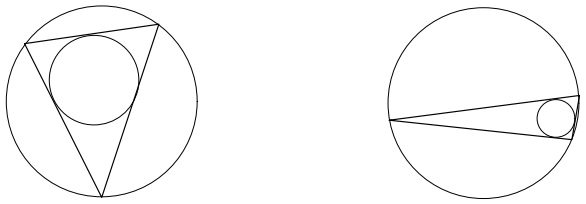


Fig. 2.6. Condition géométrique dans le cas des triangles

(2.62) implique une condition sur les angles solides au sommet, ainsi qu'une condition sur les angles diédriques entre deux faces[9] :

$$\begin{cases} \exists \omega_0 > 0,\ \forall h,\ \forall K_\ell \in \mathcal{T}_h,\ \omega_{K_\ell} \geq \omega_0 \quad (\omega_{K_\ell} \text{ plus petit angle solide de } K_\ell) \\ \exists \iota_0 > 0,\ \forall h,\ \forall K_\ell \in \mathcal{T}_h,\ \iota_{K_\ell} \geq \iota_0 \quad (\iota_{K_\ell} \text{ plus petit angle diédrique de } K_\ell) \end{cases}.$$

• Estimation d'erreur de la méthode des éléments finis

De l'estimation de l'erreur d'interpolation locale (2.56) on déduit immédiatement le théorème de convergence de l'approximation par éléments finis de Lagrange :

Théorème 2.4. (convergence de la méthode des éléments finis) *Supposons que, pour tout h, le maillage $\mathcal{T}_h$ composé d'éléments finis de Lagrange d'ordre k satisfasse aux hypothèses (2.35), (2.36), (2.44), (2.45), (2.46), et que la famille de maillages $(\mathcal{T}_h)_h$ soit régulière (voir (2.62)). Si la solution u du problème (2.51) appartient à $H^{k+1}(\Omega)$ avec $k \geq 1$, alors on a, pour tout h :*

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^k |u|_{k+1,\Omega}. \tag{2.65}$$

[9] Un angle solide au sommet est compris entre 0 et 4π ; un angle diédrique entre deux faces est compris entre 0 et 2π.

Démonstration : d'après (2.56) et l'hypothèse (2.62) on a :

$$|u - \pi_{K_\ell} u|_{L^2(K_\ell)} \leq C_1 h_{K_\ell}^{k+1} |u|_{k+1,K_\ell}$$

$$|u - \pi_{K_\ell} u|_{1,K_\ell} \leq C_2 \sigma h_{K_\ell}^{k} |u|_{k+1,K_\ell}$$

qui conduit à l'estimation :

$$\begin{aligned} \|u - \pi_{K_\ell} u\|_{H^1(K_\ell)} &\leq \left(C_1^2 h_{K_\ell}^2 + C_2^2 \sigma^2\right)^{1/2} h_{K_\ell}^k |u|_{k+1,K_\ell} \\ &\leq C(\sigma) h^k |u|_{k+1,K_\ell}, \end{aligned}$$

puisque par définition $h = \max_{K_\ell \in \mathcal{T}_h} h_{K_\ell}$ (cf. (2.37)).
En sommant sur tous les polyèdres on obtient :

$$\|u - \pi_h u\|_{H^1(\Omega)} \leq C(\sigma) h^k \left(\sum_{\ell=1,L} |u|_{k+1,K_\ell}^2 \right)^{1/2} \leq C(\sigma) h^k |u|_{k+1,\Omega}$$

qui donne (2.65) en vertu de (2.55) avec $w = u$. ∎

Même si $u \notin H^2(\Omega)$, on peut établir la convergence de l'approximation par éléments finis dans l'espace $H^1(\Omega)$, en s'appuyant sur la proposition 2.7. Typiquement, pour le problème de Neumann (cf. §1.4.1), on sait que la solution cherchée appartient à Ψ_N, voir (1.28). On aboutit alors au résultat *minimal* suivant.

Théorème 2.5. (convergence de la méthode des éléments finis) *Soit $(\mathcal{T}_h)_h$ une famille de maillages $(\mathcal{T}_h)_h$ vérifiant les hypothèses du théorème 2.4. Soient u la solution du problème de Neumann avec second membre f dans $L^2(\Omega)$ et u_h la solution du problème discret. Alors, pour tout $s < \sigma_N$, il existe $C_s > 0$ telle que, pour tout h :*

$$\|u - u_h\|_{H^1(\Omega)} \leq C_s h^s \|f\|_{L^2(\Omega)}. \tag{2.66}$$

Démonstration : On reprend celle du théorème 2.4, en utilisant cette fois l'estimation (2.64) pour la norme L^2 et pour la semi-norme H^1. On arrive donc dans un premier temps à

$$\|u - u_h\|_{H^1(\Omega)} \leq C_s h^s |u|_{s+1,\Omega}.$$

Ensuite, l'inclusion continue de Ψ_N dans $H^{1+s}(\Omega)$ (voir le §1.5.2) permet d'écrire

$$|u|_{s+1,\Omega} \leq C_1 \|u\|_\Psi,$$

avec une constante C_1 indépendante de u. Enfin, on note que, par définition,

$$\|u\|_\Psi^2 = \|u\|_{H^1(\Omega)}^2 + \|\Delta u\|_{L^2(\Omega)}^2 \leq C_2 \|f\|_{L^2(\Omega)}^2,$$

avec une constante C_2 indépendante de u. En effet, d'une part d'après le théorème de Lax-Milgram, $\|u\|_{H^1(\Omega)}$ dépend continûment de $\|f\|_{L^2(\Omega)}$, ce qui permet de borner le premier terme, et d'autre part, on a la relation $-\Delta u = f - u$, ce qui permet cette fois de borner le second terme.
∎

Remarque 2.18. Le paramètre σ_N appartient à $]\frac{1}{2}, 1]$, si l'on reprend la discussion du §1.5.2, selon que l'ouvert Ω est convexe ou non-convexe. D'après la proposition précédente, l'utilisation d'éléments finis de Lagrange d'ordre 1 suffit pour arriver à l'ordre minimal de convergence.

L'utilisation du lemme de Céa ne conduit qu'à une estimation en norme H^1 et ne permet pas d'obtenir une meilleure estimation en norme L^2 comme le suggère l'estimation portant sur l'erreur d'interpolation. Toutefois, moyennant une hypothèse supplémentaire, on peut obtenir une estimation d'erreur en norme L^2 d'un ordre supérieur en utilisant un argument de type dualité.

On introduit le problème suivant, dit adjoint :

$$\begin{cases} \text{Trouver } \varphi \in H^1(\Omega) \text{ tel que :} \\ a(v,\varphi) = \displaystyle\int_\Omega g v d\Omega \quad \forall v \in H^1(\Omega) \end{cases} \tag{2.67}$$

avec $g \in L^2(\Omega)$. D'après le théorème de Lax-Milgram le problème (2.67) admet une unique solution.

Définition 2.5. *Le problème (2.67) est dit régulier si l'application $g \mapsto \varphi$ est linéaire continue de $L^2(\Omega)$ dans $H^2(\Omega)$, c'est-à-dire :*

$$g \in L^2(\Omega) \implies \varphi \in H^2(\Omega) \text{ et } \|\varphi\|_{H^2(\Omega)} \leq C\|g\|_{L^2(\Omega)}.$$

(Ci-dessus, la constante C est indépendante de g).

Lorsque le problème (2.67) est régulier on obtient alors le résultat de convergence en norme L^2 suivant :

Théorème 2.6. *Sous les hypothèses du théorème 2.4, si le problème (2.67) est régulier* [10]*, alors il existe une constante C, indépendante de h, telle que :*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{k+1}|u|_{k+1,\Omega}. \tag{2.68}$$

Démonstration : Soit $g \in L^2(\Omega)$, on a :

$$\begin{aligned} \left|\int_\Omega g(u-u_h)d\Omega\right| &= |a(u-u_h,\varphi)| && \text{d'après (2.67)} \\ &= |a(u-u_h,\varphi-\psi_h) + a(u-u_h,\psi_h)| && \forall \psi_h \in V_h \\ &= |a(u-u_h,\varphi-\psi_h)| && \text{car } a(u,\psi_h) = \ell(\psi_h) = a(u_h,\psi_h) \\ &\leq M\,\|u-u_h\|_{H^1(\Omega)}\,\|\varphi-\psi_h\|_{H^1(\Omega)} && \text{par continuité de } a(\cdot,\cdot). \end{aligned}$$

D'où

$$\left|\int_\Omega g(u-u_h)d\Omega\right| \leq M\,\|u-u_h\|_{H^1(\Omega)} \inf_{\psi_h \in V_h} \|\varphi-\psi_h\|_{H^1(\Omega)}.$$

Maintenant, comme le problème (2.67) est régulier, on a $\varphi \in H^2(\Omega)$ et $\|\varphi\|_{H^2(\Omega)} \leq C\,\|g\|_{L^2(\Omega)}$. En vertu du théorème 2.4, on obtient l'estimation :

$$\inf_{\psi_h \in V_h} \|\varphi-\psi_h\|_{H^1(\Omega)} \leq \|\varphi-\varphi_h\|_{H^1(\Omega)} \leq Ch\|\varphi\|_{H^2(\Omega)} \leq Ch\|g\|_{L^2(\Omega)}.$$

[10] Par exemple, pour la forme bilinéaire $a(u,v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v\, d\Omega$ associée au problème de Laplace avec condition de Neumann homogène, on renvoie au §1.5.2 pour des résultats de régularité.

Ci-dessus, φ_h est la solution du problème discret associé à (2.67).
Par ailleurs, si $u \in H^{k+1}(\Omega)$, on a d'après l'estimation d'erreur (2.65) :

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

Ce qui conduit, compte-tenu de ce qui précède, à l'estimation :

$$\left| \int_\Omega g(u - u_h) d\Omega \right| \leq Ch^{k+1} \|g\|_{L^2(\Omega)} |u|_{k+1,\Omega} \quad \forall g \in L^2(\Omega).$$

Pour conclure, il suffit de choisir $g = u - u_h$:

$$\|u - u_h\|^2_{L^2(\Omega)} \leq Ch^{k+1} \|u - u_h\|_{L^2(\Omega)} |u|_{k+1,\Omega} \implies \|u - u_h\|_{L^2(\Omega)} \leq Ch^{k+1} |u|_{k+1,\Omega}. \quad \blacksquare$$

Le théorème 2.6 montre ainsi que les éléments finis d'ordre 1 (i.e. P^1 ou Q^1) convergent à l'ordre 1 en norme H^1 et à l'ordre 2 en norme L^2 dès lors que le problème est régulier.

Remarque 2.19. Dans le cas où la solution appartient à Ψ_N (et que $\Psi_N \not\subset H^2(\Omega)$), la démonstration précédente permet de gagner un facteur $C_s h^s$, avec $s < \sigma_N$, entre l'estimation en norme $H^1(\Omega)$ et celle en norme $L^2(\Omega)$, puisque la solution du problème adjoint appartient à Ψ_N.

L'analyse que nous venons de mener dans le cas $V = H^1(\Omega)$ s'adapte sans difficulté au cas de l'espace $V = H^1_0(\Omega)$ (cas du problème de Dirichlet), et on remplace alors Ψ_N par Ψ_D (et σ_N par σ_D). Elle s'adapte également au cas des problèmes mixtes. Dans les cas plus généraux, par exemple avec des données moins régulières, ou pour un problème mixte, il n'est plus garanti que la solution appartienne à un espace de type $H^{1+s}(\Omega)$, avec $s > \frac{1}{2}$, permettant d'obtenir des estimations d'erreurs locales (voir les propositions 2.5 et 2.7), puis de conclure à l'aide du théorème 2.4. Examinons deux cas concrets. Tout d'abord, lorsqu'on a un second membre $f \in H^{-1}(\Omega)$ pour le problème de Dirichlet homogène, on sait uniquement que la solution appartient à $H^1(\Omega)$! Ensuite, en ce qui concerne les problèmes mixtes Dirichlet-Neumann, ceux-ci ne seront pas nécessairement réguliers [18, 19] au sens où, même posés dans un ouvert Ω convexe, la solution peut ne pas appartenir automatiquement à $H^2(\Omega)$; ceci est dû au changement de condition aux limites. La solution appartiendra toutefois à un espace du type $H^{1+s}(\Omega)$, pour $s < \sigma_M$, avec $\sigma_M > 0$ (à comparer aux résultats du §1.5.2). Dans ces cas plus généraux, on pourra malgré tout établir la convergence de l'approximation par éléments finis, en s'appuyant sur les estimations du théorème 2.4 pour des fonctions régulières et en raisonnant par densité à l'aide du théorème général de convergence, le théorème 2.1 (voir e.g. [16]).

2.3.2 Estimateurs d'erreur et raffinement de maillage

Les théorèmes de convergence 2.4 et 2.5 fournissent des estimations d'erreurs globales. En réalité, comme le suggèrent les estimations d'erreur d'interpolation

(2.64), l'erreur sur la solution n'est pas uniforme sur le domaine de calcul. Elle dépend du pas local de maillage, de la régularité locale ou de la variation locale de la solution. Il paraît clair que dans une zone où la solution varie rapidement on a intérêt à utiliser un maillage plus fin que dans une zone où elle varie peu. On parle alors de raffinement local du maillage. On peut mettre en œuvre des techniques de raffinement de maillage soit à partir d'informations *a priori*, connaissance de la singularité de la solution par exemple, soit à partir d'informations *a posteriori* sur la solution calculée, obtenues par des estimateurs d'erreur.

Par exemple, si l'on reprend le §1.5.2 (encore lui...), on constate que la solution est de régularité H^2, sauf dans un voisinage ω des coins (et arêtes) rentrants de $\partial\Omega$. Si l'on veut améliorer la vitesse de convergence, on peut alors faire la remarque suivante : dans tous les triangles (ou tétraèdres) K_ℓ tels que $K_\ell \cap \omega = \emptyset$, on peut utiliser l'estimation standard (2.63) et, dans tous les autres K_ℓ, on utilise l'estimation affinée (2.64). Utilisée sur le maillage uniforme de la figure 2.7, cette seconde estimation (en $h_{K_\ell}^s$) est a priori moins bonne que la première (resp. en h_{K_ℓ}).

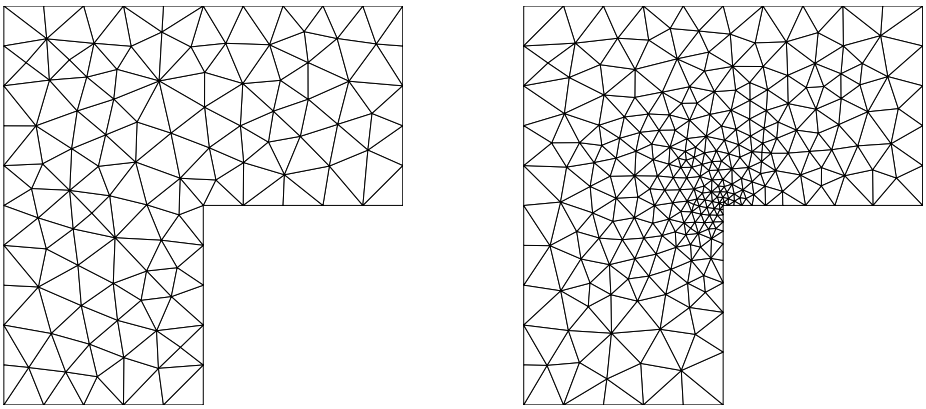


Fig. 2.7. Maillages uniforme et localement raffiné

L'idée est ici de construire, pour la seconde estimation, des triangles ou tétraèdres de plus en plus petits au voisinage des coins (ou arêtes) rentrants, de sorte que h_{K_ℓ} décroisse lorsque l'on s'en approche. En faisant décroître les rayons h_{K_ℓ} suffisamment vite (voir [3]), on arrive au résultat

$$\sum_{\ell \text{ t.q. } K_\ell \cap \omega \neq \emptyset} h_{K_\ell}^{2s} |u|_{s,K_\ell}^2 \leq C h^2 \|f\|_{L^2(\omega)}^2.$$

Ceci permet finalement de recouvrir une vitesse de convergence globalement en $O(h)$. Sur la figure 2.7 nous donnons un exemple de maillage du même ouvert, raffiné au voisinage du coin rentrant.
Dans les cas où il est difficile d'obtenir des informations *a priori* sur la solution, on recourt à des estimateurs d'erreurs qui permettent d'estimer localement la qualité de la solution discrète obtenue. Si cette qualité est jugée, au regard de l'estimateur, insuffisante, on raffine localement le maillage et on recalcule sur ce nouveau maillage une nouvelle solution approchée. On itère ce procédé jusqu'à trouver une solution approchée satisfaisant un critère de qualité donné au sens de l'estimateur. Un procédé simple de raffinement local consiste à subdiviser un triangle (resp. un tétraèdre) en 4 triangles (resp. 8 tétraèdres).

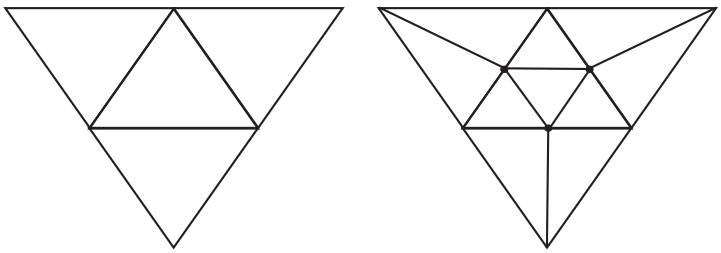


Fig. 2.8. Raffinement local d'un triangle

Il existe plusieurs estimateurs d'erreurs. Le plus simple et un des plus utilisés est celui fondé sur le résidu volumique. Nous allons le décrire dans le cas du problème de Laplace-Poisson posé dans un ouvert borné polygonal ou polyédrique Ω de $\mathbb{R}^n$ ($n = 2$ ou 3) :

$$\begin{cases} -\triangle u = f \text{ sur } \Omega \\ u = 0 \qquad \text{sur } \partial\Omega \end{cases}$$

Soit u_h la solution approchée obtenue par approximation par éléments finis de Lagrange P^k sur un maillage $\mathcal{T}_h$ composé des polyèdres $(K_\ell)_{\ell=1,L}$. On note $\mathcal{E}_h$ l'ensemble de toutes les faces internes $e = K_\ell \cap K_m$ du maillage et on définit les résidus locaux ci-dessous (n_e désigne le vecteur unitaire normal à e, orienté de K_ℓ vers K_m, et $[\cdot]_e$ le saut de valeur à la traversée de e).

$$\begin{cases} R_\ell(u_h) = f + \triangle u_h \text{ sur } \overset{\circ}{K}_\ell & \ell = 1, L \\ R_e(u_h) = \left[\dfrac{\partial u_h}{\partial n_e}\right]_e = [\nabla u_h \cdot n_e]_e & \forall e \in \mathcal{E}_h. \end{cases}$$

Notons que dans le cadre de l'approximation de Lagrange, les fonctions de base étant seulement H^1 et non H^2, $\triangle u_h$ n'est défini (au sens classique) qu'à l'intérieur

de chaque triangle ou tétraèdre. Remarquons que pour le cas de l'approximation P^1, $R_\ell(u_h) = f$ car u_h est une fonction affine. Le résidu $R_e(u_h)$ mesure bien un "défaut" d'approximation car la solution exacte u vérifie $\left[\frac{\partial u}{\partial n_e}\right]_e = 0$ dès lors que la donnée f appartient à $L^2(\Omega)$. En pratique, ces résidus sont relativement faciles à calculer. A partir de ces résidus, on construit l'estimateur local :

$$E_\ell(u_h) = \left(h_{K_\ell}^n \left\| R_\ell(u_h) \right\|_{L^2(K_\ell)}^2 + \sum_{e \in \partial K_\ell \cap \mathcal{E}_h} h_e^{n-1} \left\| R_e(u_h) \right\|_{L^2(e)}^2 \right)^{\frac{1}{2}}$$

où h_e désigne le rayon de la face e.
On a les estimations suivantes (voir par exemple [7] pour la preuve) qui garantissent que ces estimateurs caractérisent *asymptotiquement* l'erreur d'approximation. Il existe c une constante dépendant uniquement de Ω et du facteur de forme σ du maillage (cf. (2.62)) :

$$\left\| u - u_h \right\|_{H^1(\Omega)} \leq c \left(\sum_{\ell=1,L} E_\ell^2(u_h) \right)^{\frac{1}{2}}$$

$$E_\ell(u_h) \leq c \left\| u - u_h \right\|_{H^1(V_\ell)}, \quad \ell = 1, L$$

où V_ℓ désigne le domaine géométrique constitué de K_ℓ et de ses voisins K_m qui partagent une face avec K_ℓ.
Ces estimations montrent, d'une part, que si les résidus locaux tendent vers 0 alors l'erreur globale sur la solution tend vers 0 et, d'autre part, que les résidus locaux tendent bien vers 0 lorsque l'erreur d'approximation locale tend vers 0. *Attention*, ils ne prouvent pas que l'erreur d'approximation locale est proportionnelle à l'estimateur local E_ℓ et ne garantissent donc pas qu'une étape de raffinement local va améliorer la précision locale de la solution ! Par contre, c'est vrai asymptotiquement, c'est-à-dire si on raffine de plus en plus le maillage. Signalons que la constante c qui intervient dans ces estimations est liée aux constantes de continuité et de coercivité du problème. En particulier, dans le cas d'un problème faiblement coercif, cette constante est grande et la qualité de l'estimateur peut être médiocre. Néanmoins, dans la pratique, cet estimateur simple à mettre en œuvre donne des résultats satisfaisants. Il existe d'autres estimateurs ; citons-en deux à titre d'exemple : ceux basés sur la résolution de problèmes locaux et utilisant une approximation par éléments finis d'un ordre plus élevé que celle utilisée pour résoudre le problème initial, et ceux basés sur le calcul d'une solution approchée sur un maillage plus fin.

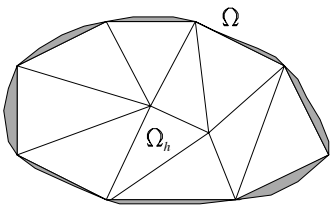
2.3.3 Domaines non polyédriques et approximation des données

Dans la présentation de l'analyse d'erreur que nous venons de faire, nous avons volontairement exclu, dans un souci de simplification, des situations plus compliquées mais qui sont celles que l'on rencontre dans la pratique : le cas des ouverts non polyédriques (i.e. les polyèdres curvilignes) et le cas des approximations des données (second membre). Nous allons les évoquer brièvement dans cette sous-section.

• Ouverts non polyédriques

Nous avons toujours supposé que l'ouvert Ω était polyédrique de telle sorte que $\overline{\Omega} = \bigcup_{\ell=1,L} K_\ell$. Or les objets que l'on maille en pratique sont rarement polyédriques. Le maillage de tels objets à l'aide d'éléments finis à face plane conduit à un ouvert Ω_h qui diffère de Ω.

$$\overline{\Omega}_h = \bigcup_{\ell=1,L} K_\ell \neq \overline{\Omega}$$



Cet écart[11] entre l'ouvert Ω et l'ouvert Ω_h introduit une approximation de la forme bilinéaire $a(\cdot,\cdot)$ et de la forme linéaire $\ell(\cdot)$; le problème discrétisé (2.52) prenant alors la forme suivante :

$$\begin{cases} \text{Trouver } \tilde{u}_h \in V_h \text{ tel que :} \\ a_h(\tilde{u}_h, v_h) = \ell_h(v_h) \qquad \forall v_h \in V_h \end{cases} \tag{2.69}$$

Evidemment, la solution $\tilde{u}_h$ du problème (2.69) est différente de la solution u_h du problème (2.52). La question est de savoir si l'estimation d'erreur (2.65) reste vraie.

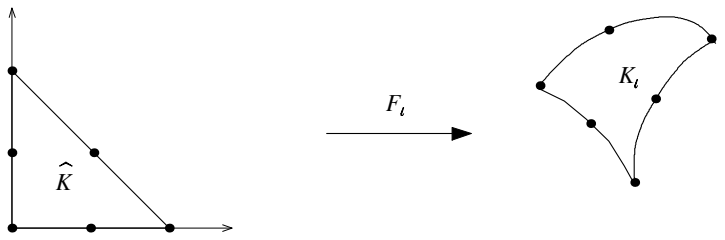
On trouvera dans [21] la démonstration du fait que l'erreur due à l'approximation par des faces planes est d'ordre $h^{3/2}$ dans le cas de l'approximation du problème de Dirichlet homogène. Ceci indique que l'utilisation[12] d'éléments finis de Lagrange d'ordre supérieur à 1 n'est pas optimale puisque l'on "perd" tout ordre de convergence meilleur que $h^{3/2}$.

[11] Si l'ouvert Ω est convexe, on a alors $\Omega_h \subset \Omega$. Si au contraire Ω n'est convexe, on a alors $\Omega_h \not\subset \Omega$.

[12] Par exemple, le choix de l'élément fini de Lagrange P^2 permet d'arriver à l'ordre h^2 lorsque $u \in H^3(\Omega)$.

Afin de "récupérer" les ordres attendus avec des éléments finis d'ordre k il faut utiliser des éléments finis courbes, alliés à une technique dite isoparamétrique.

Comme on peut l'imaginer, un élément fini courbe est un élément dont les faces ne sont pas planes et il approche évidemment mieux un ouvert non polyédrique qu'un élément à faces planes. Une technique standard pour construire des éléments finis courbes consiste à transformer l'élément de référence, dont la frontière est composée de faces planes, à l'aide d'une application F_ℓ qui n'est plus affine. Soit $\left(\widehat{K}, \widehat{\Sigma}, \widehat{P}\right)$ un élément fini de référence d'ordre k de fonctions de base $(\widehat{\tau}_i)_{i=1,n_k}$ et notons $(K_\ell, \Sigma_\ell, P_\ell)$ son image par la transformation F_ℓ.



Elément de référence *Elément "courbe"*

Parmi toute les transformations possibles, on utilise celles construites à partir de polynômes : l'ordre de la transformation est alors égal au degré du polynôme choisi. Lorsque la transformation géométrique F_ℓ est
– du même ordre que l'élément fini de référence, l'élément fini $(K_\ell, \Sigma_\ell, P_\ell)$ est dit isoparamétrique.
– d'un ordre inférieur, il est dit hypoparamétrique,
– d'un ordre supérieur, il est dit hyperparamétrique.

Ce dernier cas ne présente aucun intérêt (sauf dans le cadre de la discrétisation par éléments finis des équations intégrales).

• **Exemple** *:* élément fini P^2-Lagrange isoparamétrique en 2D

On note $(\widehat{M}_i)_{i=1,3}$ les sommets du triangle de référence $\widehat{K}$, et $(\widehat{M}_{ij})_{i,j=1,3,\ i\neq j}$ les milieux de $[\widehat{M}_i, \widehat{M}_j]$. Pour revenir à une notation mono-indice, on choisit par exemple $\widehat{M}_4 = \widehat{M}_{12}$, $\widehat{M}_5 = \widehat{M}_{23}$ et $\widehat{M}_6 = \widehat{M}_{31}$, ainsi que $\widehat{\Sigma} = \{\widehat{M}_1, \cdots, \widehat{M}_6\}$. De la même façon, $(M_i^\ell)_{i=1,3}$ désignent les sommets de K, et $(M_{ij}^\ell)_{i,j=1,3,\ i\neq j}$ les "milieux" de $[M_i, M_j]$.

On cherche alors à construire l'application $F_\ell : \mathbb{R}^2 \to \mathbb{R}^2$ inversible, de degré 2, associant à chaque sommet de $\widehat{K}$ un sommet de K, et à chaque milieu de côté $\widehat{M}_{ij}$ le "milieu" de côté correspondant M_{ij} :

$$F_\ell(\widehat{M}_i) = M_i^\ell,\ i = 1,3, \quad F_\ell(\widehat{M}_{ij}) = M_{ij}^\ell,\ i,j = 1,3,\ i \neq j.$$

On posera alors $\Sigma_\ell = F_\ell(\widehat{\Sigma})$ et $K_\ell = F_\ell(\widehat{K})$.

Reprenons la notation mono-indice. Comme la famille $(\widehat{\tau}_i)_{i=1,6}$ engendre $\widehat{P} = P^2(\widehat{K})$, on a la relation :

$$F_\ell(\widehat{x},\widehat{y}) = \sum_{i=1}^{6} \widehat{\tau}_i(\widehat{x},\widehat{y})\, M_i^\ell, \qquad \forall (\widehat{x},\widehat{y}) \in \widehat{K},$$

ce qui permet finalement de construire l'élément fini $(K_\ell, \Sigma_\ell, P_\ell)$ à partir de $(\widehat{K}, \widehat{\Sigma}, \widehat{P})$.

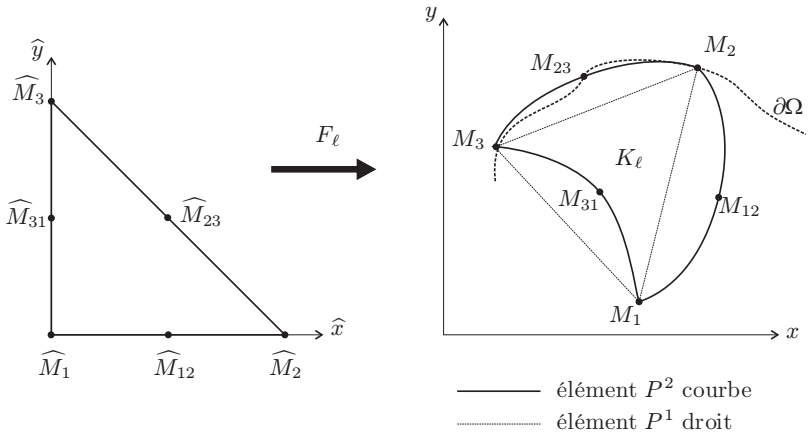


Fig. 2.9. Transformation P^2-isoparamétrique

L'avantage réside dans le fait qu'une parabole épouse mieux une frontière courbe qu'une droite. Plus précisément, on montre que l'ouvert Ω_h approche l'ouvert Ω à un ordre équivalent à celui de l'approximation éléments finis. L'utilisation d'éléments finis isoparamétriques d'ordre k fournit donc une approximation par éléments finis qui converge également à l'ordre k (voir [21]).

• Approximation des données

Lorsqu'on doit calculer des intégrales pour construire la matrice et le second membre du système linéaire à résoudre, deux situations peuvent se produire. Soit on

sait calculer ces intégrales exactement, soit ce n'est pas le cas et on doit alors utiliser des formules d'intégration approchées.

Supposons que l'on sache exprimer la quantité à intégrer sous la forme d'un produit de puissances entières des coordonnées barycentriques, alors on peut utiliser une formule générale (voir [11]). On se place dans un n-simplexe S_n de $\mathbb{R}^n$, c'est-à-dire dans l'enveloppe convexe de $n+1$ points n'appartenant pas à un même hyperplan : en pratique, il s'agit d'un segment en 1D, d'un triangle en 2D ou d'un tétraèdre en 3D. Soient $(\lambda_m)_{1\leq m\leq n+1}$ les coordonnées barycentriques associées à ces $n+1$ points. Alors, pour tous les entiers positifs ou nuls $k_1, \cdots, k_{n+1}$, on a la formule

$$\int_{S_n} \left(\prod_{m=1}^{m=n+1} \lambda_m^{k_m} \right) dx_1 \cdots dx_n = \frac{n! \prod_{m=1}^{m=n+1} k_m!}{(n + \sum_{m=1}^{m=n+1} k_m)!} \operatorname{mes}(S_n). \tag{2.70}$$

Par exemple, si on utilise les éléments finis de Lagrange P^1, les matrices de rigidité et de masse élémentaires sont écrites sous cette forme, voir (2.27) et (2.29).

En règle générale, on ne sait pas calculer de façon exacte le second membre $\ell(v_h)$ intervenant dans le problème discret (2.52). Par exemple, lorsque

$$\ell(v_h) = \int_\Omega f\, v_h d\Omega$$

on est amené à calculer des intégrales élémentaires sur chaque K_ℓ :

$$I_i^\ell(f) = \int_{K_\ell} f \tau_i^\ell d\Omega,$$

la fonction f étant donnée.

Lorque la fonction f est suffisamment régulière (continue par morceaux)[13], une technique classique consiste à approcher ces intégrales à l'aide de formules de quadrature, dites à nq_ℓ points :

$$I_i^\ell(f) \simeq \sum_{q=1}^{nq_\ell} \omega_q^\ell f\left(b_q^\ell\right) \tau_i^\ell\left(b_q^\ell\right)$$

où $\left(b_q^\ell\right)_{q=1,nq_\ell}$ sont des points de quadrature dans K_ℓ et $\left(\omega_q^\ell\right)_{q=1,nq_\ell}$ les poids positifs associés aux points de quadrature.

