La méthode des éléments finis : de la théorie à la pratique. Tome 2 : Compléments

Eliane Bécache, Patrick Ciarlet, Christophe Hazard, Éric Lunéville

Les cours

E. Bécache, P. Ciarlet, C. Hazard & E. Lunéville

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

II. Compléments

PARIS
LES PRESSES DE L'ENSTA
32, boulevard Victor, Paris 15^{e}
2010

Table des matières

Avant-propos

La simulation numérique est devenue un puissant moyen d'investigation qui tend à prendre une place croissante, à côté de l'approche expérimentale classique, dans les sciences et techniques.

Dans ce cours, nous nous proposons de présenter une des principales méthodes numériques – les *Eléments Finis* – permettant de résoudre les équations aux dérivées partielles issues de nombreux problèmes physiques. Ce cours est divisé en deux Parties, la première introduisant les concepts généraux (voir [15]), la seconde présentant un certain nombre de compléments.

Dans la première Partie, le premier chapitre a été consacré à l'étude des problèmes *elliptique*s. En particulier, la *théorie variationnelle* des équations elliptiques y a été exposée dans un cadre fonctionnel rigoureux, incluant une description élémentaire des *espaces de Sobolev*. La méthode des éléments finis a fait l'objet des deux chapitres suivants. Nous avons introduit le cadre formel qui permet de construire une grande variété d'exemples d'éléments finis. Nous avons poursuivi par quelques résultats d'*estimations d'erreurs*. Puis, nous avons étudié les aspects pratiques et algorithmiques de cette méthode, liés à sa *mise en œuvre*. Enfin, dans une annexe, nous avons présenté quelques considérations élémentaires sur la *résolution des systèmes linéaires*.

La seconde Partie, qui fait l'objet de cet ouvrage, débute par une présentation de la *théorie spectrale* des opérateurs elliptiques qui constitue un outil indispensable à l'analyse des phénomènes de *vibrations*. Dans le chapitre 2, nous abordons la théorie des *problèmes mixtes*, qui permet de résoudre des systèmes d'équations faisant intervenir plusieurs inconnues de natures différentes. Ensuite, dans les chapitres 3 et 4, nous nous intéressons d'une part à l'*équation de la chaleur* et d'autre part à l'*équation des ondes*, comme modèles d'*équations d'évolution*. Cette partie ne comporte pas de nouvelles notions mathématiques et peut être considérée comme

une application des concepts exposés précédemment, en particulier la *théorie variationnelle* et la *théorie spectrale*. La technique de discrétisation présentée est plurielle puisqu'elle utilise une approximation par éléments finis en espace et par différences finies en temps.

Afin d'appréhender les aspects pratiques de la méthode des éléments finis, nous présentons à la fin de chaque chapitre quelques illustrations numériques significatives des concepts traités. Nous considérons d'une part des applications "académiques" et d'autre part des exemples plus élaborés, plus proches de situations "réelles". Nous fournissons la plupart des codes Matlab [1] correspondants.

Enfin, les auteurs tiennent à remercier Anne-Sophie Bonnet-Ben Dhia et Patrick Joly pour leurs contributions. Les chapitres 1 et 4 ont été construits à partir de certains de leurs supports de cours.

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

1

Analyse spectrale des problèmes elliptiques

Dans ce chapitre, on s'intéresse aux valeurs λ telles que le problème :

$$-\Delta u = \lambda u \text{ dans } \Omega, \tag{1.1}$$

($\Omega \subset \mathbb{R}^n$) assorti de conditions aux limites *homogènes* prescrites sur la frontière $\partial\Omega$ de Ω, admette des solutions u non nulles. On dit alors que λ est une *valeur propre* et u une *fonction propre* associée.

Tout au long du chapitre, nous supposerons que Ω est *borné* et cette hypothèse est essentielle. Dans ce contexte, l'ensemble des valeurs propres est appelé *spectre* du problème (1.1).

On peut remarquer que si u est solution de (1.1), alors la fonction $v(x,t) = \mathbf{Re}\left\{u(x)e^{i\omega t}\right\}$, où $\omega^2 = \lambda$, est solution de l'équation des ondes (voir aussi le chapitre 4) :

$$\frac{\partial^2 v}{\partial t^2} - \Delta v = 0 \qquad x \in \Omega,\ t \in \mathbb{R}. \tag{1.2}$$

Cette équation modélise par exemple les petits mouvements d'un fluide compressible dans une cavité acoustique (v représente dans ce cas le potentiel des vitesses ou la pression). La résolution du problème aux valeurs propres (1.1) assorti de conditions aux limites appropriées correspond alors au calcul des *fréquences propres* de la cavité et des *"harmoniques"* associées, bien connues en musique.

Ce type de problème se rencontre également en mécanique lorsque l'on cherche à calculer les *modes de vibration* propres d'un corps élastique. On retrouve en particulier l'équation (1.1) en dimension $n = 1$ dans le cas d'une corde vibrante et en dimension $n = 2$ pour une membrane vibrante (un tambour !).

Dans le même ordre d'idée, on peut s'intéresser à l'équation des ondes avec un second membre non-nul :

$$\frac{\partial^2 v}{\partial t^2} - \Delta v = g \qquad x \in \Omega,\ t \in \mathbb{R}, \tag{1.3}$$

lorsque la dépendance en temps est connue, en $e^{i\omega t}$ (régime harmonique). Dans ce cas, on peut écrire $g(x,t) = \mathbf{Re}\left\{f(x)e^{i\omega t}\right\}$ et $v(x,t) = \mathbf{Re}\left\{u(x)e^{i\omega t}\right\}$, avec u solution de l'équation de Helmholtz :

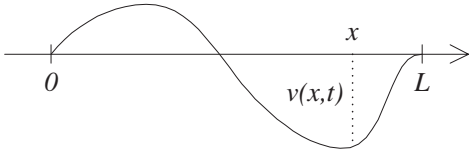
$$-\Delta u - \omega^2 u = f \text{ dans } \Omega. \tag{1.4}$$

A la première section, nous mettons en évidence sur deux exemples (la corde vibrante et l'équation de Helmholtz) les principales propriétés spectrales en s'appuyant soit sur la connaissance analytique du spectre (corde vibrante) soit sur la formulation variationnelle de l'équation de Helmholtz. La section suivante a pour objet l'extension de ces résultats dans le contexte d'un problème aux valeurs propres abstrait qui offre un cadre général pour l'étude des *vibrations libres* d'un système conservatif. On y énonce et démontre le théorème de décomposition spectrale d'un opérateur compact auto-adjoint, qui fournit les propriétés fondamentales des valeurs et vecteurs propres du système. On donne ensuite une caractérisation très utile des valeurs propres, connue sous le nom de *principe du Min-Max*. Enfin, on s'intéresse à un problème abstrait similaire à (1.4) qui modélise cette fois les vibrations du système *entretenues* par une excitation périodique. L'existence et l'unicité de la solution entrent dans le cadre de *l'alternative de Fredholm* que nous déduirons ici du théorème spectral. Nous abordons, à la troisième section, la discrétisation et l'analyse numérique des valeurs propres et des fonctions propres par une méthode d'approximation interne. Le problème discret consiste à calculer des couples vecteur propre – valeur propre de matrices. L'exposé de la méthode de la puissance inverse, qui permet le calcul effectif de ces couples, suit. Nous achevons cette section par l'analyse numérique du problème des vibrations entretenues. Enfin, nous concluons ce chapitre par des illustrations numériques.

1.1 Exemples de problèmes aux valeurs propres

1.1.1 Un exemple mono-dimensionnel

Considérons une corde homogène occupant au repos le segment $[0, L]$ de l'axe des x, et maintenue fixe à ses deux extrémités $x = 0$ et $x = L$. On note $v(x,t)$ le "petit" déplacement transversal de la corde en x à l'instant t.



En appliquant le principe fondamental de la dynamique, on peut montrer que, en l'absence de sollicitations extérieures (on néglige la pesanteur), v est solution de :

$$\begin{cases} \dfrac{\partial^2 v}{\partial t^2} - c^2 \dfrac{\partial^2 v}{\partial x^2} = 0 & x \in]0, L[,\ t > 0, \\ v(0,t) = v(L,t) = 0 & t > 0, \end{cases} \tag{1.5}$$

avec $c = \sqrt{\tau/\rho}$, τ étant la tension de la corde et ρ sa masse linéique.

On cherche les solutions harmoniques en temps de (1.5), c'est-à-dire les solutions de la forme :

$$v(x,t) = \mathbf{Re}\left\{u(x)e^{i\omega t}\right\},$$

où u est maintenant à valeurs complexes, ce qui nous conduit au problème suivant :

$$\begin{cases} -u'' = \lambda u & x \in]0, L[, \\ u(0) = u(L) = 0, \end{cases} \tag{1.6}$$

avec $\lambda = \omega^2/c^2$. Plus exactement, on cherche toutes les valeurs complexes de λ telles que le problème (1.6) admette des solutions u non nulles.

Ici, l'équation différentielle satisfaite par u est particulièrement simple. Regardons tout d'abord le cas où λ est réel et positif. La forme générale de la solution est :

$$u(x) = A\cos\left(\frac{\omega}{c}x\right) + B\sin\left(\frac{\omega}{c}x\right).$$

D'après les conditions aux limites, on a de plus :

$$A = 0 \text{ et } B\sin\left(\frac{\omega}{c}L\right) = 0.$$

Comme la solution nulle ne nous intéresse pas, on en déduit l'équation satisfaite par les pulsations propres, appelée également *équation de dispersion* :

$$\sin\left(\frac{\omega}{c}L\right) = 0, \quad \omega \neq 0. \tag{1.7}$$

Cette équation admet une suite de solutions :

$$\omega_k = \frac{k\pi}{L}c, \quad k \in \mathbb{N}^*. \tag{1.8}$$

Par des calculs similaires, on montre facilement que le problème (1.6) n'admet pas de solution non triviale si $\lambda \leq 0$, ou si $\mathbf{Im}(\lambda) \neq 0$.

Finalement, les solutions du problème aux valeurs propres (1.6) sont données par :

$$\lambda_k = \frac{k^2\pi^2}{L^2},\ u_k(x) = \sin\left(\frac{k\pi x}{L}\right),\quad k \in \mathbb{N}^*. \tag{1.9}$$

L'ensemble $\{\lambda_k,\ k \geq 1\}$ est appelé *spectre* du problème (1.6). Bien entendu, les fonctions propres u_k sont déterminées à une constante multiplicative non-nulle près, et on remarque que $\lim_{k\to\infty} \lambda_k = +\infty$.

Un calcul classique montre que ces fonctions propres sont orthogonales 2 à 2 pour le produit scalaire $L^2(]0, L[)$:

$$\int_0^L \sin\left(\frac{k\pi x}{L}\right) \sin\left(\frac{k'\pi x}{L}\right) dx = 0 \qquad \text{si } k \neq k'. \tag{1.10}$$

Par ailleurs, les résultats usuels sur les séries de Fourier (cf. [7])) montrent que toute fonction f continûment dérivable sur $[0, L]$ et telle que $f(0) = f(L) = 0$ admet une décomposition de la forme :

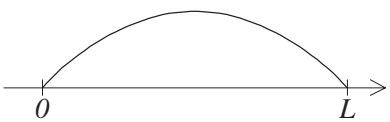
$$f(x) = \sum_{k=1}^{+\infty} \alpha_k \sin\left(\frac{k\pi x}{L}\right) \tag{1.11}$$

où la série converge uniformément sur $[0, L]$. Autrement dit, toute fonction f satisfaisant aux mêmes conditions limites que les fonctions propres, se décompose sur ces dernières.

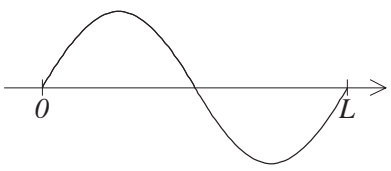
Nous allons voir dans la suite que les diverses propriétés (suite de valeurs propres tendant vers $+\infty$, orthogonalité et complétude des fonctions propres) restent vraies dans des cas beaucoup plus généraux.

Remarque 1.1 *Attention ! Les solutions du problème (1.6) dépendent fondamentalement des conditions aux limites. Le lecteur s'en convaincra en reprenant tous les calculs dans le cas des conditions suivantes : $u'(0) = u'(L) + u(L) = 0$.*

Remarque 1.2 *On peut remarquer que la $k^{ème}$ fonction propre u_k admet exactement $(k-1)$ zéros (ou "nœuds") dans l'intervalle $]0, L[$. Cette propriété est en fait assez générale, elle permet de reconnaître l'ordre k d'une valeur propre au vu de la fonction propre associée.*



Premier mode



Second mode

Remarque 1.3 *On voit (cf. (1.8)) que les fréquences propres décroissent quand la longueur L de la corde augmente. C'est pourquoi un guitariste obtient un son plus aigu en raccourcissant la longueur de la corde.*

1.1.2 Problème de Helmholtz

Soit Ω un ouvert borné, de frontière "suffisamment régulière", contenu dans $\mathbb{R}^n$. On se donne une partition de la frontière $\partial\Omega$:

$$\partial\Omega = \overline{\Gamma}_1 \cup \overline{\Gamma}_2 \quad \text{avec } \Gamma_1 \cap \Gamma_2 = \emptyset.$$

Supposons que pour une fonction donnée $f \in L^2(\Omega)$ à valeurs complexes, et pour un nombre complexe λ donné [1], on veuille résoudre le problème suivant :

$$\begin{cases} -\Delta u - \lambda u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \Gamma_1, \\ \dfrac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_2. \end{cases} \tag{1.12}$$

Pour ce faire, on applique la théorie variationnelle habituelle (cf. [15]).

Notons V le sous-espace de $H^1(\Omega)$ défini par :

$$V := \left\{ v \in H^1(\Omega) \ \text{ tel que } v_{|\Gamma_1} = 0 \right\}. \tag{1.13}$$

En utilisant la continuité de l'application trace sur Γ_1, on note que V est fermé dans $H^1(\Omega)$: on peut donc le munir du produit scalaire de $H^1(\Omega)$, à savoir

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

On vérifie alors classiquement que résoudre le problème de Helmholtz ci-dessus est équivalent à résoudre l'équation variationnelle suivante :

$$\begin{cases} \text{trouver } u \in V \text{ tel que} \\ \displaystyle\int_\Omega (\nabla u \cdot \overline{\nabla v} - \lambda\, u\, \overline{v})\, d\Omega = \int_\Omega f\, \overline{v}\, d\Omega \quad \forall v \in V. \end{cases} \tag{1.14}$$

On souhaite appliquer au problème (1.14) le théorème de Lax-Milgram :

– Si $\lambda < 0$, la coercivité de la forme bilinéaire est évidente et l'on peut conclure : le problème (1.14) est bien posé. On peut montrer qu'il en est de même si $\mathbf{Im}(\lambda) \neq 0$.

1. Pour l'équation de Helmholtz (1.4), on a $\lambda = \omega^2$.

Lorsque $a(\cdot,\cdot)$ est une forme sesquilinéaire à valeurs dans $\mathbb{C}$, on dit qu'elle est coercive si il existe une constante $\alpha > 0$ telle que :

$$|a(v,v)| \geq \alpha \|v\|_V^2, \quad \forall v \in V.$$

A partir de là, on peut appliquer le théorème de Lax-Milgram pour résoudre le problème abstrait :

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

sous des hypothèses similaires (cf. [38]) à celles du cas de formes à valeurs dans $\mathbb{R}$: une forme $a(\cdot,\cdot)$ sesquilinéaire, continue et coercive sur V, et une forme $\ell(\cdot)$ antilinéaire et continue sur V.

– Rien n'est moins sûr en revanche si $\lambda \geq 0$.

En effet, posons

$$a(u,v) = \int_\Omega (\nabla u \cdot \overline{\nabla v} - \lambda\, u\, \overline{v})\, d\Omega$$

et plaçons-nous par exemple dans le cas où $\Gamma_1 = \emptyset$ ($V = H^1(\Omega)$), avec $\lambda \neq 0$ (si $\lambda = 0$ les fonctions constantes annulent $v \mapsto a(v,v)$) :
– si on choisit $v_0 \in \mathbb{R}$ une constante non nulle, alors $a(v_0, v_0) < 0$;
– si on choisit $v_1(x) = e^{\mu x_1}$ avec $\mu > \sqrt{\lambda}$, alors $a(v_1, v_1) > 0$.
Puis, on remarque que $g : \beta \mapsto a(\beta v_0 + (1-\beta)v_1, \beta v_0 + (1-\beta)v_1)$ est continue et telle que $g(0) > 0$ et $g(1) < 0$. D'après le théorème des valeurs intermédiaires, il existe $\beta' \in]0,1[$ tel que $g(\beta') = 0$: or $v' = \beta' v_0 + (1-\beta')v_1$ est non-nulle, ce qui permet de conclure à la non-coercivité de $a(\cdot,\cdot)$.

Nous allons justement montrer qu'il n'y a pas unicité pour certaines valeurs de λ. Autrement dit, nous allons chercher les valeurs λ (nécessairement positives d'après ce qui précède) telles que le problème :

$$\begin{cases} -\Delta u = \lambda u & \text{dans } \Omega, \\ u = 0 & \text{sur } \Gamma_1, \\ \dfrac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_2, \end{cases} \tag{1.15}$$

admette des solutions u non nulles dans $H^1(\Omega)$.

Remarque 1.4 *Le problème (1.6) est un cas particulier de (1.15) avec $\Omega =]0, L[$, $\Gamma_2 = \emptyset$, $\Gamma_1 = \partial\Omega$.*

La formulation variationnelle de (1.15) s'écrit :

$$\begin{cases} \text{trouver } \lambda \in \mathbb{C} \text{ et } u \in V \setminus \{0\} \text{ tels que} \\ \displaystyle\int_\Omega \nabla u \cdot \overline{\nabla v}\, d\Omega = \lambda \int_\Omega u\, \overline{v}\, d\Omega, \quad \forall v \in V. \end{cases} \tag{1.16}$$

Dans toute la suite, on dira que λ est une *valeur propre* si le problème (1.16) admet une solution u non nulle. Dans ce cas, u est appelée une *fonction propre*.

Proposition 1.5 *On a les résultats suivants :*
(i) Toutes les valeurs propres du problème (1.16) sont réelles positives.
(ii) Soient λ_1 et λ_2 deux valeurs propres distinctes et soient u_1 et u_2 deux fonctions propres associées. Alors

$$\int_\Omega \nabla u_1 \cdot \overline{\nabla u_2}\, d\Omega = \int_\Omega u_1\, \overline{u_2}\, d\Omega = 0. \tag{1.17}$$

(iii) Supposons que $\Gamma_1 = \emptyset$ ou $\Gamma_2 = \emptyset$ (problème de Dirichlet ou de Neumann "pur"). Alors, si la frontière $\partial\Omega$ est $\mathcal{C}^\infty$, toute fonction propre u de (1.16) est telle que :

$$u \in \mathcal{C}^\infty(\overline{\Omega}).$$

Démonstration : Effectuons celle-ci point par point.

(i) En prenant $v = u$ dans (1.16), on obtient

$$\int_\Omega |\nabla u|^2\, d\Omega = \lambda \int_\Omega |u|^2\, d\Omega.$$

Comme $u \neq 0$, il en résulte que $\lambda \in \mathbb{R}^+$.

(ii) On a :

$$\lambda_1 \int_\Omega u_1\, \overline{u_2}\, d\Omega = \int_\Omega \nabla u_1 \cdot \overline{\nabla u_2}\, d\Omega = \lambda_2 \int_\Omega u_1\, \overline{u_2}\, d\Omega.$$

Les identités (1.17) s'en déduisent aisément.
(iii) Supposons par exemple que $\Gamma_2 = \emptyset$. Le problème (1.15) s'écrit alors :

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

Posons $g = \lambda u$: la fonction propre $u \in H^1(\Omega)$ est aussi solution du problème de Laplace avec condition aux limites homogène de Dirichlet, et second membre g. D'après les résultats de régularité des problèmes elliptiques (voir [15]), comme $\partial\Omega$ est $\mathcal{C}^\infty$, on sait que :

$$g \in L^2(\Omega) \Rightarrow u \in H^2(\Omega),$$

et plus généralement :

$$g \in H^k(\Omega) \Rightarrow u \in H^{k+2}(\Omega), \quad \forall k \in \mathbb{N}.$$

Dans notre cas, nous avons donc :

$$u \in L^2(\Omega) \Rightarrow u \in H^2(\Omega) \Rightarrow u \in H^4(\Omega) \ \ ...$$

et finalement : $u \in H^k(\Omega)$, $\forall k \in \mathbb{N}$. D'après les résultats d'injection des espaces de Sobolev $H^k(\Omega)$ dans les espaces $\mathcal{C}^m(\overline{\Omega})$, (cf. [15]), il en résulte que :

$$u \in \mathcal{C}^m\left(\overline{\Omega}\right), \quad \forall m \in \mathbb{N}.$$

Le résultat se démontre identiquement dans le cas du problème de Neumann. ∎

Remarque 1.6 *Les propriétés (i) et (ii) sont tout à fait analogues aux propriétés connues pour les matrices. En effet on sait qu'une matrice hermitienne a toutes ses valeurs propres réelles et que ses vecteurs propres sont orthogonaux entre eux.*

1.2 Principaux résultats de la théorie spectrale

1.2.1 Un problème aux valeurs propres abstrait

Le problème variationnel (1.16) peut être récrit de façon abstraite sous la forme suivante, qui permet de regrouper de nombreux problèmes aux valeurs propres sous une forme unique :

$$\begin{cases} \text{trouver } \lambda \in \mathbb{C} \text{ et } u \in V \setminus \{0\} \text{ tels que} \\ a(u,v) = \lambda\ (u,v)_H \quad \forall v \in V, \end{cases} \tag{1.18}$$

où V et H désignent deux espaces de Hilbert (séparables, voir remarque 1.14) munis respectivement de produits scalaires notés $(\cdot\,,\cdot)_V$ et $(\cdot\,,\cdot)_H$ (les normes associées étant notées $\|\cdot\|_V$ et $\|\cdot\|_H$). On suppose que V s'injecte de façon continue et dense dans H, autrement dit

$$\exists C_V > 0 \text{ tel que } \forall v \in V,\ \|v\|_H \leq C_V \|v\|_V, \quad \text{et} \tag{1.19}$$

$$\forall v \in H,\ \exists (v_n)_n \subset V,\ \lim_{n\to\infty} \|v - v_n\|_H = 0. \tag{1.20}$$

Dans (1.18), $a(\cdot\,,\cdot)$ désigne une forme sesquilinéaire sur $V \times V$ que l'on suppose hermitienne, continue et coercive : il existe donc deux constantes strictement positives C_a et α telles que

$$a(v,u) = \overline{a(u,v)}, \tag{1.21}$$

$$|a(u,v)| \leq C_a\ \|u\|_V \|v\|_V, \tag{1.22}$$

$$a(u,u) \geq \alpha\ \|u\|_V^2, \tag{1.23}$$

pour tout $(u,v) \in V \times V$.
Dans le cas de l'exemple monodimensionnel (§ 1.1.1), nous avons constaté que les valeurs propres forment une suite de $\mathbb{R}^+$ qui tend vers l'infini, et que les vecteurs propres associés satisfont des propriétés d'orthogonalité et de complétude. Nous allons voir plus loin (cf. corollaire 1.15), que ceci reste valable dans le contexte général de notre problème abstrait (1.18) sous une hypothèse supplémentaire liée à la notion de *compacité*.

Rappelons qu'un sous-ensemble E d'un espace de Hilbert H est dit compact si, de toute suite de E, on peut extraire une sous-suite qui converge dans E. Si H est de dimension finie, les sous-ensembles compacts sont exactement les fermés bornés. Cette propriété caractérise en fait les espaces de dimension finie, comme l'indique la proposition suivante.

Proposition 1.7 *Soit H un espace de Hilbert et $B_H = \{v \in H;\ \|v\|_H \leq 1\}$ sa boule unité fermée. Si B_H est compacte, alors H est de dimension finie.*

Cette proposition, ainsi que la proposition 1.9, sont démontrées dans [9]. La notion d'opérateur compact permet de retrouver dans des espaces de dimension infinie des propriétés analogues au cas de la dimension finie.

Définition 1.8 *Soient H_1 et H_2 deux espaces de Hilbert. Un opérateur $T : H_1 \to H_2$ est dit compact si, de toute suite bornée de H_1, on peut extraire une sous-suite dont l'image par T converge dans H_2.*

On peut donner une autre caractérisation d'un opérateur compact en introduisant la notion de *convergence faible* : on dit qu'une suite $(u_n)_n$ d'un espace de Hilbert H tend faiblement vers u si

$$\lim_{n\to\infty} (u_n, v)_H = (u, v)_H, \quad \forall v \in H.$$

Une suite fortement convergente est évidemment faiblement convergente, mais la réciproque est fausse (sauf en dimension finie). On a la propriété fondamentale suivante.

Proposition 1.9 *Dans un espace de Hilbert H, de toute suite bornée on peut extraire une sous-suite faiblement convergente.*

Cette seconde notion nous conduit à la caractérisation suivante (très simple à vérifier) d'un opérateur compact.

Proposition 1.10 *Un opérateur $T : H_1 \to H_2$ est dit compact si*

$$u_n \to u \quad \textit{faiblement dans } H_1$$

entraine

$$Tu_n \to Tu \quad \textit{fortement dans } H_2.$$

En plus des hypothèses (1.19) à (1.23), nous supposerons donc que

$$\text{l'injection de } V \text{ dans } H \text{ est compacte.} \tag{1.24}$$

Cette hypothèse joue un rôle essentiel pour notre problème aux valeurs propres. Comme nous allons le voir, c'est elle qui assure le caractère discret de l'ensemble des valeurs propres.

Remarque 1.11 *Le problème (1.16) peut toujours se mettre sous la forme (1.18) en prenant $H = L^2(\Omega)$ et V défini par (1.13). C'est évident dans le cas où $\Gamma_1 \neq \emptyset$ car alors, d'après l'inégalité de Poincaré-Friedrichs (voir [15]), la forme sesquilinéaire*

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

est coercive sur V (elle est clairement continue et hermitienne). Dans le cas du problème de Neumann ($\Gamma_1 = \emptyset$), il suffit de poser :

$$a(u,v) = \int_\Omega (\nabla u \cdot \overline{\nabla v} + u\bar{v})\, d\Omega.$$

Si λ est solution (1.18), $(\lambda - 1)$ est solution de (1.16) et réciproquement. On a simplement effectué une translation du spectre. Notons que l'hypothèse de densité (1.20) est satisfaite puisque $\mathcal{D}(\Omega) \subset V$ est dense dans $L^2(\Omega)$. Enfin l'hypothèse de compacité (1.24) est vérifiée, puisque le théorème de Rellich [9] nous assure que lorsque Ω est borné, l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte.

1.2.2 Le théorème spectral

Le problème abstrait (1.18) peut s'interpréter comme la recherche des valeurs et vecteurs propres d'un opérateur compact T défini de H dans H. Comment définir un tel opérateur ? Soit $u \in H$ quelconque. Alors, l'application :

$$v \mapsto (u,v)_H$$

est une forme anti-linéaire continue sur V. En effet, d'après l'inégalité de Cauchy-Schwarz et (1.19),

$$|(u,v)_H| \leq \|u\|_H \, \|v\|_H \leq C_V \, \|u\|_H \, \|v\|_V.$$

Par ailleurs, d'après (1.22)–(1.23) et le théorème de Lax-Milgram (cf. [15]), il s'ensuit qu'il existe un unique élément de V, que nous noterons Tu, tel que

$$a(Tu,v) = (u,v)_H, \quad \forall v \in V. \tag{1.25}$$

Enfin, comme $V \subset H$, Tu peut être vu comme un élément de H. On a ainsi défini un opérateur continu T de H dans H.

Lemme 1.12 *Sous les hypothèses (1.19)-(1.24), l'opérateur $T : H \to H$ défini par (1.25) est compact, auto-adjoint, injectif et positif. De plus le problème (1.18) est équivalent au problème :*

$$u \in H \quad et \quad u = \lambda\, Tu. \tag{1.26}$$

Démonstration : T est clairement positif puisque d'après (1.23) et (1.25), on a

$$0 \leq \alpha \, \|Tu\|_V^2 \leq a(Tu,Tu) = (u,Tu)_H.$$

Par l'inégalité de Cauchy-Schwarz et (1.19), on en déduit de plus que

$$\alpha \|Tu\|_V^2 \leq \|u\|_H \|Tu\|_H \leq C_V \|u\|_H \|Tu\|_V,$$

et par conséquent,

$$\|Tu\|_V \leq \frac{C_V}{\alpha} \|u\|_H \quad \forall u \in H,$$

ce qui montre que T est continu de H dans V. Définir T de H dans H revient à composer par l'injection canonique de V dans H, qui est par hypothèse compacte (cf. (1.24)). La composée d'un opérateur continu par un opérateur compact est évidemment compacte. Ainsi $T : \ H \to H$ est compact.

D'autre part, comme $a(\cdot\,,\cdot)$ est hermitienne, on a

$$(u, Tv)_H = a(Tu, Tv) = \overline{a(Tv, Tu)} = \overline{(v, Tu)_H} = (Tu, v)_H,$$

ce qui montre que T est auto-adjoint.

Par ailleurs, d'après (1.25), dire que $Tu = 0$ revient à dire que $(u, v)_H = 0$ pour tout $v \in V$, donc aussi pour tout $v \in H$ d'après (1.20). Donc $u = 0$, ce qui signifie que T est injectif.

Enfin, le problème (1.18) s'écrit, d'après (1.25) :

$$\begin{cases} \text{trouver } \lambda \in \mathbb{C} \text{ et } u \in V \setminus \{0\} \text{ tels que} \\ a(u, v) = \lambda \ a(Tu, v) \quad \forall v \in V. \end{cases}$$

D'où l'identité (1.26) comme $a(\cdot\,,\cdot)$ est coercive. ∎

Notons que toute solution $u \in H$ de (1.26) est *a fortiori* dans V puisque l'image de T est contenue dans V. Nous sommes maintenant en mesure d'énoncer le résultat fondamental pour les opérateurs compacts et auto-adjoints.

Théorème 1.13 (théorème spectral) *Soit $T : \ H \to H$ un opérateur compact, auto-adjoint, injectif et positif.*
(i) Les valeurs propres de T sont toutes de multiplicité finie et constituent une suite de réels positifs qui tend vers 0 (si H est de dimension infinie).
(ii) Il existe une base hilbertienne $(u_n)_{n\geq 1}$ de H formée de fonctions propres de T.

Démonstration : Commençons par remarquer que si u et u' sont des fonctions propres associées respectivement à deux valeurs propres μ et μ' de T (forcément non nulles puisque T est injectif), alors

$$\mu \ (u, u')_H = (Tu, u')_H = (u, Tu')_H = \overline{\mu'} \ (u, u')_H,$$

puisque T est auto-adjoint. Ainsi,

$$\mu \neq \overline{\mu'} \quad \Longrightarrow \quad (u, u')_H = 0.$$

Dans le cas où $\mu = \mu'$, on en déduit que $\mu = (Tu, u)_H / \|u\|_H^2$, qui est réel et positif. Il s'ensuit d'une part que l'on peut remplacer $\overline{\mu'}$ par μ' dans l'implication précédente, et d'autre part que

$$\mu \leq \mu_* = \sup_{v \in H, \ \|v\|_H = 1} (Tv, v)_H.$$

Nous allons montrer que cette borne supérieure est atteinte, autrement dit que μ_* est valeur propre de T. Remarquons tout d'abord que μ_* coïncide ici avec

$$\|T\|_H = \sup_{u, v \in H \ \ \|u\|_H = \|v\|_H = 1} (Tu, v)_H.$$

En effet, il est clair que $\mu_* \leq \|T\|_H$. Par ailleurs, les hypothèses sur T impliquent que la forme $(u,v) \mapsto (Tu,v)_H$ est positive et hermitienne. On peut donc lui appliquer l'inégalité de Cauchy-Schwarz, qui s'écrit

$$(Tu,v)_H \leq (Tu,u)_H{}^{1/2} (Tv,v)_H{}^{1/2}.$$

En prenant le sup sur u et v dans H tels que $\|u\|_H = \|v\|_H = 1$, on en déduit que $\|T\|_H \leq \mu_*$, d'où finalement $\|T\|_H = \mu_*$ comme annoncé.

Par définition de μ_*, il existe une suite $(v_n)_n$ de H telle que $\|v_n\|_H = 1$ et $\lim_{n\to\infty}(Tv_n, v_n)_H = \mu_*$. Comme T est compact, on peut en extraire une sous-suite, encore notée $(v_n)_n$, telle que $(Tv_n)_n$ converge, soit vers $u \in H$. En remarquant que $\|Tv_n\|_H \leq \mu_*$, on a

$$\|Tv_n - \mu_*\, v_n\|_H^2 = \|Tv_n\|_H^2 - 2\mu_*\,(Tv_n, v_n)_H + \mu_*^2 \leq 2\mu_*^2 - 2\mu_*\,(Tv_n, v_n)_H,$$

où le terme de droite tend vers 0. Ainsi $Tv_n - \mu_*\, v_n \to 0$. On a donc $\mu_*\, v_n \to u$, soit $v_n \to \mu_*^{-1}\, u$. Il s'ensuit que $Tu = \mu_*\, u$ où $u \neq 0$ puisque $\|v_n\|_H = 1$ pour tout n. Ainsi $\mu_* = \|T\|_H$ est bien valeur propre de T.

Vérifions maintenant que si μ est une valeur propre de T, le sous-espace propre associé V_μ est de dimension finie. Soit $(u_n)_n$ une suite de V_μ telle que $\|u_n\|_H \leq 1$. Comme T est compact, on peut en extraire une sous-suite, encore notée $(u_n)_n$, telle que $(Tu_n)_n$ converge. Mais ici, $Tu_n = \mu\, u_n$, et comme $\mu \neq 0$, la sous-suite $(u_n)_n$ converge. Ceci signifie que dans V_μ, la boule unité fermée est compacte. D'après la proposition 1.7, V_μ est de dimension finie.

Montrons maintenant que, si H est de dimension infinie, les valeurs propres de T ne peuvent s'accumuler qu'en 0. Raisonnons par l'absurde en supposant qu'il existe une suite $(\mu_n)_n$ de valeurs propres de T, toutes distinctes, qui tend vers $\mu > 0$. On peut donc construire une suite $(u_n)_n$ telle que $\|u_n\|_H = 1$ et $Tu_n = \mu_n\, u_n$. Les μ_n étant tous distincts, on a la relation d'orthogonalité : $(u_n, u_m)_H = \delta_{nm}$. Comme $\mu \neq 0$, la suite $(v_n)_n$, avec $v_n = u_n/\mu_n$, est bornée. On peut donc en extraire une sous-suite telle que $(Tv_n)_n$ converge, ce qui est évidemment impossible puisque $Tv_n = u_n$ et que la suite $(u_n)_n$ n'est pas de Cauchy (d'après la relation d'orthogonalité $\|u_n - u_m\|_H^2 = 2$ pour $n \neq m$).

Il nous reste à montrer (ii) et le fait que les valeurs propres sont en nombre infini (on sait pour l'instant qu'il en existe au moins une, à savoir $\|T\|_H$. Nous avons vu que deux sous-espaces propres associés à deux valeurs propres distinctes sont orthogonaux. Dans chacun d'eux, on peut choisir une base de vecteurs propres. En considérant l'ensemble des valeurs propres, on construit une famille orthonormale de H. Dire que cette famille est une base hilbertienne de H revient à dire que le sous-espace G de H engendré par cette famille est H tout entier, ou encore que $G^\perp = \{0\}$. Par construction, G est stable par T (c'est-à-dire $G(T) \subset G$), donc $G^\perp$ aussi, ce qui permet de définir la restriction $\tilde{T}$ de T à $G^\perp$. Nous allons montrer que $\tilde{T} = 0$, ce qui entraine que $G^\perp = \{0\}$ puisque T étant injectif, $\tilde{T}$ l'est aussi. En effet, nous avons vu plus haut que $\|\tilde{T}\|_{G^\perp}$ est valeur propre de $\tilde{T}$. Si cette quantité était strictement positive, ce serait une valeur propre de T, ce qui est impossible par construction. Il s'ensuit que $\|\tilde{T}\|_{G^\perp} = 0$ soit $\tilde{T} = 0$. ∎

Remarque 1.14 *Les espaces de Hilbert que nous considérons ici sont supposés* séparables*, ce qui signifie qu'on peut trouver pour chacun d'eux une famille dénombrable qui constitue une base hilbertienne. C'est le cas des espaces usuels intervenant dans les formulations variationnelles des équations aux dérivées partielles* ($L^2(\Omega)$, $H^1(\Omega)$, *etc.). Le théorème ci-dessus nous montre comment construire une*

telle base à partir d'un opérateur compact auto-adjoint injectif et positif. Il s'ensuit que dans un espace non séparable, il n'existe pas de tel opérateur.

Corollaire 1.15 *Sous les hypothèses (1.19)-(1.24), on a :*
(i) Les valeurs propres λ du problème (1.18) sont toutes de multiplicité finie et constituent une suite de réels positifs $(\lambda_n)_{n\geq 1}$ qui tend vers $+\infty$.
(ii) Il existe une base hilbertienne $(u_n)_{n\geq 1}$ de H formée de fonctions propres de (1.18).

Le point (ii) signifie que pour tout élément $u \in H$, on a

$$u = \sum_{n\geq 1} (u, u_n)_H \, u_n \quad \text{avec} \quad \|u\|_H^2 = \sum_{n\geq 1} |(u, u_n)_H|^2. \tag{1.27}$$

Plus précisément, si on pose

$$\tilde{u}_N = \sum_{n=1}^{N} (u, u_n)_H \, u_n,$$

on a $\tilde{u}_N \longrightarrow u$ dans H, et $\|\tilde{u}_N\|_H^2 = \sum_{1\leq n\leq N} |(u, u_n)_H|^2 \longrightarrow \|u\|_H^2$, ce qui se vérifie facilement à partir de la relation d'orthogonalité : $(u_n, u_m)_H = \delta_{n,m}$ pour tout $n, m \in \mathbb{N}$.

Réciproquement, si $(\alpha_n)_{n\geq 1}$ est une suite de nombres complexes telle que :

$$S = \sum_{n\geq 1} |\alpha_n|^2 < +\infty,$$

alors la série

$$\sum_{n\geq 1} \alpha_n \, u_n$$

définit un élément u de H de norme $\sqrt{S}$.

Dans le cas où $u \in V$, la série (1.27) converge dans V. Pour le voir, remarquons que $a(u_n, u_m) = \lambda_n \, (u_n, u_m)_H = \lambda_n \, \delta_{n,m}$, pour tout $n, m \in \mathbb{N}$. On a alors

$$\sum_{n=1}^{N} \lambda_n |(u, u_n)_H|^2 = a\,(\tilde{u}_N, \tilde{u}_N) \leq a(u, u).$$

En effet, par orthogonalité des $(u_n)_{n\leq N}$ et des $(u_m)_{m>N}$:

$$a(u, u) - a\,(\tilde{u}_N, \tilde{u}_N) = a\,(u - \tilde{u}_N, u - \tilde{u}_N) \geq 0.$$

Donc la série $\sum_{n\geq 1} \lambda_n |(u,u_n)_H|^2$, qui est croissante, converge. Cela prouve que pour tout $\varepsilon > 0$, on a pour M et N assez grands, $N < M$:

$$a\left(\tilde{u}_M - \tilde{u}_N, \tilde{u}_M - \tilde{u}_N\right) = \sum_{n=N+1}^{M} \lambda_n |(u,u_n)_H|^2 < \varepsilon.$$

Par coercivité de $a(\cdot,\cdot)$ dans V, il en résulte que la suite $(\tilde{u}_N)$ est de Cauchy dans V, d'où par unicité de la limite (cf. proposition 1.10), $\tilde{u}_N \longrightarrow u$ dans V et

$$a(u,u) = \sum_{n\geq 1} \lambda_n |(u,u_n)_H|^2. \tag{1.28}$$

Autrement dit, la famille $(u_n/\sqrt{\lambda_n}\,)_{n\geq 1}$ constitue une base hilbertienne de V, lorsque celui-ci est muni du produit scalaire $a(\cdot,\cdot)$ (et de la norme associée $\|v\|_{V,a} = (a(v,v))^{1/2}$). Notons comme précédemment que si $(\alpha_n)_{n\geq 1}$ est une suite de nombres complexes telle que

$$S' = \sum_{n\geq 1} \lambda_n |\alpha_n|^2 < +\infty,$$

alors la série $\sum_{n\geq 1} \alpha_n\, u_n$ converge dans V vers une fonction u telle que $a(u,u) = S'$.