Il existe de nombreux types de quadrature (Gauss-Lobatto...), voir par exemple [14, 12], qui fournissent des approximations des intégrales :

[13] Attention aux fonctions singulières ou hyperoscillantes pour lequelles il convient d'utiliser des techniques de quadrature numérique particulières.

$$\int_{K_\ell} g \, d\Omega.$$

Ces méthodes de quadrature se différencient par le fait qu'elles intégrent exactement les polynômes d'un ordre donné k_{quad}, ce qui caractérise leur précision.

En dimension 2, on pourra utiliser, par exemple, la formule de quadrature à 7 points (de Gauss-Lobatto) sur le triangle de référence $\widehat{T}$, qui est d'ordre 5 :

$\widehat{b}_q$	(s_0, s_0)	(s_1, s_1)	(s_1, s_3)	(s_3, s_1)	(s_2, s_2)	(s_2, s_4)	(s_4, s_2)
$\widehat{\omega}_q$	η_0	η_1	η_1	η_1	η_2	η_2	η_2

avec $s_0 = \frac{1}{3}$, $s_1 = \frac{6-\sqrt{15}}{21}$, $s_2 = \frac{6+\sqrt{15}}{21}$, $s_3 = \frac{9+2\sqrt{15}}{21}$, $s_4 = \frac{9-2\sqrt{15}}{21}$, $\eta_0 = \frac{9}{80}$, $\eta_1 = \frac{155-\sqrt{15}}{2400}$, $\eta_2 = \frac{155+\sqrt{15}}{2400}$.

En dimension 1 (termes de bord), on utilisera, par exemple, la formule de quadrature à 3 points (de Gauss-Legendre) sur le segment $[-1, 1]$, qui est d'ordre 3 :

$\widehat{b}_q$	$-\sqrt{3/5}$	0	$\sqrt{3/5}$
$\widehat{\omega}_q$	5/9	8/9	5/9

L'utilisation de telles méthodes conduit à des problèmes discrétisés de la forme :

$$\begin{cases} \text{Trouver } \tilde{u}_h \in V_h, \text{ tel que} \\ a\left(\tilde{u}_h, v_h\right) = \tilde{\ell}_h(v_h) \qquad \forall v_h \in V_h \end{cases}$$

On demeure cette fois dans le cadre de l'approximation interne, mais une erreur supplémentaire est introduite pour l'évaluation approchée du second membre.

On trouvera dans [25] la démonstration du résultat suivant :

Proposition 2.8. *Supposons que la formule de quadrature utilisée intègre exactement les polynômes d'ordre k_{quad} avec $k_{quad} \geq 2k-2$, où k est l'ordre des éléments finis associés au maillage $\mathcal{T}_h$ de l'ouvert Ω. Alors, si $u \in \mathcal{C}^{k+2}\left(\overline{\Omega}\right)$:*

$$\|u - \tilde{u}_h\|_{H^1(\Omega)} \leq C h^k \|u\|_{\mathcal{C}^{k+2}\left(\overline{\Omega}\right)}.$$

Dans [21] on trouvera des résultats plus généraux.

En particulier, il ne faut pas surestimer la précision d'intégration car alors le calcul devient inutilement coûteux. Lorsque l'on utilise des éléments finis d'ordre 1, on observe ainsi qu'il suffit d'une formule de quadrature qui intègre exactement les constantes (i.e. $k_{quad} = 0$).

3

Aspects concrets de la méthode des éléments finis

Nous consacrons ce chapitre aux aspects pratiques de la méthode des éléments finis. Nous commençons par des questions de mise en œuvre informatique, puis nous poursuivons par des considérations algorithmiques, et nous concluons par quelques illustrations numériques obtenues à l'aide de Matlab.

3.1 Mise en œuvre

La réalisation d'un code d'éléments finis présente quelques particularités, liées, d'une part, aux propriétés des fonctions de base (support et aspect local) et d'autre part, à la formulation variationnelle, en particulier le traitement des conditions essentielles.

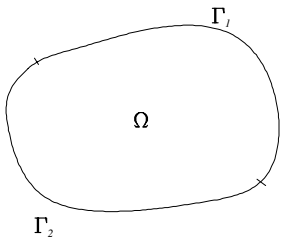
La résolution numérique d'un problème elliptique par une méthode d'éléments finis se déroule suivant 7 grandes étapes :

- la génération d'un maillage du domaine de calcul $\overline{\Omega}$,
- la fabrication des matrices élémentaires,
- l'assemblage des matrices globales,
- la constitution du second membre,
- l'élimination des conditions essentielles (s'il y en a !),
- la résolution du système linéaire,
- des post-traitements numériques et/ou graphiques.

Notons que la génération automatique de maillage, particulièrement ardue en dimension 3, est généralement du ressort de logiciels spécialisés. Nous n'abordons pas ici les principes de la génération et indiquons seulement quelles informations sur le maillage sont requises par une méthode d'éléments finis. En annexe, sont donnés des codes Matlab permettant de réaliser des maillages de domaines 2D élémentaires.

Au cours de cette section, nous allons décrire ces différentes étapes avec plus ou moins de détails dans le cas du problème mixte Dirichlet-Neumann suivant :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \dfrac{\partial u}{\partial n} = g & \text{sur } \Gamma_1 \\ u = \alpha & \text{sur } \Gamma_2 \end{cases} \qquad (3.1)$$



avec : $\Gamma_1 \cap \Gamma_2 = \emptyset$ (Γ_1 et Γ_2 sont des ouverts) et $\overline{\Gamma}_1 \cup \overline{\Gamma}_2 = \partial\Omega$, $f \in L^2(\Omega)$, $g \in L^2(\Gamma_1)$ et $\alpha \in L^2(\Gamma_2)$ telle qu'il existe un relèvement $\tilde{\alpha} \in H^1(\Omega)$ tel que $\tilde{\alpha} = \alpha$ sur Γ_2. On peut aussi écrire $\alpha \in H^{1/2}(\Gamma_2)$, comme au chapitre 1.

La formulation variationnelle du problème (3.1) dans l'espace :

$$H^1_{0,\Gamma_2}(\Omega) = \left\{ v \in H^1(\Omega),\ v_{|\Gamma_2} = 0 \right\}$$

est :

$$\left| \begin{array}{l} \text{Trouver } u = \tilde{u} + \tilde{\alpha} \text{ avec } \tilde{u} \in H^1_{0,\Gamma_2}(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla\tilde{u} \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega + \int_{\Gamma_1} g v \, d\Gamma - \int_\Omega \nabla\tilde{\alpha} \cdot \nabla v \, d\Omega \qquad \forall v \in H^1_{0,\Gamma_2(\Omega)} \end{array} \right. \qquad (3.2)$$

D'après le théorème 1.2, $H^1_{0,\Gamma_2}(\Omega)$ est un fermé de $H^1(\Omega)$. Dans la suite, on note Σ l'interface entre Γ_1 et Γ_2 : $\Sigma = \partial\Gamma_1 \cap \partial\Gamma_2$.

Nous nous restreindrons au cas des éléments finis de Lagrange.

3.1.1 Maillage pour les éléments finis

Un outil de maillage doit fournir essentiellement les informations suivantes :

- les coordonnées des degrés de liberté de Lagrange $(M_I)_{I=1,N}$,
- pour chaque élément fini $E_\ell = (K_\ell, \Sigma_\ell, P_\ell)$, $\ell = 1, L$:
  a) son type (triangle, quadrangle, cube, prisme...)
  b) son ordre (0,1,2...)
  c) la liste des degrés de liberté globaux attachés à cet élément dans un ordre établi à l'avance (par exemple pour un triangle P^2 en 2D : les 3 sommets puis les 3 nœuds intermédiaires)
  d) des numéros dits de référence pour les domaines K_ℓ, pour les faces, arêtes et sommets.

Les informations a) et b) permettent de connaître le nombre de degrés de liberté locaux n_ℓ, ainsi que les fonctions de base locales $\left(\tau_i^\ell\right)$, *via* la transformation géométrique F_ℓ construite à partir des coordonnées des degrés de liberté.

L'information c) est ni plus ni moins que la donnée de la fonction ℓ_g définie en (2.39).

L'information d) est facultative, elle permet néanmoins de repérer rapidement des objets spécifiques : des faces, des arêtes, des nœuds situés sur une frontière (traitement des conditions aux limites) ou encore des domaines situés dans une partie d'un domaine (traitement des coefficients discontinus par exemple). En particulier, il est souhaitable de connaître explicitement l'ensemble des nœuds de la frontière situés sur l'interface Σ.

La donnée des informations a), b) et c) constitue le maillage $\mathcal{T}_h : (K_\ell, \Sigma_\ell, P_\ell)_{\ell=1,L}$ qui doit évidemment satisfaire à toutes les hypothèses que nous avons énoncées au §2.2.2.

On notera par la suite : $\overline{\Omega}_h = \cup_{\ell=1,L} K_\ell$ (on se place ici dans le cas d'un ouvert Ω non nécessairement polyédrique). Pour découper la frontière $\partial\Omega_h$, on considère Γ_1^h (respectivement Γ_2^h) égal à l'ensemble des faces frontière[1] reliant des nœuds situés sur $\overline{\Gamma}_1$ (respectivement $\overline{\Gamma}_2$). Avec cette définition, Γ_1^h et Γ_2^h sont des fermés, d'intersection $\Sigma = \Gamma_1^h \cap \Gamma_2^h$, et $\partial\Omega_h = \Gamma_1^h \cup \Gamma_2^h$.

Par ailleurs, on introduit les ensembles d'indices :

$$\mathcal{N} = \left\{ I \text{ tel que } M_I \in \overset{\circ}{\Gamma_1^h} \right\} \qquad (N_1 = \text{card}\mathcal{N})$$

$$\mathcal{D} = \left\{ I \text{ tel que } M_I \in \Gamma_2^h \right\} \qquad (N_2 = \text{card}\mathcal{D})$$

$$\mathcal{I} = \{1, 2, ..., N\} \setminus \mathcal{D} \qquad (M = \text{card}\mathcal{I})$$

qui se construisent, dans la pratique, à partir des informations d). L'ensemble d'indices $\mathcal{N}$ repère les nœuds appartenant à la frontière Γ_1^h qui porte la condition de Neumann. En réalité, comme on a à calculer une intégrale sur cette frontière, le repérage des faces appartenant à cette frontière est suffisant. Bien évidemment, la connaissance des nœuds situés sur la frontière Γ_1^h permet de retrouver les faces situées sur Γ_1^h. L'ensemble d'indice $\mathcal{D}$ repère les nœuds de la frontière Γ_1^h portant la condition de Dirichlet et on introduit l'ensemble d'indices $\mathcal{I}$, représentant l'ensemble des nœuds qui ne portent pas une condition de Dirichlet, car il corres-

[1] Pour repérer rapidement les frontières Γ_1^h et Γ_2^h (maillées, si Ω n'est pas polyédrique), commençons par l'observation suivante. Chaque domaine K_ℓ possède une frontière constituée d'un ensemble de faces. Parmi ces faces, celles qui font partie de la frontière $\partial\Omega_h$ sont exactement celles qui n'appartiennent qu'à un unique domaine K_ℓ : on les appelle faces frontière. Au contraire, toute face interne appartient exactement à deux domaines K_{ℓ_1} et K_{ℓ_2}, avec $\ell_1 \neq \ell_2$.

pond à l'ensemble des fonctions de base globales servant à engendrer l'espace d'approximation dans lequel la condition de Dirichlet est prise en compte.

Remarque 3.1. On inclut *a priori* les nœuds de l'interface Σ dans la partie de la frontière avec une conditions aux limites de type Dirichlet.

Exemple : maillage constitué de deux éléments finis P^1-Lagrange

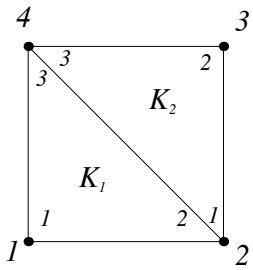


Fig. 3.1. Maillage à deux triangles

Dans cet exemple, on construit informatiquement les tableaux :

de coordonnées des sommets :

coor(1,1)=0	coor(1,2)=0
coor(2,1)=1	coor(2,2)=0
coor(3,1)=1	coor(3,2)=1
coor(4,1)=0	coor(4,2)=1

de numérotation locale→globale :

lg(1,1)=1	lg(2,1)=2
lg(1,2)=2	lg(2,2)=3
lg(1,3)=4	lg(2,3)=4

Il est aisé de réaliser des programmes de maillage d'objets simples tels que des rectangles, des cercles, des cubes, des sphères ou encore des transformations géométriques de tels objets. Pour mailler des objets plus complexes, comme ceux que l'on rencontre dans l'industrie, on utilise des "mailleurs" généraux qui, à partir de quelques points de la surface de l'objet[2], réalisent un maillage à l'aide d'algorithmes sophistiqués, dont la description sort du cadre de ce cours. Malheureusement, sur le marché actuel, peu de "mailleurs" sont capables de générer automatiquement des maillages **isoparamétriques** d'ordre k ($k \geq 2$).

Nous sommes maintenant en mesure de discrétiser le problème (3.2). Introduisons les espaces :

[2] Dans notre cadre "abstrait", ce seront des points situés sur la frontière $\partial\Omega$.

$$V_h = Vect\,(w_I)_{I=1,N}$$

$$V_h^0 = Vect\,(w_I)_{I\in\mathcal{I}}$$

et la formulation variationnelle discrète dans V_h^0 :

$$\left|\begin{array}{l}\text{Trouver } u_h = \tilde{u}_h + \tilde{\alpha} \text{ avec } \tilde{u}_h \in V_h^0 \text{ tel que :}\\ \displaystyle\int_{\Omega_h} \nabla\tilde{u}_h\cdot\nabla\tilde{v}_h d\Omega = \int_{\Omega_h} f\tilde{v}_h d\Omega + \int_{\Gamma_1^h} g\tilde{v}_h d\Gamma - \int_{\Omega_h}\nabla\tilde{\alpha}\cdot\nabla\tilde{v}_h\,d\Omega \qquad \forall\tilde{v}_h\in V_h^0.\end{array}\right. \tag{3.3}$$

D'après le lemme 2.1, cette formulation variationnelle est équivalente au système linéaire d'ordre M :

$$\boxed{\mathbb{K}_{\mathcal{I}\mathcal{I}}\overrightarrow{U}_{\mathcal{I}} = \overrightarrow{S}_{\mathcal{I}}} \tag{3.4}$$

avec :

$$\begin{aligned}(\mathbb{K}_{\mathcal{I}\mathcal{I}})_{IJ} &= \int_{\Omega_h}\nabla w_I\cdot\nabla w_J d\Omega \qquad I,J\in\mathcal{I},\\ (S_{\mathcal{I}})_I &= \int_{\Omega_h} f\,w_I d\Omega + \int_{\Gamma_1^h} g\,w_I d\Gamma - \int_{\Omega_h}\nabla\tilde{\alpha}\cdot\nabla w_I d\Omega \qquad I\in\mathcal{I},\end{aligned}$$

et on a :

$$u_h = \sum_{I\in\mathcal{I}}(U_{\mathcal{I}})_I w_I + \tilde{\alpha}. \tag{3.5}$$

3.1.2 Calculs élémentaires

On a vu qu'il est préférable de calculer sur chaque domaine K_ℓ les matrices élémentaires (cf. (2.26) et (2.27)) :

$$\left.\begin{aligned}\mathbb{K}_{ij}^\ell &= \int_{K_\ell}\nabla\tau_i^\ell\cdot\nabla\tau_j^\ell d\Omega\\ \mathbb{M}_{ij}^\ell &= \int_{K_\ell}\tau_i^\ell\tau_j^\ell d\Omega\end{aligned}\right. \qquad \forall i,j=1,n_\ell.$$

Les matrices élémentaires $\mathbb{K}^\ell$ servent à la construction de la matrice (globale) $\mathbb{K}_{\mathcal{I}\mathcal{I}}$ et les matrices élémentaires $\mathbb{M}^\ell$ à la fabrication du second membre (une façon particulière de constituer une approximation du second membre).

Rappelons que l'élément fini $(K_\ell,\Sigma_\ell,P_\ell)$ est lié à l'élément fini de référence $\left(\widehat{K},\widehat{\Sigma},\widehat{P}\right)$ par la transformation géométrique F_ℓ qui envoie $\widehat{K}$ dans K_ℓ. En effectuant le changement de variable $M = F_\ell\left(\widehat{M}\right)$, les intégrales précédentes deviennent des intégrales sur $\widehat{K}$. On trouve (cf. (2.28) et (2.29)) :

$$\mathbb{K}_{ij}^{\ell} = \int_{\widehat{K}} (\left\{ [dF_\ell(\widehat{M})]^{-1} ([dF_\ell(\widehat{M})]^{-1})^t \right\} \nabla\widehat{\tau}_i(\widehat{M})) \cdot \widehat{\nabla}\widehat{\tau}_j(\widehat{M}) \, |J_{F_\ell}(\widehat{M})| \, d\widehat{\Omega},$$
$$\mathbb{M}_{ij}^{\ell} = \int_{\widehat{K}} \widehat{\tau}_i(\widehat{M})\widehat{\tau}_j(\widehat{M}) \, |J_{F_\ell}(\widehat{M})| \, d\widehat{\Omega},$$

qui n'utilisent que la connaissance des fonctions de base sur l'élément fini de référence.

Ces intégrales sont :

- soit calculées analytiquement et de façon exacte (cf. (2.70)) dans les cas simples (éléments finis de bas degré) ;
- soit évaluées (coefficients des matrices éléments finis) ou estimées (calcul d'intégrands éléments finis présentant un coefficient variable ou calcul du second membre) à l'aide de formules de quadrature numérique ; rappelons que dans le cas du calcul du second membre, il suffit d'intégrer exactement les polynômes de degré $2k-2$ si l'élément fini est d'ordre k (cf. proposition 2.8).

Notons que les points de quadrature $((\widehat{x}_q^{\ell})_{q=1,nq_\ell})$ et leurs poids $((\omega_q^{\ell})_{q=1,nq_\ell})$ sont définis une fois pour toute sur le domaine de référence $\widehat{K}$.

Lorsque l'on utilise une formule de quadrature, pour chaque domaine K_ℓ on effectue l'algorithme suivant :

$\mathbb{K}^{\ell} = 0$
pour $q = 1, nq_\ell$ *boucle points de quadrature*
  $DF = [dF_\ell(\widehat{x}_q^{\ell})]$ *calcul de la matrice jacobienne en* $\widehat{x}_q^{\ell}$
  $J = |\det DF|$ *calcul du jacobien en* $\widehat{x}_q^{\ell}$
  $\mathrm{C} = J\,(DF^{-1}\,(DF^{-1})^t)$
  pour $i, j = 1, n_\ell$ *boucle degrés de liberté*
    $\mathbb{K}_{ij}^{\ell} = \mathbb{K}_{ij}^{\ell} + \omega_q^{\ell}(\mathrm{C}\widehat{\nabla}\widehat{\tau}_i(\widehat{x}_q^{\ell})) \cdot \widehat{\nabla}\widehat{\tau}_j(\widehat{x}_q^{\ell})$ *"calcul" de l'intégrale*
  fin
fin

Remarque 3.2. Lors de ces calculs de matrices élémentaires, on constate que le choix d'éléments finis isoparamétriques se traduit par la présence explicite de la transformation géométrique F_ℓ.

3.1.3 Assemblage des matrices globales et du second membre

Pour construire le système linéaire (3.4), il faut fabriquer, d'une part, la matrice (globale) $\mathbb{K}_{\mathcal{I}\mathcal{I}}$ à partir des matrices élémentaires $\mathbb{K}^{\ell}$ et d'autre part, le vecteur second membre $\overrightarrow{S}_{\mathcal{I}}$.

• Assemblage des matrices

Pour des raisons qui apparaîtront ultérieurement (prise en compte des conditions aux limites de Dirichlet), il est préférable et plus simple d'assembler, dans un premier temps, la matrice $\mathbb{K}$ d'ordre N s'appuyant sur tous les degrés de liberté du maillage, y compris les degrés de liberté situés sur Γ_2^h (interface Σ comprise). Il est intéressant de noter que $\mathbb{K}$ "contient" $\mathbb{K}_{\mathcal{I}\mathcal{I}}$. Dans un deuxième temps, on "extraira" la matrice $\mathbb{K}_{\mathcal{I}\mathcal{I}}$.

Rappelons que par définition des fonctions de base globales (w_I) on a :

$$w_I = \tau_i^\ell \quad \text{sur} \quad K_\ell \text{ si } \exists i \text{ t.q. } \ell_g(\ell, i) = I,$$

l'indice i, si il existe, étant unique. On introduit l'ensemble :

$$\mathcal{C}_{IJ} = \{\ell \text{ tel qu'il existe } i_\ell, j_\ell \text{ tels que } \ell_g(\ell, i_\ell) = I \text{ et } \ell_g(\ell, j_\ell) = J\}$$

qui représente l'ensemble des éléments du maillage qui possèdent *simultanément* les nœuds M_I et M_J pour degrés de liberté.

On a alors :

$$\mathbb{K}_{IJ} = \int_{\Omega_h} \nabla w_I \cdot \nabla w_J d\Omega = \sum_{\ell \in \mathcal{C}_{IJ}} \int_{K_\ell} \nabla \tau_{i_\ell}^\ell \cdot \nabla \tau_{j_\ell}^\ell d\Omega = \sum_{\ell \in \mathcal{C}_{IJ}} \mathbb{K}_{i_\ell j_\ell}^\ell,$$

et de la même façon,

$$\mathbb{M}_{IJ} = \int_{\Omega_h} w_I w_J d\Omega = \sum_{\ell \in \mathcal{C}_{IJ}} \int_{K_\ell} \tau_{i_\ell}^\ell \tau_{j_\ell}^\ell d\Omega = \sum_{\ell \in \mathcal{C}_{IJ}} \mathbb{M}_{i_\ell j_\ell}^\ell.$$

On ne construit jamais explicitement les ensembles $(\mathcal{C}_{IJ})_{I,J=1,N}$ car cela demande "d'inverser" l'application ℓ_g par rapport à son second argument i, ce qui est très coûteux d'un point de vue algorithmique ou informatique. C'est pourquoi, on procède dans l'autre sens. En effet, on peut facilement vérifier que chaque terme élémentaire $\mathbb{K}_{ij}^\ell$ intervient exactement *une fois et une seule* dans la construction de la matrice globale $\mathbb{K}$ (et de même pour la construction de $\mathbb{M}$), puisque se donnant (i, j) dans un domaine K_ℓ, on connaît explicitement la position du terme élémentaire $\mathbb{K}_{ij}^\ell$ dans la matrice globale : $(I, J) = (\ell_g(\ell, i), \ell_g(\ell, j))$.

Après initialisation de la matrice $\mathbb{K}$ à 0, on effectuera donc l'opération :

$$\boxed{\mathbb{K}_{\ell_g(\ell,i),\ell(_g\ell,j)} = \mathbb{K}_{\ell_g(\ell,i),\ell_g(\ell,j)} + \mathbb{K}_{ij}^\ell} \qquad \begin{array}{l} \ell = 1, L \\ i, j = 1, n_\ell \end{array}$$

qui constitue la phase d'assemblage de la matrice globale $\mathbb{K}$.

Dans la pratique, on réalise simultanément le calcul des matrices élémentaires et l'assemblage, ce qui conduit à l'algorithme d'assemblage suivant :

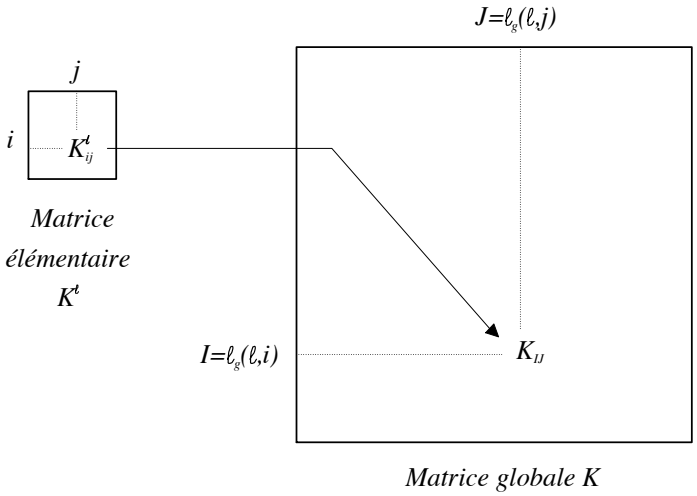


Fig. 3.2. Principe d'assemblage

$\mathbb{K} = 0,\ \mathbb{M} = 0$ *initialisation*
pour $\ell = 1, L$ *boucle sur les domaines*
  calcul des matrices élémentaires $\mathbb{K}^\ell$, $\mathbb{M}^\ell$
  pour $i, j = 1, n_\ell$ *boucle sur les dl locaux*
    $I = \ell_g(\ell, i),\ J = \ell_g(\ell, j)$ *n° global*
    $\mathbb{K}_{IJ} = \mathbb{K}_{IJ} + \mathbb{K}^\ell_{ij}$ *assemblage*
    $\mathbb{M}_{IJ} = \mathbb{M}_{IJ} + \mathbb{M}^\ell_{ij}$
  fin
fin

• Constitution du second membre

Le second membre de la formulation variationnelle (3.3) est constitué de trois termes :

- le terme volumique : $\displaystyle\int_{\Omega_h} f w_I d\Omega$,
- le terme surfacique: $\displaystyle\int_{\Gamma_1^h} g w_I d\Gamma$, pour $I \in \mathcal{N}$,
- le terme volumique : $\displaystyle\int_{\Omega} \nabla\tilde{\alpha} \cdot \nabla w_I d\Omega$

Il y a essentiellement deux méthodes de calcul de ces termes.

• *Méthode d'interpolation*

Si les fonctions f, g et $\tilde{\alpha}$ sont régulières ($\in \mathcal{C}^0(\overline{\Omega})$), on peut les approcher par leurs interpolées respectives (cf. (2.54)) $\pi_h f$, $\pi_h g$ et $\pi_h \tilde{\alpha}$, c'est-à-dire :

$$\pi_h f = \sum_{I=1,N} f(M_I)\, w_I$$
$$\pi_h g = \sum_{I \in \mathcal{N}} g(M_I)\, w_I$$
$$\pi_h \tilde{\alpha} = \sum_{I=1,N} \tilde{\alpha}(M_I)\, w_I$$

Les intégrales de volume sont alors approchées de la façon suivante, pour $I = 1, N$:

$$\int_{\Omega_h} f\, w_I d\Omega \simeq \int_{\Omega_h} \pi_h f\, w_I d\Omega \left(= \sum_{J=1,N} f(M_J) \int_{\Omega_h} w_I w_J d\Omega \right)$$
$$\int_{\Omega_h} \nabla\tilde{\alpha} \cdot \nabla w_I d\Omega \simeq \int_{\Omega_h} \nabla(\pi_h\tilde{\alpha}) \cdot \nabla w_I d\Omega \left(= \sum_{J=1,N} \tilde{\alpha}(M_J) \int_{\Omega_h} \nabla w_I \cdot \nabla w_J d\Omega \right)$$

et l'intégrale de surface :

$$\int_{\Gamma_1^h} g\, w_I d\Gamma \simeq \int_{\Gamma_1^h} \pi_h g\, w_I d\Gamma \left(= \sum_{J \in \mathcal{N}} g(M_J) \int_{\Gamma_1^h} w_I w_J d\Gamma \right) \qquad \forall I \in \mathcal{N}.$$

Cette approximation conduit à l'expression suivante du second membre apparaissant dans (3.4) :

$$\boxed{\overrightarrow{S} = \mathbb{M}\overrightarrow{F} + P_{\mathcal{N}}\left(\mathbb{M}_{\mathcal{N}}\overrightarrow{G}\right) - \mathbb{K}\overrightarrow{\alpha}} \tag{3.6}$$

avec :

- $\overrightarrow{F}$ le vecteur de $\mathbb{R}^N$ de composantes $(f(M_I))_I$,
- $\overrightarrow{G}$ le vecteur de $\mathbb{R}^{N_1}$ de composantes $(g(M_I))_I$,
- $\overrightarrow{\alpha}$ le vecteur de $\mathbb{R}^N$ de composantes $(\tilde{\alpha}(M_I))_I$,
- $\mathbb{M}_{\mathcal{N}}$ la matrice carrée d'ordre N_1 définie par :

$$(\mathbb{M}_{\mathcal{N}})_{IJ} = \int_{\Gamma_1^h} w_I w_J d\Gamma \qquad I, J \in \mathcal{N},$$

  dite matrice de masse de frontière,
- $P_{\mathcal{N}}$ l'opérateur de prolongement canonique de $\mathbb{R}^{N_1}$ dans $\mathbb{R}^N$.

On construit la matrice de frontière par le processus d'assemblage que nous avons décrit précédemment, en prenant soin de ne conserver que les degrés de liberté situés sur la frontière $\overset{\circ}{\Gamma}{}_1^h$.

Dans cette technique, comme pour passer de $\mathbb{K}$ à $\mathbb{K}_{\mathcal{I}\mathcal{I}}$, il est nécessaire d'extraire le second membre réel $\overrightarrow{S}_{\mathcal{I}}$ à partir du vecteur $\overrightarrow{S}$ en éliminant les termes correspondant à des nœuds situés sur Γ_2^h (incluant encore une fois ceux de l'interface Σ).

• *Méthode de quadrature*

L'autre technique consiste à fabriquer le second membre au niveau local à l'aide de formules de quadrature. En effet, on a :

$$\int_{\Omega_h} f\, w_I d\Omega = \sum_{\ell \in \mathcal{C}_{II}} \int_{K_\ell} f\, \tau_{i_\ell}^\ell d\Omega$$

On se ramène à l'élément de référence à l'aide de la formule :

$$F_i^\ell = \int_{K_\ell} f\, \tau_i^\ell d\Omega = \int_{\widehat{K}} f\left(F_\ell\left(\widehat{x}\right)\right) \widehat{\tau}_i \left|J_{F_\ell}\left(\widehat{x}\right)\right| d\widehat{\Omega},$$

et on calcule cette intégrale à l'aide des points de quadrature $(\widehat{x}_q^\ell)_{q=1,nq_\ell}$ avec les poids $(\widehat{\omega}_q^\ell)_{q=1,nq_\ell}$.

En effectuant l'opération d'assemblage suivante, après initialisation à 0 :

$$\boxed{(S_f)_{\ell_g(\ell,i)} = (S_f)_{\ell_g(\ell,i)} + F_i^\ell} \qquad \begin{array}{l} \ell = 1, L \\ i = 1, n_\ell \end{array}$$

on construit ainsi la contribution volumique $\overrightarrow{S}_f$ au second membre. On construit, de même, les autres contributions $\overrightarrow{S}_g$ et $\overrightarrow{S}_\alpha$.

Comme on l'a indiqué lors de l'étude d'erreur, il faut que les formules de quadrature utilisées intègrent exactement les polynômes d'ordre $k_{quad} \geq 2k-2$.

Remarque 3.3. Pour cette seconde technique, on peut choisir de calculer $\overrightarrow{S}_{\mathcal{I}}$ directement, ou bien choisir de calculer $\overrightarrow{S}$, puis extraire $\overrightarrow{S}_{\mathcal{I}}$. La seconde approche n'introduit pas de manipulations supplémentaires, puisqu'on doit de toute façon extraire $\mathbb{K}_{\mathcal{I}\mathcal{I}}$ de $\mathbb{K}$.

Ces deux techniques de calcul sont concurrentes. La première occupe plus de place mémoire dès lors que l'on n'a pas besoin de la matrice de masse par ailleurs. Par contre, la seconde requiert plus de calculs. Lorsque les données f ou g présentent des singularités (discontinuités, ou caractère non borné), il est préférable d'utiliser la seconde méthode en adaptant les formules de quadrature à la nature des singularités.

Remarque 3.4. Si l'on utilise pour approcher la matrice de masse $\mathbb{M}$ la formule de quadrature de Lobatto :

$$\int_{K_\ell} v d\Omega \simeq \frac{mes(K_\ell)}{n+1} \sum_{i=1,n+1} v(\bar{M}_i^\ell), \qquad (v \text{ fonction régulière sur } K_\ell)$$

où $\left(\bar{M}_i^\ell\right)_{i=1,n+1}$ sont les sommets du segment ($n=1$), du triangle ($n=2$) ou du tétraèdre ($n=3$), on obtient une matrice approchée $\widetilde{\mathbb{M}}$ diagonale.

Par ailleurs, on peut également construire une autre approximation $\widehat{\mathbb{M}}$ de la matrice de masse $\mathbb{M}$ en utilisant une technique, appelée "lumping" :

$$\widehat{\mathbb{M}}_{II} = \sum_{J=1}^{N} \mathbb{M}_{IJ} \text{ et } \widehat{\mathbb{M}}_{IJ} = 0 \text{ si } I \neq J$$

qui conduit également à une matrice diagonale.
Ces deux techniques permettent un gain substantiel de place mémoire et de temps calcul et on montre qu'elles ne détériorent pas les estimations d'erreurs pour les éléments finis de Lagrange d'ordre 1 [12].

3.1.4 Elimination des conditions essentielles

Rappelons que les conditions aux limites essentielles sont celles qui interviennent dans l'espace de la formulation variationnelle du problème E.D.P. que l'on traite. Dans le cas du problème (3.1), on a à traiter la condition de Dirichlet :

$$u = \alpha \text{ sur } \Gamma_2.$$

Comme on a pu l'observer, au risque de perdre toute l'efficacité du processus d'assemblage des matrices et du second membre, il n'est pas possible d'éliminer à ce stade les degrés de liberté de Dirichlet. Il est donc nécessaire, afin de constituer le système linéaire (3.4), d'extraire la matrice $\mathbb{K}_{\mathcal{I}\mathcal{I}}$ de la matrice $\mathbb{K}$ et le second membre $S_{\mathcal{I}}$ du vecteur $\overrightarrow{S}$ après que l'assemblage a été réalisé.

• Technique d'élimination réelle

Cette opération consiste à éliminer les lignes et colonnes d'indices $I, J \in \mathcal{D}$ dans la matrice $\mathbb{K}$ ainsi que les lignes d'indices $I \in \mathcal{D}$ dans le vecteur $\overrightarrow{S}$. Dans la pratique, on construit une nouvelle numérotation des degrés de liberté n'appartenant pas à Γ_2^h (rappel : $\Sigma \subset \Gamma_2^h$). On utilise un pointeur $\mathcal{P}_{\mathcal{I}} : i \mapsto I$ qui associe au numéro i (compris entre 1 et M=card$\mathcal{I}$) son ancien numéro I compris entre 1 et N. On a alors :

$$(\mathbb{K}_{\mathcal{I}\mathcal{I}})_{ij} = \mathbb{K}_{\mathcal{P}_{\mathcal{I}}(i)\mathcal{P}_{\mathcal{I}}(j)} \qquad i,j = 1,M$$

$$(S_{\mathcal{I}})_i = S_{\mathcal{P}_{\mathcal{I}}(i)} \qquad i = 1,M$$

Cette technique présente des inconvénients car les matrices éléments finis ont des structures très particulières de stockage dues au fait qu'elles sont creuses. Il s'avère compliqué et coûteux d'effectuer cette élimination des lignes et colonnes des matrices éléments finis. C'est pourquoi, il est souvent préférable d'utiliser la technique de pseudo-élimination que nous allons présenter maintenant.

• Technique de pseudo-élimination

Nous traitons le cas où le second membre est obtenu par interpolation, son expression étant alors donnée par (3.6).

Partitionnons la matrice $\mathbb{K}$ et le second membre $\overrightarrow{S}$ – par blocs – suivant $\mathcal{I}$ et $\mathcal{D}$ (avec $N = M + N_2$ et card$\mathcal{D} = N_2$) :

$$\mathbb{K} = \begin{bmatrix} \mathbb{K}_{\mathcal{I}\mathcal{I}} & \mathbb{K}_{\mathcal{I}\mathcal{D}} \\ \mathbb{K}_{\mathcal{D}\mathcal{I}} & \mathbb{K}_{\mathcal{D}\mathcal{D}} \end{bmatrix} \qquad \overrightarrow{S} = \begin{pmatrix} \overrightarrow{S}_{\mathcal{I}} \\ \overrightarrow{S}_{\mathcal{D}} \end{pmatrix}.$$

On procède de même pour toute matrice $N \times N$, et tout vecteur de $\mathbb{R}^N$.
Posons :

$$\overrightarrow{T} = \mathbb{M}\overrightarrow{F} + P_{\mathcal{N}}\left(\mathbb{M}_{\mathcal{N}}\overrightarrow{G}\right)$$

On a alors, en vertu de (3.6) :

$$\overrightarrow{S} = \overrightarrow{T} - \mathbb{K}\overrightarrow{\alpha} = \begin{pmatrix} \overrightarrow{T}_{\mathcal{I}} \\ \overrightarrow{T}_{\mathcal{D}} \end{pmatrix} - \begin{pmatrix} \mathbb{K}_{\mathcal{I}\mathcal{I}}\overrightarrow{\alpha}_{\mathcal{I}} + \mathbb{K}_{\mathcal{I}\mathcal{D}}\overrightarrow{\alpha}_{\mathcal{D}} \\ \mathbb{K}_{\mathcal{D}\mathcal{I}}\overrightarrow{\alpha}_{\mathcal{I}} + \mathbb{K}_{\mathcal{D}\mathcal{D}}\overrightarrow{\alpha}_{\mathcal{D}} \end{pmatrix}$$

d'où on déduit que :

$$\overrightarrow{S}_{\mathcal{I}} = \overrightarrow{T}_{\mathcal{I}} - \mathbb{K}_{\mathcal{I}\mathcal{I}}\overrightarrow{\alpha}_{\mathcal{I}} - \mathbb{K}_{\mathcal{I}\mathcal{D}}\overrightarrow{\alpha}_{\mathcal{D}}.$$

Introduisons maintenant la matrice diagonale d'ordre N_2 :

$$\mathbb{D} = \text{Diag}\,(\mathbb{K}_{\mathcal{D}\mathcal{D}})$$

qui est inversible car les éléments diagonaux de $\mathbb{K}$ sont strictement positifs.

Le vecteur de $\mathbb{R}^N$:

$$\overrightarrow{U} = \begin{pmatrix} \overrightarrow{U}_{\mathcal{I}} \\ \overrightarrow{U}_{\mathcal{D}} \end{pmatrix} \qquad \text{avec } \overrightarrow{U}_{\mathcal{I}} \text{ la solution de (3.4) et } \overrightarrow{U}_{\mathcal{D}} = \overrightarrow{\alpha}_{\mathcal{D}},$$

vérifie, compte-tenu de ce qui précède :

$$\begin{bmatrix} \mathbb{K}_{\mathcal{I}\mathcal{I}} & 0 \\ 0 & \mathbb{D} \end{bmatrix} \begin{pmatrix} \overrightarrow{U}_{\mathcal{I}} \\ \overrightarrow{U}_{\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \overrightarrow{S}_{\mathcal{I}} \\ \mathbb{D}\overrightarrow{\alpha}_{\mathcal{D}} \end{pmatrix}. \tag{3.7}$$

Par définition de $\mathcal{D}$ et des fonctions de base w_I des éléments finis de Lagrange, on a :

$$(\alpha_{\mathcal{D}})_I = \alpha(M_I) \quad \forall I \in \mathcal{D},$$

ce qui montre que le vecteur $\overrightarrow{U}_{\mathcal{D}}$ est le vecteur des valeurs nodales de la donnée de Dirichlet $\tilde{\alpha}$ sur la frontière $\overline{\Gamma}_2$.

La technique de pseudo-élimination consiste donc à substituer au système linéaire (3.4) d'ordre M, le système linéaire (3.7) d'ordre N. Ceci ne nécessite aucune renumérotation et par conséquent aucune modification des structures de stockage de la matrice $\mathbb{K}$ et du vecteur $\overrightarrow{S}$. Seule une nouvelle affectation de certaines composantes est opérée.

Avant de décrire cette opération de façon détaillée, nous allons introduire une nouvelle simplification.

• Choix du relèvement

Nous allons choisir le relèvement suivant :

$$\alpha_h = \sum_{I \in \mathcal{D}} \alpha(M_I) w_I \tag{3.8}$$

qui appartient à $H^1(\Omega_h)$ car $w_I \in H^1(\Omega_h)$ et qui vérifie :

$$\alpha_h(M_I) = \begin{cases} \alpha(M_I) & \text{si } I \in \mathcal{D} \\ 0 & \text{si } I \notin \mathcal{D} \quad \text{(i.e. } I \in \mathcal{I}) \end{cases} \tag{3.9}$$

Ce relèvement dépend de l'approximation, mais cela n'a aucune importance. En effet, la formulation variationnelle (3.2) est équivalente à la formulation suivante :

$$\left| \begin{array}{l} \text{trouver } u \in H^1(\Omega) \text{ tel que :} \\ \quad \displaystyle\int_\Omega \nabla u \cdot \nabla v d\Omega = \int_\Omega f v d\Omega + \int_{\Gamma_1} g v d\Gamma \qquad \forall v \in H^1_{0,\Gamma_2}(\Omega) \\ \quad u = \alpha \quad \text{sur } \Gamma_2 \end{array} \right.$$

qui se discrétise sous la forme :

$$\left| \begin{array}{l} \text{trouver } u_h \in V_h \text{ tel que :} \\ \quad \displaystyle\int_{\Omega_h} \nabla u_h \cdot \nabla \tilde{v}_h d\Omega = \int_{\Omega_h} f \tilde{v}_h d\Omega + \int_{\Gamma_1^h} g \tilde{v}_h d\Gamma \qquad \forall \tilde{v}_h \in V_h^0 \\ \quad u_h(M_I) = \alpha(M_I) \quad \forall I \in \mathcal{D} \end{array} \right.$$

montrant ainsi que le relèvement peut-être choisi sous la forme (3.8).

Pour le choix (3.9), on déduit immédiatement que le bloc supérieur de $\overrightarrow{\alpha}$ est égal à :

$$\overrightarrow{\alpha}_{\mathcal{I}} = 0$$

d'où l'on déduit que :

$$\overrightarrow{S}_{\mathcal{I}} = \overrightarrow{T}_{\mathcal{I}} - \mathbb{K}_{\mathcal{I}\mathcal{D}}\overrightarrow{\alpha}_{\mathcal{D}} = \left(\overrightarrow{T} - \mathbb{K}\overrightarrow{\alpha}\right)_{\mathcal{I}} \quad \text{avec } \overrightarrow{\alpha} = \begin{pmatrix} 0 \\ \overrightarrow{\alpha}_{\mathcal{D}} \end{pmatrix}.$$

Il n'est donc pas nécessaire de partitionner effectivement la matrice $\mathbb{K}$ pour calculer le second membre !

D'après (3.5) et l'interpolation de la donnée de Dirichlet, la solution discrète u_h est donnée par :

$$u_h = \sum_{I\in\mathcal{I}} (U_{\mathcal{I}})_I w_I + \alpha_h = \sum_{I\in\mathcal{I}} (U_{\mathcal{I}})_I w_I + \sum_{I\in\mathcal{D}} \alpha(M_I) w_I \quad (\text{d'après (3.8)})$$

qui montre que :

$$u_h = \sum_{I=1,N} U_I w_I \quad \text{avec } U_I = \begin{cases} (U_{\mathcal{I}})_I & \text{si } I \in \mathcal{I} \\ \alpha(M_I) & \text{si } I \in \mathcal{D} \end{cases}$$

et finalement que la solution du système linéaire (3.7) est bien le vecteur $\overrightarrow{U}$, dont les composantes sont les coordonnées de la solution discrète u_h dans la base éléments finis.

• Réalisation pratique de la pseudo-élimination

Afin de réaliser cette pseudo-élimination, on peut, soit construire directement le pointeur de numérotation des degrés de liberté de Dirichlet :

$$\begin{array}{rccl} \mathcal{P}_{\mathcal{D}} : & \{1,\cdots,N_2\} & \to & \{1,\cdots,N\} \\ & i & \mapsto & I \ : \begin{array}{l}\text{numéro du } i^{ème} \text{ degré} \\ \text{de liberté de Dirichlet}\end{array} \end{array}$$

soit, utiliser un test permettant de savoir si un degré de liberté est un degré de liberté de Dirichlet ou non, par exemple en référençant tous les degrés de liberté, avec un numéro spécifique pour ceux de type Dirichlet.

Remarque 3.5. Si on oublie de prendre en compte les conditions de Dirichlet, on aboutit à un système linéaire non inversible car on se souvient que le problème de Neumann pur est mal posé (cf. §1.4.3); la matrice $\mathbb{K}$ ayant pour noyau l'espace vectoriel engendré par le vecteur $\overrightarrow{1}$ de composantes 1 (fonctions constantes de V_h). Signalons que la propriété $\mathbb{K}\overrightarrow{1} = \overrightarrow{0}$ peut être exploitée comme élément de validation du calcul de la matrice $\mathbb{K}$. Dans le même état d'esprit, le calcul $(\mathbb{M}\overrightarrow{1}\,|\,\overrightarrow{1})$ doit donner l'aire ou le volume du domaine de calcul car

$$\text{mes}(\Omega_h) = \int_{\Omega_h} d\Omega = \int_{\Omega_h} \sum_I w_I \sum_J w_J d\Omega = (\mathbb{M}\overrightarrow{1}\,|\,\overrightarrow{1}).$$

Dans le cas où on utilise le pointeur de Dirichlet $\mathcal{P}_\mathcal{D}$ on réalise un algorithme de pseudo-élimination du type suivant :

$\overrightarrow{\alpha} = \begin{pmatrix} 0 \\ \overrightarrow{\alpha}_\mathcal{D} \end{pmatrix}$ *donnée de Dirichlet*
$\overrightarrow{S} = \overrightarrow{T} - \mathbb{K}\overrightarrow{\alpha}$ *calcul du second membre*
pour $i = 1, N_2$ *boucle sur les noeuds Dirichlet*
  $I = \mathcal{P}_\mathcal{D}(i)$
  $\left(\overrightarrow{S}\right)_I = \mathbb{K}_{II}\alpha(M_I)$
  pour $J = 1, N$ *boucle sur tous les dl*
    si $J \neq I$ **alors**
      $\mathbb{K}_{IJ} = 0$ et $\mathbb{K}_{JI} = 0$ *"élimination"*
    fin
  fin
fin

Il existe d'autres procédés de prise en compte des conditions de Dirichlet. Ainsi, dans de nombreux codes de calcul, on utilise la technique qui consiste à rendre "très grands" les coefficients diagonaux $\mathbb{K}_{II}$ pour des indices I correspondants à des nœuds Dirichlet. Cette technique présente l'avantage d'être facile à implémenter et rapide. Elle présente néanmoins deux défauts : d'une part, il faut définir ce que veut dire "très grand" et, d'autre part, la présence de très grand coefficients détériore le conditionnement de la matrice. Ce qui dans certaines situations peut se révéler ennuyeux. Il est également possible de traiter la condition de Dirichlet via une approche de type pénalisation-régularisation, en introduisant la condition limite "approchée" :

$$u + \varepsilon \frac{\partial u}{\partial n} = \alpha \text{ sur } \Gamma_2.$$

On peut montrer que lorsque $\varepsilon \to 0$, la solution u_ε obtenue tend en norme H^1 vers la solution du problème initial (convergence en $\sqrt{\varepsilon}$). Cette approche ne nécessite plus aucun traitement algébrique particulier des matrices mais présente des défauts de même nature que ceux de l'approche précédente : choix de ε et déconditionnement du système linéaire.

3.2 Considérations algorithmiques

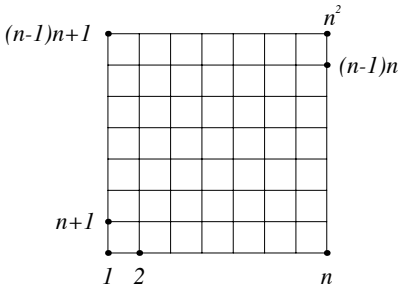
Outre le fait que les méthodes d'éléments finis permettent d'approcher les solutions faibles avec une grande généralité, elles présentent l'avantage d'aboutir à des systèmes linéaires creux. En effet, rappelons que :

$$\mathbb{K}_{IJ} = 0$$

dès que M_I et M_J n'appartiennent pas à un même domaine.

3.2.1 Un cas particulier

Considérons le cas d'un maillage régulier du carré unité, comportant $N = n^2$ degrés de liberté. Un sommet appartient (au plus) à 4 carrés (exactement à 4 carrés pour tous les sommets intérieurs) dont la réunion comprend (au plus) 9 sommets distincts.



On choisit de d'utiliser une approximation par élément fini Q^1. De la structure du maillage, on en déduit que, pour I donné, $\text{card}\{J \text{ t.q. } \mathbb{K}_{IJ} \neq 0\} \leq 9$. Il y a donc moins de $9N$ termes non nuls parmi les N^2 termes de la matrice, soit un taux de remplissage plus petit que $9N^{-1}$. Or, dans le cas que nous étudions, le pas du maillage est $h \sim n^{-1}$, et ainsi $N = n^2 \to \infty$ quand $h \to 0$. Le taux de remplissage de la matrice tend donc vers 0 quand $h \to 0$: on parle alors de matrice creuse. Notons que la donnée du pointeur ℓ_g, qui permet de passer de la numérotation locale des degrés de liberté à leur numérotation globale (cf. (2.39)), permet de connaître *a priori* la structure de la matrice.

Les principales techniques de stockage et les principaux algorithmes de résolution de systèmes linéaires sont décrits dans l'annexe.

3.2.2 Le cas général

Dans cette sous-section, nous allons montrer que, d'une façon générale, les matrices issues de la discrétisation par éléments finis sont *creuses*, toujours selon l'idée que leur taux de remplissage tend vers 0 quand le pas du maillage h tend vers 0. A cette fin, nous allons traiter en détail le cas de la discrétisation par *éléments finis de Lagrange* P^1 d'un problème défini sur un ouvert *polygonal* de $\mathbb{R}^2$ (maillé à l'aide de triangles), avant de passer à d'autres cas dans $\mathbb{R}^2$ ou $\mathbb{R}^3$. Pour fixer les idées, on suppose que l'on résout un problème de Neumann. Soit donc $(\mathcal{T}_h)_h$ la suite de triangulations. Fixons h, la triangulation $\mathcal{T}_h$ comprend alors :

$$N \text{ sommets, } A \text{ arêtes et } L \text{ triangles.}$$

Par construction, la dimension de l'espace discret V_h est égale à N.

Dans un premier temps, examinons le nombre d'éléments Z_1 non-nuls de la matrice de rigidité $\mathbb{K}$ (ou de la matrice de masse $\mathbb{M}$), qui est de taille $N \times N$. Le taux de remplissage τ_1 est alors défini par :

$$\tau_1(N) = \frac{Z_1}{N^2}. \tag{3.10}$$

Nous avons vu au §2.2.1 (cf. (2.19)) que si M_I et M_J sont deux sommets n'appartenant pas à un même triangle de $\mathcal{T}_h$, alors $\mathbb{K}_{IJ} = 0$. Donc, pour que $\mathbb{K}_{IJ}$ soit non-nul, il est *nécessaire*[3] que :

$$M_I = M_J \quad \text{ou} \quad [M_I, M_J] \text{ côté d'un triangle.}$$

Le premier cas se produit exactement N fois, et le second, $2A$ fois, une arête étant représentée soit par $[M_I, M_J]$, soit par $[M_J, M_I]$. En conclusion,

$$Z_1 \leq N + 2A. \tag{3.11}$$

Pour évaluer le taux de remplissage $\tau_1(N)$, nous allons exprimer A (ainsi que L) en fonction de N. Dans la suite, on notera N_b (resp. A_b) le nombre de sommets (resp. d'arêtes) sur $\partial\Omega$, et $N_i = N - N_b$ (resp. $A_i = A - A_b$).

Proposition 3.1. *Soit Ω un polygone de $\mathbb{R}^2$ sans trous, et $\mathcal{T}_h$ une triangulation. Alors, on a les relations*

$$L = 2N - N_b - 2 \tag{3.12}$$

$$A = 3N - N_b - 3 \tag{3.13}$$

Démonstration : *Etape 1* : on commence par le *comptage des angles*.

(i) Dans chaque triangle, la somme des angles internes est égale à π radians : on compte $\pi \times L$ radians en sommant les angles internes de tous les triangles.
(ii) Si on raisonne par sommet dans le décompte ci-dessus :
- La somme des angles autour d'un sommet *interne* est égale à 2π radians ;
- Pour les sommets de la frontière, on *ne connaît pas* la valeur de la somme des angles pour un sommet *donné*. Par contre, on peut prouver par récurrence que la somme des angles internes aux sommets d'un polygone de K côtés ou sommets est égale à $(K-2)\pi$!

Récapitulons : $\pi L = 2\pi N_i + (N_b - 2)\pi$, ce qui donne $L = 2N_i + N_b - 2 = 2N - N_b - 2$.
Etape 2 : *comptage des arêtes*.

(i) Dans chaque triangle, il y a *trois* arêtes : on compte $3 \times L$ arêtes en parcourant tous les triangles.
(ii) Si on raisonne par arête dans le décompte ci-dessus :
- Une arête *de la frontière* est comptée *une fois* exactement, car elle appartient à un seul triangle ;
- Une arête *interne* est comptée *deux fois* exactement, car elle appartient à deux triangles.

Récapitulons : $3L = A_b + 2A_i$, ou $3L = 2A - A_b$.

On remarque que $N_b = A_b$, puisqu'un polygone possède le même nombre de sommets et de côtés ! On reprend alors les résultats précédents, et on élimine le nombre de triangles L pour trouver :

[3] La condition *n'est pas suffisante*... En effet, sur certains maillages (par ex. composés de triangles rectangles isocèles), il peut arriver que $[M_I, M_J]$ soit un côté d'un triangle, mais que $\nabla w_I(x) \perp \nabla w_J(x)$, pour tout $x \in \Omega$, et dans ce cas $\mathbb{K}_{IJ} = 0$. Par contre, $\mathbb{K}_{II} > 0$ pour tout indice I.

$$2A - A_b \overset{\text{Etape 2}}{=} 3L \overset{\text{Etape 1}}{=} 6N - 3N_b - 6 \ ;$$

d'où finalement $A = 3N - N_b - 3$. ∎

Remarque 3.6. La présence de trous modifie uniquement l'étape de comptage des angles. On vérifie facilement que les formules (3.12-3.13) n'évoluent que marginalement, les termes constants (resp. -2 et -3) incorporant le nombre de trous.

Corollaire 3.1. *Lorsque h tend vers 0, la dimension de V_h tend vers l'infini :*

$$\lim_{h \to 0} N = \infty. \tag{3.14}$$

Remarque 3.7. Ce résultat est bien sûr fortement souhaité (!) pour pouvoir appliquer le cadre général des méthodes d'approximation (voir le §2.1.1), mais il convient de le prouver...

Démonstration du corollaire 3.1 : D'après (2.37), on sait que $\text{Aire}(T_\ell) \leq \pi h^2$, pour $\ell = 1, L$. Comme $\text{Aire}(\Omega) = \sum_\ell \text{Aire}(T_\ell)$, on en déduit que $\text{Aire}(\Omega) \leq \pi h^2 L$ et, par voie de conséquence,

$$\lim_{h \to 0} L = \infty.$$

D'après (3.12), $L < 2N$, et la conclusion suit. ∎

Théorème 3.1. *Soit Ω un polygone de $\mathbb{R}^2$. Pour une discrétisation par éléments finis P^1, les matrices $\mathbb{K}$ et $\mathbb{M}$ sont creuses.*

Démonstration : D'après (3.11) et (3.13), on a $N < Z_1 < 7N$. Le taux de remplissage est donc encadré par

$$\frac{1}{N} < \tau_1(N) < \frac{7}{N}.$$

Or, d'après le corollaire 3.1, N tend vers l'infini lorsque h tend 0, et la conclusion suit. ∎

De façon remarquable, ce résultat se généralise lorsqu'on utilise des éléments finis de Lagrange d'ordre plus élevé, ainsi qu'à d'autres types d'éléments finis. Les résultats fondamentaux qui permettent de parvenir à cette conclusion portent sur le dénombrement, et sont contenus dans la proposition 3.1 et son corollaire 3.1. Prenons deux exemples : l'élément fini de Lagrange P^2 et l'élément fini d'Hermite P^3.

Pour l'*élément fini de Lagrange P^2*, il y a six degrés de liberté par triangle : un pour chaque sommet et un pour chaque milieu d'arête. Le nombre total de degrés de liberté (et la dimension de V_h) est donc égal à $N_2 = N + A$, et $\mathbb{K}$ et $\mathbb{M}$ sont des matrices $N_2 \times N_2$. Pour que $\mathbb{K}_{IJ}$ soit non-nul, il est nécessaire que

- Si M_I et M_J sont deux sommets : $M_I = M_J$, ou $[M_I, M_J]$ côté d'un triangle ;
- Si M_I et M_J sont deux milieux d'arête : $M_I = M_J$, sinon M_I, M_J sont dans un même triangle ;
- Si M_I et M_J sont de types différents : M_I, M_J appartiennent à un même côté, sinon ils sont dans un même triangle.

On trouve alors, en dénombrant chaque ensemble de configurations, un nombre d'éléments non-nuls Z_2 borné par

$$Z_2 \leq (N + 2A) + (A + 6L) + (4A + 6L) = N + 7A + 12L.$$

D'après (3.12) et (3.13), $Z_2 < 46N$ et $N_2 > 3N$. Ainsi, le taux de remplissage $\tau_2(N_2)$ pour l'élément fini de Lagrange P^2 est plus petit que

$$\tau_2(N_2) < \frac{46N}{9N^2} = \frac{46}{9} \times \frac{1}{N},$$

et il tend bien vers 0 lorsque h tend vers 0.

Remarque 3.8. Avant de passer à l'élément fini d'Hermite, nous comparons les taux de remplissage τ_1 et τ_2 (pour une *même* taille de matrice). Pour cela, nous faisons l'hypothèse – raisonnable – selon laquelle $N_b = o(N)$ lorsque $N \to \infty$, c'est-à-dire que le nombre de sommets de la frontière est négligeable par rapport au nombre total de sommets lorsque $N \to \infty$. On a alors, si le cas de la note de bas de page[3] ne se produit pas, les résultats d'équivalence suivants :

$$\begin{aligned} \tau_1(N) &\sim \frac{7}{N} && \text{pour l'élément fini } P^1\,; \\ \tau_2(N_2) &\sim \frac{46N}{N_2^2} \sim \frac{11.5}{N_2} && \text{pour l'élément fini } P^2. \end{aligned}$$

Ainsi, pour une dimension de l'espace de discrétisation donnée, le taux de remplissage augmente quand l'ordre augmente (passage du facteur multiplicatif de 7 à 11.5). C'est une propriété générale.

Pour l'*élément fini d'Hermite* P^3, il y a dix degrés de liberté par triangle : un pour chaque sommet, deux par d'arête, et un pour le milieu. Le nombre total de degrés de liberté est cette fois égal à $N_3 = N + 2A + L$, et $\mathbb{K}$ et $\mathbb{M}$ sont des matrices $N_3 \times N_3$. En effectuant le dénombrement des configurations dans lesquelles $\mathbb{K}_{IJ}$ est non-nul, on trouve un nombre total d'éléments non-nuls Z_3 borné par

$$Z_3 \leq N + 10A + 31L.$$

D'après (3.12) et (3.13), $Z_3 < 93N$ et $N_3 > 6N$, et le taux de remplissage $\tau_3(N_3)$ pour l'élément fini d'Hermite P^3 est plus petit que

$$\tau_3(N_3) < \frac{31}{12} \times \frac{1}{N}.$$

Ainsi, τ_3 tend bien vers 0 lorsque h tend vers 0.

Dans le cas général, en recensant les degrés de liberté par triangle, on en compte $N_{tot} = c_N N + c_A A + c_L L$. Les matrices $\mathbb{K}$ et $\mathbb{M}$ sont de taille $N_{tot} \times N_{tot}$. Après

dénombrement des configurations dans lesquelles $\mathbb{K}_{IJ} \neq 0$, on en déduit qu'il existe C_N, C_A et C_L positives et indépendantes de la triangulation $\mathcal{T}_h$, telles que

$$Z_{tot} \leq C_N N + C_A A + C_L L.$$

D'après (3.12) et (3.13),

$$Z_{tot} \leq (C_N + 3C_A + 2C_L)N \quad \text{et} \quad N_{tot} > (c_N + 2c_A + c_L)N,$$

d'où un taux de remplissage majoré par

$$\tau(N_{tot}) < \frac{(C_N + 3C_A + 2C_L)}{(c_N + 2c_A + c_L)^2} \times \frac{1}{N},$$

qui tend vers 0 lorsque h tend vers 0.

Bien sûr, on pourrait raisonner de la même façon sur des maillages quadrangulaires, les seules différences venant de l'expression des nombres de quadrilatères et d'arêtes pour ce type de maillage (ce qui modifierait (3.12) et (3.13)). D'où finalement le résultat général ci-dessous.

Théorème 3.2. ***(matrices creuses pour les éléments finis 2D)*** *Soit Ω un polygone de $\mathbb{R}^2$. Les matrices $\mathbb{K}$ et $\mathbb{M}$ obtenues par une discrétisation de type éléments finis sont creuses.*

On peut raisonner selon les mêmes lignes, pour un problème (par ex. de Neumann) défini sur un ouvert *polyédrique* de $\mathbb{R}^3$ (maillé à l'aide de tétraèdres). Soit donc $(\mathcal{T}_h)_h$ la suite de triangulations. Fixons h, la triangulation $\mathcal{T}_h$ comprend alors :

$$N \text{ sommets, } A \text{ arêtes, } F \text{ faces et } L \text{ tétraèdres.}$$

Pour établir que les matrices éléments finis $\mathbb{K}$ et $\mathbb{M}$ sont creuses, ou voudrait déterminer les nombres d'arêtes, de faces et de tétraèdres en fonction du nombre de sommets, pour obtenir les équivalents 3D de (3.12) et (3.13). Comme on va le voir par la suite, on ne peut pas trouver des formules exactes, mais plutôt des encadrements, qui permettent néanmoins de conclure. On note N_b (resp. A_b, F_b) le nombre de sommets (resp. d'arêtes, de faces) sur $\partial\Omega$, et $N_i = N - N_b$ (resp. $A_i = A - A_b$, $F_i = F - F_b$).

Dans un premier temps, on détermine A_i, A_b, F_i et F_b en fonction de N_i, N_b et L. Pour cela, on utilise les quatre relations ci-dessous :

- Comptage des arêtes de la frontière : $3F_b = 2A_b$;
- Comptage des faces : $4L = 2F_i + F_b$;
- Formule d'Euler sur la frontière : $N_b - A_b + F_b = 2$;
- Formule d'Euler dans l'ouvert : $N - A + F - L = 1$.

Pour fixer les idées, on a supposé que le polyèdre est convexe, sans trous et simplement connexe. Si tel n'est pas le cas, il fait modifier les valeurs des constantes dans les formules. On en déduit alors les relations :

$$\begin{aligned} A_b &= 3N_b - 6 \\ F_b &= 2N_b - 4 \\ A_i &= N_i - N_b + L + 3 \\ F_i &= 2L - N_b + 2. \end{aligned}$$

Il reste maintenant à éliminer L en fonction de N_i et N_b... Pour cela, une idée naturelle est de compter les *angles solides* aux sommets des tétraèdres, notés $(\omega^i_{K_\ell})_{i=1,4}$. Par tétraèdre, on trouve

$$0 \leq \sum_{i=1,4} \omega^i_{K_\ell} \leq 2\pi. \tag{3.15}$$

En effet, la somme des angles solides d'un tétraèdre n'est ni constante, ni minorée par une constante strictement positive !

La somme des angles solides en un sommet intérieur de $\mathcal{T}_h$ vaut 4π. Pour un sommet de $\mathcal{T}_h$ situé sur la frontière, elle est supérieure à $\omega_{\partial\Omega}$, la valeur du plus petit angle solide en un sommet de $\partial\Omega$. En sommant les inégalités (3.15) sur tous les tétraèdres, on arrive à la *minoration*

$$4N_i + \frac{\omega_{\partial\Omega}}{\pi} N_b \leq 2L.$$

Malheureusement, on ne sait pas *majorer* L en fonction de N_i et N_b !? Pour y parvenir, il faut faire une *hypothèse supplémentaire* sur les triangulations : on suppose que la famille de triangulations $(\mathcal{T}_h)_h$ est *régulière*. En effet, si l'on revient à la discussion qui suit l'énoncé de la condition (2.62), on constate que, dans ce cas,

$$\exists \omega_0 > 0,\ \forall h,\ \forall K_\ell \in \mathcal{T}_h,\ \omega_{K_\ell} \geq \omega_0 \quad (\omega_{K_\ell} \text{ plus petit angle solide de } K_\ell).$$

Dans ce cas, (3.15) devient

$$4\omega_0 \leq \sum_{i=1,4} \omega^i_{K_\ell} \leq 2\pi. \tag{3.16}$$

En sommant les inégalités (3.16) sur tous les tétraèdres, on arrive à

$$L \leq \frac{\pi}{\omega_0}(N_i + N_b),$$

ce qui correspond à la majoration cherchée ! Comme annoncé, on ne sait plus exprimer les relations entre les diverses quantités par des égalités, mais par des minorations et des majorations. Néanmoins, à partir de là, on peut raisonner comme dans le cas 2D, pour arriver au résultat général ci-dessous.