Remarque 1.16 *Les résultats énoncés ici généralisent ce que l'on sait sur les matrices hermitiennes pour lesquelles il existe toujours une base de vecteurs propres, orthogonaux deux à deux pour le produit scalaire usuel et pour le produit scalaire associé à la matrice.*
Tout comme les bases de fonctions harmoniques (voir exemple ci-dessous), les diverses familles de polynômes orthogonaux (polynômes de Legendre, de Laguerre, de Hermite...) peuvent être construites comme bases de fonctions propres d'opérateurs différentiels (cf. [21, 23]).

Exemple d'application

Revenons à l'exemple de la corde vibrante traité au §1.1.1. Comme c'est un problème de Dirichlet, il s'écrit directement sous la forme (1.18) avec :

$$\Omega =]0,L[, \quad H = L^2(\Omega), \quad V = H_0^1(\Omega) \ \text{ et } \ a(u,v) = \int_\Omega u'v'\,dx.$$

Le corollaire 1.15 s'applique donc.

Les valeurs propres λ_k et les fonctions propres u_k sont données par les formules (1.9). On vérifie aisément que :

$$\int_\Omega u_k(x)^2\,dx = \frac{L}{2} \quad \text{et} \quad \int_\Omega u_k'(x)^2\,dx = \frac{\lambda_k L}{2}.$$

On pose alors :

$$v_k = \sqrt{\frac{2}{L}} u_k.$$

D'après le corollaire 1.15, la famille $(v_k)_{k\geq 1}$ est une base Hilbertienne de $L^2(\Omega)$. Autrement dit, toute fonction v de $L^2(\Omega)$ s'écrit :

$$v = \sum_{k=1}^{+\infty} (v, v_k)_{L^2(\Omega)} v_k, \tag{1.29}$$

la série convergeant au sens de $L^2(\Omega)$, et l'on a :

$$\|v\|^2_{L^2(\Omega)} = \sum_{k=1}^{+\infty} |(v, v_k)_{L^2(\Omega)}|^2.$$

De plus, si $v \in H_0^1(\Omega)$, la série (1.29) converge dans $H_0^1(\Omega)$ et l'on a

$$a(v,v) = \sum_{k=1}^{+\infty} \lambda_k |(v, v_k)_{L^2(\Omega)}|^2, \text{ d'où } \|v\|^2_{H^1(\Omega)} = \sum_{k=1}^{+\infty} (1+\lambda_k) |(v, v_k)_{L^2(\Omega)}|^2.$$

1.2.3 Le principe du Min-Max

Nous allons maintenant montrer, à l'aide du corollaire 1.15, que les valeurs propres du problème (1.18) admettent une caractérisation très utile : les formules dites de "Min-Max". Nous présentons ici les formules les plus classiques mais il existe de nombreuses autres caractérisations des valeurs propres (cf. [21, 51, 5]).

Pour tout $u \in V$ non nul, nous introduisons le quotient de Rayleigh :

$$\mathcal{R}(u) = \frac{a(u,u)}{\|u\|^2_H}. \tag{1.30}$$

On désigne toujours par $(u_n)_{n\in\mathbb{N}}$ la base hilbertienne de H constituée des fonctions propres et on suppose que la suite des valeurs λ_n, telles que u_n est associée à λ_n, est ordonnée en une suite croissante.

On a alors le résultat suivant :

Lemme 1.17 *Les valeurs propres vérifient :*

$$\lambda_1 = \min_{u\in V\setminus\{0\}} \mathcal{R}(u) \tag{1.31}$$

$$\lambda_n = \min_{u\in \mathcal{O}_n\setminus\{0\}} \mathcal{R}(u) \qquad \text{si } n > 1, \tag{1.32}$$

avec $\mathcal{O}_n = \{u \in V,\ (u, u_i)_H = 0 \ \ \forall i = 1, n-1\}$.

Démonstration : D'après (1.27) et (1.28), on a

$$\mathcal{R}(u) = \frac{\displaystyle\sum_{n\geq 1} \lambda_n |(u,u_n)_H|^2}{\displaystyle\sum_{n\geq 1} |(u,u_n)_H|^2} \qquad \forall u \in V \setminus \{0\}$$

qui, joint aux identités :

$$\mathcal{R}(u_n) = \lambda_n \qquad \forall n \geq 1,$$

permet d'établir facilement les formules (1.31) et (1.32). ∎

La formule (1.31) est très pratique car elle fournit une expression de la première valeur propre, sans qu'il soit nécessaire de calculer de fonction propre explicitement. Les formules (1.32), en revanche, peuvent être améliorées. C'est l'objet du théorème suivant.

Théorème 1.18 (Min-Max) *Pour tout $m \geq 1$, on a :*

$$\lambda_m = \min_{E_m \in \mathcal{V}_m} \max_{u \in E_m \setminus \{0\}} \mathcal{R}(u) \tag{1.33}$$

où $\mathcal{V}_m$ désigne l'ensemble des sous-espaces de V de dimension m.

Démonstration : Soit $m \geq 1$. Appelons μ_m le membre de droite de l'égalité (1.33) et notons V_m l'espace engendré par $u_1, ...u_m$. On vérifie aisément que :

$$\max_{u \in V_m \setminus \{0\}} \mathcal{R}(u) = \mathcal{R}(u_m) = \lambda_m.$$

Par conséquent : $\mu_m \leq \lambda_m$.

Réciproquement, soit $E_m \in \mathcal{V}_m$ quelconque. Alors il existe un élément $v \in V$ tel que

$$\begin{cases} v \in E_m \setminus \{0\} \\ (v, u_i) = 0 \qquad i = 1, 2..., m-1 \end{cases}$$

D'après (1.32), on a alors :

$$\lambda_m \leq \mathcal{R}(v) \leq \max_{u \in E_m \setminus \{0\}} \mathcal{R}(u)$$

Ceci étant vrai pour tout $E_m \in \mathcal{V}_m$, on en déduit finalement : $\lambda_m \leq \mu_m$. ∎

Exemples d'application du principe du Min-Max

Le principe du Min-Max est très utile pour établir des résultats de comparaison.

– Supposons que l'on considère deux problèmes aux valeurs propres du type (1.18) pour deux formes bilinéaires $a(\cdot\,,\cdot)$ et $\tilde{a}(\cdot\,,\cdot)$ agissant sur le même espace de Hilbert V. On note $(\lambda_m)_{m\geq 1}$ les valeurs propres associées à $a(\cdot,\cdot)$ et $(\tilde{\lambda}_m)_{m\geq 1}$ les valeurs propres associées à $\tilde{a}(\cdot,\cdot)$, les deux suites étant ordonnées par valeurs croissantes. Alors, si les formes $a(\cdot,\cdot)$ et $\tilde{a}(\cdot,\cdot)$ sont telles que :

$$a(u,u) \leq \tilde{a}(u,u), \qquad \forall u \in V,$$

on déduit immédiatement du principe de Min-Max le résultat suivant :

$$\lambda_m \leq \tilde{\lambda}_m, \ \forall m \geq 1.$$

– Supposons cette fois que l'on considère deux problèmes du type (1.18) pour deux espaces V et $\tilde{V}$, et la même forme sesquilinéaire $a(\cdot\,,\cdot)$. On note $(\lambda_m)_{m\geq 1}$ et $(\tilde{\lambda}_m)_{m\geq 1}$ les valeurs propres obtenues dans chacun des cas, les deux suites étant également ordonnées par valeurs croissantes. Alors, si V et $\tilde{V}$ sont tels que

$$V \subset \tilde{V}$$

on déduit du principe du Min-Max l'inégalité suivante :

$$\lambda_m \geq \tilde{\lambda}_m, \qquad \forall m \geq 1.$$

Ce raisonnement nous servira dans la section suivante pour l'approximation des valeurs propres par éléments finis.

1.2.4 Alternative de Fredholm

Le problème aux valeurs propres abstrait (1.18) peut s'interpréter comme la recherche de vibrations *libres* d'un système conservatif. On s'intéresse ici au cas de vibrations *entretenues* par une excitation périodique, ce qui nous amène à considérer le problème suivant :

$$\begin{cases} \text{trouver } u \in V \text{ tel que} \\ a(u,v) - \lambda\ (u,v)_H = (f,v)_H \qquad \forall v \in V, \end{cases} \tag{1.34}$$

où λ est un complexe *donné* et $f \in H$ est également *donnée* (les notations et hypothèses étant celles du problème (1.18)).

– Si λ est négatif ou nul, ou si $\mathbf{Im}(\lambda) \neq 0$, le théorème de Lax-Milgram permet d'affirmer que le problème (1.34) est bien posé.

– En revanche, si λ est strictement positif, la forme bilinéaire $a(u,v) - \lambda(u,v)_H$ n'est généralement pas coercive. Au vu de la formulation (1.34), on parlera ici de problème de type *coercif+compact*. Cependant, le corollaire 1.15 nous permet de savoir si le problème (1.34) est ou n'est pas bien posé. En effet, on a le résultat suivant :

Théorème 1.19 (alternative de Fredholm) *De deux choses l'une :*
(i) Soit λ n'est pas une valeur propre de (1.18) et dans ce cas le problème (1.34) est bien posé.

(ii) Soit λ est une valeur propre de (1.18) de multiplicité m_λ et dans ce cas, pour que le problème (1.34) admette une solution, il faut et il suffit que la fonction f vérifie :

$$(f, v)_H = 0 \qquad \forall v \in V_\lambda \tag{1.35}$$

où V_λ est l'espace vectoriel (de dimension m_λ) engendré par les fonctions propres associées à λ. De plus, sous la condition (1.35), l'ensemble des solutions de (1.34) est un espace affine de dimension m_λ.

Remarque 1.20 *Supposons que nous ayons démontré l'unicité pour le problème (1.34) (c'est-à-dire que pour $f = 0$, la seule solution est $u = 0$), alors le point (i) de l'alternative de Fredholm nous permet d'affirmer qu'il y a existence d'une solution pour tout second membre f. La réciproque (existence pour tout $f \Rightarrow$ unicité) est également vraie d'après le point (ii). Autrement dit, l'alternative de Fredholm signifie qu'il y a équivalence entre l'injection et la surjection : la situation est donc analogue à la dimension finie.*

Démonstration du théorème 1.19 : D'après le corollaire 1.15, il existe $(u_n)_{n\geq 1}$ une base hilbertienne de H, constituée de fonctions propres de (1.18), u_n étant associée à la valeur propre λ_n. Alors, si u est solution de (1.34), on a en particulier :

$$a(u, u_n) - \lambda(u, u_n)_H = (f, u_n)_H, \quad \forall n \geq 1. \tag{1.36}$$

On sait que $a(u, u_n) = \lambda_n(u, u_n)_H$. L'identité (1.36) s'écrit donc de façon équivalente :

$$(\lambda_n - \lambda)(u, u_n)_H = (f, u_n)_H, \qquad \forall n \geq 1. \tag{1.37}$$

On en déduit que, si λ n'est pas une valeur propre, la solution du problème (1.34) si elle existe est donnée par :

$$u = \sum_{n\geq 1} \frac{(f, u_n)_H}{\lambda_n - \lambda} u_n.$$

Réciproquement, on vérifie aisément que la formule ci-dessus définit bien une fonction $u \in V$. En effet, comme $f \in H$, on a :

$$\sum_{n\geq 1} |(f, u_n)_H|^2 = \|f\|_H^2.$$

Par conséquent, on déduit de (1.28) :

$$a(u, u) = \sum_{n\geq 1} \lambda_n |(u, u_n)_H|^2 = \sum_{n\geq 1} \lambda_n \left(\frac{|(f, u_n)_H|}{\lambda_n - \lambda} \right)^2 \leq K(\lambda) \|f\|_H^2$$

avec $K(\lambda) = \sup_{n\geq 1} \dfrac{\lambda_n}{(\lambda_n - \lambda)^2}$.

En revanche, si λ est une valeur propre, alors on déduit de (1.37) la relation de compatibilité suivante :

$$(f, u_i)_H = 0 \ \text{ pour tout } i \in \mathbb{N} \text{ tel que } \lambda_i = \lambda.$$

Si f satisfait à ces conditions, une solution de (1.34) est :

$$u = \sum_{n\geq 1, \lambda_n \neq \lambda} \frac{(f, u_n)_H}{\lambda_n - \lambda} u_n.$$

De plus, si v est une fonction propre associée à λ, $(u + v)$ est aussi solution de (1.34). ∎

On voit que le module de continuité $K(\lambda)$ peut être très grand, dès lors que λ est proche d'une valeur propre. Ce comportement aura également des conséquences numériques, voir §1.4.2.

Remarque 1.21 *On peut facilement généraliser l'alternative de Fredholm au cas d'un second membre exprimé sous la forme $f(v)$, avec $f \in V'$. Dans ce cas, la même démonstration s'applique, modulo l'écriture de la norme $\|\cdot\|_{V'}$ par rapport aux valeurs propres $(\lambda_n)_{n\geq 1}$: $\|f\|^2_{V'} = \sum_{n\geq 1}(\lambda_n)^{-1} |\langle f, u_n\rangle_{V',V}|^2$. Par ailleurs, on trouvera dans [9] un énoncé plus général de l'alternative de Fredholm, valable pour les problèmes non hermitiens.*

Remarque 1.22 *Nous avons vu à la remarque 1.11 que la formulation variationnelle (1.16) entre dans le cadre de notre problème abstrait (1.18). Il s'ensuit que le problème de Helmholtz (1.14) relève de l'alternative de Fredholm.*

1.3 Approximation des problèmes spectraux

On s'intéresse maintenant à l'approximation numérique des valeurs propres et des fonctions propres du problème (1.18).

1.3.1 Discrétisation

Pour réaliser la discrétisation, on introduit un sous-espace V_h de V de dimension finie :

$$\dim V_h = N.$$

Comme $V_h \subset V$, il s'agit d'une méthode d'approximation interne.
On considère le problème discret suivant :

$$\begin{cases} \text{trouver } \lambda_h \in \mathbb{C} \text{ et } u_h \in V_h \setminus \{0\} \text{ tels que} \\ a(u_h, v_h) = \lambda_h (u_h, v_h)_H, \qquad \forall v_h \in V_h. \end{cases} \tag{1.38}$$

On cherche donc les valeurs λ_h telles que le problème (1.38) admette des solutions non nulles.

Remarquons tout d'abord qu'il s'agit d'un problème aux valeurs propres matriciel. Pour cela, nous considérons une base $(w_i)_{i=1,N}$ de l'espace V_h. On a alors le lemme :

Lemme 1.23 *Le problème (1.38) est équivalent au problème aux valeurs propres matriciel suivant :*

$$\begin{cases} \text{trouver } \lambda_h \in \mathbb{C} \text{ et } \overrightarrow{U} \neq 0 \text{ tels que} \\ \mathbb{A}\overrightarrow{U} = \lambda_h \mathbb{M}\overrightarrow{U} \end{cases} \tag{1.39}$$

où $\overrightarrow{U} \in \mathbb{C}^N$ *et où les matrices hermitiennes* $\mathbb{A}$ *et* $\mathbb{M}$ *sont respectivement définies par*

$$\mathbb{A}_{IJ} = a(w_J, w_I) \text{ et } \mathbb{M}_{IJ} = (w_J, w_I)_H, 1 \leq I, J \leq N.$$

Démonstration : On cherche u_h sous la forme :

$$u_h = \sum_{J=1}^{N} U^J w_J.$$

Le problème (1.38) s'écrit alors :

$$\sum_{J=1}^{N} a(w_J, w_I) U^J = \lambda_h \sum_{J=1}^{N} (w_J, w_I)_H U^J, \qquad I = 1, N,$$

qui conduit à (1.39), avec $\overrightarrow{U}$ le vecteur de composantes $(U^J)_{J=1,N}$. ∎

Remarque 1.24 *A priori, on doit se placer dans* $\mathbb{C}$ *pour résoudre les problèmes aux valeurs propres exacts et discrets. Néanmoins, lorsque la forme* $a(\cdot,\cdot)$ *est à valeurs réelles pour des fonctions-tests à valeurs réelles, on peut se placer dans* $\mathbb{R}$*, ce qui permet de réduire significativement le coût du calcul des valeurs propres discrètes.*

A l'aide du lemme 1.23, on démontre le théorème suivant :

Théorème 1.25 *Il existe une base* $(u_{m,h})_{m=1,N}$ *de* V_h*, orthonormale dans* H*, telle que les vecteurs* $(\overrightarrow{V}_m)_{m=1,N}$*, définis par :*

$$u_{m,h} = \sum_{I=1}^{N} V_m^I w_I,$$

sont des vecteurs propres du problème (1.38).

Notons que la démarche du §1.2.2 reste bien sûr valable en dimension finie, mais on propose ici une autre démonstration qui repose sur le fait qu'une matrice hermitienne est diagonalisable.

Démonstration : La matrice $\mathbb{M}$ étant hermitienne définie-positive, elle admet une décomposition de Cholesky :

$$\mathbb{M} = \mathbb{L}\mathbb{L}^*$$

où $\mathbb{L}$ est une matrice triangulaire inversible et $\mathbb{L}^* = \overline{\mathbb{L}}^t$. Le problème (1.39) équivaut donc au problème :

$$\mathbb{B}\overrightarrow{X} = \lambda_h \overrightarrow{X} \tag{1.40}$$

où l'on a posé :

$$\begin{cases} \mathbb{B} = \mathbb{L}^{-1}\mathbb{A}\,(\mathbb{L}^*)^{-1} \\ \overrightarrow{X} = \mathbb{L}^*\overrightarrow{U} \end{cases}.$$

La matrice $\mathbb{B}$ étant hermitienne, elle est diagonalisable, et l'on peut de plus choisir les vecteurs propres de sorte qu'ils forment une base orthonormale de $\mathbb{C}^N$: il existe donc $(\overrightarrow{X}_m)_{m=1,N}$ une base orthonormale de $\mathbb{C}^N$ et une suite de réels $(\lambda_{m,h})_{m=1,N}$, tels que $\mathbb{B}\overrightarrow{X}_m = \lambda_{m,h}\overrightarrow{X}_m$, pour $m = 1, N$. Si on revient au problème de départ, les vecteurs $(\overrightarrow{V}_m)_{m=1,N}$, définis par $\overrightarrow{V}_m = (\mathbb{L}^*)^{-1}\overrightarrow{X}_m$, vérifient les relations :

$$\begin{aligned} &\mathbb{A}\overrightarrow{V}_m = \mathbb{L}\mathbb{B}\mathbb{L}^*(\mathbb{L}^*)^{-1}\overrightarrow{X}_m = \mathbb{L}(\mathbb{B}\overrightarrow{X}_m) = \mathbb{L}(\lambda_{m,h}\overrightarrow{X}_m) = \lambda_{m,h}\mathbb{L}\mathbb{L}^*\overrightarrow{V}_m = \lambda_{m,h}\mathbb{M}\overrightarrow{V}_m\,; \\ &(\mathbb{M}\overrightarrow{V}_m|\overrightarrow{V}_p) = (\mathbb{L}\mathbb{L}^*\overrightarrow{V}_m|\overrightarrow{V}_p) = (\mathbb{L}^*\overrightarrow{V}_m|\mathbb{L}^*\overrightarrow{V}_p) = (\overrightarrow{X}_m|\overrightarrow{X}_p) = \delta_{m,p}. \end{aligned}$$

Ci-dessus, on a noté $(\cdot|\cdot)$ le produit scalaire hermitien de $\mathbb{C}^N$.
Le fait que les $(u_{m,h})_{m=1,N}$, avec

$$u_{m,h} = \sum_{J=1}^{N} V_m^J w_J, \quad m = 1, N$$

soient des fonctions propres de (1.38), s'en déduit par retour aux éléments de V_h, comme dans la démonstration du lemme précédent. Enfin, la famille d'éléments $(u_{m,h})_{m=1,N}$ forme une base orthonormale de H. En effet

$$(u_{m,h}, u_{p,h})_H = \sum_{J=1}^{N}\sum_{I=1}^{N}(w_J, w_I)_H V_m^J V_p^I = \sum_{I=1}^{N}\sum_{J=1}^{N}\mathbb{M}_{IJ}V_m^J V_p^I = (\mathbb{M}\overrightarrow{V}_m|\overrightarrow{V}_p) = \delta_{m,p},$$

ce qui permet de conclure. ∎

L'analogue en dimension finie du théorème 1.18 fournit la caractérisation suivante des valeurs propres discrètes $(\lambda_{m,h})_{m=1,N}$, que l'on supposera dorénavant ordonnées par valeurs croissantes :

$$\lambda_{m,h} = \min_{E_m \subset \mathcal{V}_m^h} \max_{u \in E_m \setminus \{0\}} \mathcal{R}(u). \tag{1.41}$$

où $\mathcal{V}_m^h$ est l'ensemble des sous-espaces de V_h de dimension m.

1.3.2 Analyse d'erreur des éléments propres

Supposons un instant que l'on ait construit l'espace V_h par une méthode d'éléments finis (le paramètre h mesure alors la finesse de la discrétisation). On dispose donc d'une famille de sous-espaces $(V_h)_h$ de V, et d'un ensemble de problèmes discrets (1.38). Lorsque h tend vers 0, l'espace V_h tend (en un certain sens) vers l'espace V. On espère donc que pour h assez petit, les valeurs propres et vecteurs propres du problème discret seront peu différents de ceux du problème continu. Un premier résultat élémentaire est donné par le lemme suivant :

Lemme 1.26 *Pour $m = 1, N$, on a :*

$$\lambda_{m,h} \geq \lambda_m.$$

Démonstration : Ceci résulte directement des formules (1.33) et (1.41). En effet, comme $V_h \subset V$, on a : $\mathcal{V}_m^h \subset \mathcal{V}_m$, pour tout $m \in \mathbb{N}$. ∎

On veut de plus disposer d'une estimation a priori de l'écart $(\lambda_{m,h} - \lambda_m)$, dont on sait maintenant qu'il est positif, et de l'erreur $\|u_{m,h} - u_m\|_V$. *A titre illustratif*, nous commencerons par l'approximation du premier couple d'éléments propres (λ_1, u_1) pour le Laplacien (1.15), avant de passer au cas général.

Pour $f \in H$, notons z l'unique élément de V tel que

$$a(z, v) = (f, v)_H \quad \forall v \in V. \tag{1.42}$$

(L'existence et l'unicité de z découlent de l'application du théorème de Lax-Milgram).
Définissons également z_h l'unique élément de V_h tel que

$$a(z_h, v_h) = (f, v_h)_H \quad \forall v_h \in V_h. \tag{1.43}$$

Rappelons le lemme de Céa (cf. [15]) : puisque $a(\cdot, \cdot)$ est continue et coercive, il existe une constante C'_a, indépendante de h et de f, telle que

$$\|z - z_h\|_V \leq C'_a \, \varepsilon_f(h), \text{ avec } \varepsilon_f(h) = \inf_{v_h \in V_h} \|z - v_h\|_V.$$

Dans la suite, on fait une hypothèse d'approximabilité "uniforme", c'est-à-dire

$$\lim_{h \to 0} \left(\sup_{f \in H} \frac{\varepsilon_f(h)}{\|f\|_H} \right) = 0. \tag{1.44}$$

Remarque 1.27 *Prenons l'exemple de l'équation de Laplace dans un ouvert borné Ω de $\mathbb{R}^n$ de frontière "suffisamment régulière" et polyédrique, avec condition aux limites homogène de Dirichlet ($V = H_0^1(\Omega)$), et second membre f dans $H = L^2(\Omega)$. On sait (cf. [15]) que toute solution appartient en fait à $H^{1+s}(\Omega)$, pour $s \geq 1/2$ dépendant de la géométrie, avec stabilité de la solution en norme $\|\cdot\|_{H^{1+s}(\Omega)}$. Pour ce qui est de l'approximation par éléments finis de Lagrange P^1 de ce problème, on sait que $\varepsilon_f(h) \leq C_D h^s \|f\|_{L^2(\Omega)}$, avec $C_D > 0$ indépendante de f et h, ce qui garantit (1.44) dans ce cas.*

Le cas du premier couple d'éléments propres du Laplacien

Proposition 1.28 *Supposons que la première valeur propre λ_1 du problème (1.15) soit simple, et que le problème (1.42) associé soit régulier, à savoir que pour tout $f \in L^2(\Omega)$, on ait $z \in H^2(\Omega)$. Alors :*

$$\forall h \in]0, h_0[, \; 0 \leq \lambda_{1,h} - \lambda_1 \leq C \, h^2, \tag{1.45}$$

avec h_0 et $\underline{C}$ des constantes indépendantes de h.
Pour un choix judicieux de $u_{1,h}$ ($(u_1, u_{1,h})_{L^2(\Omega)} > 0$), on a de plus :

$$\forall h \in]0, h_0[,\ \|u_1 - u_{1,h}\|_{H^1(\Omega)} \leq \underline{C}\, h, \tag{1.46}$$

avec h_0 et $\underline{C}$ des constantes indépendantes de h.

Démonstration : D'après le lemme 1.26, on sait déja que $0 \leq \lambda_{1,h} - \lambda_1$. Pour majorer l'écart $\lambda_{1,h} - \lambda_1$, introduisons z_h la solution du problème discret (1.43) correspondant à $f = \lambda_1 u_1$. Comme le problème (1.42) associé est régulier, on sait d'après [15] que :

$$\|u_1 - z_h\|_{L^2(\Omega)} \leq C_1\, h^2,\ C_1 \text{ dépendant seulement de } \lambda_1.$$

Comme par ailleurs $\|u_1\|_{L^2(\Omega)} = 1$, on en déduit au passage que $\lim_{h\to 0} \|z_h\|_{L^2(\Omega)} = 1$. Soit donc $h_0 > 0$ tel que $\|z_h\|_{L^2(\Omega)} \geq 1/2$, pour tout $h < h_0$. D'après le principe du Min-Max discret (1.41), on a

$$\lambda_{1,h} - \lambda_1 \leq \frac{a(z_h, z_h)}{\|z_h\|^2_{L^2(\Omega)}} - \lambda_1.$$

Or, par définition de z_h, on sait que $a(z_h, z_h) = \lambda_1 (u_1, z_h)_{L^2(\Omega)}$. En appliquant Cauchy-Schwarz, on trouve finalement, pour $h < h_0$,

$$\begin{aligned}\lambda_{1,h} - \lambda_1 &\leq \frac{\lambda_1}{\|z_h\|^2_{L^2(\Omega)}} (u_1 - z_h, z_h)_{L^2(\Omega)} \leq \frac{\lambda_1}{\|z_h\|_{L^2(\Omega)}} \|u_1 - z_h\|_{L^2(\Omega)} \\ &\leq C\, h^2, \text{ avec } C = 2\lambda_1\, C_1.\end{aligned}$$

Ceci prouve le résultat (1.45). Dans la suite, on suppose que $h < h_0$.

Pour l'estimation d'erreur (1.46), soit maintenant $u_{1,h}$, une fonction propre discrète associée à $\lambda_{1,h}$, de norme $\|u_{1,h}\|_{L^2(\Omega)} = 1$. D'après le principe du Min-Max discret (1.41), on a $\lambda_{1,h} = a(u_{1,h}, u_{1,h})$. L'idée est d'estimer $a(u_1 - u_{1,h}, u_1 - u_{1,h})$, puis de conclure par coercivité.
On écrit la décomposition de $u_{1,h}$ sur la base hilbertienne $(u_m)_m$ de $L^2(\Omega)$:

$$u_{1,h} = \gamma u_1 + w, \text{ avec } \gamma = (u_{1,h}, u_1)_{L^2(\Omega)} \text{ et } w = \sum_{m\geq 2} (u_{1,h}, u_m)_{L^2(\Omega)} u_m.$$

Par orthogonalité, on en déduit les relations

$$a(u_1 - u_{1,h}, u_1 - u_{1,h}) = (1-\gamma)^2 \lambda_1 + a(w, w), \quad 1 = \gamma^2 + \|w\|^2_{L^2(\Omega)}, \quad \lambda_{1,h} = \lambda_1\, \gamma^2 + a(w, w).$$

On va utiliser ensuite la majoration sur l'écart $\lambda_{1,h} - \lambda_1$. On trouve, à l'aide du principe du Min-Max pour λ_2 :

$$\lambda_{1,h} - \lambda_1 = -(1-\gamma^2)\lambda_1 + a(w, w) \geq -(1-\gamma^2)\lambda_1 + \lambda_2 \|w\|^2_{L^2(\Omega)} = (1-\gamma^2)(\lambda_2 - \lambda_1).$$

Le fait que λ_1 soit simple nous fournit ici une majoration ! Dans le cas contraire ($\lambda_1 = \lambda_2$), on a simplement retrouvé le résultat $0 \leq \lambda_{1,h} - \lambda_1$. On écrit donc :

$$0 \leq (1-\gamma^2) \leq \frac{1}{\lambda_2 - \lambda_1} (\lambda_{1,h} - \lambda_1) \leq \frac{C}{\lambda_2 - \lambda_1} h^2.$$

En particulier, lorsque h est petit, γ^2 est proche de 1, c'est-à-dire que γ est proche de ± 1. Dans la suite, si γ est proche de -1, on choisit de remplacer $u_{1,h}$ par $-u_{1,h}$.

On peut maintenant estimer $a(w, w)$:

$$a(w,w) = \lambda_{1,h} - \lambda_1\,\gamma^2 \le \lambda_{1,h} - \lambda_1 + \frac{C\,\lambda_1}{\lambda_2-\lambda_1}h^2 = \frac{C\,\lambda_2}{\lambda_2-\lambda_1}h^2,$$

et en conclure :

$$\begin{aligned} a(u_1-u_{1,h},u_1-u_{1,h}) &= (1-\gamma)^2\lambda_1 + a(w,w) \le (1-\gamma^2)^2\lambda_1 + a(w,w) \\ &\le \frac{C'}{(\lambda_2-\lambda_1)^2}h^2, \text{ avec } C' = C\lambda_2(\lambda_2-\lambda_1) + C^2\,\lambda_1\,h_0^2. \end{aligned}$$

D'après la propriété (1.23) de coercivité, il existe $\alpha > 0$ tel que, pour tout $u \in H^1(\Omega)$, on ait $\alpha\|u\|^2_{H^1(\Omega)} \le a(u,u)$: ainsi, pour $h < h_0$,

$$\|u_1 - u_{1,h}\|_{H^1(\Omega)} \le \left(\frac{C'}{\alpha(\lambda_2-\lambda_1)^2}\right)^{1/2} h.$$

Ceci prouve le second résultat. ∎

Remarque 1.29 *Les constantes C et $\underline{C}$ qui apparaissent dans les estimations (1.45)-(1.46) sont indépendantes de h. Par contre, $\underline{C}$ croît comme $(\lambda_2-\lambda_1)^{-1}$: lorsque ces deux valeurs propres sont très proches, cette constante devient très grande.*

Le cas général

Si λ_m est simple, notons $\varepsilon_m(h) = \inf_{v_h\in V_h}\|u_m - v_h\|_V$ pour une fonction propre u_m de norme $\|u_m\|_V = 1$. Si au contraire λ_m est de multiplicité $m_\lambda > 1$ avec $\lambda_m = \cdots = \lambda_{m+m_\lambda-1}$, soit V_λ l'espace vectoriel engendré par les fonctions propres $u_m,\cdots,u_{m+m_\lambda-1}$, et notons $\varepsilon_m(h) = \sup_{u_\lambda\in V_\lambda,\ \|u_\lambda\|_V=1}\inf_{v_h\in V_h}\|u_\lambda - v_h\|_V$. Le résultat suivant est tiré de [5], et sa démonstration fait appel à la théorie de la résolvante (calcul fonctionnel holomorphe).

Théorème 1.30 *On suppose que l'hypothèse (1.44) est satisfaite. Soit $m \ge 1$. Si λ_m est une valeur propre simple, alors pour un choix judicieux de $u_{m,h}$ (précisément, $(u_m,u_{m,h})_V > 0$), on a (h_0 et $\underline{C}$ cstes indépendantes de h) :*

$$\forall h \in]0,h_0[,\ \|u_m - u_{m,h}\|_V \le \underline{C}\,\varepsilon_m(h). \tag{1.47}$$

Si $\lambda = \lambda_m$ est valeur propre de multiplicité $m_\lambda > 1$ ($\lambda_m = \cdots = \lambda_{m+m_\lambda-1}$), on obtient le résultat ci-dessous. Soit V_λ l'espace vectoriel engendré par les fonctions propres $u_m,\cdots,u_{m+m_\lambda-1}$, et soit $V_{\lambda,h}$ l'espace vectoriel engendré par les fonctions propres discrètes $u_{m,h},\cdots,u_{m+m_\lambda-1,h}$. Alors on a ($h_0$ et $\underline{C}$ cstes indépendantes de h) :

$$\begin{gathered} \forall h \in]0,h_0[,\ \forall v_\lambda \in V_\lambda,\ \|v_\lambda\|_V = 1,\ \exists v_{\lambda,h} \in V_{\lambda,h}, \|v_{\lambda,h}\|_V = 1, \\ \|v_\lambda - v_{\lambda,h}\|_V \le \underline{C}\,\varepsilon_m(h). \end{gathered} \tag{1.48}$$

Remarque 1.31 *La constante $\underline{C}$ qui apparaît dans les estimations (1.47)-(1.48) est indépendante de h. Par contre, elle dépend de m. En particulier, $\underline{C}$ croît comme $\max_{\lambda' \ v.p.,\ \lambda' \neq \lambda_m}(\mid \lambda' - \lambda_m \mid^{-1})$. Lorsqu'une valeur propre est très proche de λ_m, cette constante devient très grande.*

Corollaire 1.32 *On suppose que l'hypothèse (1.44) est satisfaite. Soit $m \geq 1$, on a :*

$$\forall h \in]0, h_0[,\ 0 \leq \lambda_{m,h} - \lambda_m \leq \underline{C}'\ (\varepsilon_m(h))^2\,, \tag{1.49}$$

avec h_0 et $\underline{C}'$ des constantes indépendantes de h.

Démonstration : On omet les indices $_m$. Soient donc λ la valeur propre et u une fonction propre associée. Soit λ_h la valeur propre discrète correspondante, et soit u_h la fonction propre discrète telle que (1.47) ou (1.48) soit vérifié. D'après le lemme 1.26, on a le premier résultat $0 \leq \lambda_h - \lambda$. D'après (1.18) avec $v = u_h$ et (1.38) avec $v_h = u_h$, on sait que

$$a(u, u_h) = \lambda(u, u_h)_H \text{ et } a(u_h, u_h) = \lambda_h(u_h, u_h)_H.$$

Par différence, on trouve

$$a(u - u_h, u_h) = \lambda(u, u_h)_H - \lambda_h(u_h, u_h)_H = (\lambda - \lambda_h)(u_h, u_h)_H + \lambda(u - u_h, u_h)_H,$$

ce qui s'écrit encore

$$(\lambda_h - \lambda)\|u_h\|_H^2 = \lambda(u - u_h, u_h)_H - a(u - u_h, u_h).$$

Or, en reprenant (1.18) avec $v = u - u_h$, on note que $a(u, u - u_h) - \lambda(u, u - u_h)_H = 0$. En prenant le conjugué ($\lambda \in \mathbb{R}$) et en se souvenant que $a(\cdot,\cdot)$ est hermitienne, on en déduit que $a(u - u_h, u) - \lambda(u - u_h, u)_H = 0$. Si on ajoute cette contribution, on aboutit à

$$(\lambda_h - \lambda)\|u_h\|_H^2 = -\lambda\|u - u_h\|_H^2 + a(u - u_h, u - u_h) \leq a(u - u_h, u - u_h) \leq C_a\,\|u - u_h\|_V^2.$$

(Ci-dessus, C_a est le module de continuité de $a(\cdot,\cdot)$, cf. (1.22)).
Comme $(u_h)_h$ converge vers u, il existe $h_0 > 0$ tel que, pour tout $h < h_0$, $\|u_h\|_H \geq \frac{1}{2}\|u\|_H = \frac{1}{2}$. Ainsi, pour tout $h < h_0$, on en conclut que

$$0 \leq \lambda_h - \lambda \leq 4C_a\underline{C}^2\ (\varepsilon_m(h))^2\,. \qquad \blacksquare$$

Remarque 1.33 *Les estimations (1.47)-(1.49) dépendent de façon cruciale de l'approximabilité des fonctions propres u_m. Par exemple, choisissons une approximation par éléments finis de Lagrange d'ordre k pour calculer les modes propres dans un domaine Ω polyédrique. Supposons que les fonctions propres u_m associées à la valeur propre λ_m soient telles que $u_m \in H^{p_m+1}(\Omega)$, avec $p_m > 0$, alors $\varepsilon_m(h) = h^{\min(p_m,k)}$. Si on approche simultanément plusieurs valeurs et fonctions propres ($m = 1, M$), on peut donc obtenir des ordres de convergence différents, si les exposants $(p_m)_{m=1,M}$ sont différents (voir §1.4.1 pour des illustrations numériques).*

1.3.3 La méthode de la puissance inverse

Pour conclure ce rapide tour d'horizon sur l'approximation des problèmes aux valeurs propres, nous allons présenter un algorithme élémentaire qui permet de déterminer une valeur propre du problème matriciel (1.39) posé dans $\mathbb{C}^N$, lorsque $\mathbb{A}$ est hermitienne et inversible et $\mathbb{M}$ est hermitienne définie-positive. Ce cadre correspond précisément à celui issu de l'approximation des problèmes aux valeurs propres précédents.
La matrice $\mathbb{M}$ étant hermitienne définie-positive, elle admet une décomposition de Cholesky :

$$\mathbb{M} = \mathbb{L}\mathbb{L}^*$$

où $\mathbb{L}$ est une matrice triangulaire inférieure inversible. Le problème (1.39) équivaut donc au problème :

$$\mathbb{B}\overrightarrow{X} = \lambda\overrightarrow{X} \tag{1.50}$$

où l'on a posé $\mathbb{B} = \mathbb{L}^{-1}\mathbb{A}(\mathbb{L}^*)^{-1}$ et $\overrightarrow{X} = \mathbb{L}^*\overrightarrow{U}$. La matrice $\mathbb{B}$ étant hermitienne, elle est diagonalisable. Résoudre le problème matriciel (1.39) revient donc à calculer les valeurs propres d'une matrice $\mathbb{B}$ hermitienne définie-positive. Notons qu'il s'agit d'un problème non linéaire puisqu'il équivaut au calcul des racines du polynôme de degré N :

$$P(x) = \det(\mathbb{B} - x\mathbb{I})$$

où $\mathbb{I}$ désigne la matrice identité.

En pratique, il n'est généralement pas souhaitable de calculer explicitement la matrice $\mathbb{B}$. En effet, si le problème a été discrétisé à l'aide d'une méthode d'éléments finis, les matrices $\mathbb{A}$ et $\mathbb{M}$ sont creuses, alors que la matrice $\mathbb{B}$ est pleine. C'est pourquoi nous présentons ci-dessous un algorithme qui ne nécessite pas le calcul explicite de $\mathbb{B}$.
– Désignons par

$$0 < \mu_1 \leq \mu_2 \cdot \cdot \leq \mu_N$$

les valeurs propres du problème (1.50) et $(\overrightarrow{X}_m)_{m=1,N}$ une base orthogonale de vecteurs propres associés.

– Pour un vecteur $\overrightarrow{X} \in \mathbb{C}^N$, on note $\mathrm{Max}(\overrightarrow{X})$ la valeur de la coordonnée de $\overrightarrow{X}$ qui est la plus grande en module. Si plusieurs coordonnées de $\overrightarrow{X}$ ont la même valeur absolue, on ne retiendra que la valeur de la première d'entre elles. Par exemple, si $\overrightarrow{X} = (0; -1; .5; 1)$, $\mathrm{Max}(\overrightarrow{X}) = -1$.

– On suppose que la base de vecteurs propres est choisie de telle sorte que :

$$\mathrm{Max}(\overrightarrow{X}_m) = 1, \quad m = 1, N.$$

La méthode de la puissance inverse permet, pour une valeur ν donnée, de calculer une approximation aussi bonne que l'on veut de la valeur propre μ_m la plus proche de ν (s'il y en a une, et une seule) et du vecteur propre associé. En voici l'algorithme :

Algorithme de la puissance inverse

– *Initialisation*
On se donne $\nu \geq 0$ (qui n'est pas une valeur propre) et un vecteur initial $\overrightarrow{Y}^{(0)}$.