Théorème 3.3. ***(matrices creuses pour les éléments finis 3D)*** *Soit Ω un polyèdre de $\mathbb{R}^3$ et soit $(\mathcal{T}_h)_h$ une famille de triangulations régulière. Les matrices $\mathbb{K}$ et $\mathbb{M}$ obtenues par une discrétisation de type éléments finis sont creuses.*

En conclusion, afin de tirer parti de cet avantage (i.e. la connaissance *a priori* de la nature creuse des matrices à manipuler en 2D et en 3D), il est impératif d'utiliser des structures de stockage permettant de stocker seulement les termes non nuls, ainsi que des algorithmes de résolution de systèmes linéaires n'opérant que sur les termes non nuls afin de diminuer le temps de calcul. Enfin, pour une bonne mise en œuvre informatique, il faut disposer de logiciels et/ou de langages traitant efficacement les structures creuses ainsi que les opérations à mener sur celles-ci.

3.3 Quelques illustrations numériques

L'objet de cette section est d'illustrer les concepts de mise en œuvre qui ont été exposés précédemment. Nous présentons sucessivement deux exemples montrant la relative simplicité de la mise en œuvre des éléments finis P^1 et P^2 pour des problèmes bidimensionnels dans l'environnement Matlab. Dans un premier temps nous traitons le cas de l'équation de Laplace et nous montrons que les erreurs obtenues sont tout à fait en accord avec les estimations théoriques présentées au chapitre 2, théorèmes 2.4 et 2.6. Nous montrons également comment il est facile de généraliser le calcul au cas d'un milieu non homogène dans lequel peuvent apparaître des conditions de transmission qui sont naturellement prises en compte dans une formulation variationnelle. Nous abordons ensuite le cas d'un problème de nature *vectorielle*, celui de l'élasticité linéaire homogène isotrope. Les techniques restent fondamentalement les mêmes, la difficulté résidant dans l'aspect vectoriel du problème qui nécessite une gestion un peu plus compliquée.
Pour chacun de ces exemples nous rappelons l'arrière-plan théorique et donnons les aspects de mise en œuvre ainsi que les codes Matlab associés. La méthode des éléments finis est basée sur la donnée d'un maillage (triangulation pour des éléments finis P^1 ou P^2). Il existe des "mailleurs" généraux (*emc2*, *gmsh* par exemple) mais nous avons ici utilisé des outils élémentaires de maillage, développés en Matlab, largement suffisants pour nos objectifs. Il n'est pas nécessaire de les analyser pour comprendre les aspects éléments finis. Il faut néanmoins savoir que ces outils produisent essentiellement 4 tableaux que ce soit en P^1 ou en P^2 :

- S : tableau de dimension $ns \times 2$ contenant les coordonnées des nœuds du maillage (ns nombre total de nœuds)
- T : tableau de dimension $nt \times nd$ contenant la numérotation des triangles (nt nombre de triangles, nd nombre de nœuds par triangle : $nd = 3$ en P^1 et $nd = 6$ en P^2 (avec pour le second cas la convention que les 3 premiers numéros correspondent à des sommets d'un triangle)

- BR : tableau $nt \times 3$ fournissant un numéro de référence pour chaque arête des triangles (permet de traiter les conditions aux limites)
- RT : tableau $nt \times 1$ fournissant un numéro de référence par triangle (permet de traiter des milieux différents)

Pour les deux derniers tableaux, les conventions sont, d'une part, que toute arête de la frontière est sujette à une unique condition aux limites (Dirichlet, Neumann, etc.), et d'autre part, que tout triangle contient un matériau homogène.
Dans toute la suite, nous notons $(M_i)_{i=1,ns}$ les nœuds du maillage, $(T_\ell)_{\ell=1,nt}$ les triangles du maillages et indifféremment w_i la fonction de base globale P^1 ou P^2 attachée au nœud M_i. Nous supposons que l'ouvert Ω est polygonal, que $\overline{\Omega} = \bigcup_{\ell=1,nt} T_\ell$ (cf. (2.35)) et que le maillage est admissible (cf. (2.36)), de sorte que l'espace vectoriel de dimension ns :

$$V_h = \text{vect}(w_i)_{i=1,ns}$$

est inclus dans $H^1(\Omega)$ (approximation interne).
Enfin, on suppose que la famille de triangulations est régulière (cf. (2.62)).

3.3.1 Equation de Laplace-Poisson

Nous nous intéressons au problème de Dirichlet sur un ouvert borné Ω de $\mathbb{R}^2$ de frontière notée Γ, avec des données $f \in L^2(\Omega)$ et $g \in H^{1/2}(\Gamma)$:

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \Gamma \end{cases}$$

dont une formulation variationnelle est :

$$\left| \begin{array}{l} \text{Trouver } u \in H^1(\Omega) \text{ tel que } u = g \text{ sur } \Gamma \text{ et} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f \, v \, d\Omega \qquad \forall v \in H_0^1(\Omega). \end{array} \right.$$

Rappelons que ce problème est bien posé. De plus, si on désigne par $\mathcal{D}$ l'ensemble des indices des nœuds situés sur le bord ($\mathcal{D} = \{i \text{ tel que } M_i \in \Gamma\}$) et $\mathcal{I} = \{1, ..., ns\} \setminus \mathcal{D}$, une approximation dans V_h de cette formulation est :

$$\left| \begin{array}{l} \text{Trouver } u_h \in V_h \text{ tel que } u_h(M_i) = g(M_i) \ \forall i \in \mathcal{D} \text{ et} \\ \qquad \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h \, d\Omega = \int_\Omega f \, v_h \, d\Omega \qquad \forall v_h \in V_h^0 \end{array} \right.$$

où $V_h^0 = \text{vect}(w_i)_{i \in \mathcal{I}}$.
On peut reformuler le problème approché sous forme matricielle, en notant $u_h = \sum_{j=1,ns} U_j w_j$, $\mathbb{K}_{ij} = \displaystyle\int_\Omega \nabla w_i \cdot \nabla w_j \, d\Omega$, $G_i = g(M_i)$ et $S_i = \displaystyle\int_\Omega f \, w_i \, d\Omega$ et en

partitionnant les indices suivant $\mathcal{I}$ et $\mathcal{D}$. A l'aide de la technique de pseudo-élimination (cf. (3.7) avec $\mathbb{D} = \mathrm{Diag}\,(\mathbb{K}_{\mathcal{DD}})$), on aboutit à :

$$\begin{bmatrix} \mathbb{K}_{\mathcal{II}} & 0 \\ 0 & \mathbb{D} \end{bmatrix} \begin{bmatrix} \overrightarrow{U}_{\mathcal{I}} \\ \overrightarrow{U}_{\mathcal{D}} \end{bmatrix} = \begin{bmatrix} \overrightarrow{S}_{\mathcal{I}} - \mathbb{K}_{\mathcal{ID}} \overrightarrow{G}_{\mathcal{D}} \\ \mathbb{D} \overrightarrow{G}_{\mathcal{D}} \end{bmatrix}$$

Si la donnée f est régulière ($f \in \mathcal{C}^0(\overline{\Omega})$) il est possible d'approcher S_i en utilisant l'interpolée de f ($\pi_h f = \sum_{i=1,ns} f(M_i) w_i$) :

$$S_i \simeq \int_\Omega \pi_h f \, w_i \, d\Omega = \left(\mathbb{M}\overrightarrow{F}\right)_i \qquad \text{avec } F_i = f(M_i) \text{ et } \mathbb{M}_{ij} = \int_\Omega w_i \, w_j \, d\Omega.$$

En résumé, on doit donc :

– calculer la matrice de masse $\mathbb{M}_{ij} = \displaystyle\int_\Omega w_i \, w_j \, d\Omega$

– calculer la matrice de rigidité $\mathbb{K}_{ij} = \displaystyle\int_\Omega \nabla w_i \cdot \nabla w_j \, d\Omega$

– assembler le système $\begin{bmatrix} \mathbb{K}_{\mathcal{II}} & 0 \\ 0 & \mathbb{D} \end{bmatrix} \begin{bmatrix} \overrightarrow{U}_{\mathcal{I}} \\ \overrightarrow{U}_{\mathcal{D}} \end{bmatrix} = \begin{bmatrix} \left(\mathbb{M}\overrightarrow{F}\right)_{\mathcal{I}} - \mathbb{K}_{\mathcal{ID}} \overrightarrow{G}_{\mathcal{D}} \\ \mathbb{D} \overrightarrow{G}_{\mathcal{D}} \end{bmatrix}$

En pratique on calcule les matrices complètes $\mathbb{K}$ et $\mathbb{M}$ sans distinguer les indices parmi $\mathcal{I}$ et $\mathcal{D}$, puis on applique une procédure dite de *pseudo-élimination* qui donne le système précédent. Comme on l'a déjà vu, cette technique de *pseudo-élimination* ne modifie pas l'ordre des indices comme pourrait le faire croire la présentation précédente.
Le script Matlab suivant réalise la résolution du problème de Dirichlet dans le carré de côté 4 ayant pour solution $u(x, y) = \cos(xy)$:

```
%calcul elements finis
[S,T,BR,RT]=triangule_rectangle([0 4 0 4],40,40,1);%maillage
[K,M]=calcul_EF_2D(S,T,RT);                          %matrices EF
%test de U=cos(xy)
X=S(:,1);Y=S(:,2);
Uex=cos(X.*Y);                                       %sol. exacte
B=M*(Uex.*(X.*X+Y.*Y));                              %second membre
%prise en compte des cond. de Dirichlet
Noeud_dir=noeud_bords(S,T,BR,[2 3]);                 %liste noeuds Dirichlet
G=Noeud_dir.*Uex;                                    %donnee de Dirichlet
[Ke,Be]=cd_Dirichlet(K,B,Noeud_dir,G);               %pseudo-elimination
%resolution, calcul d'erreurs L2 et H1
U=Ke\B;                                              %resolution du systeme
E=U-Uex;                                             %ecart
el2=sqrt(E'*M*E);eh1=sqrt(E'*K*E);                   %erreur L2 et H1
%dessin
figure; trisurf(T,S(:,1),S(:,2),U);shading interp;
```

Il est intéressant de lire attentivement le code de la fonction `calcul_EF_2D` qui constitue le cœur de la méthode des éléments finis. Cette fonction construit les matrices de masse $\mathbb{M}$ et de rigidité $\mathbb{K}$ en réalisant une boucle sur les triangles du maillage. Sur chaque triangle T_ℓ sont alors calculées les matrices dites élémentaires :

$$\mathbb{K}_{ij}^{\ell} = \int_{T_\ell} \nabla\tau_i^{\ell} \cdot \nabla\tau_j^{\ell}\, d\Omega \quad \text{et} \quad \mathbb{M}_{ij}^{\ell} = \int_{T_\ell} \tau_i^{\ell}\tau_j^{\ell}\, d\Omega \qquad \forall i,j = 1, nd.$$

Dans le code proposé, nous avons adopté le point de vue général qui consiste à calculer ces intégrales, via un changement de variable qui ramène l'intégrale sur le triangle unité $\widehat{T}$ et une formule de quadrature numérique (voir le chapitre 2). Nous avons utilisé la formule de quadrature à 7 points exacte pour des polynômes de degré 5. Ce point de vue permet également de traiter le cas d'équations à coefficients variables telles que, par exemple :

$$-\operatorname{div} k(x)\nabla u + q(x)u = f$$

Bien évidemment, si on s'intéresse à des problèmes à coefficients constants approchés par éléments finis P^1, il est plus efficace d'utiliser les formules explicites des matrices élémentaires, c'est-à-dire sans passer par $\widehat{T}$.

```
function [K,M]=calcul_EF_2D(S,T,RT,fK,fM)
if(nargin==3) fK=@un;fM=@un;end,                    %fK coef. associe a K
if(nargin==4) fM=@un;end,                           %fM coef. associe a M
%formule de quadrature a 7 points
os=sqrt(15);s3=1./3.;
pp1=(6.-os)/21.;pp2=(6.+os)/21.;
pp3=(9.+2.*os)/21.;pp4=(9.-2.*os)/21}.;
pts_quadT=[s3 s3;pp1 pp1;pp1 pp3;pp3 pp1;...       %points de quadrature
           pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];         %poids de quadrature
%initialisation
nt=size(T,1);ns=size(S,1);q=size(T,2);              %tailles diverses
nbq=length(pds_quadT);
K=sparse(ns,ns);M=sparse(ns,ns);                    %def. matrices creuses
%boucle sur les triangles
for t=1:nt,                                         %t indice du triangle
    St=[S(T(t,:),:)];                               %sommets du triangle t
    S21=St(2,:)-St(1,:);S31=St(3,:)-St(1,:);
    delta=S21(1)*S31(2)-S21(2)*S31(1);
    Jfmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;%inverse du jacobien
    Mt=zeros(q,q);Kt=zeros(q,q);                    %matrices elementaires
    for k=1:nbq,                                    %boucle pts de quadrature
        x=pts_quadT(k,1);y=pts_quadT(k,2);
        if(q==3) w=[1-x-y x y];gw=[-1 1 0;-1 0 1];      %fct de base P1
        else a=1-x-y;b=2*x-1;c=2*y-1;                   %fct de base P2
             w=[a*(1-2*x-2*y),b*x*,c*y*,4*x*a,4*x*y,4*y*a];
             gw=[4*(x+y)-3 4*x-1 0 4*(1-2*x-y) 4*y -4*y;
                 4*(x+y)-3 0 4*y-1 -4*x 4*x 4*(1-x-2*y)];
        end,
    P=St'*[1-x-y;x;y];pk=pds_quadT(k)*abs(delta);%pt du plan physique
    Mt=Mt+fM(P(1),P(2),RT(t))*pk*w'*w;              %matrice masse elem.
    jg=Jfmt*gw;
    Kt=Kt+fK(P(1),P(2),R(t))*pk*jg'*jg;             %matrice rigidite elem.
end,
%assemblage de M et K
```

```
I=T(t,:);
K(I,I)=K(I,I)+Kt;                                   %assemblage de K
M(I,I)=M(I,I)+Mt;                                   %assemblage de M
function [z]=un(x,y,r) z=1}.;                       %fonction f=1
```

La prise en compte des conditions de Dirichlet constitue l'autre partie importante d'un code éléments finis. A partir de la connaissance des nœuds qui sont soumis à une condition de Dirichlet il est facile de générer le système linéaire final :

```
function [Ae,Be] = cd_Dirichlet(A,B,Noeud_dir,G)
ns=size(A,1);Ae=A;Be=B;                             %initialisation
if(nargin==3) G=zeros(ns,1);                        %cond. homogene
else Be=B-A*(Noeud_dir.*G);end,                     %correction sec. membre
%pseudo-elimination
for i=1:ns,
    if(Noeud_dir(i)==1),
       Ae(i,:)=0;Ae(:,i)=0;
       Ae(i,i)=A(i,i);Be(i)=A(i,i)*G(i);            %elimination
    end,
end
```

Nous donnons sur la figure 3.3 le résultat obtenu par le script Matlab précédent. On peut se convaincre de la qualité de l'approximation par éléments finis en représentant les normes $L^2(\Omega)$ et $H^1(\Omega)$ de l'erreur, c'est-à-dire les quantités :

$$\begin{aligned}
\|u_h - u\|_{L^2(\Omega)} &= \left(\int_\Omega |u_h - u|^2 d\Omega\right)^{1/2} \\
&\simeq \left(\int_\Omega |u_h - \pi_h u|^2 d\Omega\right)^{1/2} = \left((\mathbb{M}(\overrightarrow{U} - \overrightarrow{U}_\pi)|(\overrightarrow{U} - \overrightarrow{U}_\pi))\right)^{1/2} \\
\|u_h - u\|_{H_0^1(\Omega)} &= \left(\int_\Omega |\nabla u_h - \nabla u|^2 d\Omega\right)^{1/2} \\
&\simeq \left(\int_\Omega |\nabla u_h - \nabla \pi_h u|^2 d\Omega\right)^{1/2} = \left((\mathbb{K}(\overrightarrow{U} - \overrightarrow{U}_\pi)|(\overrightarrow{U} - \overrightarrow{U}_\pi))\right)^{1/2}
\end{aligned}$$

où $\overrightarrow{U}_\pi$ représente le vecteur des composantes de $\pi_h u$ dans la base éléments finis $(w_i)_{i=1,ns}$. L'utilisation de l'interpolée $\pi_h u$ de la solution u n'est valable que si la fonction u est suffisamment régulière (encore une fois, si $u \in \mathcal{C}^0(\overline{\Omega})$).
Pour la solution exacte $u(x,y) = \cos(xy)$, on représente sur les figures 3.4 et 3.5 les courbes d'erreurs en fonction de la finesse du maillage (représentée à l'aide de h), en représentation logarithmique. On obtient les comportements asymptotiques prévus théoriquement lorsque $h \to 0$, à savoir que l'erreur en norme $H^1(\Omega)$ est du même ordre que le degré des éléments finis et que l'erreur $L^2(\Omega)$ est d'un ordre de plus. En pointillés, les droites de pentes respectives 1 ou 2.

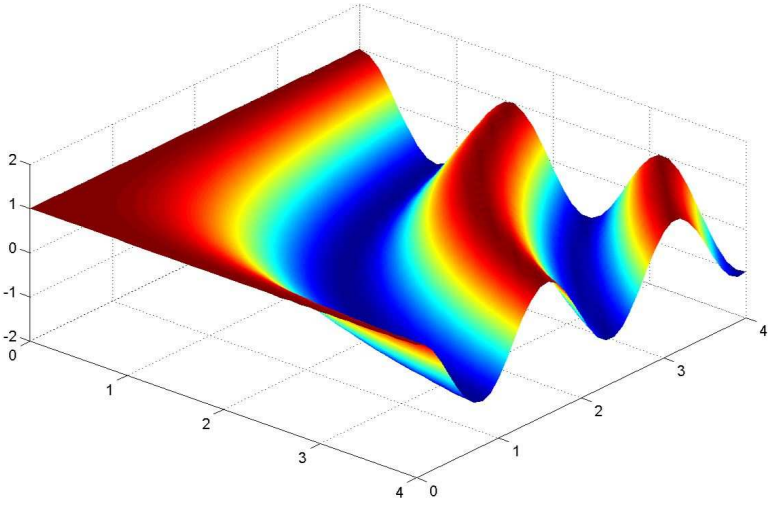


Fig. 3.3. Solution de l'équation de Laplace approchée par éléments finis

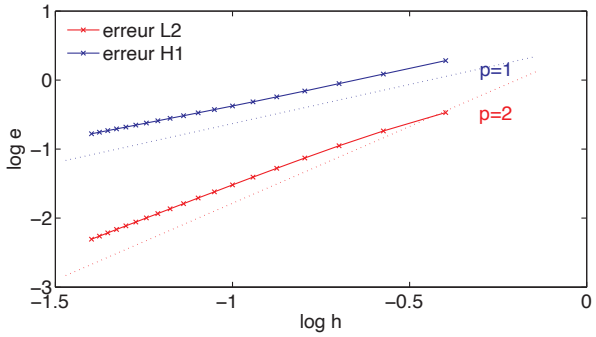


Fig. 3.4. Erreurs éléments finis P^1

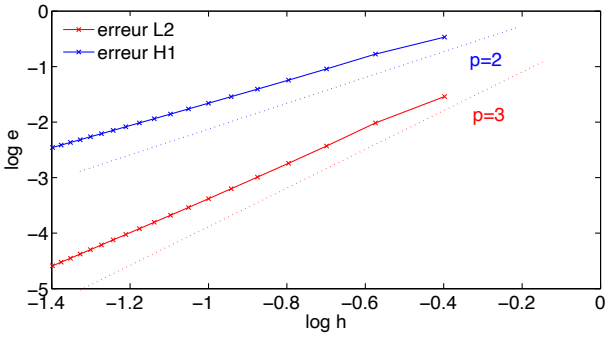


Fig. 3.5. Erreurs éléments finis P^2

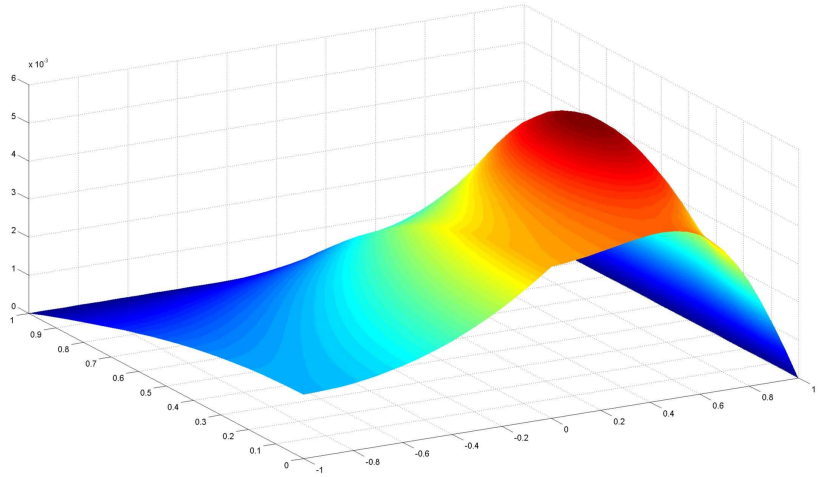


Fig. 3.6. Solution approchée par éléments finis d'un problème de transmission

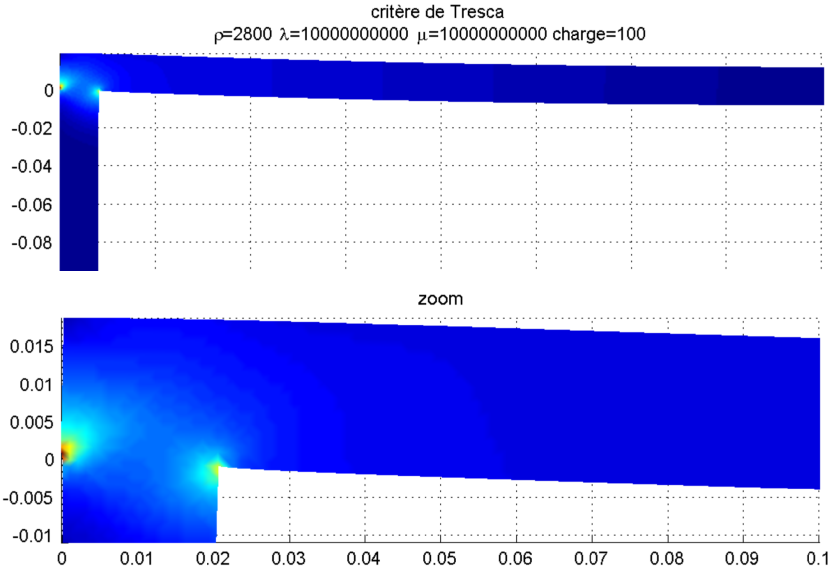


Fig. 3.7. Solution approchée par éléments finis P^2 des équations de l'élasticité

• Problème de transmission

Les fonctions Matlab précédentes permettent également de traiter un problème de Laplace à coefficient variable tel que, par exemple, le problème de transmission sur l'ouvert Ω, avec $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, $k \in L^\infty(\Omega)$ et une donnée $f \in L^2(\Omega)$:

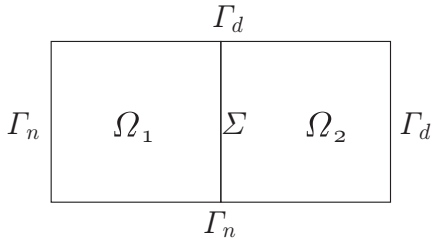


Fig. 3.8. Problème de transmission

$$\begin{cases} -\operatorname{div}(k(x)\nabla u) = f & \text{sur } \Omega_1 \cup \Omega_2 \\ u = 0 & \text{sur } \Gamma_d \\ \dfrac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_n \\ \left[k\dfrac{\partial u}{\partial n}\right] = 0 & \text{sur } \Sigma \end{cases} \qquad \text{avec } k(x) = \begin{cases} k_1 \text{ sur } \Omega_1 \\ k_2 \text{ sur } \Omega_2 \end{cases}$$

dont une formulation variationnelle est :

$$\left| \begin{array}{l} \text{trouver } u \in H^1_{0,\Gamma_d}(\Omega) \text{ et} \\ \displaystyle\int_\Omega k\nabla u \cdot \nabla v \, d\Omega = \int_\Omega f\, v \, d\Omega \qquad \forall v \in H^1_{0,\Gamma_d}(\Omega) \end{array} \right. .$$

Dès que $k(x,y) \geq k_0 > 0$ presque pour tout $(x,y) \in \Omega$, ce problème variationnel admet une unique solution. Le script suivant réalise le calcul de la solution approchée par éléments finis P^2 pour les données $k_1 = 10$, $k_2 = 1$ et la donnée f :

$$f(x,y) = \begin{cases} 1 \text{ si } \left(x - \frac{3}{4}\right)^2 + \left(y - \frac{1}{2}\right)^2 \leq \frac{1}{10} \\ 0 \text{ sinon} \end{cases}$$

```
%calcul elements finis
 [S,T,BR,RT]=triangule_rectangle([-1 1 0 1],20,10);    %maillage P1
 [S,T,BR,RT]=maillageP2(S,T,BR,RT);[T,S]=renume(T,S); %maillage P2
 [K,M]=calcul_EF_2D(S,T,RT,@k);                        %matrices EF
 %second membre
 X=S(:,1);Y=S(:,2);X1=X-0.75;Y1=Y-0.5;
 B=M*((X1.*X1+Y1.*Y1)<0.1);}                           %second membre
 %prise en compte des cond. de Dirichlet
 Noeud_dir=noeud_bords(S,T,BR,[2 3]);                  %liste dl Dirichlet
```

```
[Ke,Be]=cd_Dirichlet(K,B,Noeud_dir);                    %pseudo-elimination
%dessin
T1=isop2(T);
figure; trisurf(T,S(:,1),S(:,2),U);shading interp;
%fonction k(x,y)
function z= k(x,y,r)
z=1;if(x>0)z=10;end,
```

On a représenté sur la figure 3.6 la solution obtenue à l'aide du script Matlab précédent. On observe que la solution, si elle est globalement continue, est non dérivable au passage de l'interface Σ : $u \in \mathcal{C}^0(\overline{\Omega})$, $\nabla u \notin \mathcal{C}^0(\overline{\Omega})^2$.

3.3.2 Elasticité bidimensionnelle

Dans cet exemple, on s'intéresse à la mise en œuvre des éléments finis dans le cas d'un problème vectoriel, celui de l'élasticité linéaire homogène et isotrope en dimension 2. Rappelons que pour un solide Ω (borné de $\mathbb{R}^2$) qu'on suppose homogène isotrope, il s'agit de trouver le champ déplacement $\mathbf{u} = (u_1, u_2)$ solution des équations suivantes :

$$(\mathcal{E}) \qquad \begin{cases} -\sum_{j=1,2} \dfrac{\partial}{\partial x_j}\sigma_{ij}(\mathbf{u}) = f_i \ \text{ dans } \Omega \\ \sum_{j=1,2} \sigma_{ij}(\mathbf{u})n_j = h_i \quad \text{sur } \Gamma_c \\ \sum_{j=1,2} \sigma_{ij}(\mathbf{u})n_j = 0 \quad \text{sur } \Gamma_l \\ \mathbf{u} = 0 \quad \text{sur } \Gamma_e \end{cases} \qquad \text{pour } i = 1,2$$

où le tenseur des contraintes est (avec $\lambda > 0$ et $\mu > 0$ coefficients de Lamé) :

$$\sigma_{ij}(\mathbf{u}) = \lambda(\operatorname{div}\mathbf{u})\,\delta_{ij} + 2\mu\varepsilon_{ij}(\mathbf{u}) \quad i,j = 1,2$$

et celui des déformations est donné par :

$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad i,j = 1,2.$$

Nous avons considéré ici le cas d'un solide soumis à une force intérieure $\mathbf{f} = (f_1, f_2)$, encastré sur une partie Γ_e de sa frontière, soumis à un chargement $\mathbf{h} = (h_1, h_2)$ sur une partie Γ_c et libre de tout effort sur le reste de la frontière, noté Γ_l.
La formulation variationnelle de $(\mathcal{E})$ dans l'espace :

$$V_0 = \{\mathbf{v} \in H^1(\Omega)^2, \mathbf{v} = 0 \text{ sur } \Gamma_e\}$$

consiste à trouver $\mathbf{u} \in V_0$, tel que $a(\mathbf{u}, \mathbf{v}) = \ell(\mathbf{v})$, $\forall \mathbf{v} \in V_0$, où on a posé :

$$\left|\begin{array}{l} a(\mathbf{u},\mathbf{v}) = \lambda\displaystyle\int_\Omega \operatorname{div}\mathbf{u}\,\operatorname{div}\mathbf{v}\,d\Omega + 2\mu\sum_{i,j=1,2}\int_\Omega \varepsilon_{ij}(\mathbf{u})\,\varepsilon_{ij}(\mathbf{v})\,d\Omega \\ \ell(\mathbf{v}) = \displaystyle\int_{\Gamma_c}\mathbf{h}\cdot\mathbf{v}\,d\Gamma + \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega \end{array}\right. .$$

On peut montrer que ce problème est bien posé.
Dans le contexte des éléments finis de Lagrange P^1 ou P^2, on introduit l'espace d'approximation de dimension $2ns$ (inclus dans $H^1(\Omega)^2$) :

$$V_h = \operatorname{vect}(\mathbf{w}_i^1, \mathbf{w}_i^2)_{i=1,ns} \quad \text{où } \mathbf{w}_i^1 = \begin{pmatrix} w_i \\ 0 \end{pmatrix} \text{ et } \mathbf{w}_i^2 = \begin{pmatrix} 0 \\ w_i \end{pmatrix}, \quad i = 1, ns.$$

ainsi que l'espace ($\mathcal{D}$ désignant l'ensemble des indices des nœuds encastrés du maillage – $M_i \in \overline{\Gamma}_e$ – et $\mathcal{I}$ les nœuds non encastrés) :

$$V_h^0 = \operatorname{vect}(\mathbf{w}_i^1, \mathbf{w}_i^2)_{i\in\mathcal{I}} \subset V_0.$$

On introduit alors le problème variationnel approché :

$$\left|\begin{array}{l} \text{Trouver } \mathbf{u}_h \in V_h^0 \text{ tel que} \\ \lambda\displaystyle\int_\Omega \operatorname{div}\mathbf{u}_h\,\operatorname{div}\mathbf{v}_h\,d\Omega + 2\mu\sum_{i,j=1,2}\int_\Omega \varepsilon_{ij}(\mathbf{u}_h)\,\varepsilon_{ij}(\mathbf{v}_h)\,d\Omega = \\ \qquad\qquad \displaystyle\int_{\Gamma_c}\mathbf{h}\cdot\mathbf{v}_h\,d\Gamma + \int_\Omega \mathbf{f}\cdot\mathbf{v}_h\,d\Omega, \quad \forall \mathbf{v}_h \in V_h^0 \end{array}\right.$$

En décomposant $\mathbf{u}_h$ sur la base $(\mathbf{w}_i^1, \mathbf{w}_i^2)_{i\in\mathcal{I}}$:

$$\mathbf{u}_h = \sum_{j\in\mathcal{I}} U_j^1 \mathbf{w}_j^1 + U_j^2 \mathbf{w}_j^2$$

et en considérant toutes les équations obtenues pour $\mathbf{v}_h = \mathbf{w}_i^1$ et $\mathbf{v}_h = \mathbf{w}_i^2$ $\forall i \in \mathcal{I}$ on obtient un système linéaire de dimension $2\text{card}\mathcal{I}$, portant sur l'inconnue $\overrightarrow{U} = \begin{pmatrix} \overrightarrow{U}^1 \\ \overrightarrow{U}^2 \end{pmatrix}$. Les écritures se font par blocs $\text{card}\mathcal{I}\times 1$ pour les vecteurs, et $\text{card}\mathcal{I}\times\text{card}\mathcal{I}$ pour les matrices :

$$\begin{bmatrix} \mathbb{A}^{11} & \mathbb{A}^{12} \\ \mathbb{A}^{21} & \mathbb{A}^{22} \end{bmatrix} \begin{pmatrix} \overrightarrow{U}^1 \\ \overrightarrow{U}^2 \end{pmatrix} = \begin{pmatrix} \overrightarrow{B}^1 \\ \overrightarrow{B}^2 \end{pmatrix}, \text{ soit } \mathbb{A}\overrightarrow{U} = \overrightarrow{B}, \tag{3.17}$$

avec $k, l = 1, 2$, $\forall i, j \in \mathcal{I}$:

$$\left(\mathbb{A}^{kl}\right)_{ij} = \lambda\int_\Omega \operatorname{div}\mathbf{w}_i^k\,\operatorname{div}\mathbf{w}_j^l\,d\Omega + 2\mu\sum_{m,n=1,2}\int_\Omega \varepsilon_{mn}(\mathbf{w}_i^k)\,\varepsilon_{mn}(\mathbf{w}_j^l)\,d\Omega$$

$$\left(B^k\right)_i = \int_{\Gamma_C}\mathbf{h}\cdot\mathbf{w}_i^k\,d\Gamma + \int_\Omega \mathbf{f}\cdot\mathbf{w}_i^k\,d\Omega.$$

Lorsque les données $\mathbf{h} = \left(h^1, h^2\right)$ et $\mathbf{f} = \left(f^1, f^2\right)$ sont suffisamment régulières, i.e. $\mathbf{h} \in \mathcal{C}^0(\overline{\Gamma}_c)^2$, $\mathbf{f} \in \mathcal{C}^0(\overline{\Omega})^2$, on peut utiliser leur interpolées

$$\begin{cases} \pi_h \mathbf{h} = \displaystyle\sum_{j=1,ns} h^1(M_j)\mathbf{w}_j^1 + h^2(M_j)\mathbf{w}_j^2 \\ \pi_h \mathbf{f} = \displaystyle\sum_{j=1,ns} f^1(M_j)\mathbf{w}_j^1 + f^2(M_j)\mathbf{w}_j^2 \end{cases},$$

pour construire une approximation de $\overrightarrow{B}$. On considère pour cela $\overrightarrow{F}^k$ de composantes $(f^k(M_j))_j$ et $\overrightarrow{H}^k$ de composantes $(h^k(M_j))_j$:

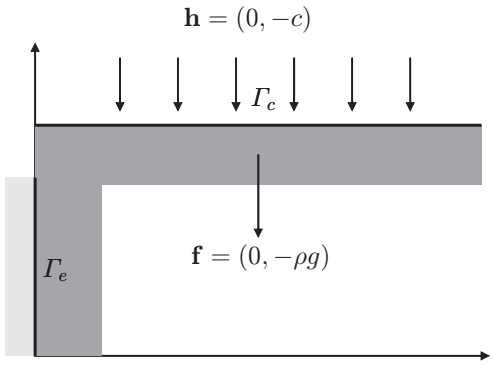
$$\widetilde{B} = \begin{pmatrix} \mathbb{M}\overrightarrow{F}^1 \\ \mathbb{M}\overrightarrow{F}^2 \end{pmatrix} + \begin{pmatrix} \mathbb{M}_c\overrightarrow{H}^1 \\ \mathbb{M}_c\overrightarrow{H}^2 \end{pmatrix} \quad \text{avec } \mathbb{M}_{ij} = \int_\Omega w_i w_j \, d\Omega, \text{et } \left(\mathbb{M}_c\right)_{ij} = \int_{\Gamma_c} w_i w_j \, d\Gamma.$$

Ci-dessus, $\mathbb{M}_c$ est la matrice de masse de bord. Pour construire le système (3.17), il suffit donc d'évaluer les intégrales :

$$\int_\Omega \operatorname{div} \mathbf{w}_i^k \operatorname{div} \mathbf{w}_j^l \, d\Omega, \ \int_\Omega \varepsilon_{mn}(\mathbf{w}_i^k)\, \varepsilon_{mn}(\mathbf{w}_j^l) \, d\Omega, \ \int_\Omega w_i w_j \, d\Omega \text{ et } \int_{\Gamma_c} w_i w_j \, d\Gamma.$$

Ces intégrales sont très voisines de celles déjà calculées dans le cas du problème de Dirichlet pour l'équation de Laplace. En pratique, on calcule ces intégrales, ainsi que la matrice $\mathbb{A}$ sans se préoccuper des conditions d'encastrement ($\forall i, j = 1, ns$). En utilisant une technique de pseudo-élimination similaire à celle précédemment utilisée, on se ramène à un système équivalent au système (3.17). Dans les faits on utilise exactement la même fonction Matlab !
Nous donnons dans ce qui suit le script Matlab permettant de résoudre le problème du chargement d'une étagère idéalisée, de longueur $40cm$, de hauteur $20cm$ et d'épaisseur $2cm$, constituée d'un matériau de type aluminium, via une approximation par éléments finis P^2 composante par composante.

```
%donnees
lambda0=10^10;c=100;                                    %facteur adim.
rho=2800;lambda=1;mu=1;                                 %parametres
g=9.81/lambda0;charge=100/lambda0;
%maillage P2 de l'etagere (plateau P1, support P1, fusion P1, P2)
[S1,T1,BR1,RT1]=triangle_rectangle([0 0.4 0 0.02],200,10,0,[1121]);
[S2,T2,BR2,RT2]=triangle_rectangle([0 0.02 -0.1 0],10,50,0,[1103]);
[SF,TF,BRF,RTF]=fusion_triangulation(S1,T1,BR1,RT1,S2,T2,BR2,RT2);
[SF,TF,BRF,RTF]=maillageP2(SF,TF,BRF,RTF);[TF,SF]=renume(TF,SF);
%calcul des matrices elements finis
[M,DivDiv,EpsEps,Dx,Dy]=matef(SF,TF);                   %matrices volume
[MB2]=calcul_EF_1D(SF,TF,BRF,2);}                       %matrice de bord
K=lambda*DivDiv+2*mu*EpsEps;                            %matrice elasticite
%calcul des chargements
ns=size(SF,1);un=ones(ns,1);ze=zeros(ns,1);
B=[ze;-c*MB2*un-rho*g*M*un];
%elimination des conditions d'encastrement
```




```
Noeud_dir=noeud_bords(SF,TF,BRF,[3]);                %noeuds encastres
ND=[Noeud_dir;Noeud_dir];
[K,B]=cd_Dirichlet(K,B,ND);                          %pseudo-elimination
%resolution et calcul des contraintes
U=K\B                                                %res. du systeme
Def=SF+5000*reshape(U,ns,2);                         %deformee
[Sigma]=cal_sigma(U,M,Dx,Dy,...
        lambda*lambda0,mu*lambda0);                  %contraintes
tr=tresca(Sigma);                                    %critere de Tresca
%visualisation
TF21=isop2(TF);figure;trisurf(TF21,Def(:,1),Def(:,2),tr);
shadinginterp;view(2);axis image;
```

Ce script s'appuie essentiellement sur la fonction **matef.m** qui construit les matrices éléments finis. Nous en donnons ici le détail afin de mettre en évidence l'aspect vectoriel du problème. Il est à noter que les principes de calcul des intégrales sont les mêmes que pour un problème scalaire (boucle sur les triangles, changement de variable et méthode de quadrature numérique) :

```
function [M,DivDiv,EpsEps,Dx,Dy]=matef(S,T)
ns=size(S,1);nt=size(T,1);                           %tailles
DivDiv=sparse(2*ns,2*ns);                            %initialisation
EpsEps=sparse(2*ns,2*ns);M=sparse(ns,ns);            %matrices creuses
Dx=sparse(ns,ns);Dy=sparse(ns,ns);
%boucle sur les triangles
for t=1:nt,                                          %t indice triangle
    I=T(t,:);J=ns+I;                                 %no des dl (ux,uy)
    [Mt,DXX,DYY,DXY,DX,DY]=matelm(S(I,:));           %calcul mat. elementaires
    M(I,I)=M(I,I)+Mt;                                %matrices scalaires
    Dx(I,I)=Dx(I,I)+DX;Dy(I,I)=Dy(I,I)+DY;
    DivDiv(I,I)=DivDiv(I,I)+DXX;                     %matrices assemblées
    DivDiv(J,J)=DivDiv(J,J)+DYY;
    DivDiv(I,J)=DivDiv(I,J)+DXY;
    DivDiv(J,I)=DivDiv(J,I)+DXY';
    EpsEps(I,I)=EpsEps(I,I)+DXX+DYY/2;
    EpsEps(J,J)=EpsEps(J,J)+DXX+DYY/2;
    EpsEps(I,J)=EpsEps(I,J)+DXY/2;
    EpsEps(J,I)=EpsEps(J,I)+DXY'/2;
end,
```

La fonction **matelm.m** réalise les calculs des intégrales élémentaires (scalaires) :

$$\int_{T_\ell} \tau_i^\ell \tau_j^\ell \, d\Omega, \int_{T_\ell} \partial_x \tau_i^\ell \partial_x \tau_j^\ell \, d\Omega, \int_{T_\ell} \partial_y \tau_i^\ell \partial_y \tau_j^\ell \, d\Omega \ , \int_{T_\ell} \partial_x \tau_i^\ell \partial_y \tau_j^\ell \, d\Omega \ ,$$

$$\int_{T_\ell} \tau_i^\ell \partial_x \tau_j^\ell \, d\Omega \ , \int_{T_\ell} \tau_i^\ell \partial_y \tau_j^\ell \, d\Omega$$

suivant le même principe que le calcul présenté au §3.3.1. Les matrices **DX** et **DY** sont utilisées pour calculer les contraintes.

```
function [M,DXX,DYY,DXY,DX,DY]=matelm(Noeuds)
%initialisation
nd=size(Noeuds,1);
DXX=zeros(nd,nd);DYY=zeros(nd,nd);DXY=zeros(nd,nd);
M=zeros(nd,nd);DX=zeros(nd,nd);DY=zeros(nd,nd);
%formule de quadrature a 7 points
nbq=7;os=sqrt(15);s3=1./3.;
pp1=(6.-os)/21.;pp2=(6.+os)/21.;
pp3=(9.+2.*os)/21.;pp4=(9.-2.*os)/21.;
pts_quadT=[s3 s3;pp1 pp1;pp1 pp3;pp3 pp1;          %pts de quad.
           pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;}}
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];        %poids de quad.
%jacobien de la transformation (affine)
S21=Noeuds(2,:)-Noeuds(1,:);S31=Noeuds(3,:)-Noeuds(1,:);
delta=S21(1)*S31(2)-S21(2)*S31(1);
Jflmt=[S31(2)-S21(2);-S31(1)S21(1)]/delta;
%boucle sur les pts de quadrature
for k=1:nbq,                                       %k no du point
    x=pts_quadT(k,1);y=pts_quadT(k,2);             %pts de quad.
    if(nd==3) w=[1-x-yxy];gw=[-1 1 0;-1 0 1];      %fonctions P1
    else a=1-x-y;b=2*x-1;c=2*y-1;                  %fonctions P2
       w=[a*(1-2*x-2*y),b*x*,c*y*,4*x*a,4*x*y,4*y*a];
       gw=[4*(x+y)-3 4*x-1 0 4*(1-2*x-y) 4*y -4*y;
        4*(x+y)-3 0 4*y-1 -4*x 4*x 4*(1-x-2*y)];
    end,
pk=pds_quadT(k)*abs(delta);
M=M+pk*w'*w;                                       %wi.wj
jx=Jflmt(1,:)*gw;jy=Jflmt(2,:)*gw;
DXX=DXX+pk*jx'*jx;                                 %dx(wi)dx(wj)
DYY=DYY+pk*jy'*jy;                                 %dy(wi)dy(wj)
DXY=DXY+pk*jx'*jy;                                 %dx(wi)dy(wj)
DX=DX+pk*w'*jx;                                    %wi.dx(wj)
DY=DY+pk*w'*jy;                                    %wi.dy(wj)
end,
```

Dans cet exemple on a également besoin de calculer la matrice de masse de bord $\mathbb{M}_c$. Cette opération est réalisée par une technique similaire à celle utilisée pour calculer les matrices éléments finis hormis le fait qu'il convient de restreindre les fonctions de bases, ainsi que le domaine d'intégration aux arêtes des triangles situées sur le bord concerné par le calcul. Elle est réalisée par la fonction **calcul_EF_1D.m** :

```
function [M]=calcul_EF_1D(S,T,BR,ref_bord)
s35=sqrt(3./5); pts_quadS=0.5*[1-s35 1 1+s35];        %quadrature 1D
os=1/18; pds_quadS=os*[5 8 5];
nt=size(T,1); ns=size(S,1); q=size(T,2);              %initialisation
nbq=length(pds_quadS);
M=sparse(ns,ns);
%boucle sur les aretes des triangles
for t=1:nt,
    for a=1:3,
        as=mod(a,3)+1; I=Numtri(t,a); J=Numtri(t,as);%num. arete
        if(ismember(BR(t,a),ref_bord))                %arete sur le bord
           L=norm(S(I,:)-S(J,:));                     %longueur de l'arete
           in=[I J]; if(q==6) in=[in T(t,3+a)];end,   %numerotation globale
           for k=1:nbq,                               %boucle pts quad.
               x=pts_quadS(k); c=L*pds_quadS(k);
               if(q==3) w=[1-x x];                    %fonctions P1
               else w=[(1-2*x)*(1-x) (2*x-1)*x 4*x*(1-x)];  % P2
               end,
               M(in,in)=M(in,in)+c*w'*w;              %assemblage
           end,
        end,
    end,
end,
```

Le calcul des contraintes et par là-même du critère de Tresca requiert la connaissance des dérivées du champ de déplacement approché $\mathbf{u}_h$. Dans le cadre de l'approximation par éléments finis de Lagrange, $\mathbf{u}_h$ n'est pas dérivable (seulement continu) aux interfaces des triangles. Pour construire une approximation de ces dérivées on a essentiellement deux possibilités : soit se contenter d'une représentation discontinue des dérivées, soit construire une approximation en chaque nœud du maillage en moyennant les dérivées obtenues sur chaque triangle possédant ce nœud. Dans le second cas, pour obtenir une moyenne pondérant correctement le poids de chaque triangle il suffit de considérer le problème suivant, en notant $\mathbf{v}_x$ l'approximation P^1 ou P^2 de $\partial_x \mathbf{u}_h$ (idem avec $\mathbf{v}_y \approx \partial_y \mathbf{u}_h$) :

$$\int_\Omega \mathbf{v}_x w_i d\Omega = \int_\Omega \partial_x \mathbf{u}_h w_i d\Omega \quad \forall i = 1, ns$$

conduisant au système linéaire :

$$\begin{bmatrix} \mathbb{M} & 0 \\ 0 & \mathbb{M} \end{bmatrix} \begin{bmatrix} \overrightarrow{V}_x^1 \\ \overrightarrow{V}_x^2 \end{bmatrix} = \begin{bmatrix} \mathbb{D}_x & 0 \\ 0 & \mathbb{D}_x \end{bmatrix} \begin{bmatrix} \overrightarrow{U}^1 \\ \overrightarrow{U}^2 \end{bmatrix}$$

avec $\mathbb{M}$ la matrice de masse et $\mathbb{D}_x$ la matrice de terme $\int_\Omega w_i \, \partial_x w_j \, d\Omega$. Rappelons que l'ordre d'approximation des dérivées est d'un ordre inférieur à celui des valeurs.
La fonction **cal_sigma.m** calcule les dérivées suivant ce principe et construit le tenseur de contraintes en tout nœud du maillage :

```
function [Sigma]=cal_sigma(U,M,Dx,Dy,lambda,mu)
ns=size(M,1); Ux=U(1:ns); Uy=U(ns+1:2*ns);          %initialisation
DxUx=M\(Dx*Ux); DxUy=M\(Dx*Uy);                     %calcul des derivees
DyUx=M\(Dy*Ux); DyUy=M\(Dy*Uy);
Sigma(:,1)=lambda*(DxUx+DyUy)+2*mu*DxUx;            %calcul de sigma
Sigma(:,4)=lambda*(DxUx+DyUy)+2*mu*DyUy;
Sigma(:,2)=mu*(DxUy+DyUx); Sigma(:,3)=Sigma(:,2);
```

Enfin, le critère de Tresca permet de représenter les efforts de cisaillement et est donné par :

$$tr = |\lambda_1 - \lambda_2|, \text{ avec } \lambda_1, \lambda_2 \text{ valeurs propres du tenseur des contraintes.}$$

La fonction **tresca.m** réalise, à partir de la connaissance du tenseur des contraintes, le calcul du critère de Tresca en tout nœud du maillage :

```
function [tr]=tresca(Sigma,S,T)
ns=size(Sigma,1);                                   %initialisation
for i=1:ns,                                         %boucle noeuds
    sp=eig([Sigma(i,1),Sigma(i,2);
            Sigma(i,3),Sigma(i,4)]);                %val. prop.
    tr(i)=abs(sp(1)-sp(2));                         %Tresca
end,
```

Nous donnons sur la figure 3.7 le résultat du calcul de l'étagère. On présente sur le même graphique la déformation de l'étagère (exagérée), ainsi que le critère de Tresca indiquant les zones de fortes contraintes de cisaillement.

3.3.3 Quelques outils élémentaires de maillage 2D

Dans cette dernière section, nous donnons les codes Matlab des outils élémentaires de maillage 2D qui ont été utilisés dans les exemples précédents. L'outil de base (**triangule_carre.m**) est celui qui réalise le maillage "structuré" du carré unité, suivant une grille comprenant m (resp. n) intervalles suivant x (resp. y), un tableau de référence de matériau ref_mat (0 par défaut) et un tableau de référence de bord ref_bord. Chaque rectangle élémentaire de cette grille est découpé en deux triangles suivant la première ou la seconde bissectrice. Ce découpage est réalisé de façon alternée afin de ne pas induire une trop forte dissymétrie du maillage qui génére un biais systématique dans les calculs. Cette fonction produit le tableau S (dimension $ns \times 2$) des cooordonnées des sommets de la triangulation, le tableau T (dimension $nt \times 3$) de la numérotation des sommets, le tableau RT (dimension nt) indiquant une référence pour chaque triangle et le tableau BR (dimension $nt \times 3$) indiquant pour chaque arête des triangles une référence de bord (0 pour une arête interne).

```
function [S,T,BR,RT]=triangule_carre(m,n,ref_mat,ref_bord)

%initialisation
```

```
dx=1./m; dy=1./n; x=0:dx:1.; y=0:dy:1.;
ns=(m+1)*(n+1); nt=2*m*n;
BR=zeros(nt,3);
if(nargin==2) ref_mat=0; ref_bord=[1 2 3 4]; end,
if(nargin==3) ref_bord=[1 2 3 4]; end,
RT=ref_mat*ones(1,nt);
%définition des sommets
[X,Y]=meshgrid(x,y);
S=[reshape(X',ns,1) reshape(Y',ns,1)];
%définition de la numérotation des triangles
N1=[]; N2=[]; mp=m+1;
%ligne découpage de type 1
for i=1:m
    if(mod(i,2)==1)  %i impair
        N1=[N1 i i+1 mp+i i+1+mp i+mp i+1];
    else             %i pair
        N1=[N1 i+1 i+1+mp i i+mp i i+1+mp];
    end,
end,
N1=reshape(N1,3,2*m)';
%ligne découpage de type 2
for i=1:m
    if(mod(i,2)==1)  %i impair
        N2=[N2 i+1 i+1+mp i i+mp i i+1+mp];
    else             %i pair
        N2=[N2 i i+1 mp+i i+1+mp i+mp i+1];
    end,
end,
N2=reshape(N2,3,2*m)';
% maillage par alternance des lignes 1 et 2 (translation de numérotation)
T=[];
for j=1:n
    if(mod(j,2)==1) T=[T;N1+(j-1)*mp];
    else T=[T;N2+(j-1)*mp];
    end,
end,
%numérotation des bords
p=2*(n-1)*m+1;
p1=ref_bord(3); p2=0; if(mod(n,2)==0) p2=ref_bord(3); p1=0; end,
for i=1:4:2*m,
    BR(i,:)=[ref_bord(1) 0 0];
    if(i+2<=2*m) BR(i+2,:)=[0 0 ref_bord(1)]; end,
    BR(p+i,:)=[p1 0 p2];
    if(p+i+2<=nt) BR(p+i+2,:)=[p2 0 p1]; end,
end,
%bords gauche et droit (4 2)
t=2*m+1; r=2*(n-1)*m+2;
s=2*(m-mod(m,2))+1; q=2*m-2+mod(m,2);
q1=ref_bord(2); q2=0; if(mod(m,2)==0) q2=ref_bord(2); q1=0; end,
for i=1:4*m:2*(n-1)*m+1,
    BR(i,:)=BR(i,:)+[0 0 ref_bord(4)];
    if(i+t<=r) BR(i+t,:)=BR(i+t,:)+[ref_bord(4) 0 0]; end,
    BR(q+i,:)=BR(q+i,:)+[q2 0 q1];
    if(q+i+s<=nt) BR(q+i+s,:)=BR(q+i+s,:)+[q1 0 q2]; end,
end,
```

A titre d'exemple, nous donnons sur la figure 3.9 le maillage P^1 obtenu par la commande :

```
[S,T,BR,RT]=triangle_carre(10,20,0,[1 1 2 1]);
plot_triangulation(S,T,BR,RT,0);
```

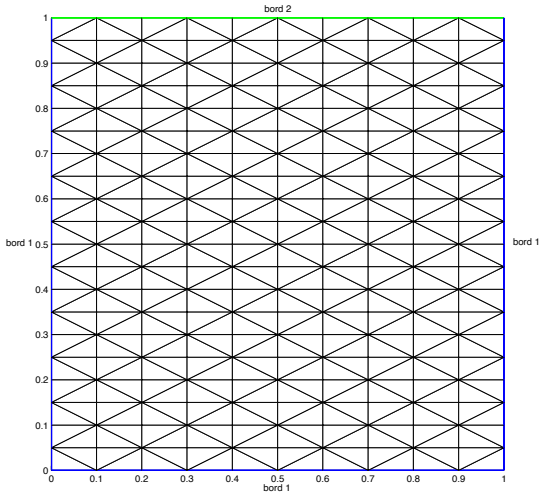


Fig. 3.9. Maillage régulier du carré unité

Le maillage d'un rectangle (pas trop étiré) se déduit par homotétie du maillage d'un carré.

```
function [S,T,BR,RT]=triangle_rectangle(rect,m,n,ref_mat,ref_bord)
if(nargin==3) ref_mat=0;ref_bord=[1 2 3 4];end,
if(nargin==4) ref_bord=[1 2 3 4];end,
[S,T,BR,RT]=triangle_carre(m,n,ref_mat,ref_bord);
%transf des coordonnées S
l=rect(2)-rect(1);S(:,1)=rect(1)+l*S(:,1);
l=rect(4)-rect(3);S(:,2)=rect(3)+l*S(:,2);
```

La commande suivante :

```
[S,T,BR,RT]=triangle_rectangle([-1 1 0 1],20,10,0,[1 1 2 1]);
plot_triangulation(S,T,BR,RT,0);
```

produit le maillage représenté sur la figure 3.10.
L'utilisation d'un outil général de fusion de maillage (**fusion_triangulation.m**) permet de générer des maillages de domaines géométriques plus complexes.

```
function [SF,TF,BRF,RTF]=fusion_triangulation(S1,T1,BR1,RT1,S2,T2,BR2,RT2)
%recherche des noeuds en coincidence
%à eps près défini par la taille du domaine /10000
xmax=max(max(S1(:,1)),max(S2(:,1)));
xmin=min(min(S1(:,1)),min(S2(:,1)));
ymax=max(max(S1(:,2)),max(S2(:,2)));
ymin=min(min(S1(:,2)),min(S2(:,2)));
eps=max(xmax-xmin,ymax-ymin)/10000.;
%fabrication du tableau d'arêtes
```

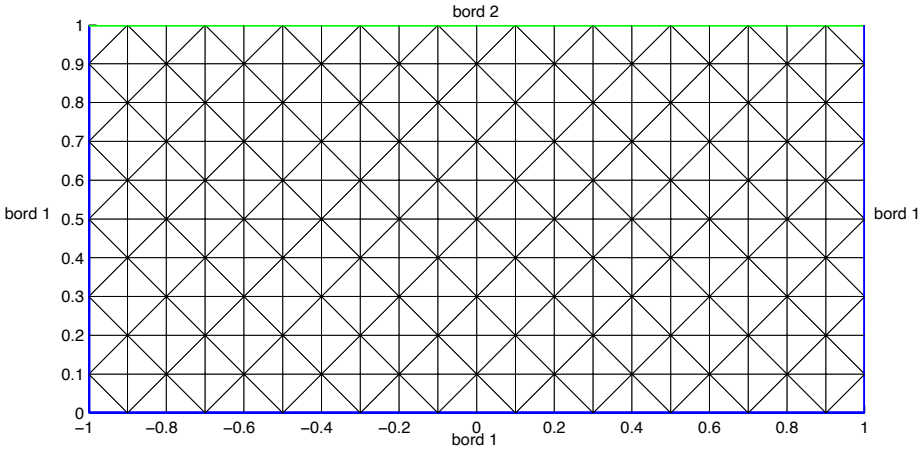


Fig. 3.10. Maillage régulier d'un rectangle

```
A1=graphe_arete(S1,T1);
L1=detecte_noeuds_bord(A1);
A2=graphe_arete(S2,T2);
L2=detecte_noeuds_bord(A2);
%noeud en coincidence (recherche brutale ...)
ns1=size(S1,1);ns2=size(S2,1);
nt1=size(T1,1);nt2=size(T2,1);
C=zeros(ns2,1);D=zeros(ns1,1);
for i=1:size(L1,1),
    I=L1(i);
    for j=1:size(L2,1),
        J=L2(j);
        if(norm(S1(I,:)-S2(J,:))<eps)    %coincidence
            C(J)=I;    %Numéro J de S2 devient I de S1
            D(I)=J;    %Numéro J de S2 devient I de S1
            %break;
        end,
    end,
end,
%fusion des triangulations
SF=S1;TF=[T1;T2];
BRF=[BR1;BR2];RTF=[RT1 RT2];
%renumérotation de la deuxième triangulation
n2=0;
for i=1:ns2,
    if(C(i)==0)
        n2=n2+1;C(i)=ns1+n2;
        SF=[SF;S2(i,:)];
    end,
end,
%maj numérotation des triangles (2 eme partie)
for t=1:nt2,
    for i=1:3,
        I=T2(t,i);
        TF(nt1+t,i)=C(I);
    end,
end,
%suppression d'arête du bord devenue interne
```

```
for t=1:nt1;
    for a=1:3,
        if(BRF(t,a)~=0)
            I=T1(t,a); if(a==3) J=T1(t,1); else J=T1(t,a+1);end,
            if(D(I)~=0 && D(J)~=0) BRF(t,a)=0;end,    %arête commune
        end,
    end,
end,
for t=1:nt2;
    for a=1:3,
        if(BRF(nt1+t,a)~=0)
            I=T2(t,a); if(a==3) J=T2(t,1); else J=T2(t,a+1);end,
            if(C(I)<=ns1 && C(J)<=ns1) BRF(nt1+t,a)=0;end,   %arête commune
        end,
    end,
end,
```

Nous illustrons l'utilisation de cette fonction, en "collant" deux maillages de rectangles (voir figure 3.11) :

```
%maillages P1
[S1,T1,BR1,RT1]=triangle_rectangle([0 1 0 2],10,20,0,[1 1 1 1]);
[S2,T2,BR2,RT2]=triangle_rectangle([1 2 0 1],10,10,0,[1 1 1 1]);
%fusion des maillages
[SF,TF,BRF,RTF]=fusion_triangulation(S1,T1,BR1,RT1,S2,T2,BR2,RT2);
plot_triangulation(SF,TF,BRF,RTF,0);
```

produit le maillage représenté sur la figure 3.10.

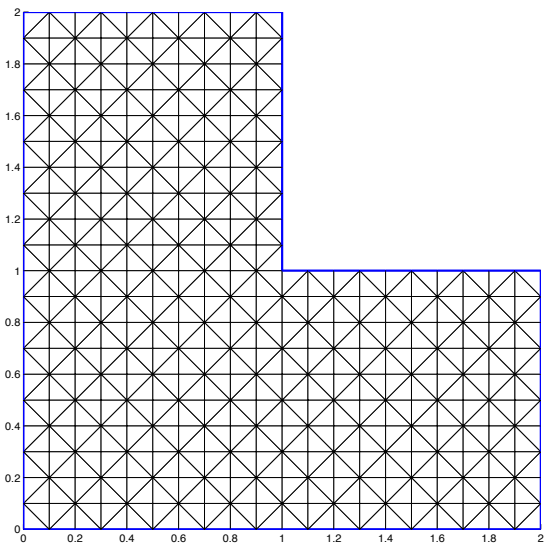


Fig. 3.11. Fusion de deux maillages (compatibles)

Il est assez facile de générer un maillage P^2 à partir d'un maillage P^1 en ajoutant les points milieux des arêtes de la triangulation P^1. Attention, toutefois, que lorsque la frontière du domaine de calcul est courbe, on perd de la précision car on n'a pas réalisé un maillage isoparamétrique P^2 du domaine (cf. §2.3.3). La fonction Matlab **maillageP2.m** réalise cette opération.

```
function [S2,T2,BR2,RT2]=maillageP2(S,T,B,R)
%Initialisation
ns=size(S,1); nt=size(T,1);
ns2=ns+3*nt;  %surestimation du nombre de points P2
S2=zeros(ns2,2); S2(1:ns)=S;
T2=zeros(nt,6); T2(:,1:3)=T;
Reftri2=Reftri; Bord2=Bord;
%graphe des arêtes
Arete=sparse(ns,ns); n=ns;
%boucle sur les triangles
for l=1:nt,
    %boucle sur les arêtes
    for k=1:3,
        I=T(l,k);
        J=T(l,mod(k,3)+1);
        if(J<I) I1=J; I2=I; else I1=I; I2=J; end,
        %mise a jour
        t=Arete(I1,I2);
        if(t==0) %ajout d'un point
            Arete(I1,I2)=l+k/10.;  %codage n° triangle + N° arête
            n=n+1;
            S2(n,:)=(S2(I1,:)+S2(I2,:))/2.;
            T2(l,3+k)=n;
        else, %point déjà ajouté
            q=floor(t); a=round(10*(t-q));
            T2(l,3+k)=T2(q,3+a);
        end,
    end,
end,
%redimensionnement de S2
S2=S2(1:n,:);
```

Lorqu'on réalise un maillage P^2 à partir d'un maillage P^1, la numérotation des nœuds n'est pas du tout optimale en terme de profil de la matrice (voir annexe suivante). Il peut être intéressant de trouver de meilleures numérotations (non optimales mais quasi optimales). Ces renumérotations sont basées sur des algorithmes de Cuthill-Mackee ou reverse Cuthill-Mackee qui sont disponibles dans Matlab. Par exemple, la fonction **renume.m** suivante s'appuie sur de tels algorithmes pour optimiser une numérotation que ce soit pour des maillages P^1 ou P^2.

```
function [T_s,S_s]=renume(T,S)
%tailles diverses
nt=size(Numtri,1); ns=size(Corneu,1); q=size(Numtri,2);
%matrice du graphe de connexion
G=sparse(ns,ns);
for t=1:nt,
    for i=1:q,
```

```
            I=Numtri(t,i);
            for j=1:q,
                J=Numtri(t,j);
                G(I,J)=1;
            end,
        end,
end,
%optimisation (reverse Cuthill-Mackee)
R = symrcm(G);
%renumérotation des noeuds
S_s=S(R,:);
T_s=zeros(size(T));
Rm=zeros(ns,1);
%permutation inverse de R
for n=1:ns,
    Rm(R(n))=n;
end,
%renumérotation des triangles
for t=1:nt,
    for i=1:q,
        T_s(t,i)=Rm(T(t,i));
    end,
end,
```

Afin de prendre en compte des conditions essentielles, par exemple des conditions de Dirichlet ou d'encastrement en élasticité, il est utile de disposer de petits outils permettant de détecter des nœuds d'un bord identifié par une référence. Ainsi, on utilise la fonction **detecte_noeuds_bord.m** basée sur une liste d'arêtes obtenues à l'aide de la fonction **graphe_arete.m** :

```
function L=detecte_noeuds_bord(A)
[r,c,v]=find(A==1);     %indices des arêtes non partagées
L=unique(r);
```

```
function A=graphe_arete(S,T)
ns=size(S,1);nt=size(T,1);A=sparse(ns,ns);
for t=1:nt,
    I=T(t,1);J=T(t,2);K=T(t,3);
    A(I,J)=A(I,J)+1;
    A(J,I)=A(J,I)+1;
    A(I,K)=A(I,K)+1;
    A(K,I)=A(K,I)+1;
    A(J,K)=A(J,K)+1;
    A(K,J)=A(K,J)+1;
end,
```

Pour terminer signalons qu'il peut-etre utile de disposer d'outils de visualisation de maillage un peu plus élaborés que ceux fournis par Matlab (*trimesh*). Ainsi, on peut utiliser par exemple la fonction **plot_triangulation.m** suivante, qui propose l'affichage des numérotations et des références de bord :

```
function plot_triangulation(S,T,BR,RT,num)
%fabrication d'une liste de couleurs par défaut (11 couleurs)
couleur=[1 1 1; 0 0 1; 0 1 0; 1 0 0; 1 1 0; 0 1 1; 1 0 1;
```

```
          0.5  0.5  0.5;  0.5  0.5  0;  0  0.5  0.5;  0.5  0  0.5];
% définition des couleurs
reft=unique(RT);n=length(reft);refc=11*ones(1,n);c=0;m=min(reft)+1;
for i=1:min(n,11) c=c+1;refc(reft(i)+m)=c;end;
%dessin du maillage
figure;hold on;rmax=0;
for l=1:size(T,1)
    X=[S(T(l,1),1) S(T(l,2),1) S(T(l,3),1)];
    Y=[S(T(l,1),2) S(T(l,2),2) S(T(l,3),2)];
    i=refc(RT(l)+m);c=couleur(i,:);     patch(X,Y,c);
    %couleur des bords
     for k=1:3,
         ks=mod(k,3)+1;
         r=BR(l,k);if(r>rmax) rmax=r;end,
         if(r>0 && r<11)
            line([X(k) X(ks)],[Y(k) Y(ks)],'LineWidth',...
                 2,'Color',couleur(r+1,:));
         end,
     end,
    if(num>0) %affichage de la numérotation
        if(num>=10) %affichage de la numérotation des triangles
            G=[X;Y]*[1;1;1]/3;   %barycentre du triangle
            text(G(1),G(2),int2str(l),'FontSize',8);
        end,
        if(num~=10) %affichage de la numérotation des noeuds P1
                for k=1:3,
                    text(X(k),Y(k),int2str(T(l,k)),...
                         'FontSize',8,'BackgroundColor',[1 1 1]);
                end,
        end,
        if((num==2 || num==12 )&& size(T,2)==6) % numérotation P2
            for k=1:3,
                k2=mod(k,3)+1;
                text((X(k)+X(k2))*.5,(Y(k)+Y(k2))*.5, ...
                      int2str(T(l,k+3)),'BackgroundColor', ...
                      [1 1 1],'FontSize',8);
            end,
        end,

    end,
end
```

Notons que si l'on souhaite utiliser la primitive *trisurf* de Matlab pour dessiner des isosurfaces, cette dernière ne fonctionnant qu'avec des triangulations P^1, il est nécessaire de convertir les triangulations P^2 en triangulations P^1. C'est l'objet de la fontion **isop2.m** suivante, basée sur un redécoupage en quatre triangles P^1 d'un triangle P^2 :

```
function [T1]=isop2(T2)
nt=size(T2,1);T1=zeros(4*nt,3);
k=0;
for t=1:nt,
    k=k+1; T1(k,1)=T2(t,1);T1(k,2)=T2(t,4);T1(k,3)=T2(t,6);
    k=k+1; T1(k,1)=T2(t,4);T1(k,2)=T2(t,2);T1(k,3)=T2(t,5);
    k=k+1; T1(k,1)=T2(t,4);T1(k,2)=T2(t,5);T1(k,3)=T2(t,6);
    k=k+1; T1(k,1)=T2(t,3);T1(k,2)=T2(t,6);T1(k,3)=T2(t,5);
end;
```

A

Résolution des systèmes linéaires

La discrétisation par éléments finis des problèmes elliptiques linéaires conduit à des systèmes linéaires de grande taille, dont la structure des matrices est très particulière (matrice "creuse", c'est-à-dire présentant un très grand nombre de termes nuls). Il est donc important de connaître les principales techniques de résolution des systèmes linéaires afin de choisir la plus judicieuse suivant la nature de la matrice du système. Cette annexe est destinée à tous ceux qui seraient confrontés à des questions de mise en œuvre de méthodes d'élements finis en Fortran, en C, C++, ... Dans l'environnement Matlab ces questions de stockage de matrice et d'algorithmes de résolution de système sont masquées par le langage (*sparse matrix)* mais il n'en demeure pas moins qu'il reste intéressant de connaître ces techniques.
Il n'est pas question de présenter, ici, toute la théorie des systèmes linéaires. Ce sujet est traité abondamment dans la littérature classique. Nous nous attacherons à présenter les principaux algorithmes (méthodes directes et méthodes itératives) et nous insisterons plus particulièrement sur l'influence de structures de stockage des matrices creuses sur le choix d'un algorithme de résolution.
Dans la suite, nous considérerons pour commencer la résolution d'un système linéaire d'ordre n :

$$\mathbb{A}X = B \tag{A.1}$$

où $\mathbb{A}$ est une matrice d'ordre n inversible, B un vecteur de $\mathbb{R}^n$ donné et X désigne le vecteur solution. Dans cette annexe, dans un souci de lisibilité, nous omettrons les flèches sur les vecteurs.
Nous commencerons par rappeler quelques éléments de la théorie de la propagation des erreurs numériques lors d'une résolution d'un système linéaire. Ensuite, nous présenterons, d'une part, les méthodes de résolution, dites *directes*, conduisant à la solution en un nombre fini d'opérations et d'autre part, les méthodes qualifiées d'*itératives*, générant une suite de vecteurs qui converge vers la solution du système linéaire. Ces méthodes sont d'abord exposées dans le cas de matrices

"pleines". Leur adaptation aux structures de stockage complexes est abordée dans une seconde partie.