– *Boucle itérative : pour* $k = 0, 1, \cdots$

(i) On calcule $\overrightarrow{Z}^{(k)}$ solution du système linéaire suivant :

$$(\mathbb{A} - \nu\mathbb{M})\overrightarrow{Z}^{(k)} = \mathbb{M}\overrightarrow{Y}^{(k)} \tag{1.51}$$

(ii) On pose

$$\overrightarrow{Y}^{(k+1)} = \frac{1}{\mathrm{Max}\left(\overrightarrow{Z}^{(k)}\right)}\overrightarrow{Z}^{(k)}$$

Le paramètre ν est appelé le "shift" de la méthode.

Pour démontrer la convergence de la méthode, nous supposerons qu'il existe un indice ℓ tel que

$$\mid \mu_\ell - \nu \mid < \mid \mu_m - \nu \mid \qquad \forall m,\ m \neq \ell.$$

On a alors le résultat de convergence :

Proposition 1.34 *Si* $\left(\mathbb{M}\overrightarrow{Y}^{(0)}, \overrightarrow{X}_\ell\right) \neq 0$ *alors la suite* $(\overrightarrow{Y}^{(k)})_k$ *converge vers* $\overrightarrow{X}_\ell$ *lorsque* k *tend vers* $+\infty$ *et la suite des scalaires* $(\mathrm{Max}(\overrightarrow{Z}^{(k)}))_k$ *tend vers* $\dfrac{1}{\mu_\ell - \nu}$. *De plus, pour ces deux suites, l'erreur tend vers 0 comme la suite géométrique de raison :*

$$r = \frac{\mid \mu_\ell - \nu \mid}{\min\limits_{m \neq \ell} \mid \mu_m - \nu \mid}.$$

Démonstration : Le vecteur $\overrightarrow{Y}^{(0)}$ se décompose sur la base des $(\overrightarrow{X}_m)_m$ comme suit :

$$\overrightarrow{Y}^{(0)} = \sum_{m=1}^{N} \alpha_m \overrightarrow{X}_m$$

et par hypothèse, α_ℓ est non nul. On vérifie par ailleurs aisément que

$$\overrightarrow{Y}^{(k)} = \frac{\left((\mathbb{A} - \nu\mathbb{M})^{-1}\mathbb{M}\right)^{-k}\overrightarrow{Y}^{(0)}}{\mathrm{Max}\left\{\left((\mathbb{A} - \nu\mathbb{M})^{-1}\mathbb{M}\right)^{-k}\overrightarrow{Y}^{(0)}\right\}}.$$

Or on a

$$\left((\mathbb{A}-\nu\mathbb{M})^{-1}\mathbb{M}\right)^{-k}\overrightarrow{Y}^{(0)}=\sum_{m=1}^{N}\frac{\alpha_m}{(\mu_m-\nu)^k}\overrightarrow{X}_m,$$

et par conséquent :

$$\overrightarrow{Y}^{(k)}=\frac{\alpha_\ell\overrightarrow{X}_\ell+\sum_{m\neq\ell}\left(\frac{\mu_\ell-\nu}{\mu_m-\nu}\right)^k\alpha_m\overrightarrow{X}_m}{\mathrm{Max}\left(\alpha_\ell\overrightarrow{X}_\ell+\sum_{m\neq\ell}\left(\frac{\mu_\ell-\nu}{\mu_m-\nu}\right)^k\alpha_m\overrightarrow{X}_m\right)}.$$

On en déduit

$$\overrightarrow{Y}^{(k)}=\overrightarrow{X}_\ell+O\left(r^k\right)$$

et

$$\begin{aligned}\mathrm{Max}\left(\overrightarrow{Z}^{(k)}\right)&=\mathrm{Max}\left((\mathbb{A}-\nu\mathbb{M})^{-1}\mathbb{M}\overrightarrow{Y}^{(k)}\right)\\&=\frac{1}{\mu_\ell-\nu}+O\left(r^k\right).\end{aligned}$$

■

Remarque 1.35 *La convergence est d'autant plus rapide que la raison r est petite, et donc que ν est proche de μ_ℓ. Cependant, dans ce cas, la matrice $(\mathbb{A}-\nu\mathbb{M})$ est "presque" singulière et le système (1.51) est a priori très mal conditionné. Heureusement, on peut montrer que l'erreur qui en résulte n'affecte pas l'efficacité de l'algorithme (cf. [42]).*

Remarque 1.36 *Si μ_ℓ n'est pas une valeur propre simple, la convergence vers un vecteur propre n'est plus assurée. Par contre, la convergence vers la quantité $\frac{1}{\mu_\ell-\nu}$ reste garantie, sous réserve que l'on ne soit pas dans la situation : $\exists\ell'$ tel que*

$$\begin{cases}\mid\mu_\ell-\nu\mid=\mid\mu_{\ell'}-\nu\mid<\mid\mu_m-\nu\mid & \forall m,\ m\notin\{\ell,\ell'\}\\ \mu_\ell\neq\mu_{\ell'}.\end{cases}$$

Il existe de nombreuses méthodes plus sophistiquées permettant de calculer *toutes* les valeurs propres d'une matrice (méthode QR) ou *plusieurs* d'entre elles (méthode des sous-espaces et méthode de Lanczos) ; voir pour cela [11, 42].

1.3.4 Approximation des problèmes coercif+compact

Considérons maintenant l'approximation numérique des problèmes de type coercif+compact (1.34), avec pour données $f\in H$ et $\lambda>0$, λ n'appartenant pas à l'ensemble des valeurs propres du problème (1.18). D'après l'alternative de Fredholm (théorème 1.19), il existe une solution et une seule u à ce problème. Comme pour les problèmes aux valeurs propres, on introduit $(V_h)_h$ une famille de sous-espaces de dimension finie N de V (méthode d'approximation interne). Comme précédemment, h est destiné à tendre vers 0, N dépend de h, et $(w_i)_{i=1,N}$ est une base de V_h. Pour h donné, il s'agit de résoudre le problème discret :

$$\begin{cases} \text{trouver } u_h \in V_h \text{ tel que} \\ a(u_h, v_h) - \lambda\ (u_h, v_h)_H = (f, v_h)_H \qquad \forall v_h \in V_h. \end{cases} \tag{1.52}$$

Sous forme matricielle, ce problème s'écrit de façon équivalente

$$\begin{cases} \text{trouver } \overrightarrow{U} \in \mathbb{C}^N \text{ tel que} \\ (\mathbb{A} - \lambda \mathbb{M})\, \overrightarrow{U} = \overrightarrow{F}, \end{cases} \tag{1.53}$$

où $\mathbb{A}$ et $\mathbb{M}$ sont définies comme dans le lemme 1.23, et $\overrightarrow{F} \in \mathbb{C}^N$ est défini par $F^I = (f, w_I)_H$, pour $I = 1, N$.

On s'attend à ce que le problème variationnel approché (1.52) (respectivement le problème matriciel (1.53)) admette une solution u_h (resp. $\overrightarrow{U}$) et une seule... *Ceci n'est pas toujours le cas !* En effet, même si λ n'est pas valeur propre du problème exact, λ peut être valeur propre du problème matriciel (1.39), et, dans ce cas, la matrice $(\mathbb{A} - \lambda\mathbb{M})$ n'est plus inversible. C'est pourquoi, on doit raisonner sur une *suite de problèmes approchés*, suite indexée par h. Nous avons besoin de l'hypothèse "faible" d'approximabilité ci-dessous :

$$\begin{array}{c} \text{il existe } W \text{ dense dans } V, \text{ et, pour tout } h,\ r_h\ :\ W \to V_h \text{ tels que} \\ \lim_{h\to 0} \|r_h w - w\|_V = 0, \quad \forall w \in W. \end{array} \tag{1.54}$$

Remarque 1.37 *Par densité [15], on en déduit aisément que l'on peut en fait approcher tout élément v de V par une suite d'éléments $(v_k)_{k\geq 1}$, chaque v_k appartenant à un des sous-espaces $(V_h)_h$ ou que* $\lim\limits_{h\to 0} \inf\limits_{v_h \in V_h} \|v - v_h\|_V = 0$.

Ceci étant précisé, on peut établir le résultat ci-dessous.

Proposition 1.38 *On suppose que l'hypothèse (1.54) est satisfaite. Alors, il existe $h_0 > 0$ tel que, pour tout $h \in]0, h_0[$, le problème variationnel approché (1.52) admette une solution et une seule.*

Remarque 1.39 *En d'autres termes, lorsque h est "suffisamment petit",* λ ne peut plus être valeur propre *du problème matriciel (1.39). Ce résultat est à comparer à celui obtenu pour l'approximation des problèmes aux valeurs propres.*

Démonstration (de la proposition 1.38) : Avant de débuter la démonstration proprement dite, rappelons que $a(\cdot\,,\cdot)$ est une forme sesquilinéaire sur $V \times V$, hermitienne, continue et coercive. Ainsi, $\|\cdot\|_{V,a} : v \mapsto (a(v,v))^{1/2}$ définit une norme équivalente à $\|\cdot\|_V$ dans V. Nous choisissons de munir V de cette norme, de produit scalaire associé $(v, w) \mapsto a(v, w)$.
Raisonnons par l'absurde, en supposant que, pour tout $h_0 > 0$, il existe $h \in]0, h_0[$ tel que le problème (1.52) n'admette pas de solution. On peut donc construire une suite de pas $(h_k)_{k\geq 1}$ tendant vers 0 telle que, pour tout h_k, le problème (1.52) n'admette pas de solution : pour cela, on choisit $h_0 = 1/k$ et $h_k \in]0, 1/k[$ tel que (1.52) n'admette pas de solution. Or, tous les problèmes approchés sont de dimension finie ($\dim V_h < \infty$) : la non-existence équivaut donc à la non-unicité. Ainsi, pour chaque h_k, il existe w_{h_k} non-nul, solution de (1.52) avec second membre nul :

$$\forall k,\ \exists w_{h_k} \in V_{h_k} \setminus \{0\} \text{ tel que } a(w_{h_k}, v_h) - \lambda (w_{h_k}, v_h)_H = 0,\ \forall v_h \in V_{h_k}. \tag{1.55}$$

Dans la suite, on choisit tous les w_{h_k} de norme égale à 1, soit $\|w_{h_k}\|_{V,a} = 1$ pour tout k. Comme la suite $(w_{h_k})_k$ est bornée dans l'espace de Hilbert V, il existe une sous-suite – toujours notée $(w_{h_k})_k$ – qui converge faiblement dans V vers une limite appelée w (voir la proposition 1.9) :

$$\lim_{k\to\infty} a(w_{h_k}, v) = a(w, v),\ \forall v \in V.$$

Qui plus est, l'injection de V dans H est compacte : on déduit de la proposition 1.10 que $(w_{h_k})_k$ converge fortement dans H vers w, c'est-à-dire $\lim_{k\to\infty} \|w_{h_k} - w\|_H = 0$. Déterminons maintenant la valeur de w, avant d'aboutir à une contradiction.
Soit y un élément quelconque de V : d'après la propriété d'approximabilité (1.54), il existe une suite $(y_{h_k})_k$ telle que d'une part $y_{h_k} \in V_{h_k}$, pour tout k, et d'autre part $\lim_{k\to\infty} \|y_{h_k} - y\|_{V,a} = 0$. A partir de là, on écrit successivement :

$$\begin{aligned} a(w, y) - \lambda (w, y)_H &= a(w - w_{h_k}, y) - \lambda (w - w_{h_k}, y)_H + a(w_{h_k}, y) - \lambda (w_{h_k}, y)_H \\ &= a(w - w_{h_k}, y) - \lambda (w - w_{h_k}, y)_H + a(w_{h_k}, y - y_{h_k}) - \lambda (w_{h_k}, y - y_{h_k})_H, \end{aligned}$$

le passage à la dernière ligne se fait en utilisant $v_h = y_{h_k}$ dans (1.55).
Ceci étant valable pour tout k, on peut en particulier passer à la limite... En utilisant les convergences forte et faible de $(w_{h_k})_k$ vers w, on en déduit tout d'abord que

$$\lim_{k\to\infty} \{a(w - w_{h_k}, y) - \lambda (w - w_{h_k}, y)_H\} = 0.$$

Puis, en utilisant la convergence forte de $(y_{h_k})_k$ vers y, et en se souvenant que $(w_{h_k})_k$ est une suite bornée dans V (et donc aussi dans H), on en déduit cette fois que

$$\lim_{k\to\infty} \{a(w_{h_k}, y - y_{h_k}) - \lambda (w_{h_k}, y - y_{h_k})_H\} = 0.$$

Ainsi, on a prouvé que

$$a(w, y) - \lambda (w, y)_H = 0,\ \forall y \in V.$$

Comme λ n'est pas valeur propre, il suit $w = 0$.
Pour aboutir à une contradiction, utilisons une dernière fois (1.55), pour $v_h = w_{h_k}$:

$$1 = \|w_{h_k}\|^2_{V,a} = \lambda (w_{h_k}, w_{h_k})_H = \lambda \|w_{h_k}\|^2_H. \tag{1.56}$$

Or, $(w_{h_k})_k$ tend vers $w = 0$ dans H, d'où la contradiction. ∎

A l'aide d'une technique similaire, on peut établir la convergence de la méthode d'approximation interne, sous la forme d'un analogue du lemme de Céa [15].

Théorème 1.40 *On suppose que l'hypothèse (1.54) est satisfaite. Alors, il existe une constante $C > 0$ qui ne dépend que de λ et de $a(\cdot,\cdot)$, et $h_1 > 0$, tels que, pour tout $h \in]0, h_1[$:*

$$\|u_h - u\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{1.57}$$

Démonstration : Comme pour la démonstration de la proposition 1.38, nous choisissons de munir V de la norme $\|\cdot\|_{V,a}$, de produit scalaire associé $a(\cdot,\cdot)$.
Dans une première étape, on reprend la démonstration du lemme de Céa. Soit donc $w_h \in V_h$. On effectue la différence entre la formulation variationnelle (1.34) avec $v = w_h$ et la formulation variationnelle discrète (1.52) avec la même fonction-test w_h :

$$a(u-u_h,w_h)-\lambda(u-u_h,w_h)_H=0,\ \forall w_h\in V_h. \tag{1.58}$$

Soit $v_h\in V_h$ quelconque. En reprenant (1.58) avec $w_h=u_h-v_h$, on peut alors écrire la suite d'(in)égalités :

$$\begin{aligned}\|u-u_h\|_{V,a}^2-\lambda\|u-u_h\|_H^2 &= a(u-u_h,u-v_h)-\lambda(u-u_h,u-v_h)_H\\ &\le \|u-u_h\|_{V,a}\|u-v_h\|_{V,a}+\lambda\|u-u_h\|_H\|u-v_h\|_H\\ &\le (1+\beta\lambda)\|u-u_h\|_{V,a}\|u-v_h\|_{V,a},\end{aligned} \tag{1.59}$$

où $\beta={C_V}^2/\alpha^2$, avec α la constante de coercivité, cf. (1.23), et C_V le module de continuité de (1.19). La différence, par rapport au lemme de Céa, est qu'il reste un terme négatif à gauche du signe $\le$, sans lequel on pourrait directement conclure.

L'objet de la seconde étape est donc de traiter ce terme négatif. Pour cela, supposons que $u\neq u_h$ (car si $u=u_h$, la propriété (1.57) est vraie !), et formons

$$z_h=\frac{u-u_h}{\|u-u_h\|_{V,a}}.$$

Par construction, la suite $(z_h)_h$ est bornée dans V. En reprenant la démonstration de la proposition 1.38, il existe $z\in V$ tel qu'une sous-suite – toujours notée $(z_h)_h$ – converge faiblement vers z dans V, et fortement vers z dans H. Par ailleurs, pour tout $y\in V$, il existe une suite $(y_h)_h$ telle que d'une part $y_h\in V_h$, pour tout h, et d'autre part $\lim_{h\to 0}\|y_h-y\|_V=0$. On écrit maintenant :

$$\begin{aligned}a(z,y)-\lambda(z,y)_H &= a(z-z_h,y)-\lambda(z-z_h,y)_H+a(z_h,y)_V-\lambda(z_h,y)_H\\ &= a(z-z_h,y)-\lambda(z-z_h,y)_H\\ &\quad +a(z_h,y-y_h)-\lambda(z_h,y-y_h)_H\\ &\quad +a(z_h,y_h)-\lambda(z_h,y_h)_H.\end{aligned}$$

Ces égalités étant valables pour tout h, on peut passer à la limite dans le membre de droite (écrit sur trois lignes).
D'après les convergences faible et forte de $(z_h)_h$ vers z, et d'après la convergence forte de $(y_h)_h$ vers y, les deux premières lignes tendent vers 0. Quant à la troisième, il suffit de revenir à la définition de z_h et utiliser (1.58) avec $w_h=y_h$ pour obtenir sa nullité (pour tout h et tout y_h). Ainsi,

$$a(z,y)-\lambda(z,y)_H=0,\ \forall y\in V.$$

Comme λ n'est pas valeur propre, il suit $z=0$. La suite $(z_h)_h$ converge donc vers 0 dans H ; on introduit alors $h_1>0$ tel que, pour tout $h\in\,]0,h_1[$,

$$\|z_h\|_H^2\le\frac{1}{2\lambda},\ \text{soit}\ -\lambda\|u-u_h\|_H^2\ge-\frac{1}{2}\|u-u_h\|_{V,a}^2.$$

Si on reprend (1.59) pour $h\in\,]0,h_1[$, on a alors :

$$\begin{aligned}&\frac{1}{2}\|u-u_h\|_{V,a}^2\le(1+\beta\lambda)\|u-u_h\|_{V,a}\|u-v_h\|_{V,a},\ \text{ou encore}\\ &\|u-u_h\|_{V,a}\le 2(1+\beta\lambda)\|u-v_h\|_{V,a}.\end{aligned}$$

Cette inégalité étant valable pour tout élément v_h de V_h, la conclusion en découle, avec la norme $\|\cdot\|_{V,a}$. Par équivalence des normes dans V, on en déduit le résultat (1.57). ∎

Remarque 1.41 *Supposons que l'on s'intéresse à l'approximation du problème de Helmholtz en utilisant des éléments finis de Lagrange. A l'aide de la borne (1.57), on peut déduire la vitesse de convergence de la méthode numérique en s'appuyant sur les estimations d'erreur à notre disposition pour la discrétisation par ces éléments finis, selon la régularité de la solution.*

1.4 Illustrations numériques

Nous donnons dans cette section quelques illustrations numériques du calcul des valeurs propres de l'opérateur de Laplace muni de conditions de Dirichlet et de la résolution du problème de Helmholtz. Les calculs numériques ont été réalisés à l'aide de Matlab, en utilisant les procédures générales de maillage et de construction des matrices éléments finis exposées dans [15]. Notons que, lorsque les formes sont à valeurs réelles (cf. remarque 1.24), les problèmes sont a priori résolus dans $\mathbb{R}$.

1.4.1 Calculs de valeurs et fonctions propres

On s'intéresse ici au calcul des valeurs propres de l'opérateur de Laplace équipé de conditions de Dirichlet homogènes dans un ouvert borné (et connexe) de $\mathbb{R}^2$:

$$\begin{cases} -\Delta u = \lambda u \text{ dans } \Omega, \\ u = 0 \quad \text{sur } \partial\Omega. \end{cases} \tag{1.60}$$

C'est le problème (1.15), avec $n = 2$ et $\Gamma_2 = \emptyset$, dont la formulation variationnelle dans $H_0^1(\Omega)$ est :

$$\begin{cases} \text{trouver } \lambda \in \mathbb{R} \text{ et } u \in H_0^1(\Omega) \setminus \{0\} \text{ tels que} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \lambda \int_\Omega uv \, d\Omega, \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{1.61}$$

Rappelons que, d'après le corollaire 1.15, ce problème admet une infinité de valeurs propres réelles $(\lambda_n)_{n\geq 1}$ avec $0 < \lambda_1 \leq \cdots \leq \lambda_n \to +\infty$ et il existe une base hilbertienne de $L^2(\Omega)$, formée de fonctions propres $(u_n)_{n\geq 1}$.
Soit V_h un espace d'approximation interne de dimension finie N de $H_0^1(\Omega)$, on introduit le problème spectral discrétisé :

$$\begin{cases} \text{trouver } \lambda_h \in \mathbb{R} \text{ et } u_h \in V_h \setminus \{0\} \text{ tels que} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h \, d\Omega = \lambda_h \int_\Omega u_h \, v_h \, d\Omega, \quad \forall v_h \in V_h. \end{cases} \tag{1.62}$$

Sur une base $(w_I)_{I=1,N}$ de l'espace V_h, le problème précédent s'écrit sous la forme matricielle :

$$\begin{cases} \text{trouver } \lambda_h \in \mathbb{R} \text{ et } \overrightarrow{U} \in \mathbb{R}^N \setminus \{0\} \text{ tels que} \\ \mathbb{K}\overrightarrow{U} = \lambda_h \, \mathbb{M}\overrightarrow{U} \end{cases} \tag{1.63}$$

avec

$$u_h = \sum_{J=1,N} U^J w_J,$$
$$\mathbb{K}_{IJ} = \int_\Omega \nabla w_I \cdot \nabla w_J \, d\Omega, \ \mathbb{M}_{IJ} = \int_\Omega w_I \, w_J \, d\Omega, \ \forall I, J = 1, \leq N.$$

Notons que le problème spectral (1.63) s'inscrit dans le cadre du théorème 1.25. Par conséquent, il existe N valeurs propres réelles strictement positives (chaque valeur propre étant comptée autant de fois que sa multiplicité) $(\lambda_{m,h})_{m=1,N}$ associées à une base de vecteurs propres $(\overrightarrow{V}_m)_{m=1,N}$, orthonormale pour le produit scalaire : $(\overrightarrow{V}, \overrightarrow{W}) \mapsto (\mathbb{M}\overrightarrow{V}|\overrightarrow{W})$, où $(\cdot|\cdot)$ dénote le produit scalaire euclidien de $\mathbb{R}^N$.

Dans le cadre de l'approximation par éléments de finis de Lagrange, les matrices étant creuses il convient d'utiliser des méthodes de recherche de valeurs propres de type itératif, Lanczos ou sous-espace par exemple. Ces méthodes permettent en outre de ne calculer que quelques valeurs propres, les plus petites par exemple ; ces dernières étant les plus pertinentes du point de vue de la physique. Dans les exemples qui suivent, a été utilisée la procédure **eigs** de Matlab qui s'appuie sur une méthode de Lanczos.

La procédure **eigs** de Matlab requiert que la matrice au second membre soit rigoureusement symétrique. Ce n'est pas nécessairement le cas lorsque les matrices éléments finis ne sont pas calculées par une méthode préservant strictement la symétrie. Une légère dissymétrie peut apparaître, due à des erreurs d'arrondis. On peut donc être amené à re-symétriser les matrices en effectuant les opérations $\frac{1}{2}\left(\mathbb{M} + \mathbb{M}^t\right)$ et $\frac{1}{2}\left(\mathbb{K} + \mathbb{K}^t\right)$.

Dans le cas de conditions de Dirichlet homogène sur la frontière, ce qui est le cas ici, il faut prendre quelques précautions. En effet, si on utilise la technique de pseudo-élimination (cf. [15]) on fait apparaître des valeurs propres "parasites". Rappelons que cette technique consiste à transformer la matrice de rigidité totale $\widetilde{\mathbb{K}}$ (définie sur tous les nœuds du maillage, y compris les nœuds portant une condition de Dirichlet sur la frontière) en la matrice où on a annulé toutes les lignes et colonnes correspondant à un nœud de Dirichlet, exceptés les termes diagonaux qui demeurent inchangés :

$$\widetilde{\mathbb{K}} \longrightarrow \begin{bmatrix} \widetilde{\mathbb{K}}_{\mathcal{I}\mathcal{I}} & 0 \\ 0 & diag(\widetilde{\mathbb{K}}_{\mathcal{D}\mathcal{D}}) \end{bmatrix} \qquad \begin{array}{l} (\mathcal{D} \text{ indice des nœuds Dirichlet ;} \\ \mathcal{I} \text{ indice des autres nœuds).} \end{array}$$

Pour la matrice de masse $\widetilde{\mathbb{M}}$, on procède de même :

$$\widetilde{\mathbb{M}} \longrightarrow \begin{bmatrix} \widetilde{\mathbb{M}}_{\mathcal{I}\mathcal{I}} & 0 \\ 0 & diag(\widetilde{\mathbb{M}}_{\mathcal{D}\mathcal{D}}) \end{bmatrix}.$$

Dans ce cas on constate que l'on fait apparaître les valeurs propres parasites $\lambda_d = \widetilde{\mathbb{K}}_{dd}/\widetilde{\mathbb{M}}_{dd}$, pour $d \in \mathcal{D}$. Or, pour un problème dans $\mathbb{R}^n$ les coefficients $\widetilde{\mathbb{K}}_{dd}$ et $\widetilde{\mathbb{M}}_{dd}$ se comportent respectivement comme h^{n-2} et h^n, montrant que les valeurs propres $(\lambda_d)_{d\in\mathcal{D}}$ se comportent comme h^{-2}. Par conséquent, elles ne "parasitent" pas les petites valeurs propres lorsque $h \to 0$. Par ailleurs, dans une technique itérative telle que la méthode de Lanczos il est possible de choisir des vecteurs initiaux n'ayant aucune composante sur le sous-espace correspondant aux nœuds de Dirichlet. Le processus itératif conservant cette propriété, les fonctions propres liées à ces valeurs propres ne sont pas atteignables.

Bien évidemment, la technique d'élimination réelle[2] des conditions de Dirichlet sur la frontière ne demande quant à elle aucune précaution !

Nous donnons ci-dessous la procédure Matlab correspondant au calcul des premières valeurs propres par une approximation par éléments finis de Lagrange d'ordre 2.

```
p=50;nbvp=10;
[S,T,BR,RT]=triangle_rectangle([0 1 0 1],p,p,1);          %maillage P2
[S2,T2,BR2,RT2]=maillageP2(S,T,BR,RT);                    %du carre unite
[T2,S2]=renume(T2,S2);                                    %renumerotation
[K,M]=calcul_EF_2D(S2,T2,RT2);                            %matrices EF
K=(K+K')/2;M=(M+M')/2;                                    %symetrisation
Noeud_dir=noeud_bords(S2,T2,BR2,[1 2 3 4]);               %noeuds Dirichlet
Z=zeros(size(S2,1),1);                                    %pseudo-elimination
[Ke,B]=cd_Dirichlet(K,Z,Noeud_dir,Z);                     %des cond. de Dirichlet
[Me,B]=cd_Dirichlet(M,Z,Noeud_dir,Z);                     %des matrices K et M
[V,D]=eigs(Ke,Me,nbvp,'sm');                              %calcul des elts propres
[lambda,Ix]=sort(diag(D));                                %tri des val. propres
V=V(:,Ix);                                                %et des vect. propres
T21=isop2(T2);                                            %dessins des modes
figure
for k=1:4,
  subplot(2,2,k);
  trisurf(T21,S2(:,1),S2(:,2),V(:,k));
  shading interp;view(2);axis image;
  title(['mode ' num2str(k) ' \lambda ' num2str(k)
         ' = ' num2str(lambda(k))] );
end,
```

Remarque 1.42 *Il est facile de réaliser l'élimination effective des conditions de Dirichlet sur la frontière à l'aide des commandes Matlab ci-dessous :*

```
Indir=find(Noeud_dir==1);Ke=K(Indir,Indir);
Me=M(Indir,Indir);
```

Néanmoins, pour des cas de très grande dimension N, la réalisation de ce calcul peut devenir prohibitive !

2. C'est-à-dire, lorsque l'on résout exactement (1.63) avec $\mathbb{K} = \widetilde{\mathbb{K}}_{\mathcal{I}\mathcal{I}}$ et $\mathbb{M} = \widetilde{\mathbb{M}}_{\mathcal{I}\mathcal{I}}$ (cf. [15]).

Nous utilisons ce code pour déterminer les fonctions propres du laplacien dans un carré unité et dans un carré unité incomplet présentant un coin rentrant, ou domaine en forme de L.

Fonctions propres du laplacien dans un carré

Nous nous plaçons sur le carré unité : $\Omega =]0,1[\times]0,1[$. Dans ce cas, on vérifie facilement (par séparation de variables) que les valeurs propres et les fonctions propres sont données par :

$$\lambda_{mn} = (m^2 + n^2)\pi^2 \qquad u_{nm}(x,y) = a_{nm} \sin m\pi x \sin n\pi y \quad \forall m \in \mathbb{N}^*, \forall n \in \mathbb{N}^*.$$

Il y a des valeurs propres doubles provenant de la symétrie du domaine $\lambda_{mn} = \lambda_{nm}$. Par ailleurs, les fonctions propres sont régulières, $u_{nm} \in \mathcal{C}^\infty(\overline{\Omega})$. Le code Matlab précédent calcule les 10 premières valeurs propres et fournit une représentation graphique des 4 premières fonctions propres (sans compter les fonctions propres symétriques) ; voir la figure 1.1.

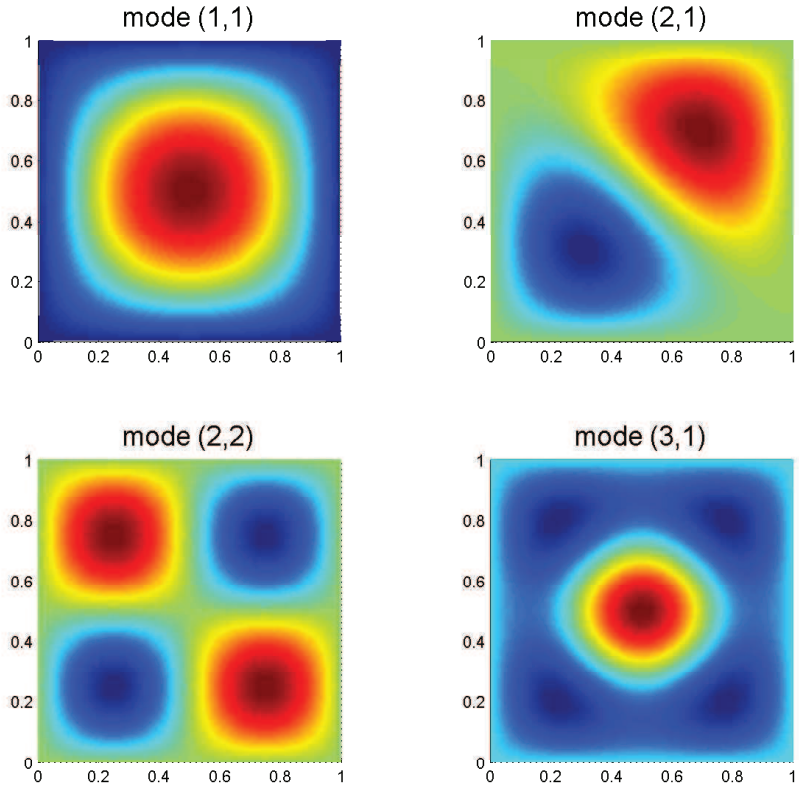


Figure 1.1. Premiers modes du laplacien dans le carré

Il est intéressant de vérifier si l'on obtient les taux de convergence prévus par la théorie. Nous représentons sur la figure 1.2 pour les 4 premières valeurs propres l'écart en valeur absolue entre la valeur théorique et la valeur calculée par une

approximation par éléments finis P^1 en fonction du pas du maillage. La représentation est en échelle log-log. On observe un accord parfait avec la théorie (cf. (1.49) et la remarque 1.33), à savoir des taux de convergence en h^2. On note que plus la valeur propre est grande plus l'écart est grand, ce qui est également en accord avec cette estimation, puisque dans ce cas $\varepsilon_m(h) = C_D\, h \|\Delta u_m\|_{L^2(\Omega)} = C_D\, h \|\lambda_m u_m\|_{L^2(\Omega)} = C_D\, h\, \lambda_m$.

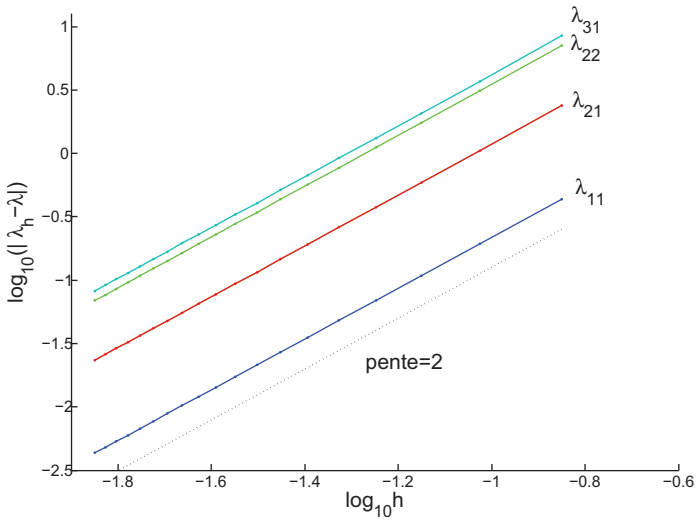


Figure 1.2. Convergence des valeurs propres du laplacien dans le carré en P^1

Sur la figure 1.3 est représentée la même variation mais pour une approximation par éléments finis P^2. On note une fois encore un excellent accord avec la théorie, à savoir un taux de convergence en h^4.

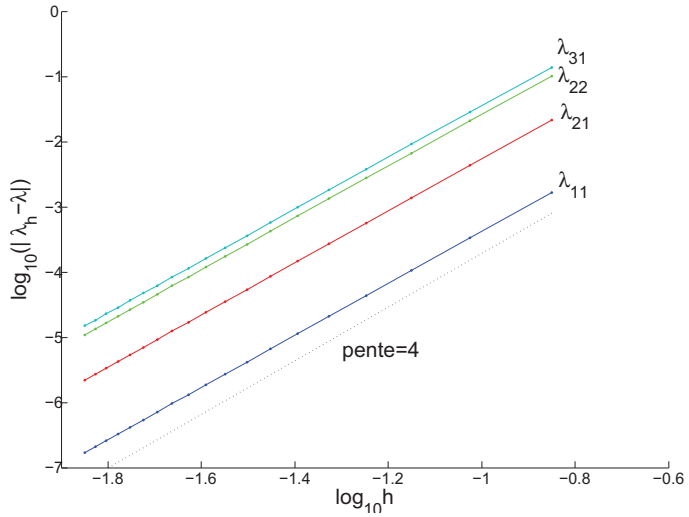


Figure 1.3. Convergence des valeurs propres du laplacien dans le carré en P^2

Fonctions propres du laplacien sur un domaine à coin rentrant

Nous considérons maintenant le domaine géométrique Ω représenté figure 1.4. La différence essentielle avec le cas précédent est l'apparition de singularités, c'est-à-dire des fonctions propres n'appartenant pas toujours à $H^2(\Omega)$, mais à $\cap_{s<1+\sigma_D} H^s(\Omega)$, avec $\sigma_D \in]1/2, 1[$ dépendant de l'angle au coin rentrant. Dans ce

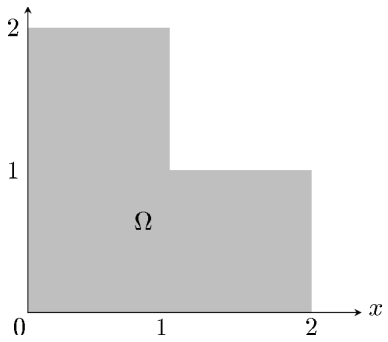


Figure 1.4. Domaine en L, ou carré incomplet

cas les valeurs propres et les fonctions propres ne sont pas toutes connues analytiquement. Sur la figure 1.5 nous avons représenté les 4 premières fonctions propres obtenues à l'aide d'une approximation par éléments finis P^2, associées aux valeurs propres : $\lambda_1 \approx 9.6418$ $\lambda_2 \approx 15.1973$ $\lambda_3 \approx 19.7392$ $\lambda_4 \approx 29.5215$.

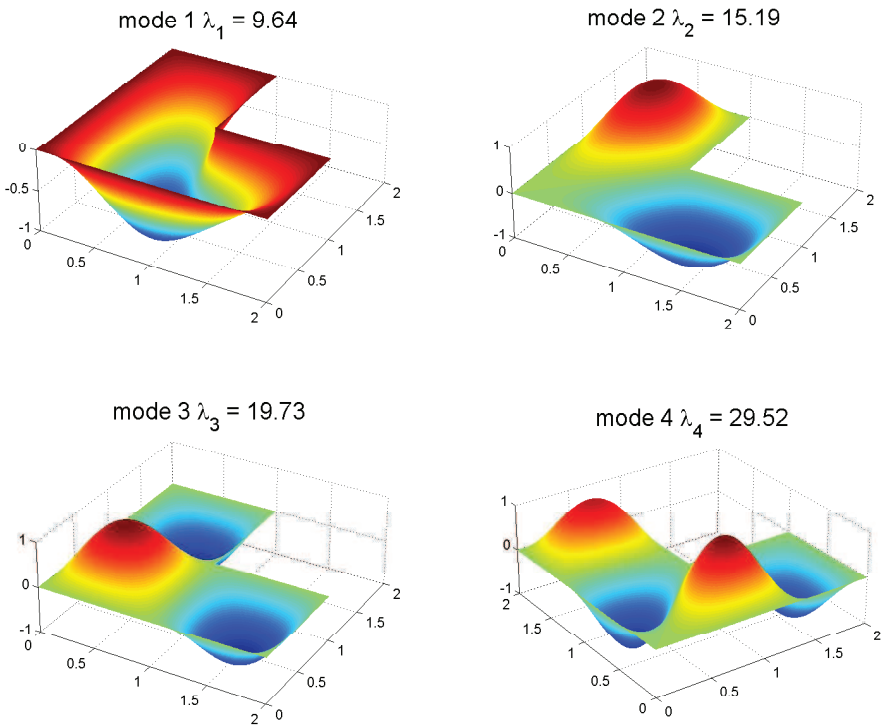


Figure 1.5. Premiers modes du laplacien dans le carré incomplet

Ces fonctions propres sont à rapprocher des fonctions propres obtenues sur un disque incomplet (voir figure 1.6), présentant un coin rentrant. Dans ce cas, on

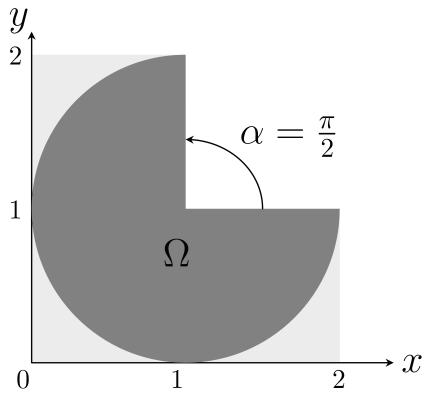


Figure 1.6. Disque incomplet

peut déterminer toutes les fonctions propres du problème (1.61) :

$$\forall m > 0, \forall n \geq 1 \quad u_{mn}(r,\theta) = a_{mn} J_{\nu_m}(\gamma_{mn} r) \sin \nu_m(\theta - \alpha) \text{ et } \lambda_{mn} = \gamma_{mn}^2,$$

avec γ_{mn} le $n^{ème}$ zéro de la fonction de Bessel J_{ν_m}de première espèce d'ordre $\nu_m = \frac{m\pi}{2\pi - \alpha}$ (cf. [1]). Dans le cas où $\alpha = \frac{\pi}{2}$, on obtient ainsi les 4 premières fonctions propres suivantes :

$$\begin{aligned}
u_{11} &= a_{11} J_{\frac{2}{3}}(\gamma_{11} r) \sin \tfrac{2}{3}(\theta - \tfrac{\pi}{2}) \quad \lambda_{11} = 11.39 \\
u_{21} &= a_{21} J_{\frac{4}{3}}(\gamma_{21} r) \sin \tfrac{4}{3}(\theta - \tfrac{\pi}{2}) \quad \lambda_{21} = 18.27 \\
u_{31} &= a_{31} J_{2}(\gamma_{31} r) \sin 2(\theta - \tfrac{\pi}{2}) \quad \lambda_{31} = 26.31 \\
u_{41} &= a_{41} J_{\frac{8}{3}}(\gamma_{41} r) \sin \tfrac{8}{3}(\theta - \tfrac{\pi}{2}) \quad \lambda_{41} = 35.63
\end{aligned}$$

que l'on représente sur la figure 1.7. On peut vérifier que $u_{11} \in \cap_{s<5/3} H^s(\Omega)$, $u_{21} \in \cap_{s<7/3} H^s(\Omega)$ et plus généralement que $u_{mn} \in \cap_{s<2m/3+1} H^s(\Omega)$.
L'allure générale (nombre de creux et de bosses) est similaire dans les 2 cas. Du point de vue du comportement local au voisinage du coin rentrant, la théorie [29] prévoit des comportements similaires en termes d'appartenance aux espaces de Sobolev. Sur la figure 1.8 (resp. 1.9) nous représentons une estimation de l'erreur sur les valeurs propres pour le carré avec un coin obtenue à l'aide d'une approximation par éléments finis P^1 (resp. P^2). Il s'agit d'une estimation car nous avons utilisé comme valeurs "exactes" les valeurs propres obtenues avec une approximation P^2 avec un maillage fin. Les pentes obtenues pour les plus grandes valeurs de h sont significatives des taux de convergence.

Dans le cas de l'approximation P^1, on note que les valeurs propres λ_2, λ_3 et λ_4 sont approchées à l'ordre 2, ordre optimal prévu par la théorie. La valeur propre λ_1 est

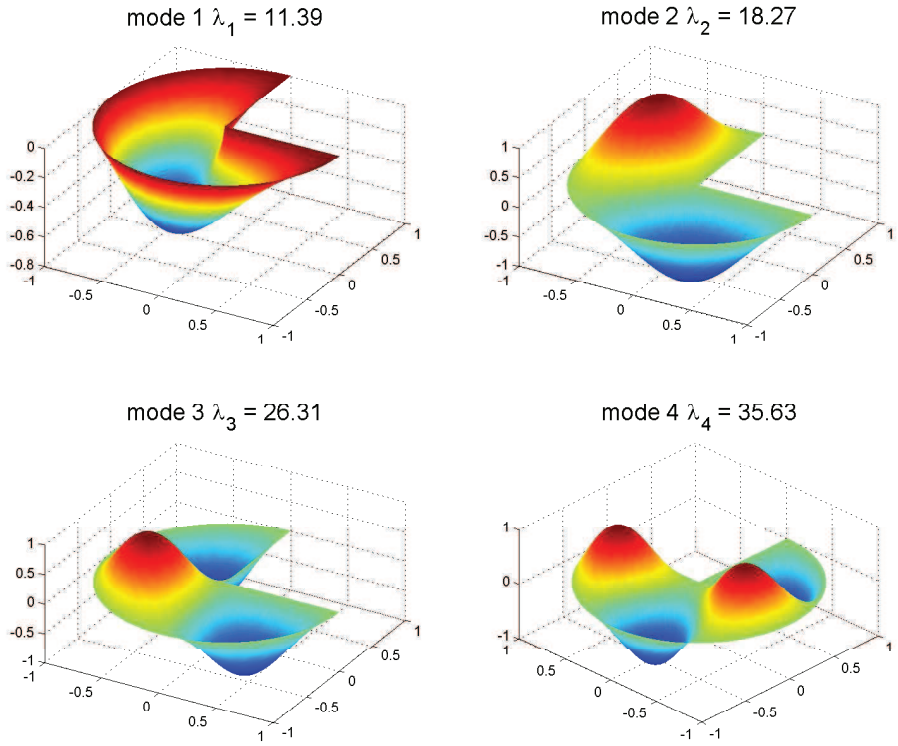


Figure 1.7. Premiers modes du laplacien dans le disque incomplet

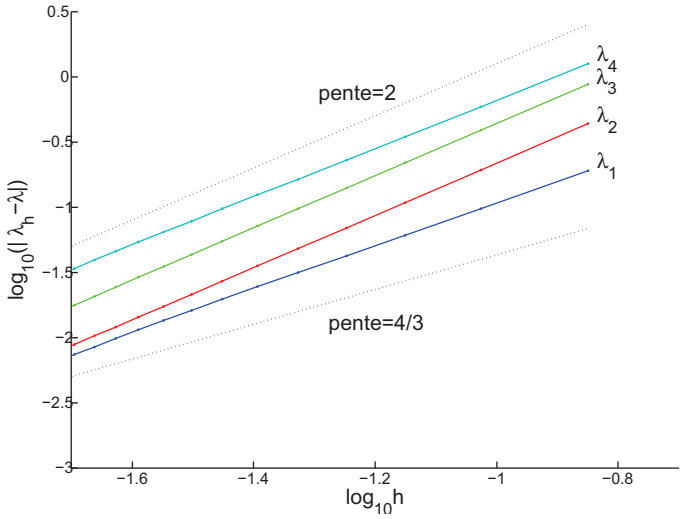


Figure 1.8. Convergence des valeurs propres du laplacien dans le carré incomplet en P^1

seulement approchée à l'ordre 4/3. Cela s'explique par le fait que la fonction propre associée à cette valeur propre présente un comportement singulier au voisinage du coin rentrant, celui de la fonction de Bessel $J_{\frac{2}{3}}$. La fonction u_{11} n'appartenant qu'à l'espace $\cap_{s<5/3} H^s(\Omega)$, on a finalement une erreur d'approximation de la fonction propre en norme $H^1(\Omega)$ d'ordre $\frac{5}{3}-1=\frac{2}{3}$ (cf. (1.47)) et d'ordre $2(\frac{5}{3}-1)=\frac{4}{3}$ pour la valeur propre correspondante (cf. (1.49)). Pour ce qui concerne l'approximation par éléments finis P^2, on observe le même phénomène. La première valeur propre converge à l'ordre $\frac{4}{3}$ et pas mieux, la seconde converge à l'ordre $\frac{8}{3}$ car u_{21}

n'appartient qu'à l'espace $\cap_{s<7/3} H^s(\Omega)$ (avec $2(\frac{7}{3}-1)=\frac{8}{3}$) et les deux dernières convergent à l'ordre 4 car les fonctions propres associées sont plus régulières. En effet, pour des éléments finis P^2, il faut une solution dans $H^3(\Omega)$ pour retrouver l'ordre optimal de convergence.

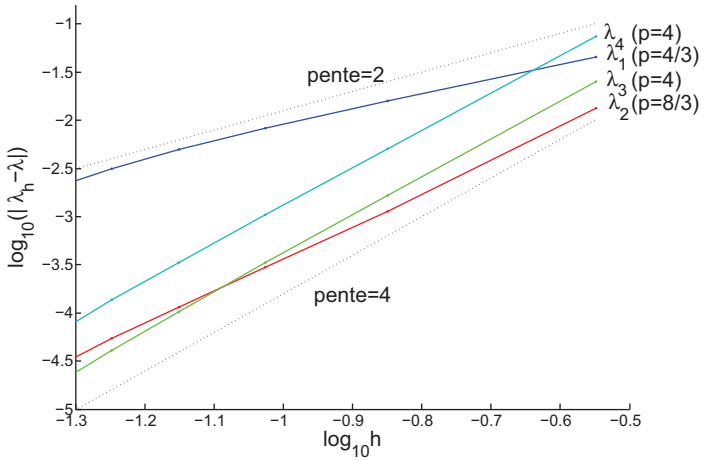


Figure 1.9. Convergence des valeurs propres du laplacien dans le carré incomplet en P^2

1.4.2 Résolution de l'équation de Helmholtz

On s'intéresse maintenant à la résolution de l'équation de Helmholtz :

$$\Delta u + \omega^2 u = 0$$

avec une condition aux limites non homogène, dans un cas où le domaine est borné (le carré unité par exemple) et dans un cas où le domaine est non borné (une bande semi-infinie par exemple), ce dernier exemple permettant d'illustrer une situation moins académique.

Problème de Helmholtz en domaine borné

On se place à nouveau sur le carré unité $\Omega =]0,1[\times]0,1[$ et on considère le problème suivant (pour $g \in L^2(\partial\Omega)$) :

$$\begin{cases} \Delta u + \omega^2 u = 0 \text{ dans } \Omega, \\ \dfrac{\partial u}{\partial n} = g \qquad \text{sur } \partial\Omega, \end{cases} \tag{1.64}$$

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

$$\left\{ \begin{array}{l} \text{trouver } u \in H^1(\Omega) \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v \, d\Omega - \omega^2 \int_\Omega uv \, d\Omega = \int_{\partial\Omega} gv \, d\Gamma, \quad \forall v \in H^1(\Omega). \end{array} \right. \tag{1.65}$$

Rappelons que le problème (1.65) est un problème de type coercif+compact (voir (1.34)) et qu'en vertu du théorème 1.19, si ω^2 est différent d'une valeur propre du laplacien avec condition de Neumann ($\lambda_{mn} = (m^2+n^2)\pi^2$ $m, n \geq 0$ associée aux fonctions propres $u_{mn}(x,y) = a_{mn} \cos m\pi x \cos n\pi y$), alors il est bien posé. Dans le cas où $\omega^2 = \lambda_{mn}$, d'après l'alternative de Fredholm (cf. remarque 1.21), il existe une solution (non unique) si et seulement si :

$$\int_{\partial\Omega} g \, u_{mn} \, d\Gamma = 0, \text{ pour toute fonction propre } u_{mn} \text{ associée à } \lambda_{mn}.$$

Dans la suite, on suppose que $\omega^2 \notin \{\lambda_{mn},\ m, n \geq 1\}$. Soient V_h un sous-espace de dimension finie N de $H^1(\Omega)$ (approximation interne) et $(w_I)_{I=1,N}$ une base de V_h, on introduit le problème discrétisé :

$$\left\{ \begin{array}{l} \text{trouver } u_h \in V_h \text{ tel que :} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h \, d\Omega - \omega^2 \int_\Omega u_h v_h \, d\Omega = \int_{\partial\Omega} g v_h \, d\Gamma, \quad \forall v_h \in V_h \,, \end{array} \right. \tag{1.66}$$

qui conduit au système linéaire :

$$\left\{ \begin{array}{l} \text{trouver } \overrightarrow{U} \in \mathbb{R}^N \text{ tel que} \\ \left(\mathbb{K} - \omega^2 \, \mathbb{M}\right) \overrightarrow{U} = \overrightarrow{G} \end{array} \right. \tag{1.67}$$

où $\mathbb{K}$ et $\mathbb{M}$ sont définies comme précédemment, et $\overrightarrow{G} \in \mathbb{R}^N$ est tel que $G^I = \displaystyle\int_{\partial\Omega} g w_I \, d\Gamma$, $1 \leq I \leq N$.

Si ω^2 n'appartient pas à l'ensemble $\{\lambda_{m,h},\ 1 \leq m \leq N\}$ (ce qui est vrai pour h suffisamment petit d'après la proposition 1.38), alors le système linéaire (1.67) admet une unique solution [3]. Evidemment, les valeurs propres du problème spectral discret ne coïncident pas avec celles du problème spectral continu.

Nous avons mis en œuvre la résolution de ce problème académique en utilisant les outils de maillage et de calculs des matrices éléments finis présentés dans [15]. Plutôt que de calculer le second membre $\overrightarrow{G}$, nous avons utilisé l'interpolation par éléments finis de g, à savoir ($(M_I)_{I=1,N}$ nœuds du maillage) :

$$\widetilde{g}_h = \sum_{I \in \mathcal{N}} g(M_I) w_I, \quad \mathcal{N} \text{ indice des nœuds Neumann (sur la frontière)},$$

3. Si au contraire ω^2 appartient à l'ensemble des valeurs propres discrètes, il existe une solution (non unique) si et seulement si $(\overrightarrow{G} | \overrightarrow{V}_m) = 0$ pour tout vecteur propre $\overrightarrow{V}_m$ associé à la valeur propre atteinte.

qui constitue une approximation d'ordre compatible avec l'approximation par éléments finis dès lors que la fonction g est suffisamment régulière. Nous résolvons donc le système linéaire :

$$\left(\mathbb{K} - \omega^2\,\mathbb{M}\right)\overrightarrow{U} = \mathbb{M}_{\partial\Omega}\overrightarrow{G}_{\partial\Omega},$$

avec $G^I_{\partial\Omega} = \widetilde{g}_h(M_I)$ et $(\mathbb{M}_{\partial\Omega})_{IJ} = \displaystyle\int_{\partial\Omega} w_I\, w_J\, d\Gamma$, $1 \le I, J \le N$.

Pour calculer la solution par éléments finis P^1 on utilise la procédure Matlab suivante :

```
w=20;p=50;
[S,T,BR,RT]=triangule_rectangle([0 1 0 1],p,p,1);     %maillage P1 du carre
[K,M]=calcul_EF_2D(S,T,RT);                                %matrices EF
[Mb]=calcul_EF_1D(S,T,BR,[4]);                  %masse sur le cote 4 (x=0)
A=K-w*w*M;                                              %matrice Helmholtz
Ng=noeud_bords(S,T,BR,[4]);                    %noeuds sur le cote 4 (x=0)
G=Mb*(Ng.*g(S));                                          %second membre
U=A\G;                                                  %resolution du systeme
trisurf(T,S(:,1),S(:,2),U);shading interp;              %dessin des isovaleurs
```

On suppose que $g = 0$ sur les trois côtés de la frontière autres que $\{(x,y) \in \partial\Omega,\ x = 0\}$. Si on choisit par exemple la donnée $g = 1$ sur le côté $\{(x,y) \in \partial\Omega,\ x = 0\}$, la solution exacte est donnée par $u(x,y) = -\frac{\cos\omega(1-x)}{\omega\sin\omega}$ et le résultat de la procédure Matlab est représenté sur la figure 1.10. On note que la solution est correctement approchée hormis une petite oscillation transverse dont la longueur d'onde est de l'ordre de la taille d'un triangle. En traçant les courbes de convergence pour les

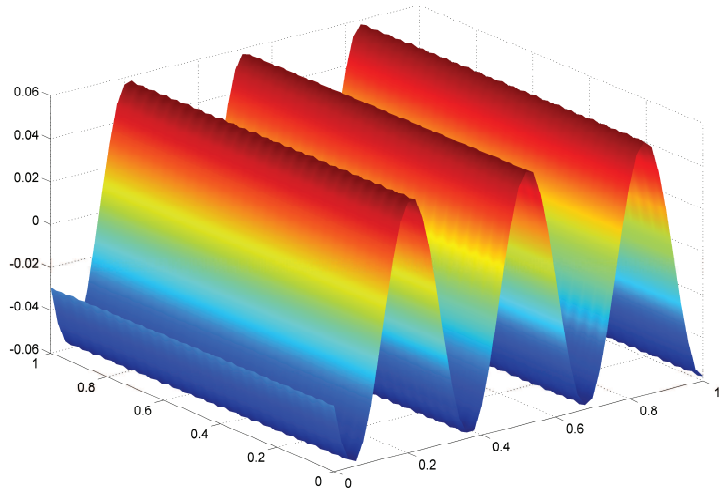


Figure 1.10. Solution du problème de Helmholtz en P^1

approximations P^1 et P^2, voir la figure 1.11 où on représente l'erreur en norme

$L^2(\Omega)$ en fonction du pas h du maillage en échelle logarithmique, on retrouve les ordres de convergence attendus [4]. On note que pour des pas trop grands ($h > 0.04$), l'approximation par éléments finis P^1 donne de très mauvais résultats. Ce qui n'est pas surprenant, puisque pour $\omega = 20$ la longueur d'onde (avec une vitesse de propagation unitaire) vaut $\Lambda = 2\pi/\omega \approx 0.31$ et il y a donc moins de 8 points de discrétisation par oscillation, ce qui ne permet pas d'approcher correctement une période d'une sinusoïde par interpolation linéaire. Dans le cas de l'approximation P^2, on obtient une convergence légèrement meilleure (ordre 3.7) que celle prévue par la théorie (ordre 3).

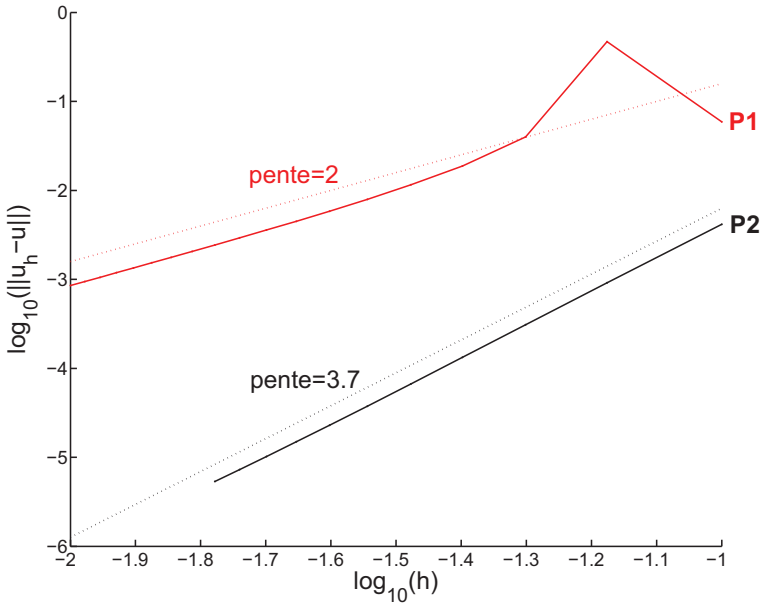


Figure 1.11. Courbes de convergence pour le problème de Helmholtz dans le carré

Il est intéressant de faire varier la pulsation ω et d'observer ce qui arrive lorsque ω^2 passe au voisinage d'une fréquence propre du laplacien. Pour ce faire nous représentons sur la figure 1.12 la variation du logarithme de l'erreur en norme $L^2(\Omega)$ en

4. L'estimation (1.57) du théorème 1.40 fournit une estimation de l'erreur en norme $H^1(\Omega)$. Pour obtenir un résultat en norme $L^2(\Omega)$, on introduit un problème adjoint du type :

$$\begin{cases} \text{trouver } \varphi \in H^1(\Omega) \text{ tel que} \\ \displaystyle\int_\Omega \nabla v \cdot \nabla \varphi \, d\Omega - \omega^2 \int_\Omega v\varphi \, d\Omega = \int_\Omega f v \, d\Omega, \quad \forall v \in H^1(\Omega), \end{cases}$$

avec $f \in L^2(\Omega)$. Dans un carré, ce problème est régulier, à savoir que $f \in L^2(\Omega) \Longrightarrow \varphi \in H^2(\Omega)$, avec condition de stabilité $\|\varphi\|_{H^2(\Omega)} \leq C \, \|f\|_{L^2(\Omega)}$, C indépendant de f. On en conclut (en suivant par exemple [15] où le cas du laplacien a été traité) que l'on gagne un ordre entre la vitesse de convergence en norme $H^1(\Omega)$ – h^t – et celle en norme $L^2(\Omega)$ – h^{t+1} –.

fonction de ω dans l'intervalle $[0, 8]$. Dans cet intervalle, il y a les valeurs propres $\lambda_{00} = 0$, $\lambda_{10} = \lambda_{01} = \pi^2$, $\lambda_{11} = 2\pi^2$, $\lambda_{20} = \lambda_{02} = 4\pi^2$ et $\lambda_{21} = \lambda_{12} = 5\pi^2$. Ces courbes ont été obtenues avec une approximation par éléments finis P^2 et un pas de maillage de 0.02. On représente également le logarithme de l'estimation du conditionnement [5] en norme 1 de la matrice du système linéaire. On observe bien évidemment que le conditionnement "explose" lorsque ω se rapproche de $\sqrt{\lambda_{mn}} = \sqrt{m^2 + n^2}\pi$, $m, n \geq 0$; les oscillations qui apparaissent sur la courbe du conditionnement sont liées à l'algorithme itératif d'estimation du conditionnement (procédure **condest**) utilisé dans Matlab. Par contre, l'erreur sur la solution n'explose que lorsque ω se rapproche des valeurs $\sqrt{\lambda_{00}}$, $\sqrt{\lambda_{10}}$ et $\sqrt{\lambda_{20}}$. Cela est dû au fait que la donnée $g = 1$ sur le côté $\{x = 0\}$ n'excite pas les modes $(m = 1, n = 1)$, $(m = 2, n = 1)$ ou $(m = 1, n = 2)$ car :

$$\int_{x=0} u_{11} dy = \int_{x=0} u_{21} dy = \int_{x=0} u_{12} dy = 0.$$

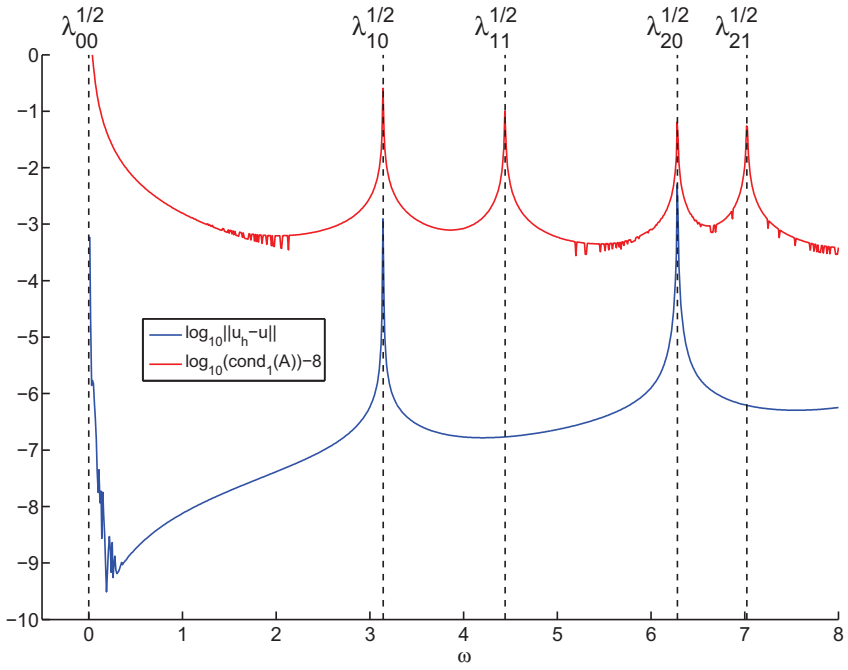


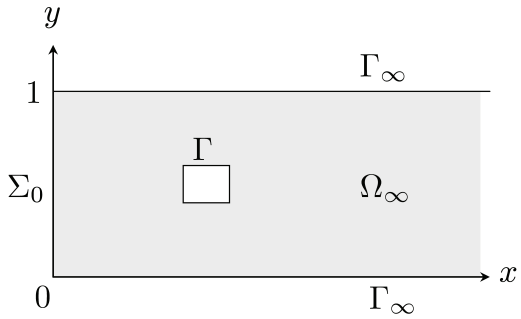
Figure 1.12. Variations de l'erreur en norme $L^2(\Omega)$ en fonction de ω, à h fixé

Problème de Helmholtz dans un guide semi-infini

Afin de traiter un exemple moins académique, nous nous intéressons maintenant au problème de la diffraction d'une onde en régime harmonique dans un guide d'onde

5. $\max_{||x||_1=1} ||\mathbb{A}^{-1}x||_1$ où $\mathbb{A} = \mathbb{K} - \omega^2\mathbb{M}$.

semi-infini. On rappelle qu'un guide est dit infini lorsque la coordonnée longitudinale varie de $-\infty$ à $+\infty$. Le guide est dit semi-infini lorsque cette coordonnée varie d'une valeur finie (par exemple 0) à $+\infty$. On choisit ici le guide semi-infini de hauteur unitaire :



sur lequel on considère le problème modèle de diffraction par un objet (de frontière Γ) :

$$\begin{cases} \Delta u + \omega^2 u = 0 & \text{dans } \Omega_\infty \\ \partial_n u = g & \text{sur } \Sigma_0 \\ \partial_n u = 0 & \text{sur } \Gamma \cup \Gamma_\infty \\ \text{cond. d'ondes sortantes à l'infini} \end{cases} \tag{1.68}$$

La condition sur les frontières Γ et Γ_∞ signifie que celles-ci sont constituées de parois parfaitement réfléchissantes. La condition sur Σ_0 traduit l'effet d'une membrane acoustique (source de l'émission). Enfin, le choix d'une condition d'onde sortante (entre "onde sortante" et "onde entrante") permet de sélectionner, parmi les solutions possibles, celles qui se propagent dans le sens des x croissants.

Dans un guide infini, et en l'absence d'objet diffractant, il existe des solutions particulières appelées modes du guide :

$$\Psi_n^{\pm}(x,y) = \theta_n(y) e^{\pm i\beta_n x} \tag{1.69}$$

avec :

$$\begin{cases} \left(\theta_n, \lambda_n^2\right)_{n\geq 0} \text{ éléments propres de } \begin{cases} -v'' = \mu^2 v \text{ sur }]0,1[\\ v'(0) = v'(1) = 0 \end{cases} . \\ \beta_n = \sqrt{\omega^2 - \lambda_n^2} \text{ avec } \mathbf{Im}\,\beta_n \geq 0 \text{ et } \mathbf{Re}\,\beta_n \geq 0 \end{cases}$$

Dans le cas présent, on a $\lambda_n = n\pi$ et $\theta_n = a_n \cos n\pi y$ (a_n coefficient de normalisation de θ_n dans $L^2(]0,1[)$). D'après le corollaire 1.15, la famille $(\theta_n)_{n\geq 0}$ est une base orthonormale de $L^2(]0,1[)$.
Posons $c = \frac{\omega}{\pi}$. La famille des modes $(\Psi_n^{\pm})_{n\geq 0}$ se décompose en des modes :
– propagatifs à droite : $(\Psi_n^+)_{0\leq n<c}$ (oscillant),
– évanescents à droite : $(\Psi_n^+)_{n>c}$ (décroissance exponentielle quand $x \to +\infty$),
– propagatifs à gauche : $(\Psi_n^-)_{0\leq n<c}$ (oscillant),
– évanescents à gauche : $(\Psi_n^-)_{n>c}$ (décroissance exponentielle quand $x \to -\infty$).

Lorsque n coincide avec c, c'est-à-dire lorsque $\omega = n\pi$, on parle de modes à la coupure qui ont un comportement à l'infini constant, ou croissant linéairement. Nous excluons cette situation dans ce qui suit.

La connaissance de ces modes, permet de résoudre explicitement le problème de Helmholtz dans le guide semi-infini $\Omega_L^{np} =]L, +\infty[\times]0,1[$ "non perturbé", c'est-à-dire sans objet à l'intérieur :

$$\begin{cases} \Delta v + \omega^2 v = 0 & \text{dans } \Omega_L^{np} \\ v = w & \text{sur } \{(x,y) \in \partial\Omega_L^{np},\ x = L\} \\ \partial_n v = 0 & \text{sur } \{(x,y) \in \partial\Omega_L^{np},\ y = 0 \text{ ou } y = 1\} \\ \text{cond. d'ondes sortantes à l'infini} \end{cases} . \tag{1.70}$$

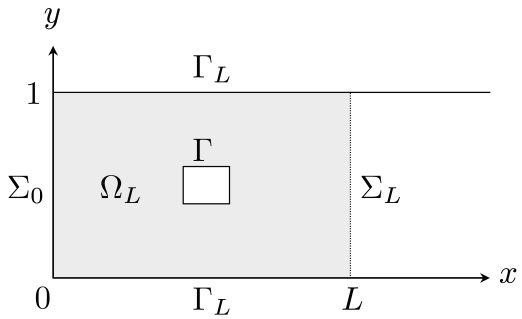
Comme on cherche une solution de régularité H^1 (sur toute partie de dimension finie du guide), sa trace sur la frontière doit appartenir à $H^{1/2}$ (sur toute partie de dimension finie de la frontière). Ainsi, on doit prescrire $w \in H^{1/2}(]0,1[)$. Or, comme $(\theta_n)_{n\geq 0}$ est une base orthonormale de $L^2(]0,1[)$, et sachant que $H^{1/2}(]0,1[) \subset L^2(]0,1[)$, on a $w = \sum_{n\geq 0}(w,\theta_n)\theta_n$ ($(\cdot,\cdot)$ produit scalaire de $L^2(]0,1[)$) conduisant à la solution sortante (dont on peut montrer l'unicité) de la forme :

$$v(x,y) = \sum_{n\geq 0}(w,\theta_n)\theta_n(y)e^{i\beta_n(x-L)}. \tag{1.71}$$

A partir de cette solution, on construit l'opérateur T_{DtN} dit de Dirichlet to Neumann :

$$\begin{aligned} T_{DtN} :\ H^{\frac{1}{2}}(]0,1[) &\longrightarrow H^{-\frac{1}{2}}(]0,1[) \\ w \qquad &\mapsto \frac{\partial v}{\partial x}_{|x=L} = \sum_{n\geq 0} i\beta_n(w,\theta_n)\theta_n. \end{aligned} \tag{1.72}$$

Notons que l'opérateur T_{DtN} est en fait indépendant de L. On peut de plus démontrer que l'opérateur T_{DtN} est continu de $H^{\frac{1}{2}}(]0,1[)$ – l'espace des traces – dans $H^{-\frac{1}{2}}(]0,1[) = \left(H^{\frac{1}{2}}(]0,1[)\right)'$ – l'espace des traces de la dérivée normale –. Cet opérateur permet de construire une condition aux limites transparente et de borner ainsi le domaine de calcul. On introduit maintenant, sur le domaine borné[6] Ω_L :



6. On a décomposé le guide semi-infini "perturbé" Ω_∞ en $\overline{\Omega}_\infty = \overline{\Omega}_L \cup \overline{\Omega}_L^{np}$, avec $\Omega_L \cap \Omega_L^{np} = \emptyset$.

le problème de Helmholtz :

$$\begin{cases} \Delta u + \omega^2 u = 0 \text{ dans } \Omega_L \\ \partial_n u = g \qquad \text{sur } \Sigma_0 \\ \partial_n u = 0 \qquad \text{sur } \Gamma \cup \Gamma_L \\ \partial_n u = T_{DtN} u \quad \text{sur } \Sigma_L \end{cases} \tag{1.73}$$

dont la formulation variationnelle dans $H^1(\Omega_L)$ est :

$$\begin{cases} \text{trouver } u \in H^1(\Omega_L) \text{ tel que} \\ \displaystyle\int_{\Omega_L} \nabla u \cdot \overline{\nabla v}\, d\Omega - \omega^2 \int_{\Omega_L} u\overline{v}\, d\Omega \\ \qquad\qquad - \displaystyle\int_{\Sigma_L} (T_{DtN} u)\overline{v}\, d\Gamma = \int_{\Sigma_0} g\overline{v}\, d\Gamma, \quad \forall v \in H^1(\Omega_L). \end{cases} \tag{1.74}$$

En décomposant l'opérateur T_{DtN} suivant les modes propagatifs et les modes évanescents :

$$T_{DtN} u = \sum_{n<c} i\beta_n (u, \theta_n)\theta_n - \sum_{n>c} |\beta_n| (u, \theta_n)\theta_n;$$

la forme sesquilinéaire $a(\cdot,\cdot)$ associée à la formulation variationnelle est de type coercif+compact [7]. En effet, on a $a(u,v) = a_{coer}(u,v) + a_{comp}(u,v)$, avec :

$$a_{coer}(u,v) = \left.\begin{array}{l} \displaystyle\int_{\Omega_L} \nabla u \cdot \overline{\nabla v}\, d\Omega + \int_{\Omega_L} u\overline{v}\, d\Omega \\ \displaystyle + \sum_{n>c} |\beta_n| \int_{\Sigma_L} (u, \theta_n)\, d\Gamma \int_{\Sigma_L} \overline{(v, \theta_n)}\, d\Gamma \end{array}\right\} \quad \text{(coercif)}$$

$$a_{comp}(u,v) = \left.\begin{array}{l} \displaystyle -(1+\omega^2) \int_{\Omega_L} u\overline{v}\, d\Omega \\ \displaystyle - \sum_{0 \le n < c} i\beta_n \int_{\Sigma_L} (u, \theta_n)\, d\Gamma \int_{\Sigma_L} \overline{(v, \theta_n)}\, d\Gamma \end{array}\right\} \quad \text{(compact).}$$

Pour le premier terme de $a_{comp}(\cdot,\cdot)$, on rappelle que, d'après le théorème de Rellich, l'injection de $H^1(\Omega_L)$ dans $L^2(\Omega_L)$ est compacte, puisque Ω_L est un domaine borné de $\mathbb{R}^2$. Par ailleurs, on peut établir (cf. [39]) que $H^s(\mathcal{O})$ s'injecte également de façon compacte dans $L^2(\mathcal{O})$, pour $s > 0$ et $\mathcal{O}$ un ouvert borné de $\mathbb{R}^n$ à frontière "suffisamment régulière" : ceci s'applique en particulier pour $\mathcal{O} = \Sigma_L$, ouvert borné de $\mathbb{R}$, et $s = 1/2$, justifiant la présence du second terme de $a_{comp}(\cdot,\cdot)$. Hormis un ensemble dénombrable de pulsations ω, le problème (1.74) admet une unique solution u. Celle-ci se prolonge à l'aide de la formule (1.71) en une fonction qui

7. Attention ! Nous sommes ici dans le cas où la perturbation compacte n'est pas hermitienne, et il convient donc d'appliquer la version généralisée de l'alternative de Fredholm (cf. remarque 1.21).

satisfait les équations (1.68). Réciproquement la restriction à Ω_L de toute solution des équations (1.68) est une solution du problème (1.74).

Nous sommes maintenant en mesure de discrétiser le problème (1.74). Nous utilisons comme précédemment une approximation par éléments finis, avec V_h un sous-espace de dimension finie N de $H^1(\Omega_L)$ (approximation interne une fois encore), et nous tronquons la série représentant l'opérateur T_{DtN} à un ordre p :

$$T_{DtN}^{(p)} u = \sum_{0 \leq n \leq p} i\beta_n (u, \theta_n)_1 \theta_n.$$

En pratique p est choisi plus grand que c. En effet, l'utilisation d'un ou deux modes modes évanescents est souvent suffisante. Le problème discrétisé est :

$$\begin{cases} \text{trouver } u_h \in V_h \text{ tel que} \\ \displaystyle\int_{\Omega_L} \nabla u_h \cdot \overline{\nabla v_h}\, d\Omega - \omega^2 \int_{\Omega_L} u_h \overline{v_h}\, d\Omega \\ \displaystyle - \sum_{0 \leq n \leq p} i\beta_n \int_{\Sigma_L} (u_h, \theta_n)\, d\Gamma \int_{\Sigma_L} \overline{(v_h, \theta_n)}\, d\Gamma = \int_{\Sigma_0} g \overline{v_h}\, d\Gamma, \quad \forall v_h \in V_h. \end{cases} \tag{1.75}$$

A l'exception de certaines pulsations (au plus dénombrables), ce problème est bien posé et est équivalent au système linéaire :

$$\begin{cases} \text{trouver } \overrightarrow{U} \in \mathbb{C}^N \text{ tel que} \\ \left(\mathbb{K} - \omega^2\, \mathbb{M} - \mathbb{T}^{(p)}\right) \overrightarrow{U} = \overrightarrow{G} \end{cases} \tag{1.76}$$

avec la matrice $\mathbb{T}^{(p)}$ de $\mathbb{C}^{N \times N}$ définie par :

$$\mathbb{T}_{IJ}^{(p)} = \sum_{0 \leq n \leq p} i\beta_n \int_{\Sigma_L} (w_J, \theta_n)\, d\Gamma \int_{\Sigma_L} \overline{(w_I, \theta_n)}\, d\Gamma, \quad 1 \leq I, J \leq N.$$

Dans la pratique, on construit la matrice $\mathbb{S}^{(p)}$ de $\mathbb{R}^{N \times (p+1)}$ ainsi que la matrice diagonale $\mathbb{D}^{(p)}$ de $\mathbb{C}^{(p+1)\times(p+1)}$, respectivement définies par :

$$\mathbb{S}_{Jm}^{(p)} = \int_{\Sigma_L} (w_J, \theta_m)\, d\Gamma \text{ et } \mathbb{D}_{mm}^{(p)} = i\beta_m$$

qui permettent de déterminer la matrice $\mathbb{T}^{(p)}$ à partir de la formule :

$$\mathbb{T}^{(p)} = \mathbb{S}^{(p)}\, \mathbb{D}^{(p)}\, (\mathbb{S}^{(p)})^t.$$

Le calcul des coefficients de la matrice $\mathbb{S}^{(p)}$ n'est pas standard puisqu'il mêle, à une base éléments finis définie polygone par polygone (typiquement des triangles), une base spectrale définie globalement sur un domaine (ici Σ_L). En particulier, il faut

prendre garde au fait que lorsque n augmente, on doit intégrer des fonctions spectrales de plus en plus oscillantes. Il existe deux moyens de calculer ces coefficients : soit de façon exacte puisqu'il existe des formules générales donnant la primitive du produit d'un polynôme par des cosinus, soit de façon approchée en utilisant une formule de quadrature numérique.
La première méthode, si elle est applicable dans l'exemple traité ici, n'est pas généralisable. Par contre, la présence de fonctions très oscillantes n'entache pas la précision.
La seconde méthode, qui se généralise sans difficulté, est confrontée à la difficulté d'intégrer correctement les termes oscillants.
Nous utilisons la technique de quadrature numérique en ramenant le calcul de l'intégrale sur le domaine Σ_L au calcul de l'intégrale sur chacun des segments qui le composent (pour un maillage donné). Si l'on veut conserver une précision suffisante, il convient de choisir une discrétisation du domaine Σ_L de telle sorte qu'il y ait au moins trois éléments par période des fonctions spectrales considérées. Il suffit d'appliquer ce critère à la dernière fonction spectrale considérée (i.e. θ_p), car c'est la plus oscillante. La procédure Matlab ci-dessous décrit la procédure permettant de calculer les matrices $\mathbb{S}^{(p)}$, $\mathbb{D}^{(p)}$ et $\mathbb{T}^{(p)}$:

```
function [S,D,T]=calcul_dtn(Corneu,Numtri,Bord,ref_bord)
global wf h nbm                          %pulsation, hauteur, nombre de modes
s35=sqrt(3./5);                           %formule de quadrature 1D, ordre 3
pts_quadS=0.5*[1-s35 1 1+s35];
os=1/18;pds_quadS=os*[5 8 5];
nbq=length(pds_quadS);
nt=size(Numtri,1);ns=size(Corneu,1);      %nombre de triangles, de noeuds
q=size(Numtri,2);                                  %3 en P1 ou 6 en P2
A=ones(nbm,1)*sqrt(2/h);                      %coefs. de normalisation L2
A(1)=sqrt(1/h);                             %des fonctions spectrales
ph=pi*pi/(h*h);w2=wf*wf;
D=sparse(nbm,nbm);
S=sparse(ns,nbm);
for t=1:nt,                               %boucle principale sur les éléments
    for a=1:3,                                        %boucle sur les aretes
        as=mod(a,3)+1;
        I=Numtri(t,a);J=Numtri(t,as);                  %numerotation globale
        if(q==6) IJ=Numtri(t,3+a);end,
        if(ismember(Bord(t,a),ref_bord)),               %arete sur ref_bord
            L=norm(Corneu(I,:)-Corneu(J,:));
            for k=1:nbq,                                %boucle quadrature 1D
                x=pts_quadS(k);
                xp=(1-x)*Corneu(I,2)+x*Corneu(J,2);
                c=L*pds_quadS(k);
                if(q==3) wI=1-x;wJ=x;                   %fonctions de base P1
                else wI=(1-2*x)*(1-x);                  %fonctions de base P2
                     wJ=x*(2*x-1);
                     wIJ=4*x*(1-x); end,
                for n=1:nbm;    %boucle sur les modes
                    cn=A(n)*cos((n-1)*pi*xp/h)*c;
                    S(I,n)=S(I,n)+cn*wI;
                    S(J,n)=S(J,n)+cn*wJ;
```

```
                    if(q==6)S(IJ,n)=S(IJ,n)+cn*wIJ; end,
                end,
            end,
        end,
    end,
end,
i=sqrt(-1);
for n=1:nbm,                                        %calculs de la matrice D
    m=n-1; f=w2-m*m*ph;
    if(f>=0) D(n,n)=i*sqrt(f);
    else D(n,n)=-sqrt(-f);end,
end,
T=S*D*S';                                           %calcul de la matrice T=S*D*S'
```

Nous avons mis en œuvre la résolution du problème (1.76) à l'aide de la procédure Matlab suivante :

```
clear all
global wf h nbm
wf=40;h=1;nbm=15;              %pulsation, hauteur, nombre de modes (>wh/pi)
q=50;
[S,T,BR,RT]=triangle_rectangle([0 0.5 0 h],q/2,q,1);          %maillage P1
[S,T,BR,RT]=def_trou(S,T,BR,RT,@objet);                  %def. de l'obstacle
[S,T,BR,RT]=maillageP2(S,T,BR,RT);                              %maillage P2
[T,S]=renume(T,S);                                          %renumerotation
[K,M]=calcul_EF_2D(S,T,RT);                                    %matrices EF
[Sp,Dp,Tp]=calcul_dtn(S,T,BR,[2]);                      %matrices DtN (cote 2)
A=K-wf*wf*M-Tp;                                          %matrice du probleme
[Mb]=calcul_EF_1D(S,T,BR,[4]);                           %masse sur le cote 4
Ng=noeud_bords(S,T,BR,[4]);                             %noeuds sur le cote 4
G=Mb*(Ng.*g(S));                                            %second membre
U=A\G;                                                  %resolution du systeme
if(size(T,2)==6) T=isop2(T);end,                        %dessin des isovaleurs
figure;trisurf(T,S(:,1),S(:,2),real(U));
shading interp;view(2);axis image;
```

Nous avons choisi le domaine $\Omega_{(L=0.5)} =]0, 0.5[\times]0, 1[$ dans lequel se situe un objet carré $]0.2, 0.3[\times]0.2, 0.3[$. Le maillage est réalisé à partir du maillage d'un rectangle que l'on "troue" à l'aide de la fonction Matlab **def_trou.m** ; cette dernière prenant la fonction Matlab **objet.m** comme critère pour définir le trou :

```
function c=objet(M)
a=0.2;b=0.3;c=0.2;d=0.3;e=0.01;
x=M(1);y=M(2);
c=(x>a-e) && (x<b+e) && (y>c-e) && (y<d+e);
```

Pour la pulsation considérée $\omega = 40$, il y a 13 modes propagatifs et nous avons choisi d'en prendre $p = 15$ pour le calcul de l'opérateur DtN. Le maillage possède 50 segments sur la frontière de couplage Σ_L ($L = 0.5$), ce qui est la limite de la méthode de quadrature numérique utilisée pour le calcul des coefficients de la matrice $\mathbb{S}^{(p)}$. On a choisi la fonction $g = 1$ comme donnée qui correspond précisément à une excitation par le mode plan ($n = 0$). Enfin, les calculs ont été réalisés avec une approximation éléments finis P^2. Nous donnons sur la figure 1.13(a) les

isovaleurs de la partie réelle de la solution obtenue. Afin de valider la résolution du problème, nous avons traité le même problème mais en prenant comme domaine (de calcul) de référence $\Omega_{(L=1)} =]0,1[\times]0,1[$ avec l'objet placé au même endroit. On déplace ainsi la frontière de couplage $\Sigma_{(L=1)}$; les résultats devraient rester inchangés, car la solution du problème continu (1.74) est indépendante de la frontière de couplage. C'est donc un excellent moyen de valider le calcul. Nous indiquons sur la figure 1.13(b) les isovaleurs de la partie réelle de la solution ainsi obtenue. On constate qu'il n'y a pas de différences perceptibles. Notons pour finir

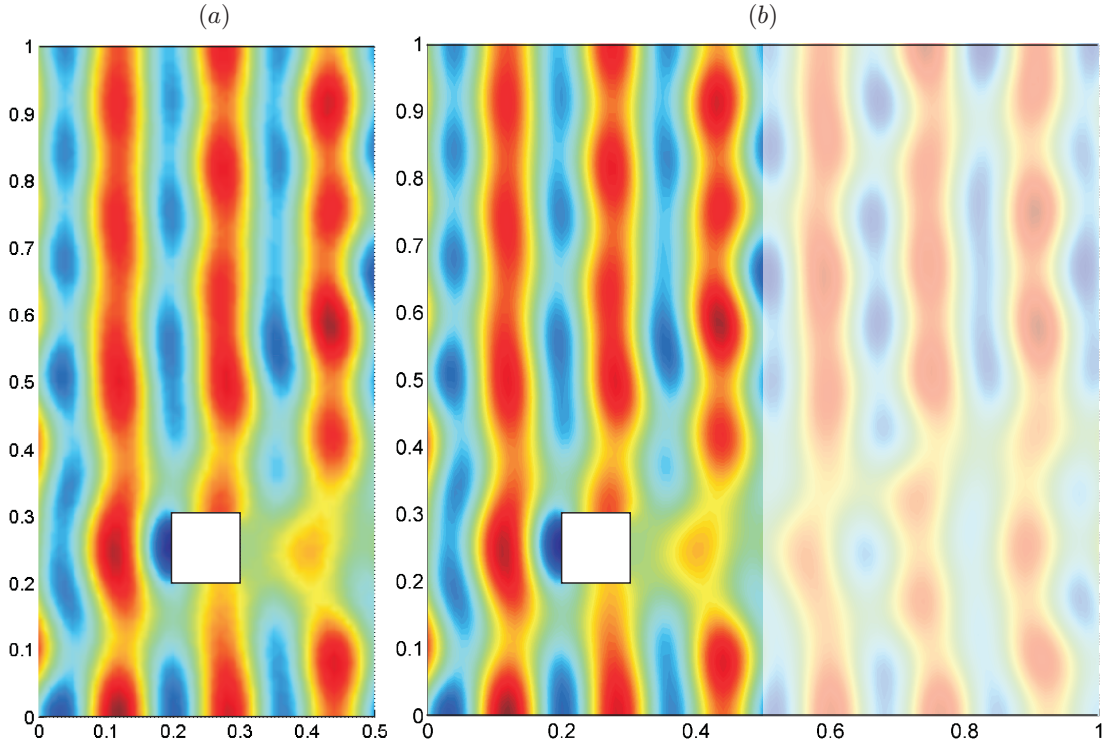


Figure 1.13. Solutions dans une boîte de calcul (a) de longueur 0.5, (b) de longueur 1

que l'étude mathématique de la convergence de l'approximation par éléments finis vers la solution exacte sort du cadre de cet ouvrage. Bien que le problème soit de type coercif+compact, cette étude est compliquée par le fait que l'opérateur DtN a été également discrétisé (car tronqué au rang p).