A.1 Propagation des erreurs numériques

Il s'agit d'estimer l'erreur que l'on commet sur la solution lorsque l'on résout, non pas le système linéaire (A.1), mais un système perturbé :

$$(\mathbb{A} + \delta\mathbb{A})Y = B + \delta B \tag{A.2}$$

où $\delta\mathbb{A}$ et δB représentent respectivement des erreurs sur la matrice du système et le second membre. Dans la pratique, ces erreurs proviennent, d'une part, d'écarts sur les données initiales dus aux arrondis effectués par la machine et d'autre part de la propagation de ces erreurs d'arrondi lors du déroulement de la résolution du système linéaire.

A.1.1 Estimations d'erreur

Si on considère qu'il n'y a pas d'erreur d'arrondi dans la méthode de résolution, mais uniquement une erreur initiale sur la donnée B, alors le problème perturbé (A.2) prend la forme :

$$\mathbb{A}Y = B + \delta B$$

pour laquelle il est facile de montrer que :

$$| Y - X | \leq \| \mathbb{A}^{-1} \| \| \delta B |$$

qui coïncide avec la notion de stabilité habituelle. Dans toute la suite, $| \, . \, |$ désigne une norme quelconque sur $\mathbb{R}^n$ et $\|.\|$ la norme matricielle subordonnée à cette norme. Par contre, si on prend en compte les erreurs d'arrondi, qui existent toujours lors d'une résolution sur ordinateur, l'erreur sur la solution est alors liée au *conditionnement* de la matrice $\mathbb{A}$, noté $\gamma(\mathbb{A})$ et défini par :

$$\gamma(\mathbb{A}) = \| \mathbb{A} \| \| \mathbb{A}^{-1} \| \tag{A.3}$$

qui lorsque la matrice $\mathbb{A}$ est hermitienne est aussi le rapport de la plus grande valeur propre sur la plus petite valeur propre (en module), si l'on a choisi la norme euclidienne $| \cdot |_2$. On a le résultat fondamental suivant :

Proposition A.1. *Soient $\mathbb{A}$ une matrice d'ordre n inversible, $\delta\mathbb{A}$ une matrice d'ordre n, B et δB deux vecteurs de $\mathbb{R}^n$.*
i) Si $\|\delta\mathbb{A}\| < \dfrac{1}{\|\mathbb{A}^{-1}\|}$ alors $(\mathbb{A} + \delta\mathbb{A})$ est inversible.

ii) Si X désigne la solution du système linéaire $\mathbb{A}X = B$ et $Y = X+\delta X$ la solution du système linéaire $(\mathbb{A}+\delta\mathbb{A})Y = B+\delta B$ alors on a l'estimation d'erreur :

$$\frac{|\delta X|}{|X|} \leq \frac{\gamma(\mathbb{A})}{1-\gamma(\mathbb{A})\frac{\|\delta\mathbb{A}\|}{\|\mathbb{A}\|}}\left(\frac{|\delta B|}{|B|}+\frac{\|\delta\mathbb{A}\|}{\|\mathbb{A}\|}\right) \tag{A.4}$$

Démonstration : i) Comme $\mathbb{A}$ est inversible, on a :

$$\mathbb{A}+\delta\mathbb{A} = \mathbb{A}\left(\mathbb{I}+\mathbb{A}^{-1}\delta\mathbb{A}\right)$$

qui d'après l'hypothèse sur $\delta\mathbb{A}$ montre que :

$$\left(\mathbb{I}+\mathbb{A}^{-1}\delta\mathbb{A}\right) = \mathbb{I}-\mathbb{C} \quad \text{avec } \|\mathbb{C}\| < 1.$$

Posons :

$$\mathbb{S}_n = \sum_{j=0}^{n}\mathbb{C}^j.$$

$\mathbb{S}_n$ converge vers une matrice $\mathbb{S}$ car la série est normalement convergente. En outre, on a :

$$\mathbb{S}_n(\mathbb{I}-\mathbb{C}) = \mathbb{I}-\mathbb{C}^{n+1} \quad \text{et} \quad \lim_{n\to\infty}\mathbb{C}^n = 0$$

ce qui prouve que :

$$\mathbb{S}(\mathbb{I}-\mathbb{C}) = \mathbb{I}.$$

On montre de même que $(\mathbb{I}-\mathbb{C})\mathbb{S} = \mathbb{I}$ et par conséquent que $(\mathbb{I}-\mathbb{C})$ est inversible, d'inverse :

$$\mathbb{S} = \sum_{j=0}^{\infty}\mathbb{C}^j.$$

ii) Il est facile de montrer que :

$$\|\mathbb{S}\| \leq \frac{1}{1-\|\mathbb{C}\|}.$$

En appliquant cette estimation $\left(\mathbb{C} = -\mathbb{A}^{-1}\delta\mathbb{A}\right)$, on obtient :

$$\left\|(\mathbb{A}+\delta\mathbb{A})^{-1}\right\| \leq \frac{\left\|\mathbb{A}^{-1}\right\|}{1-\|\mathbb{A}^{-1}\|\,\|\delta\mathbb{A}\|}$$

car on a :

$$(\mathbb{A}+\delta\mathbb{A})^{-1} = \left(\mathbb{I}+\mathbb{A}^{-1}\delta\mathbb{A}\right)^{-1}\mathbb{A}^{-1}.$$

Or on a :

$$(\mathbb{A}+\delta\mathbb{A})\delta X = B+\delta B-\mathbb{A}X-\delta\mathbb{A}X = \delta B-\delta\mathbb{A}X,$$

soit :

$$\delta X = (\mathbb{A}+\delta\mathbb{A})^{-1}(\delta B-\delta\mathbb{A}X)$$

d'où on déduit l'estimation :

$$|\delta X| \leq \frac{\left\|\mathbb{A}^{-1}\right\|}{1-\|\mathbb{A}^{-1}\|\,\|\delta\mathbb{A}\|}(|\delta B|+\|\delta\mathbb{A}\||X|),$$

qui conduit à (A.4). ∎

La condition $\|\delta\mathbb{A}\| < \frac{1}{\|\mathbb{A}^{-1}\|}$ assure que $1 > 1-\gamma(\mathbb{A})\frac{\|\delta\mathbb{A}\|}{\|\mathbb{A}\|} > 0$. L'erreur relative sur la solution X est donc proportionnelle aux erreurs relatives sur la matrice et sur le second membre avec un coefficient de proportionnalité supérieur au conditionnement de la matrice. Par conséquent, si le conditionnement est grand, la solution du système linéaire est entachée d'une erreur importante. Nous allons illustrer ce phénomène sur un exemple simple.

A.1.2 Exemple

Considérons l'exemple suivant :

$$\mathbb{A} = \begin{bmatrix} 2.23 & 1.2 \\ 1.3 & 0.7 \end{bmatrix} \quad \text{et} \quad B = \begin{bmatrix} -2.06 \\ -1.2 \end{bmatrix}.$$

Le système linéaire $\mathbb{A}X = B$ admet pour solution :

$$X = \begin{bmatrix} -2 \\ 2 \end{bmatrix}.$$

Considérons maintenenant la perturbation :

$$\delta B = \begin{bmatrix} 10^{-3} \\ 10^{-3} \end{bmatrix}.$$

Le système linéaire $\mathbb{A}Y = B + \delta B$ admet pour solution :

$$Y = \begin{bmatrix} -2.5 \\ 2.93 \end{bmatrix}!$$

Calculons le conditionnement de la matrice $\mathbb{A}$ à l'aide de la norme matricielle :

$$\|\mathbb{A}\|_\infty = \max_i \sum_j |\mathbb{A}_{ij}|$$

On a :

$$\|\mathbb{A}\|_\infty = 3.43 \quad \text{et} \quad \left\|\mathbb{A}^{-1}\right\|_\infty = 3.53\,10^3$$

qui donne pour le conditionnement (norme ∞) :

$$\gamma_\infty(\mathbb{A}) \simeq 1.2\,10^4.$$

Il est facile de construire de nombreux exemples. Signalons celui de Rutihauser :

$$\mathbb{A} = \begin{bmatrix} 10 & 1 & 4 & 0 \\ 1 & 10 & 5 & -1 \\ 4 & 5 & 10 & 7 \\ 0 & -1 & 7 & 9 \end{bmatrix} \quad \text{avec } B = \begin{bmatrix} 15 \\ 15 \\ 26 \\ 15 \end{bmatrix}$$

dont la solution de $\mathbb{A}X = B$ est $X = (1,1,1,1)^t$ avec $\gamma_\infty(\mathbb{A}) = 6.26\,10^4$. Les matrices de Hilbert $\left(\mathbb{A}_{ij} = \frac{1}{i+j}\right)$ fournissent un exemple de matrices inversibles ayant un conditionnement qui tend vers l'infini lorsque l'ordre de la matrice tend vers l'infini.

Lorsque le conditionnement d'une matrice est grand, on parle de matrice *mal conditionnée*. Aucune méthode numérique ne permet de s'affranchir de cette difficulté.

Il faut augmenter la précision de la machine en travaillant en double précision, voire en quadruple précision, ou utiliser des méthodes de correction des erreurs d'arrondi afin d'obtenir la solution du système linéaire avec la précision désirée. Par contre, il faut prendre garde au fait qu'une méthode numérique peut renforcer l'effet d'un mauvais conditionnement dès lors que cette méthode est très peu stable (amplification des erreurs d'arrondi).

A.2 Méthodes directes de résolution

Rappelons que ces méthodes conduisent à la solution du système linéaire en un nombre *fini* d'opérations. Ce qui va les différencier l'une de l'autre est bien évidemment le nombre d'opérations qu'il est nécessaire d'effectuer pour aboutir à la solution du système linéaire. On entend par "opérations" : les additions, les soustractions, les multiplications, les divisions et les extractions de racines. Sur les ordinateurs, les temps d'évaluation de ces opérations sont certes différents, mais comme nous nous intéresserons à ce qui se passe asymptotiquement suivant n (ordre de la matrice) nous supposerons que ces opérations sont équivalentes.

A.2.1 Méthode du déterminant

Cette méthode, bien connue, est basée sur le calcul des déterminants des cofacteurs d'ordre $n-1$ et permet de calculer l'inverse d'une matrice en un nombre fini d'opérations. Elle nécessite le calcul d'un déterminant d'ordre n et n^2 déterminants d'ordre $n-1$. Or le calcul d'un déterminant d'ordre p nécessite p calculs de déterminants d'ordre $p-1$ et $p-1$ opérations, que l'on résume par la formule suivante (N_{op} signifie nombre d'opérations) :

$$N_{op}(\mathrm{Det}_p) = p \times N_{op}(\mathrm{Det}_{p-1}) + (p-1)$$

d'où on tire par récurrence que : $N_{op}(\mathrm{Det}_p) = O(p!)$. Ce calcul montre que le nombre d'opérations nécessaire à l'inversion d'une matrice d'ordre n par la méthode du déterminant est :

$$N_{op}(\mathbb{A}^{-1}) = O((n+1)!).$$

C'est inutilisable en pratique ($n = 50 \Longrightarrow N_{op} \simeq 1.5\,10^{66}$)!

A.2.2 Résolution d'un système triangulaire

La méthode du déterminant permet de calculer l'inverse. Or pour résoudre un système linéaire, il n'est pas nécessaire de calculer l'inverse de la matrice du système. Nous allons voir maintenant des méthodes dont le nombre d'opérations est en n^3 (à comparer avec $n!$) qui sont basées sur la méthode d'élimination de

Gauss (combinaison d'équations du système linéaire) qui permettent de se ramener à un système triangulaire :

$$\mathbb{L}X = \tilde{B} \quad \text{ou} \quad \mathbb{U}X = \tilde{B}$$

où $\mathbb{L}$ est une matrice triangulaire inférieure, i.e. :

$$\mathbb{L}_{ij} = 0 \qquad \text{si } i < j,$$

et $\mathbb{U}$ est une matrice triangulaire supérieure, i.e. :

$$\mathbb{U}_{ij} = 0 \qquad \text{si } i > j.$$

La résolution d'un système triangulaire est une opération évidente. En effet, supposons que l'on ait à résoudre le système triangulaire inférieur $\mathbb{L}X = \tilde{B}$. On a les équations suivantes :

$$\begin{cases} \mathbb{L}_{11}X_1 & = \tilde{B}_1 \\ \mathbb{L}_{21}X_1 + \mathbb{L}_{22}X_2 & = \tilde{B}_2 \\ \dots & \\ \mathbb{L}_{i1}X_1 + \ldots + \mathbb{L}_{ii}X_i & = \tilde{B}_i \\ \dots & \\ \mathbb{L}_{n1}X_1 + \ldots + \mathbb{L}_{nn}X_i & = \tilde{B}_n \end{cases}$$

La première équation fournit X_1, la seconde X_2 (car X_1 est connu) et ainsi de suite jusqu'à la détermination de X_n. On écrit donc l'algorithme suivant, dit de *descente* :

pour $i = 1, n$
  $S = \tilde{B}_n$
  pour $j = 1, i-1$
    $S = S - \mathbb{L}_{ij}X_j$
  fin
  si $\mathbb{L}_{ii} = 0$ **alors**
    STOP : matrice non inversible
  fin
  $X_i = \dfrac{S}{\mathbb{L}_{ii}}$
fin

Cet algorithme permet de détecter les matrices non inversibles (un terme nul sur la diagonale) et on peut même diminuer l'occupation mémoire en "écrasant" le second membre, c'est-à-dire en remplaçant X_i et X_j dans l'algorithme par $\tilde{B}_i$ et $\tilde{B}_j$.

Le décompte du nombre d'opérations est facile, on a :

$$N_{op}(\text{Descente}) = O(n^2). \tag{A.5}$$

De façon similaire, l'algorithme de résolution des systèmes triangulaires supérieurs, dit de *remontée*, s'écrit :

pour $i = n, 1, -1$
  $S = \tilde{B}_i$
  pour $j = n, i+1, -1$
    $S = S - \mathbb{U}_{ij} X_j$
  fin
  si $\mathbb{U}_{ii} = 0$ **alors**
    STOP : matrice non inversible
  fin
  $X_i = \dfrac{S}{\mathbb{U}_{ii}}$
fin

qui requiert le même nombre d'opérations que l'algorithme de descente. Il existe une autre version de ces algorithmes où les *boucles* en i et j sont inversées. Cette version présente des avantages dans des situations particulières (structure de stockage des matrices, vectorisation) sur lesquelles nous reviendrons ultérieurement. Ainsi l'algorithme de descente prend la forme suivante :

pour $i = 1, n$
  si $\mathbb{L}_{ii} = 0$ **alors**
    STOP : matrice non inversible
  fin
  $X_i = \dfrac{\tilde{B}_i}{\mathbb{L}_{ii}}$
  pour $j = i+1, n$
    $\tilde{B}_j = \tilde{B}_j - \mathbb{L}_{ji} X_i$
  fin
fin

Nous allons voir maintenant quels sont les algorithmes qui permettent de se ramener à la résolution d'un système triangulaire.

A.2.3 Factorisation sous forme triangulaire

La méthode d'*élimination de Gauss* consiste, en effectuant des combinaisons linéaires des équations, à se ramener à un système triangulaire supérieur. Pour le système linéaire $\mathbb{A}X = B$, l'algorithme suivant transforme la matrice $\mathbb{A}$ en une matrice triangulaire supérieure en modifiant également le second membre.

```
pour k = 1, n − 1
  si 𝔸_kk = 0 alors
    STOP : matrice non factorisable
  fin
  pour i = k + 1, n
    p_i = 𝔸_ik / 𝔸_kk
    pour j = k + 1, n
      𝔸_ij = 𝔸_ij − p_i 𝔸_kj
    fin
    𝔸_ik = 0
    B_i = B_i − p_i B_k
  fin
fin
```

Le fait qu'un pivot $\mathbb{A}_{kk}$ soit nul ne signifie pas que la matrice $\mathbb{A}$ n'est pas inversible ! Pour s'en convaincre, prendre la matrice d'ordre 3 :

$$\mathbb{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

qui après la première itération de l'algorithme d'élimination de Gauss devient :

$$\mathbb{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \mathbf{0} & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Sur la figure A.1, on représente la modification que subit la matrice lors de l'itération k.

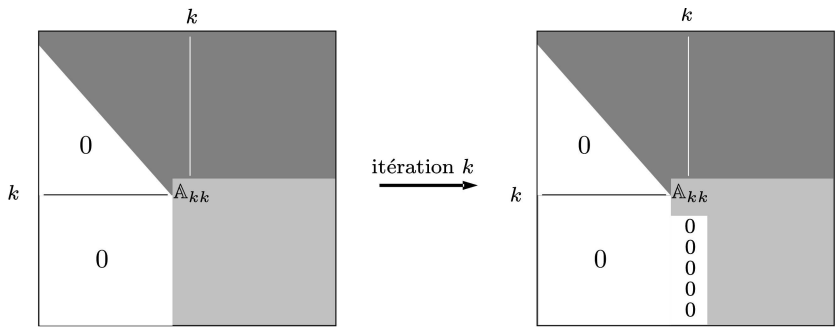


Fig. A.1. Itération k de la méthode de Gauss

L'algorithme d'élimination de Gauss est en réalité un algorithme de factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}$ où $\mathbb{L}$ est une matrice triangulaire inférieure et $\mathbb{U}$ la matrice

triangulaire supérieure obtenue par la méthode d'élimination. Sous certaines hypothèses la factorisation est possible et est unique, et lorsque la matrice $\mathbb{A}$ est symétrique on a des factorisations symétriques. On énonce le résultat suivant :

Proposition A.2.
a) Toute matrice $\mathbb{A}$, inversible et dont les sous-matrices fondamentales[1] *sont inversibles, est factorisable sous la forme $\mathbb{LU}$.*
b) La factorisation $\mathbb{A} = \mathbb{LU}$ avec $\mathbb{L}$ à diagonale unité est unique.
c) Si la matrice $\mathbb{A}$ est symétrique et factorisable $\mathbb{LU}$ alors la matrice $\mathbb{A}$ est factorisable sous la forme :

$$\mathbb{A} = \mathbb{LDL}^t$$

où $\mathbb{D}$ est une matrice diagonale.
d) Si $\mathbb{A}$ est symétrique définie-positive alors la matrice $\mathbb{A}$ est factorisable sous la forme :

$$\mathbb{A} = \mathbb{LL}^t.$$

Les démonstrations de ces propriétés ne sont pas très difficiles, mais quelques peu techniques. Les algorithmes de factorisation $\mathbb{LU}$ s'obtiennent à partir des égalités suivantes :

$$\begin{cases} \mathbb{A}_{ij} = \displaystyle\sum_{k=1}^{\min(i,j)} \mathbb{L}_{ik}\mathbb{U}_{kj} \ \forall i,j = 1,n \\ \mathbb{L}_{ii} = 1 \qquad\qquad \forall i = 1,n \qquad\qquad \text{(cond. d'unicité)} \end{cases}$$

traduisant le produit d'une matrice triangulaire inférieure par une matrice triangulaire supérieure. De ces formules, on déduit les relations :

$$\begin{cases} \mathbb{A}_{ij} = \displaystyle\sum_{k=1}^{j-1} \mathbb{L}_{ik}\mathbb{U}_{kj} + \mathbb{L}_{ij}\mathbb{U}_{jj} \ \forall\, j < i \\ \mathbb{A}_{ij} = \displaystyle\sum_{k=1}^{i-1} \mathbb{L}_{ik}\mathbb{U}_{kj} + \mathbb{U}_{ij} \qquad \forall\, j \geq i \end{cases}$$

qui permettent d'en déduire l'algorithme de factorisation $\mathbb{LU}$ suivant, également appelé algorithme de *Doolittle* :

[1] Matrice fondamentale d'ordre k : $\begin{bmatrix} \mathbb{A}_{11} & \dots & \mathbb{A}_{1k} \\ \vdots & \vdots & \vdots \\ \mathbb{A}_{k1} & \dots & \mathbb{A}_{kk} \end{bmatrix}$.

pour $i = 1, n$
  $\mathbb{L}_{ii} = 1$
fin
pour $p = 1, n$
  pour $j = p, n$
    $\mathbb{U}_{pj} = \mathbb{A}_{pj} - \sum_{k=1}^{p-1} \mathbb{L}_{pk}\mathbb{U}_{kj}$
  fin
  si $\mathbb{U}_{pp} = 0$ **alors**
    STOP : matrice non factorisable
  fin
  pour $i = p + 1, n$
    $\mathbb{L}_{ip} = \left(\mathbb{A}_{ip} - \sum_{k=1}^{p-1} \mathbb{L}_{ik}\mathbb{U}_{kp} \right) / \mathbb{U}_{pp}$
  fin
fin

Dans le cas où la matrice $\mathbb{A}$ est symétrique définie-positive, l'algorithme précédent prend alors la forme plus simple suivante, dite de *Cholesky* :

pour $p = 1, n$
  si $\mathbb{L}_{pp} = 0$ **alors**
    STOP : matrice non déf.-pos.
  fin
  $\mathbb{L}_{pp} = \left(\mathbb{A}_{pp} - \sum_{k=1}^{p-1} |\mathbb{L}_{pk}|^2 \right)^{1/2}$
  pour $i = p + 1, n$
    $\mathbb{L}_{ip} = \left(\mathbb{A}_{ip} - \sum_{k=1}^{p-1} \mathbb{L}_{ik}\mathbb{L}_{pk} \right) / \mathbb{L}_{pp}$
  fin
fin

Notons que cet algorithme permet de détecter facilement le caractère non définipositif d'une matrice (racine carrée d'un nombre négatif ou nul).

L'algorithme précédent est dit de la forme *colonne* car à chaque itération p on calcule les $n-p$ termes $(\mathbb{L}_{ip})_{i=p+1,n}$ de la colonne p. Il existe également une forme, dite *ligne*, où à chaque itération p on calcule les $p-1$ termes $(\mathbb{L}_{pj})_{j=1,p-1}$ de la ligne p :

pour $p = 1, n$
  pour $j = 1, p-1$
    $\mathbb{L}_{pj} = \left(\mathbb{A}_{pj} - \sum_{k=1}^{j-1} \mathbb{L}_{pk}\mathbb{L}_{jk} \right) / \mathbb{L}_{jj}$
  fin
  $\mathbb{L}_{pp} = \left(\mathbb{A}_{pp} - \sum_{k=1}^{p-1} |\mathbb{L}_{pk}|^2 \right)^{1/2}$
  si $\mathbb{L}_{pp} = 0$ **alors**
    STOP : matrice non déf.-pos.
  fin
fin

• Méthode de pivotage partiel
Afin de remédier au problème de l'apparition de pivot nul dans les algorithmes de factorisation, on utilise une technique de pivotage des lignes de la matrice qui s'appuie sur la propriété suivante :

Proposition A.3. *Soit* $\mathbb{A}$ *une matrice inversible, il existe une matrice de permutation* $\mathbb{P}$ *telle que* $\mathbb{PA}$ *soit factorisable sous la forme* $\mathbb{LU}$.

A titre d'exemple, nous allons décrire la méthode d'élimination de Gauss avec pivotage partiel. Dans l'algorithme d'élimination de Gauss, lorsqu'un pivot nul apparaît sur la colonne k, on cherche alors dans cette colonne un terme $\mathbb{A}_{\ell k}$ non nul avec $\ell > k$. S'il en existe un, on permute les lignes ℓ et k et on applique l'élimination. S'il n'en existe pas, cela veut dire que la matrice n'est pas inversible, car on a trouvé une sous-matrice d'ordre $(n - k + 1)$ non inversible. Dans la pratique, on utilise un vecteur q de dimension n pour représenter la permutation de n lignes. Initialisé à $q(i) = i,\ i = 1, n$, ce vecteur est mis à jour lors de chaque permutation de lignes par échanges des valeurs qu'il contient. Cette technique permet de ne pas échanger effectivement deux lignes de la matrice, la permutation étant virtuelle. Dans l'algorithme que nous présentons ci-après, on a utilisé une technique de pivotage avec recherche du pivot maximal en colonne, qui atténue la propagation des erreurs numériques pour un surcoût faible.
Notons que cette version de l'algorithme de Gauss avec pivotage partiel permet de détecter la non inversibilité d'une matrice. En effet, l'apparition d'un pivot nul correspond au fait qu'une colonne complète d'une sous-matrice est nulle et donc que la matrice n'est pas inversible. En pratique, à cause des erreurs d'arrondi, il n'est pas possible d'observer la nullité d'un pivot, mais seulement la quasi-nullité du pivot au sens d'une tolérance ε que l'on se donne. L'obtention d'un pivot très petit est la trace d'une matrice mal conditionnée et donc de la qualité du résultat.

Par conséquent, le paramètre de tolérance est à choisir en fonction de la précision avec laquelle on travaille et de la qualité du résultat attendu.

pour $i = 1, n$
  $q(i) = i$
fin
pour $k = 1, n-1$
  $K = q(k)$ et $M = 0$ *recherche du pivot maximum*
  pour $i = k, n$
    si $\left|\mathbb{A}_{q(i)K}\right| > M$ **alors**
      $\ell = i$ et $M = \left|\mathbb{A}_{q(i)K}\right|$
    fin
  fin
  si $M < \varepsilon$ **alors**
    STOP : matrice non inversible
  fin
  $q(k) = q(\ell)$ et $q(\ell) = K$ *permutation des lignes*
  $K = q(k)$
  pour $i = k+1, n$
    $I = q(i)$
    $p_i = \frac{\mathbb{A}_{IK}}{\mathbb{A}_{KK}}$
    pour $j = k+1, n$
      $J = q(j)$
      $\mathbb{A}_{IJ} = \mathbb{A}_{IJ} - p_i \mathbb{A}_{KJ}$
    fin
    $B_I = B_I - p_i B_K$
  fin
fin

On peut généraliser la méthode de pivotage en autorisant également le pivotage des colonnes c'est-à-dire en recherchant le pivot maximal à l'itération k parmi les termes $\left(\mathbb{A}_{ij}^k\right)$ pour $i \geq k$ et $j \geq k$, ce qui atténue encore la propagation des erreurs numériques.

Cette technique de pivotage partiel s'applique aussi aux algorithmes de factorisation $\mathbb{LU}$. Le pivot maximal à l'étape p de l'algorithme est alors donné par :

$$\max_{j=p,n} \left| \mathbb{A}_{pj} - \sum_{k=1}^{p-1} \mathbb{L}_{pk} \mathbb{U}_{kj} \right|$$

et il est atteint pour l'indice ℓ. On pivote alors les indices p et ℓ.

• Nombre d'opérations
Tous les algorithmes que nous venons de décrire font apparaître trois boucles imbriquées. On a donc :

$$N_{op}(\text{Factorisation}) = O(n^3).$$

• Phase de descente-remontée
Lorsque l'on dispose d'une factorisation $\mathbb{L}\mathbb{U}$ (respectivement $\mathbb{L}\mathbb{L}^t$) d'une matrice $\mathbb{A}$ il est alors facile de résoudre le système linéaire $\mathbb{A}X = B$. Il suffit de résoudre le système triangulaire inférieur :

$$\mathbb{L}Y = B$$

puis le système triangulaire supérieur :

$$\mathbb{U}X = Y \quad \big(\text{resp. } \mathbb{L}^t X = Y\big)\cdot$$

A.3 Méthodes itératives de résolution

Les méthodes itératives sont fondées sur la construction d'une suite de vecteurs $\big(X^{(k)}\big)$ qui converge vers le vecteur $\bar{X}$ solution du système $\mathbb{A}\bar{X} = B$. Les méthodes itératives se scindent classiquement en deux classes :

- les méthodes de décomposition
- les méthodes de gradient

Des liens étroits les unissent toutefois. Avant d'exposer ces méthodes, présentons quelques résultats généraux relatifs à la convergence des méthodes itératives.

A.3.1 Convergence des méthodes itératives

Considérons une suite de vecteurs $(X^{(k)})$ définie par la relation de récurrence :

$$X^{(k+1)} = \mathbb{C}X^{(k)} + D \tag{A.6}$$

Si la suite $(X^{(k)})$ converge vers le vecteur $\bar{X}$, alors on a nécessairement :

$$\bar{X} = \mathbb{C}\bar{X} + D,$$

d'où on déduit :

$$D = (\mathbb{I} - \mathbb{C})\mathbb{A}^{-1}B \tag{A.7}$$

On dit que la méthode itérative (A.6) converge si la suite $(X^{(k)})$ converge pour tout vecteur initial $X^{(0)}$. On énonce alors le résultat classique de convergence :

Proposition A.4. *La méthode itérative (A.6) converge si et seulement si*

$$\rho(\mathbb{C}) < 1. \tag{A.8}$$

Si, de plus, l'égalité (A.7) est satisfaite alors la suite $(X^{(k)})$ *converge vers* $\bar{X}$*.*

Le nombre $\rho(\mathbb{C})$ désigne le rayon spectral de la matrice $\mathbb{C}$, c'est-à-dire le module de la plus grande valeur propre (*a priori* complexe) de $\mathbb{C}$. Afin de comparer les méthodes itératives, il faut disposer d'un outil caractérisant la vitesse de convergence. On définit donc :

- *le taux de convergence pour* m *itérations :*

$$R_m(\mathbb{C}) = -\log\left(\|\mathbb{C}^m\|^{\frac{1}{m}}\right). \tag{A.9}$$

- *le taux asymptotique de convergence* :

$$R(\mathbb{C}) = -\log(\rho(\mathbb{C})). \tag{A.10}$$

Le nombre $N_m = \dfrac{1}{R_m(\mathbb{C})}$ s'interprète comme une estimation du nombre d'itérations nécessaire pour réduire l'erreur d'un facteur e (2.718...). Notons que :

$$\lim_{m\to\infty} R_m(\mathbb{C}) = R(\mathbb{C}) \quad et \quad \lim_{m\to\infty} N_m = N = \frac{1}{R(\mathbb{C})}.$$

Le nombre N est généralement utilisé pour caractériser la vitesse de convergence d'une méthode itérative.

A.3.2 Méthodes de décomposition

Un premier type de méthodes est basé sur la décomposition matricielle :

$$\mathbb{A} = \mathbb{M} - \mathbb{N} \tag{A.11}$$

où la matrice $\mathbb{M}$ est supposée *inversible*.

On associe à cette décomposition le processus itératif :

$$\boxed{\mathbb{M}X^{(k+1)} = \mathbb{N}X^{(k)} + B} \tag{A.12}$$

qui est bien de la forme générale (A.6) avec :

$$\mathbb{C} = \mathbb{M}^{-1}\mathbb{N} \;\text{ et }\; D = \mathbb{M}^{-1}B \tag{A.13}$$

et vérifie la propriété (A.7).

Il s'agit maintenant de savoir si $\rho\left(\mathbb{M}^{-1}\mathbb{N}\right)$ est inférieur à 1. Nous indiquons dans les propositions suivantes, les classes de matrices pour lesquelles on sait démontrer un tel résultat (voir [10] par exemple).

Pour les *matrices hermitiennes* (symétriques dans le cas réel) on a le résultat suivant :

Proposition A.5. *Si* $\mathbb{A}$ *est une matrice auto-adjointe alors la matrice* $(\mathbb{M}^* + \mathbb{N})$ *est auto-adjointe. Si on suppose de plus que la matrice* $(\mathbb{M}^* + \mathbb{N})$ *est définie positive alors* $\rho\left(\mathbb{M}^{-1}\mathbb{N}\right) < 1$ *si et seulement si* $\mathbb{A}$ *est définie-positive.*

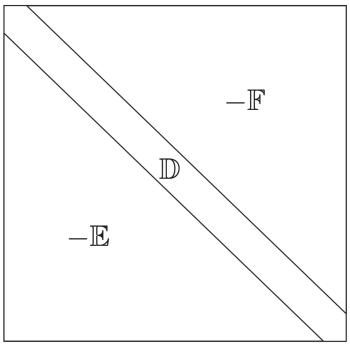
On a un résultat plus général concernant les matrices d'*inverse non-négatif*. On dit qu'un vecteur Z est non négatif ($Z \geq 0$) si toutes ses composantes sont positives ou nulles. Une matrice $\mathbb{H}$ est dite non-négative si $\mathbb{H}Z \geq 0$, $\forall Z \geq 0$.

Proposition A.6. *Si la matrice* $\mathbb{A}$*, inversible, a un inverse non négatif (*$\mathbb{A}^{-1} \geq 0$*) et si les matrices* $\mathbb{M}^{-1}$ *et* $\mathbb{N}$ *sont non négatives (*$\mathbb{M}^{-1} \geq 0$, $\mathbb{N} \geq 0$*) alors on a :*

$$\rho\left(\mathbb{M}^{-1}\mathbb{N}\right) = \frac{\rho\left(\mathbb{A}^{-1}\mathbb{N}\right)}{1 + \rho\left(\mathbb{A}^{-1}\mathbb{N}\right)} < 1.$$

Nous allons voir maintenant quelques décompositions classiques de la forme $\mathbb{A} = \mathbb{M} - \mathbb{N}$. Une classe importante de méthodes de décomposition s'appuie sur l'utilisation des sous-matrices $\mathbb{D}$, $-\mathbb{E}$ et $-\mathbb{F}$ désignant respectivement la diagonale, la partie triangulaire inférieure stricte et la partie triangulaire supérieure stricte de la matrice $\mathbb{A}$.

$\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F} =$



• Méthode de Jacobi
On utilise la décomposition suivante :

$$\boxed{\mathbb{M} = \mathbb{D} \quad \text{et} \quad \mathbb{N} = \mathbb{E} + \mathbb{F}}$$

L'itération (A.12) prend la forme algorithmique suivante :

$X^{(0)}$ donné
pour $k = 1, N_{max}$
  pour $i = 1, n$

$$X_i^{(k+1)} = \left(B_i - \sum_{j=1}^{i-1} \mathbb{A}_{ij} X_j^{(k)} - \sum_{j=i+1}^{n} \mathbb{A}_{ij} X_j^{(k)} \right) / \mathbb{A}_{ii}$$

  fin
  $R = \left| B - \mathbb{A} X^{(k+1)} \right|$
  si $R \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
fin

Ce schéma est *explicite* au sens où le calcul de l'itéré $k+1$ ne fait appel qu'à des valeurs construites aux itérations précédentes (ici k). Bien évidemment, pour appliquer la méthode de Jacobi, il faut que la matrice $\mathbb{A}$ ne présente aucun terme nul sur sa diagonale. Ici, on utilise un test de convergence de la suite $(X^{(k)})$ portant sur le *résidu* $|B - \mathbb{A}X|$ dans une norme donnée, permettant d'estimer avec quelle qualité la solution trouvée est solution du système linéaire. Rappelons que si la matrice $\mathbb{A}$ est très mal conditionnée, il faut choisir la valeur de ε en fonction de son conditionnement. N_{max} est le nombre d'itérations maximum. Ce test est obligatoire dans toute programmation de méthodes itératives.

• Méthode de Gauss-Seidel
Cette méthode consiste à *impliciter* partiellement la méthode de Jacobi et s'appuie sur le fait que dans la méthode de Jacobi, à l'itération k, lorsque l'on calcule la $i^{ème}$ composante de $X^{(k+1)}$, les $(i-1)^{èmes}$ composantes précédentes de $X^{(k+1)}$ ont déjà été évaluées. Par conséquent, on peut les utiliser. Ce qui conduit à l'algorithme suivant :

$X^{(0)}$ donné
pour $k = 1, N_{max}$
  pour $i = 1, n$

$$X_i^{(k+1)} = \left(B_i - \sum_{j=1}^{i-1} \mathbb{A}_{ij} X_j^{(k+1)} - \sum_{j=i+1}^{n} \mathbb{A}_{ij} X_j^{(k)} \right) / \mathbb{A}_{ii}$$

  fin
  $R = \left| B - \mathbb{A} X^{(k+1)} \right|$
  si $R \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
fin

Il est à noter, d'une part, que le nombre d'opérations est identique à celui de la méthode de Jacobi et, d'autre part, qu'il est possible de diminuer l'occupation mémoire en utilisant un seul vecteur. Bien que cette méthode soit *explicite*, on réalise une inversion partielle de la matrice (résolution par descente de la partie triangulaire inférieure). Par conséquent, la méthode de Gauss-Seidel constitue *a priori* une amélioration de la méthode de Jacobi. C'est bien une méthode de décomposition de la forme $\mathbb{A} = \mathbb{M} - \mathbb{N}$ avec :

$$\boxed{\mathbb{M} = \mathbb{D} - \mathbb{E} \quad \text{et} \quad \mathbb{N} = \mathbb{F}}$$

et elle est soumise aux mêmes conditions d'application que la méthode de Jacobi, à savoir, que la diagonale de la matrice $\mathbb{A}$ ne doit pas avoir de termes nuls.

• Méthode S.O.R.

On peut généraliser la méthode de Gauss-Seidel en choisissant :

$$\boxed{\mathbb{M} = \frac{1}{\omega}\mathbb{D} - \mathbb{E} \quad \text{et} \quad \mathbb{N} = \left(\frac{1}{\omega} - 1\right)\mathbb{D} + \mathbb{F}}$$

où ω est un paramètre réel, dit paramètre de *relaxation*. Lorsque $\omega = 1$, on retrouve la méthode de Gauss-Seidel et dans le cas général on a l'algorithme suivant :

$X^{(0)}$ donné
pour $k = 1, N_{max}$
  pour $i = 1, n$

$$X_i^{(k+1)} = X_i^{(k)} + \omega \left(\begin{array}{l} B_i - \displaystyle\sum_{j=1}^{i-1} \mathbb{A}_{ij} X_j^{(k+1)} \\ \\ - \displaystyle\sum_{j=i+1}^{n} \mathbb{A}_{ij} X_j^{(k)} - \mathbb{A}_{ii} X_i^{(k)} \end{array} \right) / \mathbb{A}_{ii}$$

  fin
  $R = \left| B - \mathbb{A} X^{(k+1)} \right|$
  si $R \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
fin

Lorsque $\omega \in]0, 1[$ on parle de méthode de *sous-relaxation* et lorsque $\omega > 1$ de méthode de *sur-relaxation*.

• Comparaison des méthodes

Avant toute chose, énonçons un résultat de convergence de ces méthodes de décomposition :

Proposition A.7. *Si la matrice* $\mathbb{A}$ *est hermitienne définie-positive et la matrice* $\mathbb{D}$ *est définie-positive alors les méthodes de Jacobi, de Gauss-Seidel et la méthode S.O.R avec* $\omega \in]0,2[$ *convergent vers la solution du système linéaire.*

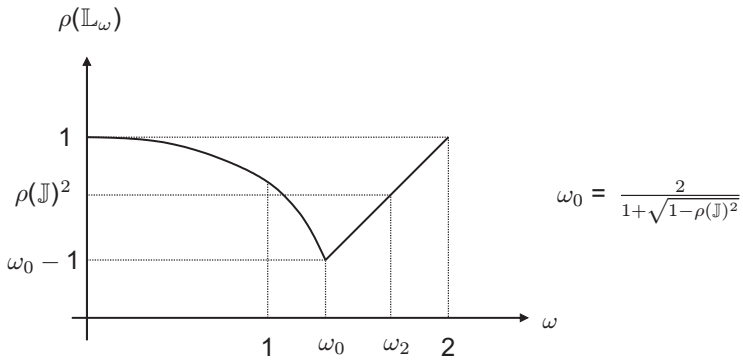
Ce résultat découle de la proposition A.5 et du fait que :

$$\mathbb{M}^* + \mathbb{N} = \frac{1}{\omega}\mathbb{D} - \mathbb{F} + \left(\frac{1}{\omega} - 1\right)\mathbb{D} + \mathbb{F} = \frac{2-\omega}{\omega}\mathbb{D}.$$

La définition du taux asymptotique de convergence (A.10) indique que la méthode itérative converge d'autant plus vite que le rayon spectral $\rho\left(\mathbb{M}^{-1}\mathbb{N}\right)$ est petit. En comparant donc le rayon spectral des matrices $\mathbb{M}^{-1}\mathbb{N}$ associées aux méthodes de Jacobi, Gauss-Seidel et S.O.R, on classe ces méthodes. Indiquons le type de résultat que l'on obtient dans le cas où la matrice $\mathbb{A}$ est tridiagonale.
On note respectivement $\mathbb{J}$, $\mathbb{G}$ et $\mathbb{L}_\omega$ les matrices $\mathbb{M}^{-1}\mathbb{N}$ associées à chacune de ces méthodes. On montre alors (voir [10]) que :

$$\boxed{\rho\left(\mathbb{G}\right) = \rho\left(\mathbb{J}\right)^2}$$

et que la fonction $\omega \longrightarrow \rho\left(\mathbb{L}_\omega\right)$ possède l'allure suivante :



Ces résultats montrent que :

- la méthode de Gauss-Seidel converge plus vite que la méthode de Jacobi,
- si $\omega \in]0,1[\cup]\omega_2, 2[$, la méthode S.O.R. converge moins vite que la méthode de Gauss-Seidel,
- si $\omega \in]1, \omega_2[$, la méthode S.O.R. converge plus rapidement que la méthode de Gauss-Seidel avec une valeur optimale du paramètre de relaxation $\omega = \omega_0$.

Il est difficile d'énoncer des résultats plus généraux concernant le classement de ces méthodes car cela requiert une connaissance assez fine du spectre de la matrice $\mathbb{A}$.

A.3.3 Méthodes de gradient

Les méthodes de gradient, fondées sur la minimisation d'une fonctionnelle quadratique associée au système linéaire, s'appliquent principalement aux matrices symétriques définies-positives. On peut néanmoins les généraliser au cas non-symétrique. Lorsque $\mathbb{A}$ est symétrique définie-positive, la résolution du système linéaire $\mathbb{A}X = B$ est équivalente à la recherche du minimum de la fonction de $\mathbb{R}^n \to \mathbb{R}$:

$$\Phi(X) = \frac{1}{2}(\mathbb{A}X \mid X) - (B \mid X), \tag{A.14}$$

car on a (cf. [9]) :

$$\nabla\Phi(X) = \mathbb{A}X - B.$$

La recherche du minimum de la fonction Φ s'effectue classiquement par des méthodes dites de *gradient* dont nous allons présenter les principales : gradient à pas constant, gradient à pas optimal et gradient conjugué.

• Méthode du gradient à pas constant
On note $W = \nabla\Phi$. Soit $\theta > 0$, on considère le processus itératif suivant :

$$\boxed{X^{(k+1)} = X^{(k)} - \theta W^{(k)}} \tag{A.15}$$

qui, compte-tenu du fait que $W^{(k)} = \mathbb{A}X^{(k)} - B$, s'inscrit dans le cadre des méthodes de décomposition (A.6) avec :

$$\mathbb{C} = \mathbb{I} - \theta\mathbb{A} \quad \text{et} \quad \mathbb{D} = \theta B,$$

et vérifie la condition (A.7).
Pour que le processus (A.15) converge vers la solution du système linéaire, il faut et il suffit que :

$$\rho(\mathbb{I} - \theta\mathbb{A}) < 1.$$

Notons λ (resp. Λ) la plus petite (resp. la plus grande) valeur propre de $\mathbb{A}$ ($0 < \lambda < \Lambda$ car $\mathbb{A}$ est supposée définie-positive). On a alors le résultat de convergence suivant :

Proposition A.8. *La méthode du gradient à pas constant converge si et seulement si :*

$$0 < \theta < \frac{2}{\Lambda}.$$

En outre, le minimum de $\rho(\mathbb{I} - \theta\mathbb{A})$ est atteint pour :

$$\theta = \theta_0 = \frac{2}{\Lambda + \lambda} \quad \text{et} \quad \rho\left(\mathbb{I} - \theta_0\mathbb{A}\right) = \frac{\Lambda - \lambda}{\Lambda + \lambda}.$$

La démonstration s'appuie sur le fait que :

$$\rho(\mathbb{I} - \theta\mathbb{A}) = \mathrm{Max}(|1 - \theta\Lambda|, |1 - \theta\lambda|).$$

En règle générale, on ne connait ni Λ, ni λ. C'est pourquoi la méthode du gradient à pas constant est aléatoire et, par voie de conséquence, peu utilisée dans la pratique. Une stratégie plus sûre consiste à adapter, à chaque itération k, le pas θ. C'est ce que nous allons voir maintenant.

• Méthode du gradient à pas optimal

A chaque itération, on choisit $\theta^{(k)}$ qui réalise le minimum de la fonction réelle :

$$\theta \mapsto h(\theta) = \Phi\left(X^{(k)} - \theta W^{(k)}\right).$$

La valeur optimale $\theta^{(k)}$ est donnée par la formule (obtenue en écrivant que $h'(\theta^{(k)}) = 0$) :

$$\theta^{(k)} = \frac{\left(\mathbb{A}X^{(k)} - B \mid W^{(k)}\right)}{\left(\mathbb{A}W^{(k)} \mid W^{(k)}\right)}. \tag{A.16}$$

Ce choix conduit à la méthode du gradient à pas optimal :

$$\boxed{X^{(k+1)} = X^{(k)} - \theta^{(k)}W^{(k)} \qquad \theta^{(k)} \text{ donné par (A.16)}}$$

Compte-tenu du fait que :

$$\theta^{(k)} = \frac{\left|W^{(k)}\right|^2}{\left(\mathbb{A}W^{(k)} \mid W^{(k)}\right)}$$

on obtient l'algorithme suivant :

$X^{(0)}$ donné
pour $k = 0, N_{max}$
  $W^{(k)} = B - \mathbb{A}X^{(k)}$
  si $\left|W^{(k)}\right| \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
  $\theta^{(k)} = \dfrac{\left|W^{(k)}\right|^2}{\left(\mathbb{A}W^{(k)}, W^{(k)}\right)}$
  $X^{(k+1)} = X^{(k)} - \theta^{(k)}W^{(k)}$
fin

Cette méthode ne s'exprime plus comme une méthode de décomposition, mais on peut montrer qu'elle converge par d'autres techniques. Plus précisément on a :

Proposition A.9. *La méthode du gradient à pas optimal converge et on a l'estimation suivante :*

$$\left|X^{(k)} - X\right|_2^2 \leq C\left(1 - \frac{\lambda}{\Lambda}\right)^k \tag{A.17}$$

L'estimation (A.17) montre que la méthode converge d'autant mieux que le conditionnement $\gamma(\mathbb{A})$ de la matrice $\mathbb{A}$ est proche de 1, i.e. que la matrice $\mathbb{A}$ approche l'identité. Les méthodes de gradient que nous venons d'exposer, utilisent une direction de déplacement opposée au gradient qui est la direction de descente de plus forte pente. On va voir maintenant une méthode qui utilise une autre direction de descente qui, bien que localement moins bonne, se révèle globalement plus efficace.

• Méthode du gradient conjugué

Cette méthode consiste à choisir des directions de descente $W^{(k)}$ qui soient mutuellement orthogonales pour le produit scalaire défini par $\mathbb{A}$, lorsque $\mathbb{A}$ est symétrique définie-positive. On parle de directions *conjuguées* pour la matrice $\mathbb{A}$. L'algorithme ci-dessous décrit le processus de construction de ces directions :

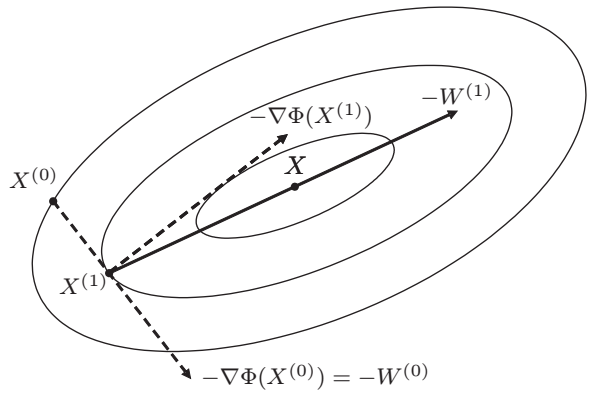
$X^{(0)}$ donné
$G^{(0)} = \mathbb{A}X^{(0)} - B$ et $W^{(0)} = G^{(0)}$
pour $k = 0, N_{max}$
  $\theta^{(k)} = \dfrac{\left(G^{(k)} \mid W^{(k)}\right)}{\left(\mathbb{A}W^{(k)} \mid W^{(k)}\right)}$ *pas optimal*
  $X^{(k+1)} = X^{(k)} - \theta^{(k)}W^{(k)}$
  $G^{(k+1)} = \nabla\Phi\left(X^{(k+1)}\right) = \mathbb{A}X^{(k+1)} - B = G^{(k)} - \theta^{(k)}\mathbb{A}W^{(k)}$
  si $\left|G^{(k+1)}\right| \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
  $\gamma^{(k)} = \dfrac{\left|G^{(k+1)}\right|^2}{\left|G^{(k)}\right|^2}$
  $W^{(k+1)} = G^{(k+1)} + \gamma^{(k)}W^{(k)}$ *nouvelle direction de descente*
fin

Dans l'algorithme, $G^{(k)}$ représente le gradient de Φ au point $X^{(k)}$ et $W^{(k)}$ les directions conjuguées. L'algorithme s'arrête dès que $G^{(k)} = \nabla\Phi\left(X^k\right) = 0$ car $\mathbb{A}X^{(k)} = B$. En réalité, l'algorithme du gradient conjugué n'est pas une méthode itérative au sens strict car il converge *en au plus n itérations*. En effet, on peut démontrer par récurrence, et c'est assez fastidieux, les relations d'orthogonalité suivantes :

$$\begin{cases} \left(\nabla\Phi\left(X^{(k+1)}\right) \mid W^{(j)}\right) = 0 \\ \left(\nabla\Phi\left(X^{(k+1)}\right) \mid \nabla\Phi\left(X^{(j)}\right)\right) = 0 \ \forall\, j < k+1 \\ \left(\mathbb{A}W^{(k+1)} \mid W^{(j)}\right) = 0 \end{cases} \tag{A.18}$$

qui prouvent qu'à l'itération n, la famille $\left(\nabla\Phi\left(X^{(j)}\right)\right)_{j=0,n}$ est une famille orthogonale de $(n+1)$ vecteurs de $\mathbb{R}^n$. Par conséquent un de ces vecteurs est nul, c'est-à-dire qu'il existe un indice $\ell \leq n$ tel que $\mathbb{A}X^{(\ell)} = B$! Ce résultat théorique n'est pas toujours garanti dans la pratique à cause des erreurs d'arrondi qui se propagent lors du processus itératif. On remarque que l'algorithme précédent ne requiert qu'une seule opération matricielle coûteuse : le produit matrice-vecteur $\mathbb{A}W^{(k)}$.

En dimension 2, on peut se faire une idée de la différence entre la méthode du gradient et celle du gradient conjugué. La fonctionnelle $X \mapsto \Phi(X)$ est représentée dans $\mathbb{R}^3$ par un paraboloïde et les courbes isovaleurs $\Phi(X) = cte$ sont des ellipses dans $\mathbb{R}^2$.



Pour les deux méthodes, à la première itération on obtient le même point $X^{(1)}$ caractérisé géométriquement par le fait que la droite passant par $X^{(0)}$ de direction $-\nabla\Phi\left(X^{(0)}\right)$ est tangente à la courbe isovaleur passant par $X^{(1)}$. La direction de descente du gradient $-\nabla\Phi\left(X^{(1)}\right)$ est orthogonale à la tangente à l'isovaleur au point $X^{(1)}$. La direction conjuguée $W^{(1)}$ est construite de telle sorte que $\left(W^{(1)} \mid \mathbb{A}\nabla\Phi\left(X^{(0)}\right)\right) = 0$ (orthogonalité dans le produit scalaire associé à $\mathbb{A}$). Cette direction passe par le centre des ellipses $\Phi = cte$ qui est le point minimum de Φ, donc la solution de $\mathbb{A}X = B$. En deux itérations, la méthode du gradient conjugué fournit la solution. Par contre, les directions de descente $-\nabla\Phi\left(X^{(k)}\right)$ construites lors de la méthode du gradient prennent alternativement d'une itération à l'autre deux directions orthogonales entre elles. De ce fait, il est impossible que la méthode du gradient converge en un nombre fini d'itérations sauf si les ellipses sont des cercles (conditionnement égal à 1, i.e. $\mathbb{A} = \alpha\mathbb{I}$) ou si le point de départ $X^{(0)}$ est situé sur une des extrémités du petit ou du grand axe d'une ellipse.

A.3.4 Préconditionnement des systèmes

Si on multiplie la matrice $\mathbb{A}$ du système linéaire à résoudre, par une matrice $\mathbb{P}$, dite de *préconditionnement*, approximation de l'inverse de la matrice $\mathbb{A}$, on obtient un nouveau système linéaire plus rapide à résoudre par une technique itérative car le conditionnement de la matrice $\mathbb{PA}$ est bien meilleur. Le choix idéal est évidemment $\mathbb{P} = \mathbb{A}^{-1}$ car dans ce cas la résolution du système est immédiate. La qualité d'une méthode itérative préconditionnée repose, par conséquent, sur le choix d'une matrice de préconditionnement qui réalise le meilleur compromis entre le degré d'approximation de l'inverse $\mathbb{A}^{-1}$ et le coût de calcul de $\mathbb{P}$. La technique générale de préconditionnement des systèmes linéaires consiste à choisir une matrice de préconditionnement $\mathbb{P}$ symétrique et définie-positive dont la factorisation de *Cholesky* est :

$$\mathbb{P} = \mathbb{L}\mathbb{L}^t.$$

On pose alors :

$$\boxed{\begin{array}{l}\tilde{\mathbb{A}} = \mathbb{L}^{-1}\mathbb{A}\mathbb{L}^{-t} \\ \tilde{B} = \mathbb{L}^{-1}B \\ \tilde{X} = \mathbb{L}^t X\end{array}}$$

ce qui conduit au nouveau système linéaire :

$$\tilde{\mathbb{A}}\tilde{X} = \tilde{B},$$

avec $\tilde{\mathbb{A}}$ une matrice symétrique définie-positive dès lors que $\mathbb{A}$ l'est. Le calcul de $\tilde{B}$ est simple et peu couteux (résolution d'un système triangulaire) et il est clair que si l'on connait $\tilde{X}$, on en déduit simplement X par résolution d'un système triangulaire. Regardons, à titre d'exemple, la forme optimale que prend l'algorithme de gradient conjugué préconditionné :

$X^{(0)}$ donné
$G^{(0)} = \mathbb{A}X^{(0)} - B$ et $W^{(0)} = \mathbb{L}^{-1}G^{(0)}$
pour $k = 0, N_{max}$
  $\theta^{(k)} = \dfrac{\left(\mathbb{P}^{-1}G^{(k)} \mid G^{(k)}\right)}{\left(\mathbb{A}W^{(k)} \mid W^{(k)}\right)}$ *pas optimal*
  $X^{(k+1)} = X^{(k)} - \theta^{(k)}W^{(k)}$
  $G^{(k+1)} = G^{(k)} - \theta^{(k)}\mathbb{A}W^{(k)}$
  si $\left|G^{(k+1)}\right| \leq \varepsilon$ **alors**
    STOP : convergence à ε près
  fin
  $\gamma^{(k)} = \dfrac{\left(\mathbb{P}^{-1}G^{(k+1)} \mid G^{(k+1)}\right)}{\left(\mathbb{P}^{-1}G^{(k)} \mid G^{(k)}\right)}$
  $W^{(k+1)} = \mathbb{P}^{-1}G^{(k+1)} + \gamma^{(k)}W^{(k)}$ *nouvelle direction de descente*
fin

Connaissant la factorisation de Cholesky $\mathbb{P} = \mathbb{L}\mathbb{L}^t$ le calcul du produit $\mathbb{P}^{-1}Y$ se ramène à la résolution des systèmes triangulaires :

$$\mathbb{L}Z = Y \quad \text{puis} \quad \mathbb{L}^t V = Z,$$

fournissant le vecteur $V = \mathbb{P}^{-1}Y$. On constate que l'algorithme précédent ne nécessite qu'une seule inversion effective et fournit directement $X^{(k)} = \mathbb{L}^{-t}\tilde{X}^{(k)}$.

Dans la pratique, de nombreuses méthodes (cf. [23]), essentiellement heuristiques, sont utilisées pour choisir la matrice de préconditionnement $\mathbb{P}$. Citons les deux plus utilisées :

- le préconditionnement SSOR : $\mathbb{P} = \left(\frac{\mathbb{D}}{\omega} - \mathbb{E}\right)\left(\frac{\mathbb{D}}{\omega}\right)^{-1}\left(\frac{\mathbb{D}}{\omega} - \mathbb{F}\right)$,
- les factorisations incomplètes de Cholesky de la matrice $\mathbb{A}$, qui consistent à omettre des termes de $\mathbb{A}$ lors de la factorisation, en ne conservant que quelques diagonales, voire la diagonale seulement.

Les méthodes itératives que nous avons présentées fonctionnent essentiellement pour des matrices symétriques définies-positives. Lorsqu'on est confronté à des matrices non symétriques il convient d'utiliser soit des méthodes directes (élimination de Gauss) soit des méthodes itératives plus sophistiquées, par exemple *GMRES (gradient minimal residue)*, ou *BICGSTAB (biconjugate gradient stabilized)*.

A.4 Structure de stockage des matrices

Les matrices issues de la discrétisation par éléments finis (ou par différences finies) sont "structurellement" creuses, c'est-à-dire qu'elles présentent un très grand nombre de termes nuls, l'emplacement de ces termes nuls étant connu à l'avance. Une idée raisonnable consiste à ne pas effectuer de calculs sur ces termes nuls afin de diminuer le temps de calcul nécessaire à l'inversion des systèmes linéaires. Par ailleurs, les matrices pouvant être d'un ordre très élevé (100 000 par exemple) il n'est pas envisageable de les "stocker" dans la mémoire de l'ordinateur sous forme habituelle : tableau à double entrée ligne-colonne ; dans le cas d'une matrice réelle d'ordre 100 000 cela fait 10^{10} termes, soit 40 gigaoctets ! C'est pourquoi l'utilisation de techniques de *compactage*, consistant à stocker la matrice sous forme d'un vecteur où le maximum de termes structurellement nuls ont été éliminés, se révèle indispensable. Nous présentons, tout d'abord, les principales techniques de stockage des matrices, dénommées *compactages*. L'adaptation des méthodes de résolution des systèmes linéaires à ces compactages est ensuite traitée.

A.4.1 Compactage des matrices

Il existe principalement quatre types de compactage : le compactage *diagonal*, le compactage *bande*, le compactage *profil* et le compactage *morse*, l'ordre précédent étant grossièrement celui de l'occupation mémoire : de la plus grande à la plus petite. Evidemment, lorsque la matrice est symétrique la partie triangulaire supérieure n'est pas stockée. Nous nous plaçons, ici, seulement dans le cas de matrices symétriques, la généralisation au cas des matrices non-symétriques étant immédiate.

• Compactage diagonal
On note D_0 la diagonale de la matrice $\mathbb{A}$ et D_k^- la $k^{ème}$ sous-diagonale de la matrice $\mathbb{A}$ définie par :

$$D_k^- = \{\mathbb{A}_{ij},\ \text{tel que } i-j=k,\ i,j=1,n\} \qquad \left(D_0^- = D_0\right)$$

où n désigne toujours l'ordre de la matrice $\mathbb{A}$. La longueur de D_k^- est $(n-k)$ et on note $\left(D_k^-\right)_\ell$ le $\ell^{ème}$ terme de la diagonale D_k^-. On a les formules suivantes :

$$\begin{cases} \left(D_k^-\right)_\ell = \mathbb{A}_{\ell+k,\ell}\ \forall k=0,n-1,\ \ell=1,n-k \\ \mathbb{A}_{ij} = \left(D_{i-j}^-\right) \quad i \geq j,\ i,j=1,n \end{cases}$$

qui permettent de passer de la représentation diagonale à la représentation matricielle et réciproquement. Dès lors que la matrice $\mathbb{A}$ présente des diagonales nulles, i.e. $\left(D_k^-\right)_\ell = 0, \forall \ell = 1, n-k$, on ne les stocke pas. On construit ainsi le vecteur compacté de la matrice $\mathbb{A}$:

$$\mathbb{C}_d = \left(D_{k_1}^-, D_{k_2}^-, \cdots, D_{k_p}^-\right)$$

où $(k_1, k_2, \cdots, k_p)$ sont ordonnés de façon croissante. Le vecteur $\mathbb{C}_d$ a pour longueur :

$$\sum_{m=1,p} (n-k_m) = pn - \sum_{m=1,p} k_m.$$

Afin de retrouver un terme $\mathbb{A}_{ij}$ (non nul) dans le vecteur $\mathbb{C}_d$ il faut disposer d'une information supplémentaire, par exemple la liste des diagonales stockées et leur position dans le vecteur $\mathbb{C}_d$. Pour ce faire, on construit l'application ℓ_c définie par :

$$\begin{cases} \ell_c(k) = 0 \text{ si } D_k^- \notin \mathbb{C}_d \\ \ell_c(k) = p \text{ si } D_k^- \in \mathbb{C}_d \end{cases}$$

où p est la position dans $\mathbb{C}_d$ du terme $\left(D_k^-\right)_1$.