Nous avons illustré le calcul des éléments propres et la résolution de problèmes de type coercif+compact dans des cas très simples (opérateur de Laplace en dimension 2). Il est bien entendu que les techniques mises en œuvre se généralisent à la dimension 3, ainsi qu'à d'autres opérateurs (opérateur de Maxwell, opérateur de l'élasticité par exemple)...

2

Les éléments finis mixtes

La méthode des éléments finis se généralise à certains systèmes d'équations aux dérivées partielles faisant intervenir plusieurs inconnues de natures différentes. Dans ce chapitre, nous nous intéressons à la résolution de problèmes dits *mixtes* dont les inconnues sont deux fonctions représentant d'une part l'*état* du système considéré, d'autre part un *multiplicateur de Lagrange* associé à une *contrainte* que doit satisfaire l'état. Toute la question est d'exhiber des *éléments finis mixtes* adaptés à la discrétisation de tel ou tel problème mixte, c'est-à-dire des couples d'éléments finis usuels (cf. [15]) permettant de discrétiser le couple (*état*, *multiplicateur*).

En mécanique des fluides, l'un des exemples les plus connus est donné par les équations dites de Stokes qui résultent d'une linéarisation des équations de Navier–Stokes modélisant l'écoulement d'un fluide visqueux incompressible. Dans ce cas, l'état du système est décrit par le champ des vitesses (inconnue vectorielle), alors que le champ de pression (inconnue scalaire) joue le rôle d'un multiplicateur associé à la contrainte d'incompressibilité du fluide. L'étude théorique d'un tel problème et de son approximation numérique est plus délicate que pour les problèmes qui relèvent du théorème de Lax–Milgram (voir [15]). Les premiers travaux sur le sujet datent des années 1970. La littérature est aujourd'hui abondante (voir par exemple [10, 27, 43, 46]). Si le cas des équations de Stokes est maintenant largement débroussaillé, il subsiste de nombreuses situations où on ne sait pas justifier *a priori* l'utilisation de certains éléments finis mixtes : les résultats numériques se chargent de les valider (ou invalider !) *a posteriori*.

Le chapitre est divisé en quatre parties. Dans la première section, on introduit à partir des équations de Stokes un problème variationnel abstrait. Pour étudier l'existence et l'unicité de sa solution, on se sert, en plus de la notion de *coercivité*, d'une condition dite "inf–sup". Cette dernière traduit la *compatibilité* de la contrainte avec les espaces fonctionnels utilisés. Nous présentons quelques exemples d'applications. Nous montrons enfin que notre problème abstrait peut être inter-

prété comme un problème d'optimisation sous contrainte dont l'étude entre dans le cadre des méthodes de dualité (ceci justifie l'utilisation du terme "multiplicateur").

La seconde section est consacrée à l'approximation numérique du problème abstrait. Nous verrons que la convergence d'une méthode d'éléments finis mixtes est soumise à une "condition inf–sup discrète". En quelque sorte, celle-ci traduit le fait que les équations qui expriment la contrainte après discrétisation restent compatibles avec les espaces de discrétisation, et ce, uniformément par rapport à la finesse du maillage. Et la vérification d'une telle condition demande... un certain travail ! Enfin, nous examinons la résolution pratique des problèmes approchés, à savoir comment résoudre les systèmes linéaires qui en sont issus. Dans cette section, nous nous intéressons principalement au cas des équations de Stokes.

La troisième section est destinée à une présentation détaillée d'un autre exemple d'application, issu des équations de l'électromagnétisme. Il s'agit d'une approximation quasi-statique des équations de Maxwell, valable lorsque les variations du déplacement électrique sont lentes. Sur les plans théorique et numérique, ce cas soulève des difficultés de nature différente du cas des équations de Stokes.

Enfin, la dernière section illustre ce chapitre par quelques résultats numériques, d'une part pour les équations de Stokes, et d'autre part pour l'approximation quasi-statique des équations de Maxwell.

2.1 La notion de problème mixte

2.1.1 Des équations de Stokes à un problème abstrait

Les équations de Navier–Stokes modélisent l'écoulement d'un fluide *visqueux incompressible* caractérisé par une loi de comportement linéaire (on parle généralement de fluide "newtonien") : le tenseur des contraintes est une fonction linéaire du couple $(\mathbf{u}, p)$ définissant les champs de vitesse et de pression (les grandeurs vectorielles sont repérées par des caractères gras). Ces équations traduisent la conservation de la quantité de mouvement et de la masse. Elles peuvent s'écrire sous la forme adimensionnelle suivante :

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} + \dfrac{1}{2}\nabla|\mathbf{u}|^2 + \mathbf{rot}\,\mathbf{u} \wedge \mathbf{u} - \nu \Delta\,\mathbf{u} + \nabla p = \mathbf{f}, \\ \operatorname{div} \mathbf{u} = 0, \end{cases}$$

où ν est appelé la *viscosité cinématique* du fluide, et $\mathbf{f}$ représente la densité des efforts extérieurs (Δ désigne l'opérateur laplacien appliqué à chacune des composantes de $\mathbf{u}$).

Les équations de Stokes s'obtiennent en négligeant d'une part les termes non linéaires (qui, lorsque la vitesse du fluide est faible, sont d'un ordre inférieur aux autres termes) et en supposant d'autre part l'écoulement stationnaire (autrement dit $\partial_t \mathbf{u} = 0$). Elles s'écrivent donc

$$\begin{cases} -\nu \Delta \mathbf{u} + \nabla p = \mathbf{f}, & (2.1) \\ \operatorname{div} \mathbf{u} = 0. & (2.2) \end{cases}$$

Nous allons nous intéresser à la résolution de ces équations dans un ouvert borné $\Omega \subset \mathbb{R}^n$ (où $n = 2$ ou 3 en pratique). Comme pour les problèmes elliptiques, l'unicité de la solution $(\mathbf{u}, p)$ ne peut être assurée que si on adjoint à ces équations des conditions aux limites sur la frontière $\partial\Omega$. La présence de l'opérateur Δ laisse penser qu'il faut imposer à chacune des composantes de $\mathbf{u}$ des conditions analogues au cas scalaire. Nous considérons ici le cas où la frontière de Ω est constituée d'une paroi parfaitement rigide (frontière que l'on supposera "suffisamment régulière", cf. [15]), représentée par une condition de Dirichlet

$$\mathbf{u} = \mathbf{0} \text{ sur } \partial\Omega. \tag{2.3}$$

Nous verrons en revanche qu'il est inutile d'imposer une quelconque condition aux limites sur p (qui est donc définie à une constante additive près).

Notre but étant de construire des méthodes d'éléments finis adaptées au traitement de ce problème, il faut commencer par en donner une formulation variationnelle. Supposons donc qu'on connaisse une solution "classique" de (2.1)–(2.3), c'est-à-dire telle que $\mathbf{u} \in (\mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega}))^n$ et $p \in \mathcal{C}^1(\Omega)$. En multipliant respectivement (2.1) et (2.2) par un champ de vitesse test $\mathbf{v} \in H_0^1(\Omega)^n$ et un champ de pression test $q \in L^2(\Omega)$, on obtient, à l'aide de la formule de Green :

$$\begin{aligned} &\nu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega - \int_\Omega p \operatorname{div} \mathbf{v} \, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\Omega, \\ &\int_\Omega q \operatorname{div} \mathbf{u} \, d\Omega = 0, \end{aligned}$$

où on a noté $\nabla \mathbf{u} : \nabla \mathbf{v} = \sum_{i,j=1,n} \partial_j u_i \, \partial_j v_i$. Il s'agit là d'une formulation faible des équations de Stokes, au sens où elle permet de considérer des champs $(\mathbf{u}, p) \in H_0^1(\Omega)^n \times L^2(\Omega)$ moins réguliers que pour une solution "classique". Notons que nous n'avons pas appliqué de formule de Green pour obtenir la seconde équation alors que nous l'avons utilisée pour faire apparaître le terme $\int_\Omega p \operatorname{div} \mathbf{v} \, d\Omega$ dans la première : ce choix permet précisément de considérer des champs de pression qui appartiennent seulement à $L^2(\Omega)$. En résumé, si l'on pose

$$a(\mathbf{u},\mathbf{v}) = \nu \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega, \quad c(\mathbf{v},q) = -\int_\Omega q \,\operatorname{div}\mathbf{v}\, d\Omega, \tag{2.4}$$

$$\ell_V(\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega, \tag{2.5}$$

les équations de Stokes prennent la forme suivante :

$$\begin{cases} \text{trouver } (\mathbf{u},p) \in H_0^1(\Omega)^n \times L^2(\Omega) \text{ tel que} \\ a(\mathbf{u},\mathbf{v}) + c(\mathbf{v},p) = \ell_V(\mathbf{v}) \quad \forall \mathbf{v} \in H_0^1(\Omega)^n, \\ c(\mathbf{u},q) = 0 \quad \forall q \in L^2(\Omega). \end{cases}$$

Remarquons que la condition aux limites (2.3) apparaît comme une condition *essentielle* : elle est incluse dans la définition de $H_0^1(\Omega)^n$. Par ailleurs, comme dans la formulation forte, le champ p ne peut être défini qu'à une constante additive près (puisque pour tout $\mathbf{v} \in H_0^1(\Omega)^n$, on a $\int_\Omega \operatorname{div}\mathbf{v}\, d\Omega = 0$ par la formule de Green). Une façon d'éliminer cette indétermination consiste à remplacer $L^2(\Omega)$ par son sous-espace constitué des fonctions à moyenne nulle :

$$L_0^2(\Omega) = \left\{ q \in L^2(\Omega) \text{ tel que } \int_\Omega q\, d\Omega = 0 \right\}. \tag{2.6}$$

Nous verrons plus loin que, dans ce cas, le problème devient bien posé.

A un petit détail près, le problème ci-dessus nous fournit la forme générale du problème abstrait que nous allons étudier dans ce chapitre, dont nous verrons plusieurs exemples d'applications, en plus des équations de Stokes (voir les paragraphes 2.1.3 et 2.3). Ce problème abstrait peut être formulé de la façon suivante. Etant donnés
– deux espaces de Hilbert V et M,
– deux formes bilinéaires $a(\cdot\,,\cdot) : V \times V \to \mathbb{R}$ et $c(\cdot\,,\cdot) : V \times M \to \mathbb{R}$,
– deux formes linéaires $\ell_V(\cdot) : V \to \mathbb{R}$ et $\ell_M(\cdot) : M \to \mathbb{R}$,
il s'agit de résoudre

$$\begin{cases} \text{trouver } (u,\lambda) \in V \times M \text{ tel que} \\ a(u,v) + c(v,\lambda) = \ell_V(v) \quad \forall v \in V, & (2.7) \\ c(u,\mu) = \ell_M(\mu) \quad \forall \mu \in M. & (2.8) \end{cases}$$

Notons que dans le cas particulier où $c(\cdot\,,\cdot) = 0$ (et $\ell_M(\cdot) = 0$), l'équation variationnelle (2.7) se ramène au problème modèle introduit dans [15] pour l'étude des équations elliptiques. Dans le cas général, la présence de cette forme bilinéaire traduit la prise en compte d'une *contrainte* formulée elle aussi de façon variationnelle par l'équation (2.8).

2.1.2 Existence et unicité de la solution

Sous quelles conditions (suffisantes) le problème variationnel mixte (2.7)–(2.8) est-il bien posé ? Pour y répondre, nous allons dans un premier temps traiter le cas d'une contrainte homogène (soit $\ell_M(\cdot) = 0$, comme pour les équations de Stokes ; voir la remarque 2.5 pour le cas non homogène). Notre problème mixte abstrait s'écrit donc

$$\begin{cases} \text{trouver } (u, \lambda) \in V \times M \text{ tel que} \\ a(u, v) + c(v, \lambda) = \ell_V(v) \quad \forall v \in V, & (2.9) \\ c(u, \mu) = 0 \quad \forall \mu \in M. & (2.10) \end{cases}$$

Commençons par donner quelques précisions sur ce problème. Les espaces de Hilbert V et M sont équipés respectivement de produits scalaires notés $(\cdot\,,\cdot)_V$ et $(\cdot\,,\cdot)_M$ (les normes associées étant désignées par $\|\cdot\|_V$ et $\|\cdot\|_M$). Le premier représente l'espace des états *non contraints* du système considéré, et nous verrons que les éléments du second jouent le rôle de *multiplicateurs de Lagrange* relatifs à la contrainte (2.10).

Dans toute la suite, les deux formes bilinéaires

$$a(\cdot\,,\cdot) \ : V \times V \to \mathbb{R} \ \text{ et } \ c(\cdot\,,\cdot) \ : V \times M \to \mathbb{R}$$

sont supposées continues, ce qui signifie qu'on peut trouver deux constantes positives m_a et m_c telles que

$$|a(u, v)| \leq m_a \|u\|_V \|v\|_V \ \text{ et } \ |c(v, \mu)| \leq m_c \|v\|_V \|\mu\|_M, \tag{2.11}$$

pour tous $u, v \in V$ et $\mu \in M$. Par ailleurs la forme linéaire $\ell_V(\cdot)$ est également supposée continue, soit

$$|\ell_V(v)| \leq m_\ell \|v\|_V \quad \forall v \in V, \tag{2.12}$$

pour une constante $m_\ell > 0$. Grâce au théorème de Riesz, ces hypothèses nous montrent qu'il existe deux opérateurs linéaires et continus

$$A \ : V \to V \ \text{ et } \ C \ : V \to M,$$

tels que, pour tous $u, v \in V$ et $\mu \in M$,

$$(Au, v)_V = a(u, v) \ \text{ et } \ (Cv, \mu)_M = c(v, \mu). \tag{2.13}$$

De plus, à la forme linéaire $\ell_V(\cdot)$ correspond un élément $f \in V$ tel que

$$\ell_V(v) = (f, v)_V \quad \forall v \in V. \tag{2.14}$$

Ainsi, les équations variationnelles (2.9) et (2.10) reviennent à dire que

$$Au + C^T\lambda = f, \tag{2.15}$$

$$Cu = 0, \tag{2.16}$$

où $C^T : M \to V$ désigne le transposé de C, défini, pour tous $v \in V$ et $\mu \in M$, par

$$(C^T\mu, v)_V = (\mu, Cv)_M = c(v, \mu). \tag{2.17}$$

On peut écrire ce système sous la forme d'un système pseudo-matriciel

$$\begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{2.18}$$

Le cas de la dimension finie

Supposons pour fixer les idées que A et C soient des matrices (respectivement $n \times n$ et $m \times n$ si $V = \mathbb{R}^n$ et $M = \mathbb{R}^m$ avec $n > m$). On peut alors facilement donner une condition d'inversibilité de ce système linéaire. On sait en effet que

$$\operatorname{Im} C^T = (\operatorname{Ker} C)^\perp. \tag{2.19}$$

Notons P la projection orthogonale de $\mathbb{R}^n$ sur $\operatorname{Ker} C$. En appliquant P et $I - P$ (c'est-à-dire la projection sur $(\operatorname{Ker} C)^\perp$) à l'équation (2.15), on obtient d'une part

$$PAu = Pf \quad \text{avec} \quad u \in \operatorname{Ker} C, \tag{2.20}$$

et d'autre part

$$C^T\lambda = f - Au \qquad (= (I - P)(f - Au)). \tag{2.21}$$

Dans (2.20), on peut remplacer PAu par $PAPu$, puisque $u \in \operatorname{Ker} C$. Pour que cette équation détermine u de façon unique, il faut et il suffit que PAP soit inversible sur $\operatorname{Ker} C$. Dans ce cas, l'équation (2.21) permet d'en déduire λ sous réserve que C^T soit inversible de $M = \mathbb{R}^m$ dans $(\operatorname{Ker} C)^\perp$, ce qui signifie que C est de rang plein (soit $\dim(\operatorname{Im} C^T) = m$). En résumé, on sait résoudre (2.18) dès que $PAP : \operatorname{Ker} C \to \operatorname{Ker} C$ et $C^T : M \to \operatorname{Im} C^T$ sont inversibles.

Le cas général

Lorsque A et C ne sont plus des matrices mais les opérateurs définis en (2.13), il nous faut trouver des conditions suffisantes pour que ces deux propriétés soient encore vérifiées. Dans le cas particulier d'un problème sans contrainte (c'est-à-dire lorsque $c(\cdot\,,\cdot) = 0$), on connaît déjà une condition assurant l'inversibilité de A (et la continuité de son inverse) : le théorème de Lax-Milgram [15] nous assure du caractère bien posé de l'équation variationnelle (2.9). En présence de la contrainte (2.10), la généralisation de ce théorème prend la forme suivante.

Théorème 2.1 *En plus des hypothèses de continuité* (2.11)–(2.12)*, on suppose qu'il existe deux constantes* $\alpha > 0$ *et* $\gamma > 0$ *telles que*

$$\inf_{v \in \operatorname{Ker} C} \frac{a(v,v)}{\|v\|_V^2} \geq \alpha, \tag{2.22}$$

$$\inf_{\mu \in M} \sup_{v \in V} \frac{c(v,\mu)}{\|v\|_V \|\mu\|_M} \geq \gamma. \tag{2.23}$$

Alors le problème (2.9)–(2.10) *est bien posé. Autrement dit, il admet une unique solution qui dépend continûment de la donnée :* $\|u\|_V^2 + \|\lambda\|_M^2 \leq K \|f\|_V^2$ *pour un certain* $K > 0$.

La condition (2.22) signifie simplement que $a(\cdot\,,\cdot)$ est coercive sur $\operatorname{Ker} C$, soit

$$a(v,v) \geq \alpha \, \|v\|_V^2 \quad \forall v \in \operatorname{Ker} C,$$

mais pas nécessairement sur tout l'espace V. Notons que $\operatorname{Ker} C$ constitue un sous-espace *fermé* de V puisque $c(\cdot\,,\cdot)$ est supposée continue. Si on note encore P la projection orthogonale sur $\operatorname{Ker} C$, cette condition assure donc l'inversibilité de PAP et la continuité de son inverse, et par conséquent, l'existence et l'unicité de l'état u et sa continuité par rapport à la donnée.

La condition (2.23) est appelée *condition inf–sup*, ou parfois condition de Babuška ou de Brezzi, qui l'ont introduite indépendamment l'un de l'autre dans les années 1970. Elle nous fournit l'existence et l'unicité du multiplicateur et sa continuité par rapport à la donnée. En effet, grâce au lemme suivant, la démonstration du théorème que nous avons donnée plus haut dans le cas de la dimension finie devient parfaitement valable.

Lemme 2.2 *La condition inf–sup* (2.23) *est équivalente à chacune des deux assertions suivantes :*
(i) $\operatorname{Im} C^T = (\operatorname{Ker} C)^\perp$*, et* C^T *est bijectif de* M *dans* $\operatorname{Im} C^T$*, d'inverse continu ; plus précisément,*

$$\|\mu\|_M \leq \gamma^{-1} \left\| C^T \mu \right\|_V \quad \forall \mu \in M, \tag{2.24}$$

où γ *est la constante figurant dans* (2.23).
(ii) C *est bijectif de* $(\operatorname{Ker} C)^\perp$ *dans* M*, d'inverse continu , et* $\|v\|_V \leq \gamma^{-1} \|Cv\|_M$ *pour tout* $v \in (\operatorname{Ker} C)^\perp$.

Démonstration : De façon générale, pour tout opérateur linéaire et continu C entre deux espaces de Hilbert, on a les relations d'orthogonalité

$$\overline{\operatorname{Im} C^T} = (\operatorname{Ker} C)^\perp \quad \text{et} \quad (\operatorname{Im} C)^\perp = \operatorname{Ker} C^T.$$

La condition inf–sup (2.23) nous fournit exactement les informations manquantes. En effet, celle-ci peut aussi s'écrire

$$\sup_{v\in V}\frac{c(v,\mu)}{\|v\|_V}\geq\gamma\|\mu\|_M\quad\forall\mu\in M,$$

où le terme de gauche n'est autre que $\left\|C^T\mu\right\|_V$ d'après la définition (2.17) de C^T. La condition inf–sup est donc équivalente à (2.24). Cette dernière montre d'une part que C^T est injectif (immédiat), donc que

$$\overline{\operatorname{Im} C}=M.$$

D'autre part, elle montre que $\operatorname{Im} C^T$ est fermée dans V : si $C^T\mu_n$ est de Cauchy dans V, alors μ_n est de Cauchy dans M par (2.24), donc convergente, soit $C^T\mu_n\to C^T\mu$. Ainsi

$$\operatorname{Im} C^T=(\operatorname{Ker} C)^\perp. \tag{2.25}$$

Le point (i) est donc démontré (la relation (2.24) traduit la continuité de l'inverse de C^T).

Pour obtenir sa forme transposée, soit (ii), il reste à remarquer que si $v\in(\operatorname{Ker} C)^\perp$,

$$\|v\|_V=\sup_{v'\in(\operatorname{Ker} C)^\perp}\frac{(v,v')_V}{\|v'\|_V}=\sup_{\mu\in M}\frac{(v,C^T\mu)_V}{\|C^T\mu\|_V}$$

d'après (2.25). En utilisant (2.24), il vient

$$\|v\|_V\leq\gamma^{-1}\sup_{\mu\in M}\frac{(Cv,\mu)_M}{\|\mu\|_M}=\gamma^{-1}\|Cv\|_M.$$

On en déduit comme précédemment que $\operatorname{Im} C$ est fermée, et par conséquent $\operatorname{Im} C=M$. ∎

Remarque 2.3 *Cette démonstration montre clairement le rôle de la condition inf–sup lorsque V et M sont de dimensions* infinies. *Dans le cas de la dimension finie, la relation* (2.19) *est toujours vérifiée, et il suffit donc de vérifier que C^T est injectif pour assurer l'existence et l'unicité du multiplicateur λ, autrement dit que C est surjectif. Dans le cas de la dimension infinie, la condition inf–sup assure à la fois l'injectivité de C^T et la relation* (2.19) *qui conduit à sa surjectivité (sur* $(\operatorname{Ker} C)^\perp$).

Par la suite, nous allons voir que les difficultés dans l'obtention des conditions (2.22) et (2.23) sont de divers ordres. Notamment, certains problèmes requièrent des résultats "pointus" pour l'établissement de la condition inf–sup (cas du problème de Stokes), alors que pour d'autres, c'est l'obtention de la coercivité sur le noyau qui est délicate (cas du problème électromagnétique quasi-statique).

En pratique, un critère utile pour vérifier la condition inf–sup est donné par la

Proposition 2.4 *La condition inf–sup* (2.23) *est satisfaite si, et seulement si, pour tout $\mu\in M$, il existe $v_\mu\in V$ tel que*

$$c(v_\mu,\mu)=\|\mu\|_M^2\quad\text{et}\quad\|v_\mu\|_V\leq\gamma^{-1}\|\mu\|_M. \tag{2.26}$$

Démonstration : La condition inf–sup est clairement une conséquence de (2.26). Réciproquement, en utilisant le lemme 2.2, si C^T vérifie (2.24), la fonction

$$v_\mu = \frac{\|\mu\|_M^2}{\|C^T\mu\|_V^2} C^T\mu \tag{2.27}$$

est en fait l'unique fonction de $(\operatorname{Ker} C)^\perp$ qui satisfait (2.26). De façon générale, v_μ n'est pas défini de façon unique dans V. ∎

Remarque 2.5 *Le théorème* 2.1 *reste valable pour des contraintes non homogènes, autrement dit lorsqu'on remplace* (2.10) *par* (2.8) *où* $\ell_M(\cdot)$ *est supposée continue sur* M*. Par le théorème de Riesz, il existe un unique* $g \in M$ *tel que*

$$(g,\mu)_M = \ell_M(\mu) \quad \forall \mu \in M. \tag{2.28}$$

Ainsi, il suffit simplement de remplacer le second membre de (2.18) *par*

$$\begin{pmatrix} f \\ g \end{pmatrix}.$$

Dans ce cas, on doit donc résoudre dans un premier temps

$$PAu = Pf \quad \text{avec} \quad Cu = g,$$

au lieu de (2.20)*, et le multiplicateur* λ *s'en déduira encore en résolvant* (2.21)*. Mais d'après le lemme* 2.2*, on sait qu'il existe* $u_0 \in (\operatorname{Ker} C)^\perp$ *tel que* $Cu_0 = g$ *(avec* $\|u_0\|_V \le \gamma^{-1}\|g\|_M$*). Ainsi, en posant* $\tilde{u} = u - u_0$*, on est conduit à résoudre*

$$PA\tilde{u} = P(f - Au_0) \quad \text{avec} \quad \tilde{u} \in \operatorname{Ker} C,$$

ce qui nous ramène exactement dans le cadre du théorème 2.1*.*

Application aux équations de Stokes

Nous avons écrit précédemment les équations de Stokes sous la forme (2.9)–(2.10) en choisissant comme espaces de travail $V = H_0^1(\Omega)^n$ et $M = L_0^2(\Omega)$ (voir (2.6)), les formes $a(\cdot\,,\cdot)$, $c(\cdot\,,\cdot)$ et $\ell_V(\cdot)$ étant données par (2.4)–(2.5). Introduisons

$$\mathcal{L}^2(\Omega) = \left\{\tau = (\tau_{ij})_{i,j=1,n}\,;\ \tau_{ij} \in L^2(\Omega),\ 1 \le i,j \le n\right\},\ \text{de produit scalaire}$$
$$(\sigma,\tau)_{\mathcal{L}^2(\Omega)} = \int_\Omega \sigma : \tau \, d\Omega = \int_\Omega \sum_{i,j=1,n} \sigma_{ij}\,\tau_{ij}\, d\Omega.$$

D'après l'inégalité de Poincaré (cf. [15]), nous savons que nous pouvons munir $H_0^1(\Omega)^n$ du produit scalaire $(\mathbf{u},\mathbf{v}) \mapsto (\nabla\mathbf{u},\nabla\mathbf{v})_{\mathcal{L}^2(\Omega)}$, et nous choisissons pour $L_0^2(\Omega)$ le produit scalaire usuel $(p,q) \mapsto (p,q)_{L^2(\Omega)}$. Ainsi, l'opérateur C associé

(par (2.13)) à la forme bilinéaire $c(\cdot\,,\cdot)$ définie en (2.4) n'est autre que $C = -\operatorname{div}$ (agissant de $H_0^1(\Omega)^n$ dans $L_0^2(\Omega)$). En effet, si

$$(C\mathbf{v}, q)_{L^2(\Omega)} = -\int_\Omega q \operatorname{div}\mathbf{v}\, d\Omega \quad \forall q \in L_0^2(\Omega),$$

l'égalité reste vraie pour tout $q \in L^2(\Omega)$; il suffit de remarquer que

$$q - \left(\int_\Omega q\, d\Omega\right) \in L_0^2(\Omega) \quad \text{et} \quad \operatorname{div}\mathbf{v} \in L_0^2(\Omega).$$

On a donc bien $C\mathbf{v} = -\operatorname{div}\mathbf{v}$.

La forme bilinéaire

$$a(\mathbf{u}, \mathbf{v}) = \nu \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega$$

est évidemment continue et coercive sur $H_0^1(\Omega)^n$, donc *a fortiori* sur

$$\operatorname{Ker} C = \left\{\mathbf{v} \in H_0^1(\Omega)^n \mid \operatorname{div}\mathbf{v} = 0\right\}.$$

La vérification de la condition inf–sup (2.23) repose sur un résultat technique dont la démonstration sort du cadre de ce cours (nous renvoyons par exemple à Girault et Raviart [27] pour les détails, voir aussi [8] pour une démonstration plus générale). Ce résultat peut prendre la forme de la caractérisation suivante de l'orthogonal de $\operatorname{Ker} C$ (pour le produit scalaire $(\mathbf{u}, \mathbf{v}) \mapsto (\nabla\mathbf{u}, \nabla\mathbf{v})_{\mathcal{L}^2(\Omega)}$) : *une fonction* $\mathbf{v} \in H_0^1(\Omega)^n$ *appartient à* $(\operatorname{Ker} C)^\perp$ *si, et seulement si, il existe* $q \in L_0^2(\Omega)$ (défini de façon unique si Ω est connexe) *tel que*

$$(\nabla\mathbf{v}, \nabla\mathbf{w})_{\mathcal{L}^2(\Omega)} = -(q, \operatorname{div}\mathbf{w})_{L^2(\Omega)} \quad \forall \mathbf{w} \in H_0^1(\Omega)^n. \tag{2.29}$$

Cette équation variationnelle correspond à une formulation faible du problème

$$\begin{cases} -\Delta\,\mathbf{v} = -\nabla q \ \text{ dans } \Omega, \\ \mathbf{v} = 0 \ \text{ sur } \partial\Omega. \end{cases}$$

Il est clair qu'une fonction qui satisfait (2.29) vérifie en particulier

$$(\nabla\mathbf{v}, \nabla\mathbf{w})_{\mathcal{L}^2(\Omega)} = 0 \quad \forall \mathbf{w} \in \operatorname{Ker} C,$$

et par conséquent, $\mathbf{v} \in (\operatorname{Ker} C)^\perp$. La démonstration de la réciproque consiste à caractériser l'image de $L_0^2(\Omega)$ par l'opérateur ∇, qui apparaît comme un sous-espace de $H^{-1}(\Omega)$, c'est-à-dire du dual de $H_0^1(\Omega)$ lorsqu'on identifie $L^2(\Omega)$ à son dual : $H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega)$.

En fait, d'après le lemme 2.2, ce résultat n'est qu'une autre façon d'exprimer la condition inf–sup (2.23) qui s'écrit ici

$$\inf_{q\in L^2_0(\Omega)} \sup_{\mathbf{v}\in H^1_0(\Omega)^n} \frac{(q, \operatorname{div}\mathbf{v})_{L^2(\Omega)}}{\|\nabla\mathbf{v}\|_{\mathcal{L}^2(\Omega)}\|q\|_{L^2(\Omega)}} \geq \gamma.$$

En effet, la solution $\mathbf{v}$ de (2.29) est par définition $C^T q$: le résultat que nous venons d'admettre signifie simplement que C^T est bijectif de $L^2_0(\Omega)$ dans $(\operatorname{Ker} C)^\perp$, d'inverse continu.

2.1.3 Quelques exemples d'applications

Nous décrivons ici quelques exemples simples de problèmes mixtes qui entrent dans le cadre du problème abstrait (2.7)–(2.8). Un exemple plus élaboré issu des équations de l'électromagnétisme fait l'objet de la section 2.3.

Le problème de Dirichlet pour le Laplacien

Soit Ω un ouvert borné de $\mathbb{R}^n$ et $f \in L^2(\Omega)$. Le problème

$$\begin{cases} -\Delta\varphi = f \text{ dans } \Omega, \\ \varphi = 0 \qquad \text{sur } \partial\Omega, \end{cases}$$

a été étudié en détail dans [15]. Nous savons en particulier que sa formulation faible, qui consiste à rechercher $\varphi \in H^1_0(\Omega)$ tel que

$$\int_\Omega \nabla\varphi\cdot\nabla\psi\, d\Omega = \int_\Omega f\,\psi\, d\Omega \quad \forall\psi \in H^1_0(\Omega), \tag{2.30}$$

définit un problème bien posé (grâce à l'inégalité de Poincaré). Nous en donnons ici une formulation mixte qui consiste simplement à introduire la variable $\mathbf{u} = \nabla\varphi$. Il s'agit donc de résoudre

$$\begin{cases} \mathbf{u} - \nabla\varphi = 0 \text{ dans } \Omega, \\ \varphi = 0 \qquad\quad \text{sur } \partial\Omega, \\ -\operatorname{div}\mathbf{u} = f \text{ dans } \Omega. \end{cases}$$

Si $(\mathbf{u}, \varphi)$ désigne une solution "classique" de ces équations, c'est-à-dire telle que $\mathbf{u} \in \mathcal{C}^1(\Omega)^n$ et $\varphi \in \mathcal{C}^1(\Omega)\cap\mathcal{C}^0(\overline{\Omega})$, on peut multiplier la première équation par un champ test $\mathbf{v} \in L^2(\Omega)^n$ et la dernière, par une fonction test $\psi \in H^1_0(\Omega)$. Il vient

$$\begin{aligned} \int_\Omega \mathbf{u}\cdot\mathbf{v}\, d\Omega - \int_\Omega \mathbf{v}\cdot\nabla\varphi\, d\Omega = 0, \\ \int_\Omega \mathbf{u}\cdot\nabla\psi\, d\Omega = \int_\Omega f\,\psi\, d\Omega. \end{aligned}$$

Nous n'avons ici utilisé la formule de Green que pour la seconde équation, ce qui permet de donner un sens au problème lorsque $\mathbf{u} \in L^2(\Omega)^n$ et $\varphi \in H^1_0(\Omega)$. Cette formulation mixte du problème de Dirichlet ne présente pas un grand intérêt : elle est parfaitement équivalente à la formulation (2.30). On peut par contre en donner une variante en utilisant cette fois la formule de Green pour la première équation :

$$\int_\Omega \mathbf{u} \cdot \mathbf{v}\, d\Omega + \int_\Omega \varphi \ \mathrm{div}\, \mathbf{v}\, d\Omega = 0, \tag{2.31}$$

$$\int_\Omega \psi \ \mathrm{div}\, \mathbf{u}\, d\Omega = -\int_\Omega f\, \psi\, d\Omega. \tag{2.32}$$

Ceci permet de diminuer la régularité présupposée de la fonction φ : on peut se contenter de la rechercher dans $L^2(\Omega)$. En contrepartie, il faut clairement augmenter celle de $\mathbf{u}$. Compte tenu des espaces fonctionnels que nous connaissons déjà, il paraît tentant de choisir $\mathbf{u}$ et $\mathbf{v}$ dans $H^1(\Omega)^n$. Malheureusement, le problème est dans ce cas mal posé : les équations ci-dessus ne font pas intervenir le gradient des champs $\mathbf{u}$ et $\mathbf{v}$, mais seulement leur divergence. Il est donc plus naturel de se placer dans l'espace

$$H(\mathrm{div};\Omega) = \left\{ \mathbf{v} \in L^2(\Omega)^n \;\middle|\; \mathrm{div}\, \mathbf{v} \in L^2(\Omega) \right\}, \tag{2.33}$$

où la divergence est comprise au sens des distributions. Cet espace est naturellement muni du produit scalaire

$$(\mathbf{u}, \mathbf{v})_{H(\mathrm{div};\Omega)} = (\mathrm{div}\, \mathbf{u}, \mathrm{div}\, \mathbf{v})_{L^2(\Omega)} + (\mathbf{u}, \mathbf{v})_{L^2(\Omega)^n}.$$

Les équations (2.31)–(2.32) entrent alors dans le cadre de notre problème abstrait (2.7)–(2.8) en posant $u = \mathbf{u}$, $\lambda = \varphi$, $V = H(\mathrm{div};\Omega)$, $M = L^2(\Omega)$ et

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \cdot \mathbf{v}\, d\Omega, \quad c(\mathbf{v}, \psi) = \int_\Omega \psi \ \mathrm{div}\, \mathbf{v}\, d\Omega, \tag{2.34}$$

$$\ell_V(\mathbf{v}) = 0 \ \text{ et } \ \ell_M(\psi) = -\int_\Omega f\, \psi\, d\Omega. \tag{2.35}$$

Il est facile de vérifier que si φ est solution de (2.30), alors $(\nabla\varphi, \varphi)$ est solution de ce problème mixte (et nous allons voir que c'est la seule possible). Pourquoi alors compliquer le problème ? L'utilisation d'une formulation mixte conduit à des méthodes numériques différentes de celles de [15], mais qui deviennent intéressantes lorsqu'on cherche à obtenir une approximation de $\nabla\varphi$. En effet, la "dérivation numérique" d'une approximation de (2.30) fournit généralement un résultat médiocre : par exemple, pour une approximation P^1, le gradient d'une fonction continue affine par morceaux est constant par morceaux. Les méthodes mixtes permettent de faire mieux !

Montrons maintenant que le problème (2.31)–(2.32) relève du théorème 2.1. Comme dans l'exemple des équations de Stokes, on a $C = \operatorname{div}$ (qui agit cette fois de $H(\operatorname{div};\Omega)$ dans $L^2(\Omega)$). La forme bilinéaire $a(\mathbf{u},\mathbf{v}) = (\mathbf{u},\mathbf{v})_{L^2(\Omega)^n}$ n'est bien sûr pas coercive dans $H(\operatorname{div};\Omega)$, mais elle l'est sur $\operatorname{Ker} C$ puisque

$$(\mathbf{u},\mathbf{v})_{H(\operatorname{div};\Omega)} = (\mathbf{u},\mathbf{v})_{L^2(\Omega)^n} \quad \text{si } \mathbf{u},\mathbf{v} \in \operatorname{Ker} C.$$

Vérifions la condition inf–sup, qu'on peut écrire ici

$$\sup_{\mathbf{v}\in H(\operatorname{div};\Omega)} \frac{(\psi, \operatorname{div}\mathbf{v})_{L^2(\Omega)}}{\|\mathbf{v}\|_{H(\operatorname{div};\Omega)}} \geq \gamma \ \|\psi\|_{L^2(\Omega)} \quad \forall \psi \in L^2(\Omega).$$

Pour cela, nous utilisons la proposition 2.4 : il s'agit de construire une application $\psi \mapsto \mathbf{v}_\psi$ de $L^2(\Omega)$ dans $H(\operatorname{div};\Omega)$ telle que

$$(\psi, \operatorname{div}\mathbf{v}_\psi)_{L^2(\Omega)} = \|\psi\|^2_{L^2(\Omega)} \quad \text{et} \quad \|\mathbf{v}_\psi\|_{H(\operatorname{div};\Omega)} \leq \gamma^{-1}\|\psi\|_{L^2(\Omega)}. \tag{2.36}$$

L'application recherchée est donnée ici par $\mathbf{v}_\psi = \nabla\theta_\psi$ où θ_ψ est la solution de

$$\begin{cases} \text{trouver } \theta_\psi \in H^1_0(\Omega) \text{ tel que} \\ (\nabla\theta_\psi, \nabla\psi')_{L^2(\Omega)^n} = -(\psi,\psi')_{L^2(\Omega)} \quad \forall \psi' \in H^1_0(\Omega), \end{cases}$$

autrement dit, une solution faible de $\Delta\theta_\psi = \psi$ avec $(\theta_\psi)_{|\partial\Omega} = 0$. Ce problème étant bien posé, on sait que $\|\nabla\theta_\psi\|_{L^2(\Omega)^n} \leq K\,\|\psi\|_{L^2(\Omega)}$. Comme $\operatorname{div}\mathbf{v}_\psi = \psi$, il s'ensuit que

$$\|\mathbf{v}_\psi\|^2_{H(\operatorname{div};\Omega)} = \|\nabla\theta_\psi\|^2_{L^2(\Omega)^n} + \|\psi\|^2_{L^2(\Omega)} \leq (1+K^2)\,\|\psi\|^2_{L^2(\Omega)}.$$

Par ailleurs,

$$(\psi, \operatorname{div}\mathbf{v}_\psi)_{L^2(\Omega)} = \|\psi\|^2_{L^2(\Omega)},$$

ce qui montre la propriété (2.36). Le théorème 2.1 s'applique donc bien.

Déformation d'une plaque mince encastrée

Pour un ouvert borné $\Omega \subset \mathbb{R}^2$, les équations suivantes modélisent les déformations d'une plaque mince homogène encastrée suivant son pourtour et soumise à une distribution f de charge normale à la plaque :

$$\begin{cases} \Delta^2\varphi = f \text{ dans } \Omega, \\ \varphi = 0 \quad \text{sur } \partial\Omega, \\ \dfrac{\partial\varphi}{\partial n} = 0 \quad \text{sur } \partial\Omega, \end{cases}$$

φ représentant le déplacement normal d'un point de la plaque. L'opérateur qui intervient ici est le bilaplacien Δ^2. L'écriture de ce problème sous forme mixte relève d'une idée analogue à l'exemple précédent. Il suffit d'introduire la nouvelle variable $u = \Delta\varphi$, ce qui nous permet de récrire le système comme suit :

$$\begin{cases} u - \Delta\varphi = 0 \text{ dans } \Omega, \\ \Delta u = f \qquad \text{dans } \Omega, \\ \varphi = 0 \qquad \text{sur } \partial\Omega, \\ \dfrac{\partial\varphi}{\partial n} = 0 \qquad \text{sur } \partial\Omega. \end{cases}$$

Si on suppose que $u \in \mathcal{C}^2(\Omega)$ et $\varphi \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1(\overline{\Omega})$, on en déduit par la formule de Green que

$$\int_\Omega u\, v\, d\Omega + \int_\Omega \nabla\varphi \cdot \nabla v\, d\Omega = 0 \quad \forall v \in H^1(\Omega),$$
$$\int_\Omega \nabla u \cdot \nabla\psi\, d\Omega = -\int_\Omega f\, \psi\, d\Omega \quad \forall \psi \in H^1_0(\Omega).$$

Ceci constitue un autre exemple d'application de notre problème abstrait avec $\lambda = \varphi$, $V = H^1(\Omega)$, $M = H^1_0(\Omega)$ et

$$a(u,v) = \int_\Omega u\, v\, d\Omega, \quad c(v,\psi) = \int_\Omega \nabla v \cdot \nabla\psi\, d\Omega, \tag{2.37}$$
$$\ell_V(v) = 0 \ \text{ et } \ \ell_M(\psi) = -\int_\Omega f\, \psi\, d\Omega. \tag{2.38}$$

Nous allons voir que contrairement aux exemples précédents, ce problème mixte est *mal posé* : il n'entre pas dans le cadre du théorème 2.1. Si on munit $M = H^1_0(\Omega)$ du produit scalaire $(\nabla\varphi, \nabla\psi)_{L^2(\Omega)^2}$ (ce qu'on ne peut évidemment pas faire pour $V = H^1(\Omega)$, pour lequel on conserve le produit scalaire usuel de $H^1(\Omega)$), l'opérateur $C : H^1(\Omega) \to H^1_0(\Omega)$ est maintenant défini de la façon suivante : pour tout $v \in H^1(\Omega)$, $Cv \in H^1_0(\Omega)$ est l'unique solution de l'équation variationnelle

$$(\nabla(Cv), \nabla\psi)_{L^2(\Omega)^2} = (\nabla v, \nabla\psi)_{L^2(\Omega)^2} \quad \forall \psi \in H^1_0(\Omega),$$

ce qui signifie que $\Delta(Cv) = \Delta v$ avec $(Cv)_{|\partial\Omega} = 0$. Ainsi,

$$\operatorname{Ker} C = \left\{ v \in H^1(\Omega) \mid \Delta v = 0 \text{ dans } \Omega \right\}.$$

Pour vérifier la condition inf–sup, on procède comme dans l'exemple précédent : il suffit de trouver une application $\psi \mapsto v_\psi$ de $H^1_0(\Omega)$ dans $H^1(\Omega)$ telle que

$$(\nabla v_\psi, \nabla\psi)_{L^2(\Omega)^2} = \|\nabla\psi\|^2_{L^2(\Omega)^2} \quad \text{et} \quad \|v_\psi\|_{H^1(\Omega)} \leq \gamma^{-1} \|\nabla v_\psi\|_{L^2(\Omega)^2}$$

Ici, on peut choisir $v_\psi = \psi$, et la continuité est assurée par l'inégalité de Poincaré.

Par contre, la forme bilinéaire $a(u,v) = (u,v)_{L^2(\Omega)}$ n'est pas coercive sur $H^1(\Omega)$, et elle ne l'est pas non plus sur $\operatorname{Ker} C$. En ce sens, il s'agit d'un problème mixte mal posé. L'existence d'une solution est en réalité soumise à une condition de régularité de la solution φ du problème initial de bilaplacien.

Les équations de l'élasticité

On s'intéresse maintenant aux (petites) déformations d'un solide élastique homogène et isotrope, occupant un domaine borné $\Omega \subset \mathbb{R}^3$. Ce solide est supposé soumis à une densité d'efforts extérieurs $\mathbf{f}$, et encastré le long de sa frontière $\partial\Omega$. Les équations portant sur le champ de déplacement $\mathbf{u}$ sont les suivantes :

$$\begin{cases} -\operatorname{\mathbf{div}} \sigma = \mathbf{f} \quad \text{dans } \Omega, & (2.39) \\ \sigma = A\, e(\mathbf{u}) = \lambda \operatorname{tr}(e(\mathbf{u}))\, \mathrm{I} + 2\mu\, e(\mathbf{u}) \quad \text{dans } \Omega, & (2.40) \\ \mathbf{u} = 0 \quad \text{sur } \partial\Omega. & (2.41) \end{cases}$$

L'équation (2.39) traduit l'équilibre du milieu, σ désignant le tenseur des contraintes, qui est relié au champ de déplacement par la loi de comportement (2.40), où $e(\mathbf{u})$ est le tenseur des déformations, donné par

$$e_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad \text{avec } \mathbf{u} = (u_1, u_2, u_3).$$

Enfin $\operatorname{tr}(e(\mathbf{u}))$ désigne la trace de $e(\mathbf{u})$, soit $\operatorname{tr}(e(\mathbf{u})) = \sum_{i=1,3} \partial_i u_i = \operatorname{div} \mathbf{u}$, et I est le tenseur identité. On peut remarquer que (2.40) entraîne

$$\operatorname{tr}(\sigma) = (3\lambda + 2\mu)\ \operatorname{tr}(e(\mathbf{u})),$$

ce qui permet d'inverser la loi de comportement, qui peut alors s'écrire de façon équivalente

$$e(\mathbf{u}) = A^{-1}\sigma = -\frac{\lambda}{2\mu(3\lambda+2\mu)} \operatorname{tr}(\sigma)\, \mathrm{I} + \frac{1}{2\mu}\sigma. \tag{2.42}$$

Nous allons considérer deux formulations variationnelles mixtes des équations ci-dessus dont l'inconnue est le couple $(\sigma, \mathbf{u})$. On introduit pour cela les espaces de tenseurs (symétriques) suivants :

$$\begin{aligned} \mathcal{L}^2_{\mathrm{sym}}(\Omega) &= \left\{\tau = (\tau_{ij})_{i,j=1,3}\,;\ \tau_{ij} = \tau_{ji} \in L^2(\Omega),\ 1 \leq i,j \leq 3\right\}, \\ \mathcal{H}_{\mathrm{sym}}(\mathbf{div};\Omega) &= \left\{\tau \in \mathcal{L}^2_{\mathrm{sym}}(\Omega);\ \mathbf{div}\,\tau \in L^2(\Omega)^3\right\}, \end{aligned}$$

qui constituent deux espaces de Hilbert lorsqu'on les munit respectivement des produits scalaires

$$(\sigma,\tau)_{\mathcal{L}^2(\Omega)} = \int_\Omega \sigma : \tau \, d\Omega \;=\; \int_\Omega \sum_{i,j=1,3} \sigma_{ij}\,\tau_{ij}\, d\Omega,$$

$$(\sigma,\tau)_{\mathcal{H}(\mathbf{div};\Omega)} = (\sigma,\tau)_{\mathcal{L}^2(\Omega)} + \int_\Omega \mathbf{div}\,\sigma \cdot \mathbf{div}\,\tau \, d\Omega.$$

Ci-dessus, $(\mathbf{div}\,\sigma)_i = \sum_{j=1,3} \partial_j \sigma_{ij}$, pour $i = 1, 2, 3$. On effectue alors d'une part le produit scalaire de $\mathcal{L}^2_{\rm sym}(\Omega)$ entre la loi de comportement (2.42) et un tenseur test symétrique τ, d'autre part le produit scalaire de $L^2(\Omega)^3$ entre l'équation d'équilibre (2.39) (qui joue ici le rôle d'une contrainte) et un champ test $\mathbf{v}$. Suivant l'équation à laquelle on applique une formule de Green, on obtient

$$\left\{\begin{array}{l} \text{trouver } (\sigma,\mathbf{u}) \in \mathcal{L}^2_{\rm sym}(\Omega) \times H^1_0(\Omega)^3 \text{ tel que} \\ \displaystyle\int_\Omega (A^{-1}\sigma) : \tau \, d\Omega - \int_\Omega \tau : e(\mathbf{u})\, d\Omega = 0 \quad \forall \tau \in \mathcal{L}^2_{\rm sym}(\Omega), \qquad (2.43) \\ \displaystyle -\int_\Omega \sigma : e(\mathbf{v})\, d\Omega = -\int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega \quad \forall \mathbf{v} \in H^1_0(\Omega)^3, \qquad (2.44) \end{array}\right.$$

ou bien

$$\left\{\begin{array}{l} \text{trouver } (\sigma,\mathbf{u}) \in \mathcal{H}_{\rm sym}(\mathbf{div};\Omega) \times L^2(\Omega)^3 \text{ tel que} \\ \displaystyle\int_\Omega (A^{-1}\sigma) : \tau \, d\Omega + \int_\Omega (\mathbf{div}\,\tau)\cdot\mathbf{u}\, d\Omega = 0 \quad \forall \tau \in \mathcal{H}_{\rm sym}(\mathbf{div};\Omega), \qquad (2.45) \\ \displaystyle\int_\Omega (\mathbf{div}\,\sigma)\cdot\mathbf{v}\, d\Omega = -\int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega \quad \forall \mathbf{v} \in L^2(\Omega)^3. \qquad (2.46) \end{array}\right.$$

Passer de l'une à l'autre de ces deux formulations revient à échanger une partie de la régularité présupposée de σ avec celle de $\mathbf{u}$. En contrepartie, d'un point de vue numérique, σ et $\mathbf{u}$ ne vont pas être discrétisés de la même façon dans les deux cas. Le premier problème autorise une discrétisation discontinue pour σ (par exemple $(P^0)^{3\times 3}$: constant par morceaux) mais pas pour $\mathbf{u}$. Dans le second cas au contraire, $\mathbf{u}$ pourra être approché par une fonction discontinue, alors que la condition $\mathbf{div}\,\tau \in L^2(\Omega)^3$ impose la continuité de la contrainte normale, à savoir $\tau\mathbf{n}$, entre deux éléments (on le vérifie de façon analogue à la contrainte de continuité d'un champ v, imposée par $v \in H^1(\Omega)$).

Notons que la condition aux limites (2.41) ne joue pas le même rôle dans les deux formulations. Dans (2.43)–(2.44), c'est une condition *essentielle* puisqu'elle est contenue dans l'espace $H^1_0(\Omega)^3$. En revanche, dans (2.45)–(2.46), elle devient *naturelle*, au sens où elle est contenue dans l'équation variationnelle (2.45). En effet,

pour la retrouver, il faut appliquer à rebours la formule de Green à la première équation variationnelle. En choisissant tout d'abord $\tau_{ij} \in \mathcal{D}(\Omega)$, on en déduit dans un premier temps que $A^{-1}\sigma = e(\mathbf{u})$ dans Ω. En prenant alors $\tau_{ij} \in \mathcal{C}^\infty(\bar{\Omega})$, il s'ensuit que

$$\int_{\partial\Omega} (\tau\mathbf{n}) \cdot \mathbf{u}\, d\Gamma = 0,$$

et ce, pour tout τ; par conséquent $\mathbf{u}_{|\partial\Omega} = 0$.

Montrons maintenant que chacun des deux problèmes ci-dessus est bien posé. Pour la première formulation (2.43)–(2.44), notons

$$a(\sigma,\tau) = \int_\Omega (A^{-1}\sigma) : \tau\, d\Omega \quad \text{et} \quad c_1(\tau,\mathbf{v}) = -\int_\Omega \tau : e(\mathbf{v})\, d\Omega.$$

La continuité de ces formes bilinéaires, définies respectivement sur $\mathcal{L}^2_{\rm sym}(\Omega) \times \mathcal{L}^2_{\rm sym}(\Omega)$ et $\mathcal{L}^2_{\rm sym}(\Omega) \times H^1_0(\Omega)^3$, ne soulève aucune difficulté (c'est précisément ce qui justifie le choix de ces espaces). La forme $a(\cdot,\cdot)$ est coercive sur $\mathcal{L}^2_{\rm sym}(\Omega)$. Il suffit en effet de remarquer que

$$a(\sigma,\sigma) = \int_\Omega \frac{1}{2\mu}\left(\sigma : \sigma - \frac{\lambda}{3\lambda+2\mu}(\mathrm{tr}(\sigma))^2\right) d\Omega \geq \frac{1}{3\lambda+2\mu}\int_\Omega \sigma : \sigma\, d\Omega, \quad (2.47)$$

puisque $(\mathrm{tr}(\sigma))^2 \leq 3\sum_{i=1,3}\sigma_{ii}^2$. Pour vérifier la condition inf–sup, on utilise un critère analogue à la proposition 2.4 : pour tout $\mathbf{v} \in H^1_0(\Omega)^3$, le tenseur $\tau_{\mathbf{v}} = -e(\mathbf{v})$ vérifie

$$\|\tau_{\mathbf{v}}\|_{\mathcal{L}^2(\Omega)} \leq \|\mathbf{v}\|_{H^1_0(\Omega)^3} \quad \text{et} \quad c_1(\tau_{\mathbf{v}},\mathbf{v}) = \int_\Omega e(\mathbf{v}) : e(\mathbf{v})\, d\Omega, \quad (2.48)$$

avec $\|\mathbf{v}\|_{H^1_0(\Omega)^3} = (\int_\Omega |\nabla\mathbf{v}|^2\, d\Omega)^{1/2} = (\sum_{i,j=1,3}\|\partial_i v_j\|^2_{L^2(\Omega)})^{1/2}$. En effet :

$$\|\tau_{\mathbf{v}}\|^2_{\mathcal{L}^2(\Omega)} = \frac{1}{4}\sum_{i,j=1,3}\left\|\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j}\right\|^2_{L^2(\Omega)} \leq \frac{1}{2}\sum_{i,j=1,3}\left(\left\|\frac{\partial v_j}{\partial x_i}\right\|^2_{L^2(\Omega)} + \left\|\frac{\partial v_i}{\partial x_j}\right\|^2_{L^2(\Omega)}\right),$$

soit $\|\tau_{\mathbf{v}}\|_{\mathcal{L}^2(\Omega)} \leq \|\mathbf{v}\|_{H^1_0(\Omega)^3}$. Par ailleurs, on a le résultat suivant.

Proposition 2.6 (Inégalité de Korn) *Pour tout* $\mathbf{v} \in H^1_0(\Omega)^3$, *on a*

$$\int_\Omega e(\mathbf{v}) : e(\mathbf{v})\, d\Omega \;\geq\; \frac{1}{2}\,\|\mathbf{v}\|^2_{H^1_0(\Omega)^3}. \quad (2.49)$$

Démonstration : Supposons tout d'abord que $\mathbf{v} = (v_1, v_2, v_3) \in \mathcal{D}(\Omega)^3$. En appliquant deux fois la formule de Green et en intervertissant l'ordre de dérivation, on remarque que

$$\int_\Omega \frac{\partial v_i}{\partial x_j}\,\frac{\partial v_j}{\partial x_i}\, d\Omega = -\int_\Omega v_i\,\frac{\partial^2 v_j}{\partial x_j \partial x_i}\, d\Omega = -\int_\Omega v_i\,\frac{\partial^2 v_j}{\partial x_i \partial x_j}\, d\Omega = \int_\Omega \frac{\partial v_i}{\partial x_i}\,\frac{\partial v_j}{\partial x_j}\, d\Omega.$$

Ainsi,

$$\int_\Omega (e_{ij}(\mathbf{v}))^2 \, d\Omega = \frac{1}{4}\int_\Omega \left(\frac{\partial v_i}{\partial x_j}\right)^2 d\Omega + \frac{1}{4}\int_\Omega \left(\frac{\partial v_j}{\partial x_i}\right)^2 d\Omega + \frac{1}{2}\int_\Omega \frac{\partial v_i}{\partial x_i}\,\frac{\partial v_j}{\partial x_j}\, d\Omega.$$

En sommant sur i et j, il vient

$$\int_\Omega e(\mathbf{v}) : e(\mathbf{v})\, d\Omega = \frac{1}{2}\sum_{i,j=1,3}\int_\Omega \left(\frac{\partial v_i}{\partial x_j}\right)^2 d\Omega + \frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{v})^2\, d\Omega,$$

d'où l'inégalité (2.49) pour $\mathbf{v} \in \mathcal{D}(\Omega)^3$. D'après la densité de $\mathcal{D}(\Omega)$ dans $H_0^1(\Omega)$, cette inégalité reste valable pour tout $\mathbf{v} \in H_0^1(\Omega)^3$. ∎

Finalement, (2.48) et (2.49) montrent que

$$c_1(\tau_\mathbf{v}, \mathbf{v}) \geq \frac{1}{2}\ \|\tau_\mathbf{v}\|_{\mathcal{L}^2(\Omega)}\ \|\mathbf{v}\|_{H_0^1(\Omega)^3},$$

ce qui entraine que la condition inf–sup est satisfaite avec $\gamma = 1/2$.

Considérons maintenant la seconde formulation (2.45)–(2.46). Notons cette fois

$$c_2(\tau, \mathbf{v}) = \int_\Omega (\mathbf{div}\,\tau)\cdot\mathbf{v}\, d\Omega,$$

la forme $a(\cdot\,,\cdot)$ étant inchangée. Cette dernière n'est pas coercive sur $\mathcal{H}_{\rm sym}(\mathbf{div};\Omega)$ mais d'après (2.47), elle l'est évidemment sur $\operatorname{Ker} C_2$ où $C_2 : \mathcal{H}_{\rm sym}(\mathbf{div};\Omega) \to L^2(\Omega)^3$ est donné par

$$(C_2\tau, \mathbf{v})_{L^2(\Omega)^3} = c_2(\tau, \mathbf{v}) \quad \text{soit ici} \quad C_2 = \mathbf{div}\,.$$

Pour vérifier la condition inf–sup, on applique la proposition 2.4. Pour tout $\mathbf{v} \in L^2(\Omega)^3$, on pose cette fois $\tau_\mathbf{v} = e(\mathbf{u_v})$ où $\mathbf{u_v} \in H_0^1(\Omega)^3$ est la solution de $\mathbf{div}(e(\mathbf{u_v})) = \mathbf{v}$. Cette équation relève du théorème de Lax-Milgram, d'après l'inégalité de Korn (c'est le système de l'élasticité dans le cas particulier où $\lambda = 0$ et $\mu = 1/2$). Sa solution dépend donc continûment de $\mathbf{v}$ au sens où il existe une constante $K > 0$ telle que

$$\|\mathbf{u_v}\|^2_{H_0^1(\Omega)^3} \leq K\ \|\mathbf{v}\|^2_{L^2(\Omega)^3},$$

ce qui montre que

$$\|\tau_\mathbf{v}\|^2_{\mathcal{H}(\mathbf{div};\Omega)} \leq (K+1)\ \|\mathbf{v}\|^2_{L^2(\Omega)^3}.$$

La conclusion découle alors du fait que

$$c_2(\tau_\mathbf{v}, \mathbf{v}) = \|\mathbf{v}\|^2_{L^2(\Omega)^3}.$$

Ainsi chacune des deux formulations (2.43)–(2.44) et (2.45)–(2.46) définit un problème bien posé au sens du théorème 2.1.

2.1.4 Le point de vue de l'optimisation

Dans ce paragraphe, nous supposerons que la forme bilinéaire $a(\cdot\,,\cdot)$ est *symétrique* et *positive*, soit

$$a(u,v) = a(v,u) \quad \text{et} \quad a(v,v) \geq 0 \quad \forall u, v \in V. \tag{2.50}$$

La notion de point selle

Pour un problème sans contrainte (c'est-à-dire lorsque $c(\cdot\,,\cdot) = 0$), nous savons (cf. par exemple [26]) que l'équation variationnelle (2.7) peut s'interpréter comme la *condition d'optimalité* du problème de minimisation

$$\inf_{v \in V} J(v) \quad \text{avec} \quad J(v) = \frac{1}{2}\, a(v,v) - \ell_V(v). \tag{2.51}$$

On vérifie en effet immédiatement que la différentielle [1] $dJ(u)$ de J en u est donnée par

$$dJ(u).v = a(u,v) - \ell_V(v), \quad \forall v \in V$$

et par conséquent, (2.7) revient à dire que $dJ(u) = 0$. Lorsque $a(\cdot\,,\cdot)$ est coercive sur V, la fonctionnelle J est strictement convexe, sa Hessienne $d^2J(u)$ étant définie par : $d^2J(u).(v,w) = a(v,w)$. La condition d'optimalité fournit donc l'unique solution du problème (2.51).

Dans le cas général, nous allons voir que le problème variationnel sous contrainte (2.7)–(2.8) correspond encore aux conditions d'optimalité du problème de minimisation de la fonctionnelle $J(v)$, mais cette fois sur l'espace des états contraints. Ce constat est très important en pratique car il inscrit notre problème modèle dans le cadre de l'*optimisation sous contrainte* d'une fonctionnelle quadratique. On dispose ainsi des méthodes générales développées dans ce contexte, en particulier les méthodes de *dualité* (voir [26]). Les hypothèses (2.50) nous permettent en effet d'interpréter le couple (u,λ) comme un *point selle* du *lagrangien* $\mathcal{L} : V \times M \to \mathbb{R}$ défini par

$$\mathcal{L}(v,\mu) = J(v) + c(v,\mu) - \ell_M(\mu), \tag{2.52}$$

où la variable μ apparaît comme un *multiplicateur de Lagrange* associé à la contrainte.

Définition 2.7 *On dit que* $(u,\lambda) \in V \times M$ *est un* point selle *de* $\mathcal{L}$ *si*

$$\mathcal{L}(u,\mu) \leq \mathcal{L}(u,\lambda) \leq \mathcal{L}(v,\lambda) \quad \forall v \in V,\ \forall \mu \in M. \tag{2.53}$$

1. Rappelons que la différentielle de J en u est l'application linéaire continue $dJ(u)$ définie par $dJ(u).v = \lim_{\varepsilon \to 0} \varepsilon^{-1}(J(u+\varepsilon v) - J(u))$ pour tout $v \in V$.

Proposition 2.8 *Sous l'hypothèse* (2.50)*, un couple* (u, λ) *est un point selle de* $\mathcal{L}$ *si, et seulement si,*

$$a(u, v) + c(v, \lambda) = \ell_V(v) \quad \forall v \in V, \tag{2.54}$$
$$c(u, \mu) = \ell_M(\mu) \quad \forall \mu \in M. \tag{2.55}$$

Démonstration : La première inégalité dans (2.53) est équivalente à la contrainte (2.55). Par définition de $\mathcal{L}$, elle s'écrit en effet

$$c(u, \mu - \lambda) \leq \ell_M(\mu - \lambda),$$

où le fait de pouvoir choisir μ quelconque impose l'égalité des deux membres.

La seconde inégalité dans (2.53) signifie en d'autres termes que u réalise le minimum de la fonctionnelle $v \mapsto \mathcal{L}(v, \lambda)$ (λ étant fixé). La condition nécessaire d'optimalité est donc

$$d_v \mathcal{L}(u, \lambda) = 0,$$

c'est-à-dire exactement (2.54). Mais cette condition est aussi suffisante puisque $v \mapsto \mathcal{L}(v, \lambda)$ est convexe car $d_v^2 \mathcal{L}(u, \lambda).(v, v) = a(v, v)$, qui est positif d'après l'hypothèse (2.50). ∎

Corollaire 2.9 *Sous les hypothèses du théorème* 2.1 *complétées par* (2.50)*, le lagrangien* $\mathcal{L}$ *admet un unique point selle* (*solution de* (2.54)–(2.55))*.*

Caractérisation d'un point selle

La notion de *point selle* constitue la clef de voûte des méthodes de dualité qui reposent sur une idée très simple : pour rejoindre un point selle (qui peut s'interpréter comme un col en montagne, les variables v et μ représentant les deux directions horizontales, la fonctionnelle considérée étant convexe suivant v et concave suivant μ), il existe deux façons canoniques de procéder.

– "Par au-dessus" : partant d'un point donné, on commence par se déplacer dans la direction μ (en laissant v constant) de façon à atteindre l'altitude maximale. Puis, en faisant varier v et μ simultanément, on descend vers le col en veillant à toujours rester à l'altitude maximale relativement à la direction μ. Ceci revient à résoudre le problème suivant, appelé *problème primal* :

$$\inf_{v \in V} \sup_{\mu \in M} \mathcal{L}(v, \mu), \tag{2.56}$$

qui peut être vu comme un problème de minimisation selon la seule variable v de la fonctionnelle $v \mapsto \sup_{\mu \in M} \mathcal{L}(v, \mu)$.

– "Par au-dessous" : en se déplaçant cette fois dans la direction v, on commence par atteindre le point le plus bas. Puis on monte jusqu'au col tout en conservant l'altitude minimale suivant v. On résout de cette façon le problème dual :

$$\sup_{\mu\in M}\inf_{v\in V}\mathcal{L}(v,\mu), \tag{2.57}$$

qui consiste quant à lui à maximiser selon μ la fonctionnelle $\mu\mapsto\inf_{v\in V}\mathcal{L}(v,\mu)$.

Le fait que ces deux chemins canoniques mènent au même point constitue en fait une caractérisation d'un point selle, comme l'exprime la proposition générale suivante [2].

Proposition 2.10 *Un couple (u,λ) est un point selle de $\mathcal{L}$ si et seulement si u est solution du problème primal, λ est solution du problème dual, et*

$$\inf_{v\in V}\sup_{\mu\in M}\mathcal{L}(v,\mu)=\sup_{\mu\in M}\inf_{v\in V}\mathcal{L}(v,\mu).$$

Problème primal, problème dual

Dans notre cas, le problème *primal* consiste à minimiser la fonctionnelle $J(v)$ sur l'espace des états satisfaisant la contrainte (2.55), autrement dit $Cv=g$ où l'opérateur C est donné par (2.13) et $g\in M$ est défini par (2.28). En effet

$$\sup_{\mu\in M}\mathcal{L}(v,\mu)=\begin{cases}J(v) & \text{si } Cv=g,\\ +\infty & \text{sinon,}\end{cases}$$

et par conséquent

$$\inf_{v\in V}\sup_{\mu\in M}\mathcal{L}(v,\mu)=\inf_{v\in V,\ Cv=g}J(v).$$

Le problème *dual* ne peut être interprété aussi simplement. Nous allons voir qu'il s'agit d'un problème *sans contrainte* analogue à ceux qu'on rencontre en *commande optimale*. L'absence de contrainte rend cette formulation particulièrement intéressante d'un point de vue numérique.

Pour $\mu\in M$ donné, notons u_μ une solution du problème de minimisation

2. Il est facile de voir qu'on a toujours

$$\sup_{\mu\in M}\inf_{v\in V}\mathcal{L}(v,\mu)\le\inf_{v\in V}\sup_{\mu\in M}\mathcal{L}(v,\mu),$$

le chemin "par au-dessous" fait toujours arriver plus bas que le chemin "par au-dessus". La définition (2.53) d'un point selle fournit l'inégalité inverse.

$$\mathcal{L}(u_\mu, \mu) = \inf_{v \in V} \mathcal{L}(v, \mu).$$

La condition (nécessaire et suffisante) d'optimalité s'écrit $d_v\mathcal{L}(u_\mu, \mu) = 0$, soit

$$a(u_\mu, v) = \ell_V(v) - c(v, \mu) \quad \forall v \in V. \tag{2.58}$$

Si on suppose que *la forme bilinéaire $a(\cdot\,,\cdot)$ est coercive sur V tout entier* (et pas seulement sur $\operatorname{Ker} C$), ce problème possède une unique solution. On a en particulier

$$a(u_\mu, u_\mu) = \ell_V(u_\mu) - c(u_\mu, \mu),$$

d'où l'on déduit que

$$\mathcal{L}(u_\mu, \mu) = -\frac{1}{2}\, a(u_\mu, u_\mu) - \ell_M(\mu). \tag{2.59}$$

En renversant le signe de cette fonctionnelle, le problème dual s'écrit donc sous forme d'un *problème de minimisation sans contrainte* :

$$\inf_{\mu \in M} \left(\frac{1}{2}\, a(u_\mu, u_\mu) + \ell_M(\mu) \right), \tag{2.60}$$

le multiplicateur μ jouant le rôle d'une commande, u_μ étant l'état du système associé à cette commande, solution du problème variationnel (2.58). Ecrivons les conditions d'optimalité de ce problème. Pour cela, remarquons qu'il peut s'écrire aussi $A\,u_\mu = f - C^T\mu$ avec A, f et C^T donnés par (2.13), (2.14) et (2.17), où l'hypothèse de coercivité de $a(\cdot\,,\cdot)$ montre que A est inversible. Ainsi, d'après (2.59),

$$-\mathcal{L}(u_\mu, \mu) = \frac{1}{2}\left(f - C^T\mu, A^{-1}(f - C^T\mu)\right)_V + (g, \mu)_M,$$

et par conséquent, $d_\mu\mathcal{L}(u_\mu, \mu) = 0$ revient à dire que μ est solution de

$$C\,A^{-1}C^T\mu = C\,A^{-1}f - g. \tag{2.61}$$

L'algorithme d'Uzawa consiste en une simple méthode de gradient appliquée à cette équation en μ qui caractérise le minimum de la fonction $\mu \mapsto \mathcal{L}(u_\mu, \mu)$. Partant d'une valeur initiale $\mu^{(0)}$, chaque itération de l'algorithme est définie par les 3 étapes ci-dessous :

(i) $\mu^{(n)}$ étant connu, résoudre $Au^{(n+1)} = f - C^T\mu^{(n)}$;

(ii) pour ρ convenablement choisi, calculer $\mu^{(n+1)} = \mu^{(n)} + \rho(Cu^{(n+1)} - g)$;

(iii) si $\left\|\mu^{(n+1)} - \mu^{(n)}\right\|_M$ est assez petit : FIN.

2.2 Approximation d'un problème mixte

2.2.1 Un résultat général

Dans [15], nous avons introduit la notion d'approximation interne (ou approximation de Galerkin) d'un problème sans contrainte, en construisant un sous-espace de dimension finie de l'espace dans lequel est posé la formulation variationnelle. L'idée que nous développons ici pour notre problème modèle sous contrainte est le prolongement naturel de cette démarche. Pour simplifier la présentation, nous nous contentons de traiter le cas d'une contrainte homogène, soit $\ell_M(\cdot) = 0$, (voir la remarque 2.15 dans le cas d'une contrainte non homogène).

Nous allons considérer deux familles de sous-espaces V_h et M_h de *dimensions finies* de V et M, le paramètre h étant destiné à tendre vers 0 (V_h et M_h sont censés approcher V et M d'autant mieux que h est plus petit). Pour chaque h, on introduit alors le problème mixte approché associé à (2.9)–(2.10) :

$$\begin{cases} \text{trouver } (u_h, \lambda_h) \in V_h \times M_h \text{ tel que} \\ a(u_h, v_h) + c(v_h, \lambda_h) = \ell_V(v_h) \quad \forall v_h \in V_h, & (2.62) \\ c(u_h, \mu_h) = 0 \quad \forall \mu_h \in M_h. & (2.63) \end{cases}$$

Comme dans le cas continu, on peut récrire ce problème variationnel sous la forme

$$\begin{pmatrix} A_h & C_h^T \\ C_h & 0 \end{pmatrix} \begin{pmatrix} u_h \\ \lambda_h \end{pmatrix} = \begin{pmatrix} f_h \\ 0 \end{pmatrix}, \tag{2.64}$$

où $A_h : V_h \to V_h$ et $C_h : V_h \to M_h$ sont les opérateurs définis par

$$(A_h u_h, v_h)_V = a(u_h, v_h) \text{ et } (C_h v_h, \mu_h)_M = c(v_h, \mu_h), \forall u_h,\ v_h \in V_h,\ \forall \mu_h \in M_h,$$

et $f_h \in V_h$ est donné par

$$(f_h, v_h)_V = \ell(v_h) \quad \forall v_h \in V_h.$$

Remarque 2.11 *Il est facile de voir que l'opérateur $C_h : V_h \to M_h$ associé à la contrainte discrète est lié à l'opérateur $C : V \to M$ par la relation*

$$C_h v_h = \Lambda_h C v_h \quad \forall v_h \in V_h,$$

où Λ_h désigne la projection orthogonale de M sur M_h, définie par

$$\Lambda_h \mu \in M_h \text{ et } (\Lambda_h \mu, \mu_h)_M = (\mu, \mu_h)_M \quad \forall \mu_h \in M_h.$$

En d'autres termes, C_h est la restriction de $\Lambda_h C$ à V_h. Lorsque $C(V_h) \subset M_h$, l'opérateur approché C_h est donc simplement la restriction de C à V_h. Dans ce

cas particulier, on a $\operatorname{Ker} C_h \subset \operatorname{Ker} C$ *: la solution* u_h *du problème approché vérifie exactement la contrainte continue. En revanche, si* $C(V_h) \not\subset M_h$*, il n'y a plus de propriété d'inclusion remarquable sur les noyaux de* C *et* C_h*. En ce sens,* u_h *apparaît comme une* approximation externe *de l'état du système, puisqu'elle ne vérifie la contrainte que de façon approchée (mais le couple* (u_h, λ_h) *reste quant à lui une approximation interne de la solution du problème mixte* (2.9)–(2.10)*).*

Désignons par P_h la projection orthogonale de V_h sur $\operatorname{Ker} C_h$. Résoudre le système (2.64) revient à résoudre successivement les deux équations

$$P_h A_h P_h u_h = P_h f_h \quad \text{avec} \quad u_h \in \operatorname{Ker} C_h, \tag{2.65}$$

$$C_h^T \lambda_h = f_h - A_h u_h. \tag{2.66}$$

La première est soluble si $P_h A_h P_h \ : \operatorname{Ker} C_h \to \operatorname{Ker} C_h$ est inversible, et la seconde, si $C_h^T \ : M_h \to (\operatorname{Ker} C_h)^\perp$ est inversible. Notons que $\operatorname{Im} C_h^T = (\operatorname{Ker} C_h)^\perp$ puisque les espaces sont de dimensions finies. Ainsi, comme dans le théorème 2.1, si on suppose d'une part que $a(\cdot\,,\cdot)$ est coercive sur $\operatorname{Ker} C_h$, d'autre part que $c(\cdot\,,\cdot)$ satisfait la condition inf–sup sur $M_h \times V_h$, on est assuré de l'existence et de l'unicité de la solution du problème approché. Pour que cette approximation puisse converger vers la solution exacte, c'est-à-dire la solution du problème continu (2.9)–(2.10), nous allons supposer que ces deux propriétés ont lieu *uniformément* par rapport à h : le théorème suivant, analogue au lemme de Céa, nous montre que l'écart entre la solution exacte (u, λ) et la solution approchée est de l'ordre de la "distance" de (u, λ) à $V_h \times M_h$.