A partir de l'application ℓ_c, il est facile de retrouver la position d'un terme $\mathbb{A}_{ij}$ dans le vecteur $\mathbb{C}_d$:

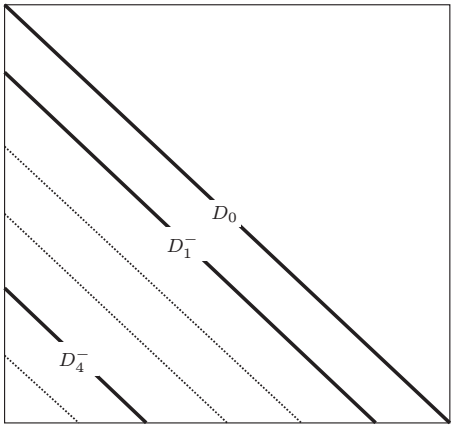


Fig. A.2. Stockage diagonal

$$\left|\begin{array}{ll} \text{si } \ell_c(i-j) = 0 \text{ alors} & \mathbb{A}_{ij} = 0 \\ \text{si } \ell_c(i-j) = p \text{ alors} & \mathbb{A}_{ij} = (\mathbb{C}_d)_{p+j-1} \end{array}\right.$$

On remarque que le descripteur de la structure compactée occupe un vecteur de longueur n, ce qui est très peu. Réciproquement, on est capable de trouver à quel terme $\mathbb{A}_{ij}$ correspond un terme $(\mathbb{C}_d)_\ell$ de la matrice compactée $\mathbb{C}_d$. Mais ceci à l'aide d'un procédé coûteux, puisqu'il faut détecter l'indice k tel que :

$$0 \neq \ell_c(k) \leq \text{ position de } (\mathbb{C}_d)_\ell < \ell_c(k+m) \neq 0.$$

L'utilisation d'un tel procédé est à bannir dans la pratique. Il est préférable d'utiliser un descripteur différent de ℓ_c. On remarque, et c'est une règle générale, que le choix d'un descripteur économique privilégie un des sens $\mathbb{A} \rightarrow \mathbb{C}_d$ ou $\mathbb{C}_d \rightarrow \mathbb{A}$. On choisit, par conséquent, le descripteur adapté aux calculs que l'on souhaite réaliser. Ainsi pour les méthodes directes il est préférable d'utiliser le sens $\mathbb{A} \rightarrow \mathbb{C}_d$, alors que pour les méthodes itératives, ne nécessitant que des produits de matrices par des vecteurs, les deux sens sont possibles.

• Compactage bande

Le compactage diagonal n'est pas compatible avec les techniques de factorisation $\mathbb{LU}$ ou $\mathbb{LL}^t$. En effet, s'il existe une sous-diagonale nulle (donc non stockée) comprise entre deux diagonales non nulles, le processus de factorisation affecte cette diagonale initialement nulle. C'est pourquoi, on introduit le compactage dit *bande*, dans lequel on conserve toutes les diagonales comprises entre deux diagonales non nulles. Evidemment, l'occupation mémoire est plus importante, mais comme on le verra par la suite, il est compatible avec les factorisations $\mathbb{LU}$ ou $\mathbb{LL}^t$.
On pose :

$$\mu = \max\{i-j,\ i \geq j \text{ tel que } \mathbb{A}_{ij} \neq 0\}$$

qui correspond à l'indice de la diagonale D_μ^- non nulle, la plus éloignée de la diagonale principale D_0. μ s'appelle la *largeur de bande* de la matrice $\mathbb{A}$. On a les propriétés immédiates :

- lorsque la matrice $\mathbb{A}$ est diagonale on a $\mu = 0$,
- lorsque $\mathbb{A}_{n1} = 0$ on a $\mu = n - 1$.

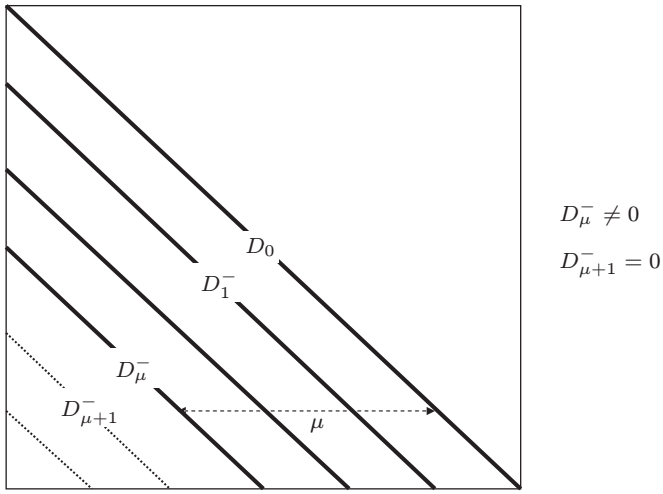


Fig. A.3. Matrice bande

Le compactage bande consiste à stocker tous les termes $\mathbb{A}_{ij}$ tels que $i \geq j$ et $(i-j) \leq \mu$. Il occupe plus de place mémoire que le compactage diagonal puisqu'on peut être amené à stocker des diagonales D_k^- qui sont nulles. Dans la pratique, on adopte soit le stockage diagonale par diagonale :

$$\mathbb{C}_b = \left(D_0, D_1^- \cdots D_\mu^-\right)$$

soit le stockage ligne par ligne :

$$\mathbb{C}_l = \left(\mathbb{A}_{11} \underbrace{\mathbb{A}_{21}\mathbb{A}_{22}}_{\text{ligne } 2} \cdots \underbrace{\mathbb{A}_{\mu+1,1} \ldots \mathbb{A}_{\mu+1,\mu+1}}_{\text{ligne } \mu+1} \cdots \underbrace{\mathbb{A}_{j,j-\mu} \ldots \mathbb{A}_{j,j}}_{\text{ligne } j} \cdots \underbrace{\mathbb{A}_{n,n-\mu} \ldots \mathbb{A}_{n,n}}_{\text{ligne } n} \right)$$

Quel que soit le stockage, la longueur du vecteur compacté $\mathbb{C}_b$ ou $\mathbb{C}_l$ est donnée par :

$$\sum_{m=0,\mu} (n-m) = n(\mu+1) - \frac{\mu(\mu+1)}{2}.$$

Le terme $\mathbb{A}_{ij}$ est facilement repéré dans le vecteur compacté $\mathbb{C}_b$ ou $\mathbb{C}_l$ à l'aide du nombre μ. En effet,

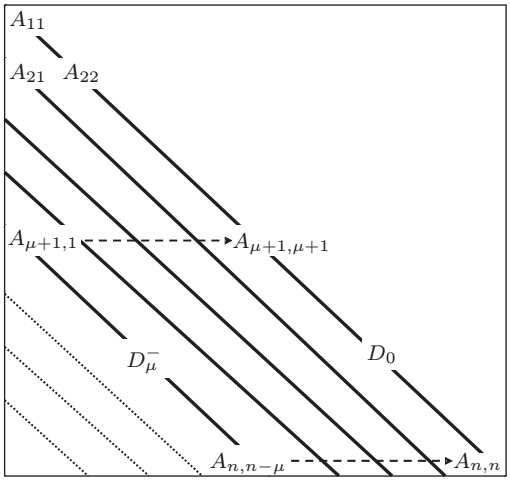


Fig. A.4. Stockage bande

- dans le cas du stockage diagonal $\mathbb{C}_b$, on a :

$$\left|\begin{array}{l} \text{si } i-j>\mu \text{ alors } \mathbb{A}_{ij}=0 \\ \text{si } i-j\le\mu \text{ alors } \mathbb{A}_{ij}=(\mathbb{C}_b)_\ell \quad \text{avec } \ell=\displaystyle\sum_{k=0}^{i-j-1}(n-k)+j \end{array}\right.$$

- dans le cas du stockage ligne $\mathbb{C}_l$, on a :

$$\left|\begin{array}{ll} \text{si } i-j<\mu & \text{alors } \mathbb{A}_{ij}=0 \\ \text{si } i-j\le\mu \text{ et } i\le\mu+1 & \text{alors } \mathbb{A}_{ij}=(\mathbb{C}_l)_\ell \text{ avec } \ell=\displaystyle\sum_{k=1}^{i-1}k+j=\tfrac{1}{2}i(i-1)+j \\ \text{si } i-j\le\mu \text{ et } i>\mu+1 & \text{alors } \mathbb{A}_{ij}=(\mathbb{C}_l)_\ell \text{ avec } \ell=\displaystyle\sum_{k=1}^{\mu}k+(\mu+1)(i-\mu)+j \end{array}\right.$$

• Compactage profil

Dès lors que μ est proche de n, le compactage bande conduit à un stockage presque plein, donc inintéressant en pratique. C'est pourquoi, on lui préfère le stockage dit *profil* (ou encore *skyline*), fondé sur le rangement par ligne et compatible avec les factorisations $\mathbb{LU}$ et $\mathbb{LL}^t$.

On commence par construire l'application profil P^- définie par :

$$P^-(i)=\min\{j,\ 1\le j<i \text{ tel que } \mathbb{A}_{ij}\neq 0\}$$

qui indique, pour chaque ligne i de la matrice, l'indice de la colonne du premier terme non nul de la ligne i. Le signe "$-$" indique que l'on ne se préoccupe que de

la partie triangulaire inférieure. L'application profil supérieur définie sur la partie triangulaire supérieure se définit de façon similaire. La définition de l'application P^- implique trivialement que :

$$\mathbb{A}_{ij} = 0 \text{ si } j < P^-(i)$$

et permet de tracer une "séparation", appelée profil de la matrice, entre les éléments nuls et ceux qui ne le sont pas :

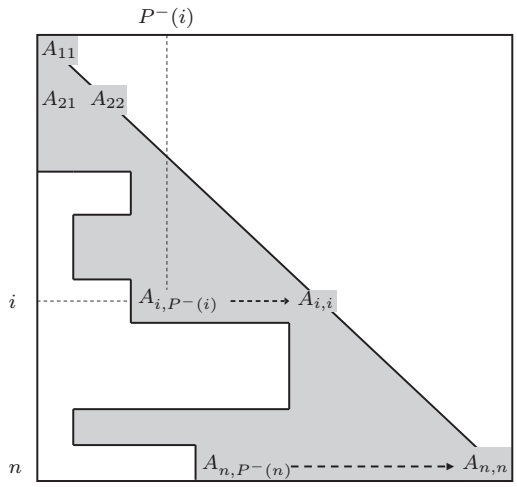


Fig. A.5. Stockage profil

On ne stocke que les termes appartenant au profil (partie grisée) en utilisant un stockage par ligne :

$$\mathbb{C}_p = \left(\mathbb{A}_{11} \underbrace{\mathbb{A}_{2,P^-(2)} \mathbb{A}_{22}}_{\text{ligne } 2} \cdots \underbrace{\mathbb{A}_{i,P^-(i)} \cdots \mathbb{A}_{ii}}_{\text{ligne } i} \cdots \underbrace{\mathbb{A}_{n,P^-(n)} \cdots \mathbb{A}_{n,n}}_{\text{ligne } n} \right)$$

Par défaut, on conserve toujours les termes diagonaux, même s'ils sont nuls. Il faut associer à ce compactage, un descripteur qui permette de retrouver un terme $\mathbb{A}_{ij}$ dans le vecteur compacté $\mathbb{C}_p$. Le plus couramment utilisé est celui qui associe au terme diagonal $\mathbb{A}_{ii}$ sa position dans le vecteur $\mathbb{C}_p$, que l'on note $\ell_p(i)$; on a donc :

$$\mathbb{A}_{ii} = (\mathbb{C}_p)_{\ell_p(i)} .$$

L'écart $e_i = \ell_p(i) - \ell_p(i-1)$ indique le nombre de termes stockés sur la ligne i et permet de retrouver le profil :

$$P^-(i) = i - e_i - 1.$$

Par conséquent, pour déterminer rapidement l'emplacement d'un terme $\mathbb{A}_{ij}$ dans le vecteur compacté $\mathbb{C}_p$, on utilise les formules suivantes :

$$\left|\begin{array}{ll} \text{si } j < P^-(i) & \text{alors } \mathbb{A}_{ij} = 0 \\ \text{si } j \geq P^-(i) & \text{alors } \mathbb{A}_{ij} = (\mathbb{C}_p)_\ell \text{ avec } \ell = \ell_p(i) + j - i \end{array}\right.$$

Stockage des matrices non symétriques
Lorsque l'on doit traiter le cas de matrices non symétriques à profil symétrique (cas des matrices éléments finis en général), on utilise le même profil P^- pour stocker la partie triangulaire supérieure suivant les colonnes. On stocke ainsi dans le vecteur compacté, à la suite de la partie triangulaire inférieure, les colonnes $\left(\mathbb{A}_{P^-(j),j} \ldots \mathbb{A}_{j-1,j}\right)_{j=1,n}$.
Si on doit traiter des matrices non symétriques quelconques, il y a deux possibilités de compactage :

- le compactage en ligne dans lequel on définit le profil supérieur $P^+(i)$ pour chaque ligne i comme le plus grand indice de colonne tel que $\mathbb{A}_{P^+(i),j} \neq 0$ et $\mathbb{A}_{ij} = 0 \ \forall j > P^+(i)$. On range alors la matrice ligne par ligne :

$$\left(\mathbb{A}_{i,P^-(i)} \ldots \mathbb{A}_{i,P+(i)}\right)_{i=1,n}.$$

- le compactage ligne/colonne dans lequel on définit le profil supérieur $P^+(j)$ pour chaque colonne j comme le plus grand indice de ligne $< j$ tel que $\mathbb{A}_{P^+(j),j} \neq 0$ et $\mathbb{A}_{ij} = 0 \ \forall i < P^+(j)$. On range alors la matrice en commençant par la partie triangulaire inférieure en utilisant le profil inférieur P^- puis la partie supérieure :

$$\left(\left(\mathbb{A}_{P^-(i)}, i \ldots \mathbb{A}_{i,i}\right)_{i=1,n}, \left(\mathbb{A}_{j,P^-(j)} \ldots \mathbb{A}_{j-1,j}\right)_{j=1,n}\right).$$

Afin de retrouver efficacement la position du terme $\mathbb{A}_{ij}$ il est nécessaire de connaître la position dans le vecteur compacté d'un terme particulier de la ligne i ou de la colonne j, en général soit le premier soit le terme diagonal. On laisse au lecteur le soin de trouver les formules permettant de retrouver efficacement la position d'un terme quelconque.

• Compactage morse
Evidemment, le compactage profil n'est pas optimal car dans une ligne tronquée il peut exister de nombreux zéros. En particulier, dans le cas où la première colonne de la matrice ne présente aucun terme nul, le stockage profil devient un stockage plein ! L'introduction du stockage *morse* qui consiste à ne stocker que les termes non nuls permet d'atteindre un stockage quasi-optimal en terme d'occupation mémoire. Il n'est bien sûr plus compatible avec les factorisations $\mathbb{LU}$ ou $\mathbb{LL}^t$ mais se révèle particulièrement efficace dans le contexte des méthodes itératives. Il existe essentiellement deux techniques de description d'un compactage morse :

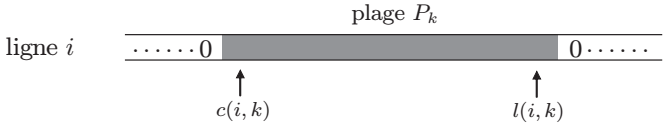
- l'une fondée sur la notion de *plages* de termes non nuls par ligne
- l'autre sur la connaissance explicite des indices de colonne (ou de ligne) des termes non nuls d'une ligne (ou d'une colonne).

En terme de gestion, la première méthode est généralement plus économique que la seconde en place mémoire, au détriment d'une performance moindre. La seconde méthode, appelée *compressed row (column) form* est la plus répandue. C'est en particulier la technique utilisée pour les matrices creuses (type *sparse* en Matlab). Nous allons décrire brièvement ces deux techniques.

Le descripteur par *plage* consiste à décrire les termes non nuls d'une ligne de la matrice comme une succession de plages de termes non nuls de longueur variable, le nombre de plages étant également variable.

Le vecteur de compactage $\mathbb{C}_m$ est alors constitué de toutes les plages de termes non nuls, rangées dans l'ordre des lignes. On utilise les descripteurs suivants, pour chaque ligne i de la matrice :

- $n_p(i)$: le nombre de plages de termes non nuls dans la ligne i, et pour chaque plage $P_k,\ k = 1, n_p(i)$:
  - $c(i,k)$: le numéro de la colonne où débute la plage P_k,
  - $\ell(i,k)$: la position dans le vecteur $\mathbb{C}_m$ du dernier terme de la plage P_k.



Notons que le descripteur utilisé ne permet pas de retrouver très simplement le rang (i,j) d'un terme quelconque de $\mathbb{C}_m$. Néanmoins, il est suffisant dès lors que l'on réalise un calcul *séquentiel* sur le vecteur $\mathbb{C}_m$, un produit matrice-vecteur par exemple. En effet, il faut savoir retrouver à quelle colonne correspond un terme $\mathbb{C}_m(q)$ situé sur une plage P_k de la ligne i ; ce qui est facile car $c(i,k)$ indique le numéro de colonne où débute la plage k.
Le stokage par *plage* peut devenir prohibitif dans certaines situations à cause de la place occupée par les tableaux servant à décrire la structure morse. Si N_p désigne le nombre total de plages, ces tableaux occupent :

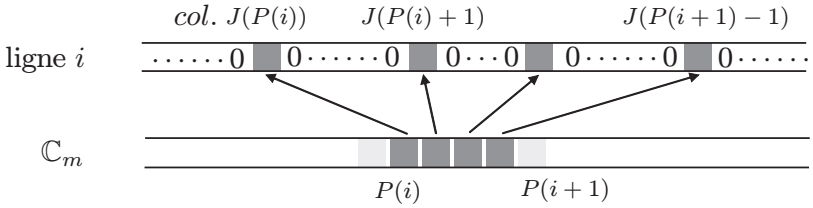
$$n + 2N_p \text{ places.}$$

Ainsi, pour une matrice *damier* (alternance de 1 et 0) on a :

$$N_p = n^2/2$$

La longueur du vecteur compacté $\mathbb{C}_m$ est également $n^2/2$, ce qui donne une occupation totale de $n + \frac{3}{2}n^2$, soit plus que le stockage plein!

Le descripteur *compressed row* utilise d'une part, un vecteur J indiquant l'indice en colonne de chaque terme non nul stocké ligne par ligne dans le vecteur compacté $\mathbb{C}_m$ et un vecteur P indiquant pour chaque ligne i la position dans le vecteur $\mathbb{C}_m$ (ou J) du premier terme non nul de la ligne.



Les termes non nuls de la ligne i sont donc donnés par :

$$\mathbb{C}_m(k) \text{ pour } P(i) \leq k < P(i+1) \text{ avec } J(k) \text{ le n° de colonne.}$$

Le vecteur J est de la même taille que le vecteur $\mathbb{C}_m$ (nombre de termes non nuls de la matrice) et le vecteur P a pour taille le nombre de lignes. Pour la matrice d'ordre 5 suivante :

$$\begin{bmatrix} 1\,7\,0\,0\,0 \\ 2\,0\,4\,0\,1 \\ 0\,2\,8\,0\,0 \\ 4\,0\,0\,3\,6 \\ 0\,0\,1\,1\,0 \end{bmatrix}$$

on a les vecteurs suivants :

$$\left|\begin{array}{ll} \mathbb{C}_m = & \big(1\,7\,2\,4\,1\,2\,8\,4\,3\,6\,1\,1\big) \\ J = & \big(1\,2\,1\,3\,5\,2\,3\,1\,4\,5\,3\,4\big) \\ P = & \big(1\,3\,6\,8\,11\big) \end{array}\right. .$$

Tout comme le descripteur par *plage*, le descripteur *compressed row* ne permet pas de trouver directement la position d'un terme (i, j) dans le vecteur compacté (il est nécessaire d'effectuer une recherche). Par contre, les algorithmes qui parcourent séquentiellement le vecteur compacté $\mathbb{C}_m$ seront efficaces. Dans le cas où la matrice n'est pas très creuse, le descripteur *compressed row* occupe plus de place que le descripteur *plage*. Par contre, il est plus rapide car l'obtention du n° de colonne d'un terme du vecteur compacté $\mathbb{C}_m$ est immédiate (vecteur J), alors que pour le descripteur *plage* un calcul est nécessaire.
Bien évidemment, si la matrice est symétrique on ne stocke que la partie triangulaire inférieure.

A.4.2 Algorithmes pour les matrices compactées

Nous venons de présenter les quatre principaux types de stockage. Dans le contexte de la résolution des problèmes d'EDP par éléments finis (ou par differences finies), le choix d'une numérotation des inconnues joue un grand rôle dans la structure des matrices auxquelles on aboutit. En particulier, les algorithmes d'optimisation de place mémoire (Cuthill-Mackee par exemple) recherchent la numérotation qui minimise la largeur de bande μ et pour certains, la taille effective du vecteur compacté profil. Le stockage morse est pratiquement insensible à la numérotation. Par exemple, dans le cas d'un maillage par éléments finis de Lagrange on cherchera, compte-tenu des propriétés de support des fonctions de base globales, à minimiser l'écart d'indice $\mid i-j \mid$ des points (M_i, M_j) appartenant au même élément K_ℓ. La largeur de bande μ est donnée dans ce cas par :

$$\mu = \max_{K_\ell \in \mathcal{T}_h} \left(\max \{ \mid i-j \mid \text{ tel que } M_i, M_j \in K_\ell \}\right)$$

Donnons maintenant quelques indications concernant les avantages et les inconvénients de ces structures de stockage, ainsi que des critères de choix.

• Occupation mémoire
Notons m_d, m_b, m_p et m_m les longueurs respectives du vecteur compacté $\mathbb{C}$ par les méthodes de stockage *diagonal*, *bande*, *profil* et *morse*. On a toujours :

$$m_b \geq m_p \geq m_m$$

car par construction : $P^-(i) \leq \mu$ et le stockage *morse* exclut tous les termes nuls, a fortiori ceux qui sont hors-profil. Par ailleurs, on a aussi :

$$m_b \geq m_d$$

car le stockage *bande* comprend toutes les diagonales du stockage *diagonal*. Par contre, on ne peut pas évaluer m_d par rapport à m_p, car cela dépend de la structure de la matrice. Mais on a $m_d \geq m_m$.

A titre d'exemple, regardons le cas du maillage éléments finis Q^1 Lagrange donné par la figure A.6. Un calcul un peu fastidieux montre que pour la matrice de rigidité, symétrique et d'ordre p^2, on a les estimations suivantes :

$$m_b = O(p^3) \quad m_p = O(p^3) \quad m_d = O(5p^2) \quad m_m = O(5p^2)$$

à comparer au stockage plein en $O(\frac{1}{2}p^4)$. Le gain que présentent les stockages *diagonal* et *morse* est donc appréciable. On se trouve dans le cas d'un maillage structuré, c'est pourquoi le stockage *diagonal* est aussi compétitif que le stockage *morse*. Si la place mémoire est un critère important, le temps de calcul des

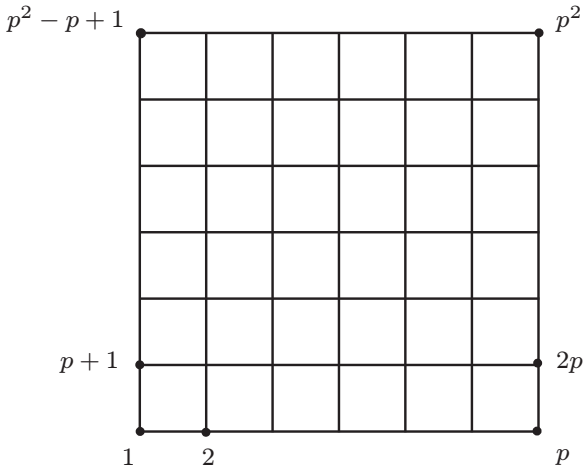


Fig. A.6. Maillage Q^1

opérations matricielles (factorisation, produit) est tout aussi important. En outre, il est primordial de savoir quelles sont les opérations matricielles qu'autorise une structure de stockage.

• Opérations matricielles et compactage
Nous ne nous intéressons qu'aux opérations matricielles qui interviennent dans les méthodes de résolution des systèmes linéaires, que nous avons présentées précédemment. Ces opérations sont de trois types :

- les factorisations (élimination de Gauss, factorisation $\mathbb{LU}$ ou $\mathbb{LL}^t$),
- les produits matrice-vecteur intervenant dans les techniques itératives,
- les extractions de sous-matrices intervenant dans les techniques itératives de Gauss-Seidel.

Ce dernier type d'opérations s'apparentant à celui des produits matrice-vecteur, il s'inscrit donc dans le cas précédent. Afin de simplifier la présentation, nous supposons toujours que la matrice $\mathbb{A}$ du système est symétrique définie-positive.

Les factorisations de type Cholesky
Rappelons qu'il s'agit de factoriser la matrice $\mathbb{A}$ sous la forme : $\mathbb{LL}^t = \mathbb{A}$. On a le résultat fondamental suivant :

Proposition A.10. *Soit $\mathbb{A}$ une matrice se factorisant sous la forme $\mathbb{LL}^t$, alors la matrice triangulaire inférieure $\mathbb{L}$ a la même structure profil (resp. bande) que la matrice $\mathbb{A}$.*

Cette proriété se démontre par récurrence. On peut tirer parti de la structure bande ou profil de la matrice $\mathbb{A}$ pour réduire le temps de calcul de la factorisation

de Cholesky, d'une part, en n'effectuant pas les calculs des termes $\mathbb{L}_{ij}$ qui sont hors-profil et, d'autre part, en n'effectuant pas les produits par les termes nuls, également hors-profil. Cela conduit à un algorithme très efficace, fondé sur la forme "ligne" de l'algorithme de Cholesky :

pour $p = 1, n$
  pour $j = 1, P^-(p), p$
    $m = \max(P^-(p), P^-(j))$
    $$\mathbb{L}_{pj} = \left(\mathbb{A}_{pj} - \sum_{k=m}^{j-1} \mathbb{L}_{pk}\mathbb{L}_{jk} \right) / \mathbb{L}_{jj}$$
  fin
  $$\mathbb{L}_{pp} = \left(\mathbb{A}_{pp} - \sum_{k=P^-(p)}^{j-1} |\mathbb{L}_{pk}|^2 \right)^{1/2}$$
fin

Remarque A.1. Il est facile d'écrire la version programmable de cet algorithme faisant seulement appel au pointeur de compactage ℓ_p (exercice). L'utilisation de la forme "colonne" de l'algorithme de Cholesky conduit à une méthode moins efficace car on doit effectuer des tests d'appartenance au profil. Il n'est pas possible d'utiliser des techniques de pivotage et ce même dans le cas des stockages bande ou profil, car elles détruisent l'application profil.

Les stockages *diagonal* et le *morse* sont incompatibles avec la factorisation de Cholesky, car cette dernière affecte les zéros inclus dans le profil. La propriété de conservation du profil et celle d'optimalité du calcul expliquent en grande partie le fait que le compactage profil est l'un des plus utilisés dans les codes d'éléments finis.

Les produits matrice-vecteur

Le produit matrice-vecteur est la principale opération utilisée par les méthodes de résolution itératives. Comme il n'y a pas de création ou de modification de la matrice lors de cette opération, tous les types de stockage supportent cette opération. Le stockage *morse* sera donc le plus compétitif si tant est que l'on sache écrire l'algorithme optimal (en nombre d'opérations) du produit. Indiquons les formes que prend cet algorithme dans le cas du stockage *profil* et dans le cas du stockage *morse* pour une matrice symétrique.
Compte-tenu de la structure du pointeur profil P^- (adressage en ligne), l'algorithme du produit matrice-vecteur en stockage *profil* est optimal si on le décompose en deux produits :

$$\mathbb{A}X = \mathbb{L}X + \mathbb{U}X$$

où $\mathbb{L}$ désigne la partie triangulaire inférieure de la matrice $\mathbb{A}$ (diagonale principale incluse), et $\mathbb{U}$ la partie triangulaire supérieure de $\mathbb{A}$ (diagonale principale exclue). Le produit $\mathbb{L}X$ est réalisé à l'aide de la forme "ligne" de l'algorithme du produit tandis que le produit $\mathbb{U}X$ est réalisé à l'aide de la forme "colonne", évitant ainsi des tests d'appartenance au profil.

```
pour i = 1, n
  Y_i = 0
  pour j = 1, P^-(i), i
    Y_i = Y_i + A_ij X_j
  fin
fin
pour j = 2, n, p
  pour i = 1, P^-(j), j - 1
    Y_i = Y_i + A_ij X_j
  fin
fin
```

Pour les matrices compactées *morse*, l'écriture d'un algorithme efficace de produit matrice-vecteur s'appuie sur le fait que les descripteurs de compactage permettent de retrouver simplement la position (i, j) du $\ell^{ème}$ élément du vecteur compacté $\mathbb{C}_m$. On effectue alors l'opération :

$$Y_i = Y_i + (\mathbb{C}_m)_\ell \times X_j.$$

Dans le cas d'une matrice quelconque stockée *morse* et utilisant le descripteur *compressed row*, on écrit l'algorithme suivant où on suppose que $P(n+1) = L = \dim(\mathbb{C}_m) + 1$:

```
pour i = 1, n
  Y_i = 0
  pour k = P(i), P(i + 1) - 1
    Y_i = Y_i + C_m(k) X_J(k)
  fin
fin
```

Les algorithmes de produit matrice-vecteur présentés ici sont optimaux au sens où ne sont effectués que les produits de termes non nuls.

Bibliographie

1. R. Abraham, J. E. Marsden, T. Ratiu, *Manifolds, tensor analysis and applications*, Coll. Applied Mathematical Science, 75, Springer Verlag (1988).
2. R. A. Adams, *Sobolev spaces*, Academic Press (1975).
3. T. Apel, *Anisotropic finite elements: local estimates and applications*, Advances in Numerical Mathematics, B. G. Teubner (1999).
4. C. Bernardi, M. Dauge, Y. Maday, *Spectral methods for axisymmetric domains*, Series in Applied Mathematics, 3, Gauthier-Villars & North Holland (1999).
5. C. Bernardi, Y. Maday, *Approximations spectrales de problèmes aux limites elliptiques*, Coll. Mathématiques et Applications, 10, Springer Verlag (1992).
6. A.-S. Bonnet-Ben Dhia, M. Lenoir, *MA102 : Outils élémentaires d'analyse pour les EDP*, Cours ENSTA (2007).
7. D. Braess, *Finite elements, theory, fast solvers and applications in solid mechanics*, Cambridge University Press (1992)
8. H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson (1983).
9. P. Ciarlet, H. Zidani, *AO101 : Optimisation quadratique*, Cours ENSTA (2007).
10. Philippe G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson (1982).
11. Philippe G. Ciarlet, *Basic error estimates for elliptic problems*, dans le Handbook of numerical analysis, Vol. II, Eds. P.G. Ciarlet and J.-L. Lions, North Holland, pp. 17–351 (1991).
12. G. Cohen, *Higher order numerical methods for transient wave equations*, Springer (2002).
13. R. Dautray, J. L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson (1987).
14. P. J. Davis, P. Rabinowitz, *Methods of numerical integration*, Academic Press (1984).
15. G. Derveaux, V. Duwig, S. Fliss, P. Joly, J.-R. Li, P. Moireau, J. Rodríguez, *MA103 : Introduction à la discrétisation des équations aux dérivées partielles*, Cours ENSTA (2007).
16. D. Gilbarg, N. S. Trudinger, *Elliptic partial differential equations of second order*, Springer Verlag (1977).
17. V. Girault, P.-A. Raviart, *Finite element methods for Navier-Stokes equations*, Coll. Springer Series in Computational Mathematics, 5, Springer Verlag (1986).
18. P. Grisvard, *Elliptic problems in nonsmooth domains*, Coll. Monographs and Studies in Mathematics, 24, Pitman (1985).
19. P. Grisvard, *Singularities in boundary value problems*, Coll. Recherche en Mathématiques Appliquées, 22, Masson (1992).
20. V. Khoan, *Disributions, analyse de Fourier, opérateurs aux dérivées partielles*, Vuibert (1972).

21. M. Lenoir, *Approximation par éléments finis des problèmes elliptiques*, Publication ENSTA 771 (1987).
22. J.-L. Lions, E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod (1968).
23. G. Meurant, *Computer solution of large linear systems*, Elsevier (1999).
24. J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson (1967).
25. P.-A. Raviart, J.-M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson (1983).

Liste des figures

Index