Théorème 2.12 *En plus des hypothèses du théorème* 2.1*, on suppose qu'il existe deux constantes* $\tilde{\alpha} > 0$ *et* $\tilde{\gamma} > 0$ *indépendantes de* h *telles que*

$$\inf_{v_h \in \operatorname{Ker} C_h} \frac{a(v_h, v_h)}{\|v_h\|_V^2} \geq \tilde{\alpha}, \tag{2.67}$$

$$\inf_{\mu_h \in M_h} \sup_{v_h \in V_h} \frac{c(v_h, \mu_h)}{\|v_h\|_V \, \|\mu_h\|_M} \geq \tilde{\gamma}. \tag{2.68}$$

Alors, pour tout h*, le problème approché* (2.62)–(2.63) *admet une solution unique* (u_h, λ_h)*, et il existe* $K > 0$ *indépendant de* h *tel que*

$$\|u - u_h\|_V + \|\lambda - \lambda_h\|_M \leq K \left\{ \inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M \right\}, \tag{2.69}$$

où (u, λ) *désigne la solution de* (2.9)–(2.10)*.*

Remarque 2.13 *A* h *fixé, la propriété* (2.67) *signifie simplement que* $a(v_h, v_h) > 0$ *pour tout* $v_h \in \operatorname{Ker} C_h$ *non nul. Le caractère uniforme de cette propriété par*

rapport à h sera assuré notamment lorsque la forme bilinéaire $a(\cdot\,,\cdot)$ est coercive sur tout l'espace V. Par contre, (2.67) *ne découle en général pas de* (2.22) *car* $\operatorname{Ker} C_h$ *n'est pas nécessairement contenu dans* $\operatorname{Ker} C$ *(voir la remarque 2.11).*

La propriété (2.68) *est appelée* condition inf–sup discrète uniforme. *A h fixé, elle exprime simplement que* $\operatorname{Ker} C_h^T = \{0\}$. *Mais l'uniformité de cette propriété par rapport à h, qui traduit une condition de compatibilité uniforme entre les espaces V_h et M_h, est souvent très délicate à vérifier. Les paragraphes 2.2.2 et 2.2.3 présentent une démarche assez générale pour l'obtenir.*

Démonstration : Commençons par remarquer qu'en choisissant $v = v_h \in V_h$ dans (2.9) et en lui retranchant l'équation approchée (2.62), on obtient

$$a(u - u_h, v_h) + c(v_h, \lambda - \lambda_h) = 0 \quad \forall v_h \in V_h. \tag{2.70}$$

Pour $w_h \in \operatorname{Ker} C_h$ quelconque, posons $v_h = u_h - w_h$. En remarquant que

$$a(v_h, v_h) = a(u - w_h, v_h) + a(u_h - u, v_h),$$

l'égalité (2.70) nous montre que

$$a(v_h, v_h) = a(u - w_h, v_h) + c(v_h, \lambda - \lambda_h),$$

où l'on peut remplacer λ_h par n'importe quel $\mu_h \in M_h$ puisque $v_h \in \operatorname{Ker} C_h$. La propriété de coercivité (2.67) jointe à la continuité de $a(\cdot\,,\cdot)$ et $c(\cdot\,,\cdot)$ (voir (2.11)) montre donc que

$$\tilde{\alpha}\,\|v_h\|_V \le m_a\,\|u - w_h\|_V + m_c\,\|\lambda - \mu_h\|_M\,.$$

Comme $u - u_h = (u - w_h) - v_h$, il s'ensuit par l'inégalité triangulaire que

$$\|u - u_h\|_V \le (1 + \tilde{\alpha}^{-1} m_a)\,\|u - w_h\|_V + \tilde{\alpha}^{-1} m_c\,\|\lambda - \mu_h\|_M \quad \forall w_h \in \operatorname{Ker} C_h,\ \forall \mu_h \in M_h,$$

et par conséquent

$$\|u - u_h\|_V \le (1 + \tilde{\alpha}^{-1} m_a) \inf_{w_h \in \operatorname{Ker} C_h} \|u - w_h\|_V + \tilde{\alpha}^{-1} m_c \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M\,. \tag{2.71}$$

Montrons maintenant que

$$\inf_{w_h \in \operatorname{Ker} C_h} \|u - w_h\|_V \le (1 + \tilde{\gamma}^{-1} m_c) \inf_{v_h \in V_h} \|u - v_h\|_V\,. \tag{2.72}$$

Comme dans le lemme 2.2, la condition inf–sup (2.68) revient à dire que C_h est bijectif de $(\operatorname{Ker} C_h)^\perp$ dans M_h, et $\|z_h\|_V \le \tilde{\gamma}^{-1}\|C_h z_h\|_M$ pour tout $z_h \in (\operatorname{Ker} C_h)^\perp$. Ainsi, pour tout $v_h \in V_h$, on sait qu'il existe un unique $z_h \in (\operatorname{Ker} C_h)^\perp$ tel que $C_h z_h = \Lambda_h C(u - v_h)$, où Λ_h désigne la projection orthogonale de M sur M_h (voir la remarque 2.11). Et on a l'estimation

$$\|z_h\|_V \le \tilde{\gamma}^{-1}\,\|\Lambda_h C(u - v_h)\|_M \le \tilde{\gamma}^{-1} m_c\,\|u - v_h\|_V.$$

La définition de z_h et le fait que $u \in \operatorname{Ker} C$ nous montrent que $C_h(z_h + v_h) = \Lambda_h C u = 0$, autrement dit que $w_h = z_h + v_h \in \operatorname{Ker} C_h$. De plus,

$$\|u - w_h\|_V \le \|u - v_h\|_V + \|z_h\|_V \le (1 + \tilde{\gamma}^{-1} m_c)\,\|u - v_h\|_V.$$

L'inégalité (2.72) en résulte puisque v_h peut être choisi quelconque dans V_h.

Pour obtenir (2.69), il reste à estimer la quantité $\|\lambda - \lambda_h\|_M$. La condition inf–sup (2.68) nous montre que

$$\|\lambda_h - \mu_h\|_M \leq \tilde{\gamma}^{-1} \sup_{v_h \in V_h} \frac{c(v_h, \lambda_h - \mu_h)}{\|v_h\|_V} \quad \forall \mu_h \in M_h,$$

où d'après (2.70),

$$c(v_h, \lambda_h - \mu_h) = a(u - u_h, v_h) + c(v_h, \lambda - \mu_h).$$

Ainsi, en vertu de la continuité de $a(\cdot\,,\cdot)$ et $c(\cdot\,,\cdot)$, on a alors

$$\|\lambda_h - \mu_h\|_M \leq \tilde{\gamma}^{-1} \left\{ m_a \|u - u_h\|_V + m_c \|\lambda - \mu_h\|_M \right\} \quad \forall \mu_h \in M_h,$$

soit, d'après l'inégalité triangulaire

$$\|\lambda - \lambda_h\|_M \leq \tilde{\gamma}^{-1} m_a \|u - u_h\|_V + (1 + \tilde{\gamma}^{-1} m_c) \|\lambda - \mu_h\|_M \quad \forall \mu_h \in M_h.$$

La conclusion découle de cette dernière estimation et de (2.71)–(2.72). ∎

Comme pour les problèmes sans contrainte, on peut alors donner une condition suffisante d'approximabilité qui assure la convergence de la solution approchée :

Corollaire 2.14 *En plus des hypothèses du théorème* 2.12*, on suppose qu'il existe un sous-espace dense $\mathcal{V}$ de V, un sous-espace dense $\mathcal{M}$ de M, et deux familles d'applications $r_h : \mathcal{V} \to V_h$ et $\rho_h : \mathcal{M} \to M_h$ tels que*

$$\lim_{h \to 0} \|r_h v - v\|_V = 0 \quad \forall v \in \mathcal{V} \quad \text{et} \quad \lim_{h \to 0} \|\rho_h \mu - \mu\|_M = 0 \quad \forall \mu \in \mathcal{M}.$$

Alors

$$\lim_{h \to 0} \{\|u - u_h\|_V + \|\lambda - \lambda_h\|_M\} = 0.$$

Remarque 2.15 *Le théorème* 2.12 *s'applique encore pour une contrainte non homogène (soit $\ell_M(\cdot) \neq 0$ dans* (2.8)*). La contrainte s'écrit dans ce cas $C_h u_h = \Lambda_h g$ où Λ_h est la projection orthogonale de M sur M_h (voir la remarque* 2.11*), et $g \in M$ est donné par* (2.28)*. La démonstration ci-dessus se généralise sans difficulté (voir par exemple [27] pour les détails) en considérant l'espace affine*

$$W_h(g) = \{v_h \in V_h \mid C_h v_h = \Lambda_h g\}.$$

On obtient ainsi à la place de (2.71)

$$\|u - u_h\|_V \leq (1 + \tilde{\alpha}^{-1} m_a) \inf_{w_h \in W_h(g)} \|u - w_h\|_V + \tilde{\alpha}^{-1} m_c \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M,$$

où le terme $\inf_{w_h \in W_h(g)} \|u - w_h\|_V$ est encore majoré par le second membre de (2.72)*.*

2.2.2 Existence et unicité du multiplicateur approché

Comme nous l'avons souligné dans la remarque 2.13, la condition inf–sup discrète uniforme (2.68) assure

– d'une part, pour chaque h, l'inversibilité de C_h^T de M_h dans $(\operatorname{Ker} C_h)^\perp$, donc l'existence et l'unicité du multiplicateur approché λ_h,
– d'autre part, le caractère *uniformément* borné (par rapport à h) de l'inverse de C_h^T, qui d'un point de vue pratique, nous fournit une propriété de stabilité du calcul numérique de λ_h.

Dans ce paragraphe, nous allons nous attacher uniquement au premier point : si on choisit deux types d'éléments finis pour u_h et λ_h respectivement, peut-on affirmer que pour un maillage donné (sur lequel repose la construction des deux espaces V_h et M_h), la matrice C_h^T est inversible ? Le second point est nettement plus délicat : il s'agit de vérifier que l'inversibilité de C_h^T ne se détériore pas lorsque h tend vers 0. Nous aborderons cette question au §2.2.3.

Le cas des équations de Stokes

Nous nous contentons ici de présenter quelques *éléments finis mixtes* pour les équations de Stokes en dimension 2. Rappelons que dans ce cas, $V = H_0^1(\Omega)^2$, $M = L_0^2(\Omega)$ et

$$a(\mathbf{u},\mathbf{v}) = \nu \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega, \quad c(\mathbf{v},q) = -\int_\Omega q \operatorname{div}\mathbf{v}\, d\Omega \quad (\text{soit } C = -\operatorname{div}).$$

Notons que la condition (2.67) est vérifiée pour toute discrétisation interne V_h de V puisque $a(\cdot\,,\cdot)$ est coercive sur $H_0^1(\Omega)^2$. La question difficile est celle de la vérification de la condition inf–sup discrète uniforme.

Nous allons décrire quelques exemples d'espaces d'approximation $V_h \times M_h$ de $V \times M$ construits à partir d'une triangulation $\mathcal{T}_h$ de Ω, que l'on supposera polygonal (on peut ainsi le recouvrir *exactement* par une triangulation). L'opérateur C_h associé à une discrétisation $V_h \times M_h$ donnée s'interprète comme une *divergence discrète* :

$$\begin{aligned} &C_h = -\operatorname{div}_h \ : V_h \to M_h \\ &\text{où} \quad (\operatorname{div}_h \mathbf{v}_h, q_h)_{L^2(\Omega)} = (\operatorname{div}\mathbf{v}_h, q_h)_{L^2(\Omega)}, \quad \text{pour tout } (v_h, q_h) \in V_h \times M_h. \end{aligned}$$

Cette dernière égalité signifie que $\operatorname{div}_h = \Lambda_h \operatorname{div}$ où Λ_h désigne la projection de $L_0^2(\Omega)$ sur M_h (voir la remarque 2.11). Deux catégories de méthodes peuvent alors être envisagées.

– Si on veut que la solution approchée vérifie *exactement* la contrainte, il faut choisir V_h et M_h de sorte que $\operatorname{div}(V_h)$ soit contenu dans M_h. Ddans ce cas,

l'opérateur div_h coïncide avec (la restriction de) div). Une discrétisation continue (mais pas dérivable) de V conduit à des fonctions discontinues dans $\mathrm{div}(V_h)$: le champ de pression approché sera donc *discontinu*. En quelque sorte, on privilégie ici la contrainte au détriment de la régularité du multiplicateur.
– On peut faire le choix opposé en cherchant à approcher le champ de pression par une fonction *continue*. L'inclusion $\mathrm{div}(V_h) \subset M_h$ n'aura en général pas lieu, et $\mathbf{u}_h$ pourra ne plus vérifier exactement la contrainte.

Verrouillage numérique : l'élément fini P^1-P^0

A h fixé, vérifier l'inversibilité de C_h^T revient à vérifier que les espaces V_h et M_h sont *compatibles* avec la contrainte discrète, au sens où les $m_h = \dim M_h$ équations qui caractérisent la contrainte $C_h \mathbf{u}_h = 0$ sont linéairement indépendantes. Ceci n'est évidemment possible que si $\dim V_h \geq \dim M_h$ (il doit y avoir plus d'inconnues que d'équations). Ce n'est pas le cas de l'exemple suivant qui décrit une discrétisation des équations de Stokes qui semble *a priori* très naturelle.

Considérons le cas d'une discrétisation par éléments finis de Lagrange, où la vitesse est approchée par l'élément fini P^1 (les deux composantes de $\mathbf{u}_h$ sont globalement continues et affines par morceaux), et la pression par l'élément fini P^0 (p_h est constant par morceaux). On se donne une triangulation de Ω comportant
– L triangles (soit $\dim M_h = L - 1$, le " -1" étant dû au fait que les fonctions de M_h sont à moyenne nulle),
– N_i sommets situés à l'intérieur de Ω (soit $\dim V_h = 2\,N_i$),
– N_b sommets situés sur la frontière de Ω.
On peut montrer [15] qu'ils sont liés par la relation

$$L = 2\,N_i + N_b - 2.$$

Comme $N_b \geq 3$, on a $L - 1 \geq 2\,N_i$. En général, la seule fonction de V_h satisfaisant la contrainte $C_h \mathbf{u}_h = 0$ est donc 0, sauf si les équations qui traduisent cette contrainte ont le bon goût d'être linéairement dépendantes. Cette situation peut se produire lorsque le maillage respecte certaines conditions géométriques. Mais on perd dans ce cas l'unicité du multiplicateur p_h (qui est défini à une fonction de $\mathrm{Ker}\, C_h^T$ près). Ce constat est à l'origine de la notion de *macro-éléments finis* formés de groupes d'éléments finis usuels possédant les propriétés géométriques recherchées, comme par exemple l'élément fini "en croix", défini sur un quadrilatère, représenté sur la figure 2.1, et $(P^1)^2$ sur chaque triangle.

On peut vérifier facilement que pour un maillage associé à N_h macro-éléments finis de ce type, la dimension de $\mathrm{Ker}\, C_h^T$ est supérieure ou égale à $N_h - 1$. Les éléments de ce noyau sont appelés des *modes parasites*. Ils peuvent être vus ici comme des

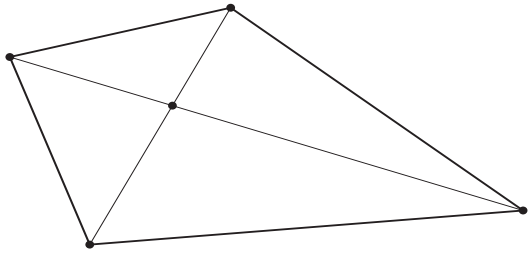


Figure 2.1. Un exemple de "macro-élément fini"

champs de pression q_h (non constants) dont le "gradient discret" s'annule, au sens où

$$\int_\Omega q_h \ \mathrm{div}\, \mathbf{v}_h \, d\Omega = 0 \quad \forall \mathbf{v}_h \in V_h.$$

Deux sortes de modes parasites peuvent être observés. Les modes *locaux* sont dus à la redondance (d'ordre 1) des 4 équations qui définissent la condition de divergence nulle sur chaque macro-élément fini. Les modes *globaux* apparaissent quant à eux pour des maillages *structurés* : ils sont liés aux propriétés globales de symétrie de tels maillages qu'on peut voir comme des "super-macro-éléments finis".

Le multiplicateur étant défini *modulo* ces modes parasites, le calcul du "bon" multiplicateur sera obtenu par un filtrage numérique destiné à les éliminer, les modes locaux étant beaucoup plus faciles à filtrer que les modes globaux. L'analyse théorique de ces macro-éléments finis nécessite une généralisation du résultat abstrait du théorème 2.12 (nous renvoyons à Brezzi et Fortin [10] pour les détails).

D'autres éléments finis mixtes font apparaître des *modes parasites* analogues à ceux de cet exemple. Parmi les éléments finis à pression discontinue, on peut citer l'élément fini quadrilatéral Q^1–P^0 pour lequel on observe des modes globaux en damiers (un champ de pression parasite peut prendre deux valeurs : l'une sur les cases blanches, l'autre sur les noires). Pour les éléments finis à pression continue, c'est le cas de tous ceux qui utilisent la même interpolation pour la vitesse et la pression : P^1–P^1, P^2–P^2, Q^1–Q^1, etc.

La question est alors : comment vérifier en pratique que pour tel ou tel élément fini mixte, ce phénomène ne peut se produire ? Le critère que nous allons énoncer ici sous une forme incomplète est dû à Fortin (voir la proposition 2.17 pour la forme complète). Comme nous le verrons dans les exemples qui suivent, pour des problèmes qui relèvent du théorème 2.1, il permet de répondre à la question en utilisant des propriétés *locales* des éléments finis.

Proposition 2.16 *On suppose que la condition inf–sup continue* (2.23) *est satisfaite. A h fixé, s'il existe un opérateur continu* $\pi_h : V \to V_h$ *tel que pour tout*

$v \in V$, on ait

$$c(v - \pi_h v, \mu_h) = 0 \quad \forall \mu_h \in M_h, \tag{2.73}$$

alors C_h est surjectif (autrement dit C_h^T est injectif).

Démonstration : Nous allons vérifier la condition inf–sup discrète uniforme (2.68), sans nous préoccuper de l'indépendance de $\tilde{\gamma}$ par rapport à h. D'après (2.73), on a

$$\sup_{v_h \in V_h} \frac{c(v_h, \mu_h)}{\|v_h\|_V} \geq \sup_{v \in V} \frac{c(\pi_h v, \mu_h)}{\|\pi_h v\|_V} = \sup_{v \in V} \frac{c(v, \mu_h)}{\|\pi_h v\|_V},$$

pour tout $\mu_h \in M_h$. La condition inf–sup continue (2.23) nous montre alors que

$$\sup_{v_h \in V_h} \frac{c(v_h, \mu_h)}{\|v_h\|_V} \geq \frac{\gamma}{K_h} \|\mu_h\|_M \quad \text{où} \quad K_h = \sup_{v \in V} \frac{\|\pi_h v\|_V}{\|v\|_V},$$

d'où le résultat. ∎

On peut récrire la condition (2.73) sous la forme

$$\Lambda_h C\,(v - \pi_h v) = 0, \tag{2.74}$$

où Λ_h désigne la projection orthogonale de M sur M_h (cf. remarque 2.11). En d'autres termes, $\pi_h v$ représente une *correction discrète* qu'il faut retrancher à un état continu $v \in V$ pour que la partie résiduelle satisfasse la contrainte discrète. Nous allons voir dans la suite comment construire un opérateur π_h pour certains éléments mixtes adaptés aux équations de Stokes.

Approximation continue de la pression : le MINI-élément fini P^1_{bul}-P^1

Pour éviter la présence de modes parasites ou le verrouillage numérique évoqués plus haut, l'idée naturelle consiste à stabiliser un élément fini défaillant (par exemple P^1–P^1) en *enrichissant* l'espace V_h, autrement dit en augmentant le nombre de degrés de liberté décrivant l'état du système sans modifier le nombre d'équations qui traduisent la contrainte que cet état doit satisfaire. L'exemple ci-dessous montre qu'on peut se contenter de rajouter un seul degré de liberté à l'élément fini usuel P^1 pour être dans le cadre de la proposition 2.16.

L'idée directrice de la construction de cet élément fini (et d'autres plus compliqués) est d'enrichir l'élément fini P^1 par une *fonction de base locale qui s'annule sur la frontière du triangle*. Ainsi, la présence de ce nouveau degré de liberté n'impose aucune condition supplémentaire de raccord entre deux triangles adjacents. Après *assemblage*, ce degré de liberté conserve son caractère *local* : il n'est couplé qu'aux autres degrés de liberté du triangle qui le contient.

Il est naturel de rechercher la fonction de base complémentaire sous la forme d'un polynôme, par exemple une fonction de base associée à un nœud intérieur d'un

élément fini P^k (défini sur un triangle K). De tels nœuds apparaissent à partir de l'ordre 3. Il y a dans ce cas un seul nœud intérieur, et la fonction de base qui lui est associée est la fonction bulle dont nous choisissons une normalisation particulière :

$$\tau_{\text{bul}} = \frac{\lambda_1\lambda_2\lambda_3}{\int_K \lambda_1\lambda_2\lambda_3}, \tag{2.75}$$

où les λ_i désignent les coordonnées barycentriques relativement aux trois sommets du triangle K. Nous noterons

$$P^1_{\text{bul}} = P^1 \oplus \text{Vect}\{\tau_{\text{bul}}\}$$

l'espace engendré par les polynômes λ_1, λ_2, λ_3 et τ_{bul}.

Pour compléter les 3 degrés de liberté de Lagrange habituels (définis par les masses de Dirac δ_i aux 3 sommets), on choisit un degré de liberté de type *moment* (d'ordre 0), associé à la fonction indicatrice 1_K du triangle K :

$$1_K(p) = \int_K p\, d\Omega.$$

Ceci définit bien un élément fini. Il est en effet facile de vérifier la propriété d'unisolvance

$$(\delta_i(p) = 0 \ i = 1,2,3 \ \text{ et } \ 1_K(p) = 0) \implies p = 0 \quad \forall p \in P^1_{\text{bul}}.$$

En fait, les fonctions de base locales de cet élément fini sont simplement données par

$$\begin{aligned} &\tau_i = \lambda_i - \left(\int_K \lambda_i\, d\Omega\right) \tau_{\text{bul}} \ \text{ pour } i = 1,2,3 \ \text{ et} \\ &\tau_4 = \tau_{\text{bul}}. \end{aligned}$$

A une triangulation $\mathcal{T}_h$ de Ω, on associe alors les deux espaces approchés

$$\begin{aligned} V_h &= \left\{ \mathbf{v}_h \in (\mathcal{C}^0(\overline{\Omega}) \cap H^1_0(\Omega))^2 \text{ tel que } \mathbf{v}_{h|K} \in P^1_{\text{bul}}(K)^2 \quad \forall K \in \mathcal{T}_h \right\}, \\ M_h &= \left\{ q_h \in \mathcal{C}^0(\overline{\Omega}) \cap L^2_0(\Omega) \text{ tel que } q_{h|K} \in P^1(K) \quad \forall K \in \mathcal{T}_h \right\}. \end{aligned}$$

Il s'agit de construire l'opérateur $\pi_h : V \to V_h$ de la proposition 2.16, qui doit satisfaire

$$\int_\Omega q_h \ \text{div}(\mathbf{v} - \pi_h\mathbf{v})\, d\Omega = 0 \quad \forall q_h \in M_h. \tag{2.76}$$

Les fonctions de M_h étant continues et régulières par morceaux, cette relation peut aussi s'écrire

$$\int_\Omega (\nabla q_h) \cdot (\mathbf{v} - \pi_h \mathbf{v})\, d\Omega = 0 \quad \forall q_h \in M_h.$$

Comme ∇q_h est constant par morceaux, il suffit que

$$\int_K (\mathbf{v} - \pi_h \mathbf{v})\, d\Omega = 0 \quad \forall K \in \mathcal{T}_h. \tag{2.77}$$

Ainsi, on peut choisir

$$(\pi_h \mathbf{v})_{|K} = \left(\int_K \mathbf{v}\, d\Omega \right) \tau_{\text{bul}} \quad \forall K \in \mathcal{T}_h, \tag{2.78}$$

ce qui définit bien une fonction de V_h qui s'annule sur toutes les arêtes de la triangulation. On voit ici que l'idée d'enrichir une discrétisation P^1 par des degrés de liberté qui conservent leur caractère local après assemblage, est liée au choix de M_h. Au contraire, en choisissant une discrétisation P^0 pour q_h, on ne peut plus interpréter localement (au sens de (2.77)) la contrainte discrète (2.76), puisque les q_h sont discontinus.

Une autre façon peut-être plus naturelle d'enrichir l'élément fini P^1 consiste simplement à "monter en ordre". Ceci conduit aux éléments finis mixtes de Taylor–Hood P^2–P^1, ou le quadrilatère Q^2–Q^1, dont l'analyse met en œuvre des techniques différentes de celles que nous présentons ici (cf. [27, 10]).

Approximation discontinue de la pression : l'élément fini P^2–P^0

Cette idée d'enrichir l'espace de discrétisation de la vitesse en augmentant l'ordre de l'élément permet de stabiliser l'élément fini P^1–P^0 présenté plus haut, en passant simplement de P^1 à P^2. Nous considérons donc le cas où les espaces associés à une triangulation $\mathcal{T}_h$ sont donnés par

$$\begin{aligned} V_h &= \left\{ \mathbf{v}_h \in (\mathcal{C}^0(\overline{\Omega}) \cap H_0^1(\Omega))^2 \text{ tel que } \mathbf{v}_{h|K} \in P^2(K)^2 \quad \forall K \in \mathcal{T}_h \right\}, \\ M_h &= \left\{ q_h \in L_0^2(\Omega) \text{ tel que } q_{h|K} \in P^0(K) \quad \forall K \in \mathcal{T}_h \right\}. \end{aligned}$$

Les fonctions q_h étant maintenant constantes par morceaux, la relation (2.76) s'écrit de façon équivalente

$$\int_K \operatorname{div}(\mathbf{v} - \pi_h \mathbf{v})\, d\Omega = 0 \quad \forall K \in \mathcal{T}_h,$$

soit encore

$$\int_{\partial K} (\mathbf{v} - \pi_h \mathbf{v}) \cdot \mathbf{n}\, d\Gamma = 0 \quad \forall K \in \mathcal{T}_h. \tag{2.79}$$

Nous allons construire π_h en imposant une propriété plus forte, à savoir :

$$\int_a (\mathbf{v} - \pi_h \mathbf{v})\, d\Gamma = 0 \quad \forall a \text{ arête de } \mathcal{T}_h.$$

Cette contrainte est plus forte, puisque d'une part les intégrales sont considérées pour toutes les arêtes de la triangulation, et d'autre part on intègre les composantes *normale et tangentielle* de $\mathbf{v} - \pi_h \mathbf{v}$. Notons $\tau_{ij} = 4\,\lambda_i \lambda_j$ la fonction de base locale associée au nœud central d'une arête a, où λ_i et λ_j sont les coordonnées barycentriques associées aux deux extrémités de l'arête. La relation ci-dessus est vérifiée si on choisit

$$(\pi_h \mathbf{v})_{|a} = \left(\frac{\int_a \mathbf{v}\, d\Gamma}{\int_a \tau_{ij}\, d\Gamma} \right) \tau_{ij}, \tag{2.80}$$

ce qui définit bien une fonction de V_h (qui s'annule cette fois sur tous les sommets des triangles, mais pas sur les arêtes). La proposition 2.16 s'applique.

Notons que la même démarche permet de traiter le cas de l'élément fini quadrilatéral Q^2–P^0.

Pour l'élément fini P^2–P^0, nous avons choisi l'opérateur π_h en résolvant une équation plus forte que (2.79). Nous disposions en effet de 2 degrés de liberté par arête. Mais il est clair qu'un seul suffirait à assurer la relation (2.79) : on peut se contenter de la *composante normale* de la vitesse au milieu de chaque arête. Ceci revient à construire un élément fini pour lequel la composante normale de la vitesse sur chaque arête est décrite par 3 degrés de liberté (approximation P^2), et la composante tangentielle, par 2 (soit P^1). Il est facile de construire un sous-espace $\mathcal{P}$ de $(P^2)^2$ qui satisfasse cette propriété. Notons

$$\mathbf{p}_1 = \lambda_2 \lambda_3 \mathbf{n}_1,\ \mathbf{p}_2 = \lambda_1 \lambda_3 \mathbf{n}_2,\ \mathbf{p}_3 = \lambda_1 \lambda_2 \mathbf{n}_3, \tag{2.81}$$

où les λ_i sont comme précédemment les coordonnées barycentriques, et $\mathbf{n}_j$ désigne la normale à l'arête a_j opposée au sommet M_j (voir figure 2.2). L'espace

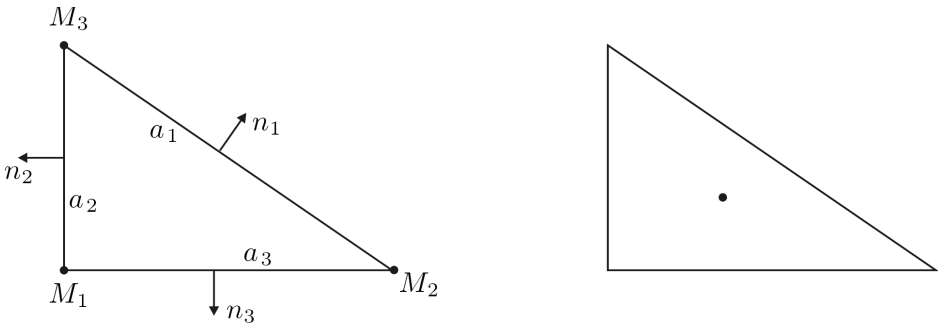


Figure 2.2. L'élément fini P^2–P^0 condensé

$$\mathcal{P} = (P^1)^2 \oplus \mathrm{Vect}\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$$

répond à la question. Il reste à compléter les 6 degrés de liberté de Lagrange (valeurs nodales aux 3 sommets) par 3 degrés de liberté à support sur chaque côté du triangle, de façon à imposer le raccord de la composante normale entre deux triangles adjacents. Plutôt que de choisir la composante normale au milieu de l'arête, on peut définir un degré de liberté de type "flux" (analogue au degré de liberté de type "moment" utilisé pour l'élément fini P^1_{bul}–P^1) :

$$\mathbf{p} \to \int_a \mathbf{p}\cdot\mathbf{n}\, d\Gamma.$$

On vérifie sans difficulté qu'avec ces 9 degrés de liberté, on a bien la propriété d'unisolvance des éléments finis.

Pour construire l'opérateur π_h, il nous reste à remarquer que si

$$\int_a (\mathbf{v}-\pi_h\mathbf{v})\cdot\mathbf{n}\, d\Gamma = 0 \quad \forall a \text{ arête de } \mathcal{T}_h,$$

la relation (2.79) est *a fortiori* satisfaite. On peut donc prendre

$$(\pi_h\mathbf{v})_{|a} = \left(\frac{\int_a \mathbf{v}\cdot\mathbf{n}\, d\Gamma}{\int_a \mathbf{p}_a\cdot\mathbf{n}\, d\Gamma}\right)\mathbf{p}_a, \tag{2.82}$$

où $\mathbf{p}_a$ désigne la fonction de type (2.81) associée à l'arête a. On note que cette expression ne dépend pas de l'orientation de la normale.

Ce nouvel élément fini possède le même taux de convergence que l'élément fini original P^2–P^0 (voir le paragraphe suivant) : par rapport à celui-ci, il a l'avantage de diminuer le nombre de degrés de libertés.

L'inconvénient d'une discrétisation P^0 pour la pression est de faire perdre le taux de convergence optimal de l'élément fini P^2 pour la vitesse. On peut retrouver ce taux en enrichissant aussi l'espace M_h. C'est le cas de l'élément fini de Crouzeix–Raviart P^2_{bul}–P^1_{disc} où $P^2_{\text{bul}} = P^2 \oplus \text{Vect}\{\tau_{\text{bul}}\}$ (cf. (2.75)) et P^1_{disc} désigne un élément fini affine associé à trois nœuds intérieurs (pour lequel il n'y a aucune condition de raccord à l'interface entre deux triangles). L'équivalent quadrilatéral de ce dernier est l'élément fini Q^2–P^1_{quad} où les 3 degrés de liberté de l'élément fini P^1_{quad} sont les valeurs de la fonction et de ses deux dérivées au centre du quadrilatère (il n'y a donc toujours pas de condition de raccord). L'analyse théorique de ces éléments finis entre dans le cadre de l'approche que nous développons ici.

2.2.3 Uniformité de la condition inf–sup discrète

Un critère général

Nous abordons maintenant le second volet de la question de la vérification de la condition (2.68) : comment peut-on affirmer que $\tilde{\gamma}$ est indépendant de h ? Le critère

proposé dans la proposition 2.16 joint à une hypothèse d'uniformité sur π_h nous fournit en fait une condition nécessaire et suffisante, comme l'exprime la

Proposition 2.17 *On suppose que la condition inf–sup continue* (2.23) *est satisfaite. La condition inf–sup discrète uniforme* (2.68) *est satisfaite si, et seulement si, il existe un opérateur* $\Pi_h : V \to V_h$ *tel que pour tout* $v \in V$, *on ait*

$$c(v - \Pi_h v, \mu_h) = 0 \quad \forall \mu_h \in M_h, \tag{2.83}$$

$$\|\Pi_h v\|_V \leq K \|v\|_V, \tag{2.84}$$

où $K > 0$ *est indépendant de* h.

Démonstration : Les conditions (2.83)–(2.84) sont suffisantes : il suffit de remarquer que dans la démonstration de la proposition 2.16, la quantité

$$K_h = \sup_{v \in V} \frac{\|\Pi_h v\|_V}{\|v\|_V}$$

est maintenant bornée par la constante K de l'inégalité (2.84).
Réciproquement, supposons que la condition inf–sup discrète uniforme soit satisfaite. D'après (2.74), la condition (2.83) peut aussi s'écrire

$$C_h \Pi_h v = \Lambda_h C\, v.$$

Mais la condition inf–sup discrète uniforme nous dit que C_h est une bijection de $(\operatorname{Ker} C_h)^\perp$ dans M_h, d'inverse continu et de norme uniformément bornée par rapport à h. Plus précisément (voir le lemme 2.2),

$$\left\|C_h^{-1} \mu_h\right\|_V \leq \tilde{\gamma}^{-1} \|\mu_h\|_M \quad \forall \mu_h \in M_h.$$

On peut donc choisir

$$\Pi_h = C_h^{-1}\, \Lambda_h\, C,$$

et on aura dans ce cas

$$\|\Pi_h v\|_V \leq \tilde{\gamma}^{-1} m_c \|v\|_V,$$

où m_c est la constante de continuité associée à la forme bilinéaire $c(\cdot\,,\cdot)$, (cf. (2.11)), soit l'inégalité (2.84) avec $K = \tilde{\gamma}^{-1} m_c$. ∎

On pourrait chercher à appliquer directement la proposition 2.17 à l'opérateur π_h construit dans l'un des exemples du paragraphe précédent (voir (2.78), (2.80) ou (2.82)). Aucun ne satisfait la condition (2.84) ! Ceci est dû au fait que dans tous les cas, $\operatorname{Im} \pi_h$ est engendré par les fonctions de base globales associées aux degrés de liberté qui ont enrichi l'élément fini original, à savoir P^1. Lorsque h tend vers 0, ces fonctions deviennent de plus en plus irrégulières et $\|\pi_h v\|_V$ va généralement tendre vers l'infini pour $v \in V$ donné.

L'idée consiste à construire un opérateur Π_h à partir de π_h en introduisant un opérateur d'interpolation $r_h : V \to V_h$ portant sur les degrés de liberté associés à l'élément fini original. On peut alors se contenter d'examiner l'action de π_h sur la partie résiduelle $v - r_h v$ d'un état, comme l'exprime le

Lemme 2.18 *Soient r_h et π_h deux opérateurs de V dans V_h tels qu'il existe deux constantes positives K_1 et K_2 telles que pour tout $v \in V$, on ait*

$$\|r_h v\|_V \leq K_1 \|v\|_V, \tag{2.85}$$
$$c(v - \pi_h v, \mu_h) = 0 \quad \forall \mu_h \in M_h, \tag{2.86}$$
$$\|\pi_h(v - r_h v)\|_V \leq K_2 \|v\|_V. \tag{2.87}$$

Alors l'opérateur

$$\Pi_h = r_h + \pi_h(\mathrm{Id} - r_h)$$

satisfait les deux conditions (2.83) *et* (2.84)*.*

Démonstration : La définition de Π_h et la condition (2.86) nous montrent que

$$\begin{aligned} c(\Pi_h v, \mu_h) &= c(r_h v, \mu_h) + c(\pi_h(v - r_h v), \mu_h), \\ &= c(r_h v, \mu_h) + c(v - r_h v, \mu_h), \\ &= c(v, \mu_h), \end{aligned}$$

d'où (2.83). Pour obtenir (2.84), il suffit de constater que

$$\|\Pi_h v\|_V \leq \|r_h v\|_V + \|\pi_h(v - r_h v)\|_V ,$$

d'après l'inégalité triangulaire. On peut donc prendre $K = K_1 + K_2$. ∎

L'opérateur d'interpolation

Dans [15], nous avons introduit la notion d'*interpolé* d'une fonction continue : c'est l'unique fonction de V_h qui prend les mêmes valeurs que v aux nœuds du maillage. Dans le cadre des équations de Stokes, cette opération d'interpolation ne peut satisfaire une inégalité du type (2.85) puisque r_h n'est pas définie sur V tout entier (les fonctions de $H^1(\Omega)$ ne sont pas toutes continues, lorsque $\Omega \subset \mathbb{R}^n$, pour $n \geq 2$). Clément [17] a construit un autre opérateur d'interpolation r_h en utilisant des moyennes locales de v au lieu de valeurs ponctuelles : celui-ci a l'avantage d'être continu sur $H^1(\Omega)$. Clément a obtenu une estimation d'erreur d'interpolation locale analogue à celles obtenues classiquement (cf. [15]). En revanche, elle n'est pas à proprement parler *locale*, puisqu'elle fait intervenir le comportement de la fonction sur l'ensemble vois(K) des triangles adjacents à $K \in \mathcal{T}_h$. Plus précisément, on a la

Proposition 2.19 *Soit (K, Σ, P) un élément fini de Lagrange d'ordre k affinement équivalent à un élément de référence $(\hat{K}, \hat{\Sigma}, \hat{P})$. Il existe une constante $C > 0$ qui ne dépend que de $(\hat{K}, \hat{\Sigma}, \hat{P})$ telle que*

$$|v - r_h v|_{m,K} \leq C \, (\sigma_{\mathrm{vois}(K)})^m \, (h_{\mathrm{vois}(K)})^{n-m} \, |v|_{n,\mathrm{vois}(K)} \quad \forall v \in H^n(\Omega),$$
$$\text{avec } 0 \leq m \leq n \quad \text{et} \quad 1 \leq n \leq k+1,$$

où $h_{\mathrm{vois}(K)}$ est le rayon du plus grand cercle circonscrit à l'un des triangles de $\mathrm{vois}(K)$, *et $\sigma_{\mathrm{vois}(K)}$ est le plus grand rapport d'applatissement de ces triangles (rapport du rayon du cercle circonscrit par celui du cercle inscrit).*

Si on suppose que la famille de triangulations $(\mathcal{T}_h)_h$ est régulière, alors en particulier $\sigma_{\mathrm{vois}(K)}$ reste borné et $h_{\mathrm{vois}(K)} \sim h_K$, uniformément sur h et K. Le résultat précédent nous montre que, pour tout $v \in H^1(\Omega)$ et pour tout K,

$$\|v - r_h v\|_{L^2(K)} \leq C\, h_K\, |v|_{1,\mathrm{vois}(K)}, \quad |v - r_h v|_{1,K} \leq C\, |v|_{1,\mathrm{vois}(K)}. \tag{2.88}$$

Soit, en sommant sur tous les triangles K de $\mathcal{T}_h$, puisque $\mathbf{v} \in V = H^1_0(\Omega)^2$,

$$\|\mathbf{v} - r_h \mathbf{v}\|_{L^2(\Omega)^2} \leq C'\, h\, \|\mathbf{v}\|_V, \quad \|\mathbf{v} - r_h \mathbf{v}\|_V \leq C'\, \|\mathbf{v}\|_V \quad \forall \mathbf{v} \in V. \tag{2.89}$$

Ainsi, *l'opérateur d'interpolation de Clément satisfait la condition* (2.85) (avec $K_1 = 1 + C'$).

L'opérateur de correction

Dans tous les exemples du paragraphe précédent, on peut retenir l'opérateur de Clément pour appliquer le lemme 2.18. Il reste donc à vérifier au cas par cas que l'*opérateur de correction* π_h satisfait la condition (2.87). En pratique, il suffit de montrer une estimation locale du type

$$\|\pi_h \mathbf{v}\|_{H^1(K)} \leq C''\, \left(h_K^{-1} \|\mathbf{v}\|_{L^2(K)^2} + |\mathbf{v}|_{1,K}\right) \quad \forall K \in \mathcal{T}_h. \tag{2.90}$$

En effet, ceci combiné avec (2.88), nous donne

$$\|\pi_h(\mathbf{v} - r_h \mathbf{v})\|_{H^1(K)} \leq C\, C''\, |\mathbf{v}|_{1,\mathrm{vois}(K)}.$$

En sommant sur K on établit ainsi (2.87).

Nous nous contentons de montrer comment obtenir (2.90) dans le cas de l'élément fini P^1_{bul}-P^1, les autres pouvant être traités de façon analogue. Dans ce cas, l'opérateur π_h défini par (2.78) vérifie

$$\|\pi_h \mathbf{v}\|_{H^1(K)} \leq \|\tau_{\mathrm{bul}}\|_{H^1(K)} \left| \int_K \mathbf{v}\, d\Omega \right|.$$

Or, d'après l'inégalité de Cauchy–Schwarz,

$$\left| \int_K \mathbf{v}\, d\Omega \right| \leq \sqrt{\pi}\; h_K\; \|\mathbf{v}\|_{L^2(K)^2}.$$

Par ailleurs, comme $\int_K \lambda_1 \lambda_2 \lambda_3\, d\Omega \sim h_K^2$, on en déduit immédiatement que

$$\|\tau_{\mathrm{bul}}\|_{H^1(K)} \leq C'''\; h_K^{-2}.$$

L'estimation (2.90) est donc vérifiée dans ce cas.

Vitesse de convergence

Les estimations locales du type de celles de la proposition 2.19 jointes au théorème 2.12 nous fournissent l'ordre de convergence d'une méthode d'éléments finis mixtes. La démarche étant rigoureusement la même, nous indiquons seulement le résultat qu'on obtient pour les éléments finis $P^1_{\rm bul}$–P^1, P^2–P^0 et P^2–P^0 condensé, lorsque la solution $(\mathbf{u}, p)$ du problème continu est supposée appartenir à $H^2(\Omega)^2 \times H^1(\Omega)$:

$$\|\mathbf{u} - \mathbf{u}_h\|_V + \|p - p_h\|_M \leq C\, h\, \left(\|\mathbf{u}\|_{H^2(\Omega)^2} + \|p\|_{H^1(\Omega)}\right).$$

Notons pour terminer que la technique présentée dans ce paragraphe peut être itérée pour traiter des éléments finis d'ordre supérieur. Par exemple, pour l'élément fini de Crouzeix–Raviart $P^2_{\rm bul}$–$P^1_{\rm disc}$ évoqué dans le paragraphe précédent, on peut utiliser comme opérateur d'interpolation r_h l'opérateur $\mathit{\Pi}_h$ construit par le lemme 2.18 pour l'élément fini P^2–P^0. En procédant comme pour l'élément fini $P^1_{\rm bul}$–P^1, on définit alors un nouvel opérateur π_h ne faisant intervenir que les fonctions bulles sur les triangles. Et on peut alors à nouveau appliquer le lemme 2.18...

2.2.4 Résolution des problèmes approchés

Nous nous intéressons maintenant à la reformulation des problèmes approchés sous la forme d'un système linéaire, puis à des moyens de résoudre celui-ci.

Le système linéaire à résoudre

Jusqu'à présent, le problème discret a été écrit soit sous forme variationnelle (2.62)–(2.63), soit à l'aide d'opérateurs discrets (2.64). Pour pouvoir le résoudre, nous allons maintenant le reformuler sous la forme d'un système linéaire. Ci-après, la forme $\ell_M(\cdot)$ peut être non-nulle (cf. le contexte de la remarque 2.15). Pour cela, nous considérons une base $(w_i)_{i=1,N}$ de V_h, ainsi qu'une base $(\theta_k)_{k=1,Q}$ de M_h. On a alors le résultat suivant.

Lemme 2.20 *Le problème* (2.62)–(2.63) *est équivalent au système linéaire :*

$$\left\{ \begin{array}{l} \text{trouver } (\overrightarrow{U}, \overrightarrow{\mathit{\Lambda}}) \in \mathbb{R}^N \times \mathbb{R}^Q \text{ tel que :} \\ \begin{pmatrix} \mathbb{A} & \mathbb{C}^T \\ \mathbb{C} & 0 \end{pmatrix} \begin{pmatrix} \overrightarrow{U} \\ \overrightarrow{\mathit{\Lambda}} \end{pmatrix} = \begin{pmatrix} \overrightarrow{F} \\ \overrightarrow{G} \end{pmatrix} \end{array} \right. \tag{2.91}$$

où les matrices $\mathbb{A}$ *et* $\mathbb{C}$ *et les vecteurs* $\overrightarrow{F}$ *et* $\overrightarrow{G}$ *sont respectivement définis par :*

$$\mathbb{A}_{IJ} = a(w_J, w_I),\ \mathbb{C}_{KJ} = c(w_J, \theta_K),\ F_I = \ell_V(w_I),\ G_K = \ell_M(\theta_K),$$

pour $\quad 1 \leq I, J \leq N,\ 1 \leq K \leq Q.$

Démonstration : On cherche (u_h, λ_h) solution de (2.62)–(2.63) sous la forme :

$$u_h = \sum_{J=1}^{N} U^J w_J, \quad \lambda_h = \sum_{L=1}^{Q} \Lambda^L \theta_L.$$

Par linéarité, le problème (2.62)–(2.63) s'écrit de façon équivalente :

$$\sum_{J=1}^{N} a(w_J, w_I) U^J + \sum_{L=1}^{Q} c(w_I, \theta_L) \Lambda^L = \ell_V(w_I),\ 1 \le I \le N,$$
$$\sum_{J=1}^{N} c(w_J, \theta_K) U^J = \ell_M(\theta_K),\ 1 \le K \le Q,$$

c'est-à-dire (2.91). ∎

Difficultés

Evidemment, dès lors que le problème discret (2.62)–(2.63) admet une solution et une seule (u_h, λ_h), le système linéaire (2.91) admet une solution et une seule $(\overrightarrow{U}, \overrightarrow{\Lambda})$. La question pratique qui se pose est : comment résoudre le système linéaire ?

Même si la forme $a(\cdot,\cdot)$ est coercive sur $V \times V$, la matrice de (2.91) n'a aucune raison d'être définie-positive ou définie-négative. En effet, on a :

$$\left(\begin{array}{c|c} \mathbb{A}\overrightarrow{V} + \mathbb{C}^T\overrightarrow{\Sigma} & \overrightarrow{V} \\ \mathbb{C}\overrightarrow{V} & \overrightarrow{\Sigma} \end{array} \right)_{\mathbb{R}^{N+Q}} = (\mathbb{A}\overrightarrow{V}\,|\,\overrightarrow{V})_{\mathbb{R}^N} + 2(\mathbb{C}\overrightarrow{V}\,|\,\overrightarrow{\Sigma})_{\mathbb{R}^Q},\ \forall (\overrightarrow{V}, \overrightarrow{\Sigma}) \in \mathbb{R}^N \times \mathbb{R}^Q.$$

Et, pour $\overrightarrow{V}$ donné tel que $\mathbb{C}\overrightarrow{V} \neq 0$, choisissons un vecteur $\overrightarrow{\Sigma}_0$ non-orthogonal à $\mathbb{C}\overrightarrow{V}$ et posons $\overrightarrow{\Sigma} = \alpha\overrightarrow{\Sigma}_0$: le produit scalaire précédent est alors une fonction affine (et non-constante) de α. Tout au plus la matrice de (2.91) est-elle symétrique, sous réserve que la forme $a(\cdot,\cdot)$ le soit.
Parmi les méthodes générales de résolution, seule celle de Gauss avec pivotage partiel ou total [13, 15] peut donc être appliquée pour résoudre le système linéaire (2.91).

Ou bien, on peut tirer profit de la structure du système linéaire. Celui-ci s'écrit :

$$\left\{ \begin{array}{lcl} \mathbb{A}\overrightarrow{U} + \mathbb{C}^T\overrightarrow{\Lambda} & = & \overrightarrow{F} \\ \mathbb{C}\overrightarrow{U} & = & \overrightarrow{G}. \end{array} \right. \tag{2.92}$$

On peut alors procéder à une résolution en deux temps, dans un cadre important en pratique (qui inclut par exemple le cas des équations de Stokes) : on va supposer que la forme bilinéaire continue $a(\cdot,\cdot)$ est coercive et symétrique sur $V \times V$. Au niveau discret, ceci entraîne en particulier que la matrice $\mathbb{A}$ est (symétrique) définie-positive, donc inversible. Pour résoudre (2.92) dans ce cadre, on note que l'on a

$\overrightarrow{U} = \mathbb{A}^{-1}(\overrightarrow{F} - \mathbb{C}^T\Lambda)$ d'après la première équation ; on peut ensuite remplacer la valeur de $\overrightarrow{U}$ dans la seconde équation, pour exprimer $\overrightarrow{\Lambda}$ en fonction des données $\overrightarrow{F}$ et $\overrightarrow{G}$. Ceci permet donc de proposer une résolution en deux temps, à savoir :

$$(\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)\overrightarrow{\Lambda} = \mathbb{C}\mathbb{A}^{-1}\overrightarrow{F} - \overrightarrow{G}, \quad \text{puis} \tag{2.93}$$

$$\mathbb{A}\overrightarrow{U} = \overrightarrow{F} - \mathbb{C}^T\overrightarrow{\Lambda}. \tag{2.94}$$

Supposons de plus que la condition inf–sup discrète (2.68) est vraie [3], sans toutefois imposer pour l'instant son uniformité en h ($\tilde{\gamma} > 0$ peut dépendre de h). Ceci signifie simplement (cf. remarque 2.13) que $\mathrm{Ker}\,\mathbb{C}^T = \{\overrightarrow{0}\}$. On en déduit aisément que la matrice $(\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)$ est (symétrique) définie-positive.
Résoudre (2.93)–(2.94) revient alors à résoudre deux systèmes linéaires mettant en jeu des matrices symétriques définies-positives, pour lesquelles de nombreuses méthodes de résolution existent [13, 15]. La difficulté est qu'on ne connaît pas explicitement la première matrice, $(\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)$.

On peut néanmoins envisager une méthode itérative de résolution pour calculer la solution de (2.93), qui requiert à chaque itération de calculer un produit matrice-vecteur de la forme $\overrightarrow{\Sigma}' = (\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)\overrightarrow{\Sigma}$, que l'on peut remplacer par les trois étapes : calculer $\overrightarrow{V} = \mathbb{C}^T\overrightarrow{\Sigma}$; déterminer $\overrightarrow{W} \in \mathbb{R}^N$ tel que $\mathbb{A}\overrightarrow{W} = \overrightarrow{V}$; calculer $\overrightarrow{\Sigma}' = \mathbb{C}\overrightarrow{W}$.
Dans ce cas, sous réserve de disposer d'un solveur pour déterminer la solution d'un système linéaire de matrice $\mathbb{A}$, on peut espérer déterminer $\overrightarrow{\Lambda}$, puis $\overrightarrow{U}$.
Mais il faut encore que l'algorithme itératif converge !

Une solution

Nous indiquons pour finir un algorithme itératif efficace pour calculer $\overrightarrow{\Lambda}$, toujours sous les deux hypothèses :
– la forme $a(\cdot,\cdot)$ est coercive et symétrique ;
– la condition inf–sup discrète est vraie, $\tilde{\gamma}$ pouvant dépendre de h dans (2.68).
La première hypothèse revient à dire que l'on peut remplacer le produit scalaire de V par $a(\cdot,\cdot)$, et la norme de V par $\|\cdot\|_a : v \mapsto \|v\|_a = (a(v,v))^{1/2}$.

Supposons que l'on utilise une méthode de type d'Uzawa (voir §2.1.4), qui consiste en une méthode de gradient appliquée à (2.93). Ci-dessous, nous choisissons la méthode du gradient conjugué préconditionné [40, 15], avec pour préconditionnement l'inverse de la matrice de masse associée au multiplicateur, $\mathbb{M} \in \mathbb{R}^{Q\times Q}$ telle que $\mathbb{M}_{KL} = (\theta_L, \theta_K)_M$, $1 \le K, L \le Q$: on résout donc un système linéaire de matrice $\mathbb{P} = \mathbb{M}^{-1}(\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)$.

3. Comme $a(\cdot,\cdot)$ est coercive, la coercivité uniforme sur le noyau discret (2.67) est automatiquement vérifiée.

On rappelle que pour un algorithme de type gradient conjugué, la réduction de l'erreur d'une itération à l'autre (selon une norme bien choisie) est majorée par $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, où κ est par définition le nombre de conditionnement du système préconditionné, égal à $\kappa = \|\mathbb{P}^{-1}\|\,\|\mathbb{P}\|$ ($\kappa \geq 1$), avec $\|\cdot\|$ la norme matricielle induite associée ici au produit scalaire $(\cdot|\cdot)_{\mathbb{R}^Q}$. Hormis les problèmes de stabilité numérique, cet algorithme doit converger puisque $0 \leq (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1) < 1$.
Pour estimer la vitesse de convergence, il faut connaître l'évolution de ce nombre de conditionnement κ en fonction de la discrétisation, c'est-à-dire en fonction de h. La situation la plus favorable se produit lorsque $\kappa = \kappa(h)$ est *majoré indépendamment de* h. En effet, dans ce cas, le nombre d'itérations pour atteindre une tolérance donnée est indépendant de h. Or, dans la situation présente, $\mathbb{M}$ et $(\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)$ sont deux matrices symétriques définies-positives. On peut vérifier à l'aide du principe du Min-Max que

$$\kappa = \frac{\nu_{max}}{\nu_{min}}, \text{ avec } \left\{ \begin{array}{l} \nu_{min} \text{ plus petite} \\ \nu_{max} \text{ plus grande} \end{array} \right. \text{valeur propre de } (\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)\overrightarrow{\Sigma} = \nu\mathbb{M}\overrightarrow{\Sigma}. \tag{2.95}$$

Il reste donc à étudier ces plus grande et plus petite valeurs propres.

Proposition 2.21 *Soient*

$$\gamma_\star(h) = \inf_{\sigma_h \in M_h} \sup_{v_h \in V_h} \frac{c(v_h,\sigma_h)}{\|v_h\|_a \|\sigma_h\|_M}, \quad \gamma^\star(h) = \sup_{\sigma_h \in M_h} \sup_{v_h \in V_h} \frac{c(v_h,\sigma_h)}{\|v_h\|_a \|\sigma_h\|_M},$$

alors $\nu_{min} = (\gamma_\star(h))^2$ *et* $\nu_{max} = (\gamma^\star(h))^2$.

Démonstration : Soit $\sigma_h \neq 0$ donné. Commençons par calculer

$$s(\sigma_h) = \sup_{v_h \in V_h} \frac{c(v_h,\sigma_h)}{\|v_h\|_a} = \sup_{v_h \in V_h, \|v_h\|_a = 1} c(v_h,\sigma_h),$$

par linéarité de $c(\cdot,\cdot)$ par rapport à son premier argument. Comme $c(\cdot,\cdot)$ est continue et comme $\{v_h \in V_h;\ \|v_h\|_a = 1\}$ est compact, le maximum est atteint (en v_h^{opt}), ce qui donne un sens à $s(\sigma_h)$. Si on note $\overrightarrow{V} \in \mathbb{R}^N$ et $\overrightarrow{\Sigma} \in \mathbb{R}^Q$ les vecteurs respectivement associés à v_h et σ_h, on peut écrire de façon équivalente que

$$s(\sigma_h) = \sup_{\overrightarrow{V} \in \mathbb{R}^N, (\mathbb{A}\overrightarrow{V}|\overrightarrow{V})_{\mathbb{R}^N} = 1} (\mathbb{C}\overrightarrow{V}|\overrightarrow{\Sigma})_{\mathbb{R}^Q}.$$

Du point de vue de l'optimisation (§2.1.4), c'est un problème de maximisation avec contraintes. On introduit le lagrangien $\mathcal{L}$ défini sur $\mathbb{R}^N \times \mathbb{R}$ par

$$\mathcal{L}(\overrightarrow{V},\mu) = (\mathbb{C}\overrightarrow{V}|\overrightarrow{\Sigma})_{\mathbb{R}^Q} - \mu\left((\mathbb{A}\overrightarrow{V}|\overrightarrow{V})_{\mathbb{R}^N} - 1\right).$$

Les conditions nécessaires d'optimalité sont

$$\begin{aligned} 0 &= d_{\overrightarrow{V}}\mathcal{L}(\overrightarrow{V}^{opt},\mu^{opt}).\overrightarrow{V} = (\mathbb{C}\overrightarrow{V}|\overrightarrow{\Sigma})_{\mathbb{R}^Q} - 2\mu^{opt}(\mathbb{A}\overrightarrow{V}^{opt}|\overrightarrow{V})_{\mathbb{R}^N} \quad \forall \overrightarrow{V} \in \mathbb{R}^N, \\ 0 &= d_\mu\mathcal{L}(\overrightarrow{V}^{opt},\mu^{opt}).\mu = (\mathbb{A}\overrightarrow{V}^{opt}|\overrightarrow{V}^{opt})_{\mathbb{R}^N} - 1 \quad \forall \mu \in \mathbb{R}, \end{aligned}$$

Ce qui correspond à

$$(\mathbb{A}\overrightarrow{V}^{opt}|\overrightarrow{V}^{opt})_{\mathbb{R}^N} = 1 \text{ et } \mathbb{C}^T\overrightarrow{\Sigma} = 2\mu^{opt}\mathbb{A}\overrightarrow{V}^{opt}.$$

Comme par hypothèse $\mathrm{Ker}\,\mathbb{C}^T = \{\overrightarrow{0}\}$, on sait que $\mathbb{C}^T\overrightarrow{\Sigma} \neq \overrightarrow{0}$ et donc $\mu^{opt} \neq 0$, ce qui permet d'écrire que $\overrightarrow{V}^{opt} = 1/(2\mu^{opt})\,\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma}$. De la contrainte $(\mathbb{A}\overrightarrow{V}^{opt}|\overrightarrow{V}^{opt})_{\mathbb{R}^N} = 1$, on tire ensuite la valeur de $|\mu^{opt}|$, d'expression

$$|\mu^{opt}| = \frac{1}{2}\left\{(\mathbb{C}^T\overrightarrow{\Sigma}\Big|\ \mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N}\right\}^{1/2}.$$

On en conclut finalement que

$$s(\sigma_h) = \left\{(\mathbb{C}^T\overrightarrow{\Sigma}\Big|\ \mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N}\right\}^{1/2}.$$

A partir de là, on peut calculer $\gamma_\star(h)$ et $\gamma^\star(h)$ en résolvant un second problème d'optimisation avec contrainte. En effet, on a

$$\begin{aligned}(\gamma_\star(h))^2 &= \inf_{\overrightarrow{\Sigma}\in\mathbb{R}^Q,\ (\mathbb{M}\overrightarrow{\Sigma}|\overrightarrow{\Sigma})_{\mathbb{R}^Q}=1} (\mathbb{C}^T\overrightarrow{\Sigma}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N},\\ (\gamma^\star(h))^2 &= \sup_{\overrightarrow{\Sigma}\in\mathbb{R}^Q,\ (\mathbb{M}\overrightarrow{\Sigma}|\overrightarrow{\Sigma})_{\mathbb{R}^Q}=1} (\mathbb{C}^T\overrightarrow{\Sigma}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N}.\end{aligned}$$

Comme précédemment, la continuité de la fonction $\overrightarrow{\Sigma} \mapsto (\mathbb{C}^T\overrightarrow{\Sigma}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N}$ associée au caractère compact de l'ensemble $\{\overrightarrow{\Sigma} \in \mathbb{R}^Q;\ (\mathbb{M}\overrightarrow{\Sigma}|\overrightarrow{\Sigma})_{\mathbb{R}^Q} = 1\}$ permet de conclure à l'existence d'extrema – minimum et maximum – atteints (en $\overrightarrow{\Sigma}^{ext}$). On considère cette fois le lagrangien $\mathcal{L}'$ défini sur $\mathbb{R}^Q \times \mathbb{R}$ par

$$\mathcal{L}'(\overrightarrow{\Sigma},\nu) = (\mathbb{C}^T\overrightarrow{\Sigma}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N} - \nu\left((\mathbb{M}\overrightarrow{\Sigma}|\overrightarrow{\Sigma})_{\mathbb{R}^Q} - 1\right).$$

Les conditions nécessaires d'optimalité sont

$$\begin{aligned}0 &= d_{\overrightarrow{\Sigma}}\mathcal{L}'(\overrightarrow{\Sigma}^{ext},\nu^{ext}).\overrightarrow{\Sigma} = 2(\mathbb{C}^T\overrightarrow{\Sigma}^{ext}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma})_{\mathbb{R}^N} - 2\nu^{ext}(\mathbb{M}\overrightarrow{\Sigma}^{ext}|\overrightarrow{\Sigma})_{\mathbb{R}^Q} \quad \forall\overrightarrow{\Sigma}\in\mathbb{R}^Q,\\ 0 &= d_\nu\mathcal{L}'(\overrightarrow{\Sigma}^{ext},\nu^{ext}).\nu = (\mathbb{M}\overrightarrow{\Sigma}^{ext}|\overrightarrow{\Sigma}^{ext})_{\mathbb{R}^Q} - 1 \quad \forall\nu\in\mathbb{R},\end{aligned}$$

soit

$$(\mathbb{M}\overrightarrow{\Sigma}^{ext}|\overrightarrow{\Sigma}^{ext})_{\mathbb{R}^Q} = 1 \text{ et } (\mathbb{C}\mathbb{A}^{-1}\mathbb{C}^T)\overrightarrow{\Sigma}^{ext} = \nu^{ext}\mathbb{M}\overrightarrow{\Sigma}^{ext}.$$

En particulier, ν^{ext} est valeur propre de (2.95). Par ailleurs, on a

$$(\mathbb{C}^T\overrightarrow{\Sigma}^{ext}|\mathbb{A}^{-1}\mathbb{C}^T\overrightarrow{\Sigma}^{ext})_{\mathbb{R}^N} = \nu^{ext}(\mathbb{M}\overrightarrow{\Sigma}^{ext}|\overrightarrow{\Sigma}^{ext})_{\mathbb{R}^Q} = \nu^{ext}.$$

Ainsi, $(\gamma_\star(h))^2 = \nu_{min}$ et $(\gamma^\star(h))^2 = \nu_{max}$ comme annoncé. ∎

On peut alors conclure à l'optimalité (en terme du nombre d'itérations) de l'algorithme d'Uzawa.

Corollaire 2.22 *Supposons que la forme $a(\cdot,\cdot)$ soit coercive et symétrique, et que la condition inf–sup discrète uniforme (2.68) soit vérifiée, alors le nombre de conditionnement $\kappa(h)$ est majoré indépendamment de h.*

Remarque 2.23 *La seconde hypothèse n'est pas vraiment contraignante ! En effet, elle intervient déjà de façon cruciale dans le résultat de convergence du théorème 2.12. Sans elle, on ne peut pas le prouver, et la question de l'optimalité de la résolution du système linéaire issu de la discrétisation devient plus secondaire. La seule hypothèse réellement contraignante est que $a(\cdot,\cdot)$ est ici coercive et symétrique.*

Démonstration : Notons pour commencer que l'on a les inégalités

$$\frac{1}{\sqrt{m_a}}\|v\|_a \leq \|v\|_V \leq \frac{1}{\sqrt{\alpha}}\|v\|_a \quad \forall v \in V,$$

où m_a et α sont respectivement le module de continuité (cf. (2.11)) de $a(\cdot,\cdot)$ et sa constante de coercivité. Puisque $V_h \subset V$ et $M_h \subset M$, on en déduit immédiatement que

$$\gamma^\star(h) \leq \frac{1}{\sqrt{\alpha}} \sup_{\sigma_h \in M_h} \sup_{v_h \in V_h} \frac{c(v_h,\sigma_h)}{\|v_h\|_V \|\sigma_h\|_M} \leq \frac{1}{\sqrt{\alpha}} \sup_{\sigma \in M} \sup_{v \in V} \frac{c(v,\sigma)}{\|v\|_V \|\sigma\|_M} = \frac{m_c}{\sqrt{\alpha}},$$

où m_c est la constante de continuité de $c(\cdot,\cdot)$ (cf. (2.11)).
Similairement, on trouve que

$$\gamma_\star(h) \geq \frac{1}{\sqrt{m_a}} \inf_{\sigma_h \in M_h} \sup_{v_h \in V_h} \frac{c(v_h,\sigma_h)}{\|v_h\|_V \|\sigma_h\|_M} \geq \frac{\tilde{\gamma}}{\sqrt{m_a}}.$$

En conclusion, on a d'après (2.95) et la proposition 2.21

$$\kappa(h) \leq \frac{m_c^2\, m_a}{\tilde{\gamma}^2\, \alpha} \quad \forall h,$$

ce qui prouve le résultat annoncé. ∎

2.3 Le cas de l'électromagnétisme quasi-statique

Les équations de Maxwell permettent de décrire de nombreux phénomènes électromagnétiques. Considérons ici un milieu matériel occupant le volume $\Omega \subset \mathbb{R}^3$, et soient ε la permittivité électrique, et μ la perméabilité magnétique du milieu (Ω, ε et μ sont indépendants du temps). Les équations de Maxwell s'écrivent dans Ω :

$$\begin{cases} \varepsilon \dfrac{\partial \mathbf{E}}{\partial t} - \mathbf{rot}\,\mathbf{H} = -\mathbf{J} & \text{(loi d'Ampère)}, \quad (2.96) \\ \mu \dfrac{\partial \mathbf{H}}{\partial t} + \mathbf{rot}\,\mathbf{E} = 0 & \text{(loi de Faraday)}, \quad (2.97) \\ \operatorname{div}(\varepsilon \mathbf{E}) = \rho & \text{(loi de Gauss)}, \quad (2.98) \\ \operatorname{div}(\mu \mathbf{H}) = 0 & \text{(absence de monopole magnétique libre)}, \quad (2.99) \end{cases}$$

avec pour données ρ et $\mathbf{J}$, respectivement la densité de charge et la densité de courant, et le champ électromagnétique $(\mathbf{E}, \mathbf{H})$ pour solution. Lorsque le milieu est entouré d'un conducteur parfait (ce que nous supposerons dans la suite), on a

en plus la condition aux limites $\mathbf{E} \times \mathbf{n} = 0$ sur $\partial\Omega$, où $\mathbf{n}$ désigne le vecteur normal unitaire sortant à $\partial\Omega$. Ces équations décrivent une évolution temporelle et spatiale du champ électromagnétique. C'est pourquoi on doit adjoindre aux équations de Maxwell des conditions initiales, du type $(\mathbf{E}, \mathbf{H})_{|t=0} = (\mathbf{E}_0, \mathbf{H}_0)$, où $(\mathbf{E}_0, \mathbf{H}_0)$ est l'état du champ électromagnétique à l'instant $t = 0$. On choisit en général $\mathbf{H}_0$ tel que $\mu\mathbf{H}_0 \cdot \mathbf{n} = 0$ sur $\partial\Omega$, et $\operatorname{div} \mu\mathbf{H}_0 = 0$ dans Ω. Si la condition sur la divergence semble naturelle (cf. la relation (2.99)), la condition aux limites est choisie pour assurer que $\mu\mathbf{H} \cdot \mathbf{n}_{|\partial\Omega} = 0$ est vérifiée à tout instant. En effet, on peut prouver (voir la proposition 2.36 plus loin) que la trace de la loi de Faraday sur $\partial\Omega$ et la condition aux limites sur $\mathbf{E}$ entrainent la relation $\partial_t(\mu\mathbf{H} \cdot \mathbf{n}_{|\partial\Omega}) = 0$; à partir de là, on en déduit que $\mu\mathbf{H}(t) \cdot \mathbf{n}_{|\partial\Omega} = \mu\mathbf{H}_0 \cdot \mathbf{n}_{|\partial\Omega} = 0$ comme annoncé. Enfin, on tire immédiatement de la loi d'Ampère la condition aux limites $(\mathbf{rot}\, \mathbf{H} - \mathbf{J}) \times \mathbf{n}_{|\partial\Omega} = 0$.

Dans toute cette section, on supposera que Ω est un ouvert borné de $\mathbb{R}^3$, simplement connexe, et dont la frontière est connexe et "suffisamment régulière" (cf. [15]) : typiquement, Ω est un polyèdre curviligne (c'est-à-dire dont les faces sont $\mathcal{C}^\infty$). Précisons les hypothèses mathématiques faites sur ε et μ : on suppose que

$$\begin{cases} \varepsilon \text{ et } \mu \text{ sont constantes par morceaux,} \\ \exists \varepsilon_\star, \varepsilon^\star, \mu_\star, \mu^\star > 0, \quad \varepsilon_\star \leq \varepsilon \leq \varepsilon^\star \text{ et } \mu_\star \leq \mu \leq \mu^\star \text{ sur } \Omega. \end{cases}$$

Les équations de Maxwell (2.96)–(2.99), complétées des conditions aux limites énoncées précédemment, modélisent la propagation d'ondes électromagnétiques et peuvent être résolues à l'aide des méthodes développées au chapitre 4. Ici, nous allons résoudre un modèle approché, obtenu en négligeant le terme $\varepsilon\partial_t\mathbf{E}$ dans la loi d'Ampère. En d'autres termes, les variations temporelles de $\varepsilon\mathbf{E}$ sont supposées être "lentes" par rapport aux autres termes. C'est ce que l'on appelle un *modèle quasi-statique*. Posons $\mathbf{f}_H = \mathbf{J}$, $\mathbf{f}_E = -\mu\partial_t\mathbf{H}$ et $g_E = \rho$, alors on peut écrire ce modèle sous la forme de deux systèmes d'équations :

$$\begin{cases} \text{trouver } \mathbf{H}(t) \text{ tel que} & \\ \mathbf{rot}\, \mathbf{H}(t) = \mathbf{f}_H(t) & \text{dans } \Omega, \\ \operatorname{div} \mu\mathbf{H}(t) = 0 & \text{dans } \Omega, \\ \mu\mathbf{H}(t) \cdot \mathbf{n} = 0 & \text{sur } \partial\Omega \end{cases} \tag{2.100}$$

et

$$\begin{cases} \text{trouver } \mathbf{E}(t) \text{ tel que} & \\ \mathbf{rot}\, \mathbf{E}(t) = \mathbf{f}_E(t) & \text{dans } \Omega, \\ \operatorname{div} \varepsilon\mathbf{E}(t) = g_E(t) & \text{dans } \Omega, \\ \mathbf{E}(t) \times \mathbf{n} = 0 & \text{sur } \partial\Omega. \end{cases} \tag{2.101}$$

Ci-dessus, on a mentionné la dépendance en temps pour préciser que, bien que les opérateurs gouvernant (2.100) et (2.101) ne dépendent pas du temps (div, **rot**

et les conditions aux limites), les données, et donc les solutions, dépendent elles *a priori* du temps. Quant à savoir pourquoi ces deux systèmes d'équations sont étudiés dans le chapitre sur les problèmes mixtes, une réponse peut être formulée de la façon suivante : on va considérer les conditions sur la divergence du champ électromagnétique (et sur la trace normale, pour $\mathbf{H}$), comme des contraintes.

Le second membre $\mathbf{f}_E$ dans (2.101) est égal à $-\mu\partial_t\mathbf{H}$, que l'on considère comme une donnée. Typiquement, on calcule $\mathbf{H}$ en résolvant (2.100), puis $\mathbf{E}$ en résolvant (2.101). La seconde condition aux limites sur $\mathbf{H}$, à savoir $(\mathbf{rot}\,\mathbf{H} - \mathbf{J}) \times \mathbf{n}_{|\partial\Omega} = 0$, est implicitement contenue dans (2.100) car elle correspond à la trace (tangentielle) sur la frontière de la première équation. La question de l'existence d'une solution aux problèmes en div-rot du type (2.100) ou (2.101), est délicate. Des réponses détaillées se trouvent notamment dans [25, 3, 14]. Il est établi rigoureusement dans [3] que le problème (2.100) admet une solution et une seule, avec stabilité par rapport à la donnée $\mathbf{f}_H$, dès lors que $\mathbf{f}_H \in L^2(\Omega)^3$ et $\operatorname{div}\mathbf{f}_H = 0$ (cette condition de compatibilité sur $\mathbf{f}_H$ est évidente, puisque $\mathbf{f}_H = \mathbf{rot}\,\mathbf{H}$). Pour le problème (2.101), si les données satisfont aux hypothèses $\mathbf{f}_E \in L^2(\Omega)^3$, $\operatorname{div}\mathbf{f}_E = 0$, $\mathbf{f}\cdot\mathbf{n}_{|\partial\Omega} = 0$, et $g_E \in H^{-1}(\Omega)$, on retrouve également l'existence, l'unicité et la stabilité [4]. Dans ce qui suit, nous allons redémontrer ces résultats, par une approche différente de celles proposées dans [25, 3, 14].

2.3.1 Un peu d'analyse fonctionnelle

Les problèmes quasi-statiques (2.100) et (2.101) font intervenir les opérateurs rotationnel et divergence, ainsi que les traces normale et tangentielle sur la frontière $\partial\Omega$. Rappelons tout d'abord comment on prend en compte une condition sur la trace normale. On sait, suivant toujours [15], que la divergence "contrôle" la trace normale dans l'espace $H^{-1/2}(\partial\Omega)$, l'espace dual de $H^{1/2}(\partial\Omega)$, d'où la prise en compte de la trace normale en tant que contrainte.

Proposition 2.24 (Trace normale) *L'espace $\mathcal{C}^\infty(\overline{\Omega})^3$ est dense dans $H(\operatorname{div};\Omega)$; et l'application trace normale*

$$\gamma_n : \left\{ \begin{array}{ccc} \mathcal{C}^\infty(\overline{\Omega})^3 & \to & H^{-1/2}(\partial\Omega) \\ \mathbf{v} & \mapsto & \gamma_n\mathbf{v} = (\mathbf{v}\cdot\mathbf{n})_{|\partial\Omega} \end{array} \right.$$

se prolonge par continuité en une application linéaire continue, encore notée γ_n, de $H(\operatorname{div};\Omega)$ dans $H^{-1/2}(\partial\Omega)$. Enfin, pour tout $\mathbf{u} \in H(\operatorname{div};\Omega)$ et tout $v \in H^1(\Omega)$, on a la formule d'intégration par parties :

$$\int_\Omega (\operatorname{div}\mathbf{u})\, v\, d\Omega = -\int_\Omega \mathbf{u}\cdot\nabla v\, d\Omega + \langle \gamma_n\mathbf{u}, v\rangle_{H^{-1/2}(\partial\Omega),H^{1/2}(\partial\Omega)}. \tag{2.102}$$

4. On rappelle que, par définition, $H^{-1}(\Omega)$ est l'espace dual de $H^1_0(\Omega)$.

Qu'en est-il de la trace tangentielle ? On peut prouver que celle-ci est "contrôlée" par le rotationnel (cf. [27]). Soit donc

$$H(\mathbf{rot}\,;\Omega) = \left\{ \mathbf{v} \in L^2(\Omega)^3 \;\middle|\; \mathbf{rot}\,\mathbf{v} \in L^2(\Omega)^3 \right\}, \tag{2.103}$$

où le rotationnel est compris au sens des distributions. On munit $H(\mathbf{rot}\,;\Omega)$ de la norme

$$\|\mathbf{v}\|_{H(\mathbf{rot}\,;\Omega)} = \left(\|\mathbf{v}\|^2_{L^2(\Omega)^3} + \|\mathbf{rot}\,\mathbf{v}\|^2_{L^2(\Omega)^3} \right)^{1/2}.$$

Proposition 2.25 (Trace tangentielle) *L'espace $\mathcal{C}^\infty(\overline{\Omega})^3$ est dense dans l'espace $H(\mathbf{rot}\,;\Omega)$; et l'application trace tangentielle*

$$\gamma_T : \begin{cases} \mathcal{C}^\infty(\overline{\Omega})^3 & \to & H^{-1/2}(\partial\Omega)^3 \\ \mathbf{v} & \mapsto & \gamma_T\mathbf{v} = (\mathbf{v}\times\mathbf{n})_{|\partial\Omega} \end{cases}$$

se prolonge par continuité en une application linéaire continue, encore notée γ_T, de $H(\mathbf{rot}\,;\Omega)$ dans $H^{-1/2}(\partial\Omega)^3$. Enfin, pour tous $\mathbf{u} \in H(\mathbf{rot}\,;\Omega)$ et $\mathbf{v} \in H^1(\Omega)^3$, on a la formule d'intégration par parties :

$$\int_\Omega \mathbf{rot}\,\mathbf{u}\cdot\mathbf{v}\,d\Omega = \int_\Omega \mathbf{u}\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega - \langle \gamma_T\mathbf{u}, \mathbf{v}\rangle_{H^{-1/2}(\partial\Omega)^3, H^{1/2}(\partial\Omega)^3}. \tag{2.104}$$

Dans la suite, il sera utile de pouvoir considérer des fonctions à trace normale, ou à trace tangentielle, nulle.

Définition 2.26 *On note $H_0(\mathrm{div};\Omega)$ la fermeture de $\mathcal{D}(\Omega)^3$ dans $H(\mathrm{div};\Omega)$:*

$$H_0(\mathrm{div};\Omega) = \overline{(\mathcal{D}(\Omega)^3)}^{\,H(\mathrm{div};\Omega)}.$$

On peut caractériser l'espace $H_0(\mathrm{div};\Omega)$ comme ci-dessous.

Théorème 2.27 (Caractérisation de $H_0(\mathrm{div};\Omega)$) *On a*

$$H_0(\mathrm{div};\Omega) = \{\mathbf{v} \in H(\mathrm{div};\Omega) \mid \gamma_n\mathbf{v} = 0\}.$$

Définition 2.28 *On note $H_0(\mathbf{rot}\,;\Omega)$ la fermeture de $\mathcal{D}(\Omega)^3$ dans $H(\mathbf{rot}\,;\Omega)$:*

$$H_0(\mathbf{rot}\,;\Omega) = \overline{(\mathcal{D}(\Omega)^3)}^{\,H(\mathbf{rot}\,;\Omega)}.$$

On remarque alors que, pour $\mathbf{u} \in H_0(\mathbf{rot}\,;\Omega)$ et $\mathbf{v} \in H(\mathbf{rot}\,;\Omega)$, on a la formule d'intégration par parties :

$$\int_\Omega \mathbf{rot}\,\mathbf{u}\cdot\mathbf{v}\,d\Omega = \int_\Omega \mathbf{u}\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega. \tag{2.105}$$

On peut caractériser l'espace $H_0(\mathbf{rot}\,;\Omega)$ comme ci-dessous.

Théorème 2.29 (Caractérisation de $H_0(\mathbf{rot}\,;\Omega)$) *On a*

$$H_0(\mathbf{rot}\,;\Omega) = \{\mathbf{v} \in H(\mathbf{rot}\,;\Omega) \mid \gamma_T \mathbf{v} = 0\}\,.$$

Nous concluons par quelques résultats d'existence de potentiels vecteur ou scalaire, bien connus en physique, énoncés ici dans une version mathématique (voir [27, 3]).

Proposition 2.30 (Existence de potentiels) *Soit* $\mathbf{v} \in L^2(\Omega)^3$ *:*
(i) si $\mathbf{rot}\,\mathbf{v} = 0$ *dans* Ω *:*
$\exists q \in H^1(\Omega)$ *tel que* $\mathbf{v} = \nabla q$ *dans* Ω *;*
(ii) si $\mathbf{rot}\,\mathbf{v} = 0$ *dans* Ω *et* $(\mathbf{v} \times \mathbf{n})_{|\partial\Omega} = 0$ *:*
$\exists q \in H^1_0(\Omega)$ *tel que* $\mathbf{v} = \nabla q$ *dans* Ω *;*
(iii) si $\operatorname{div}\mathbf{v} = 0$ *dans* Ω *:*
$\exists \mathbf{w} \in H(\mathbf{rot}\,;\Omega) \cap H_0(\operatorname{div};\Omega)$, $\operatorname{div}\mathbf{w} = 0$ *dans* Ω, *tel que* $\mathbf{v} = \mathbf{rot}\,\mathbf{w}$ *dans* Ω *;*
(iv) si $\operatorname{div}\mathbf{v} = 0$ *dans* Ω *et* $(\mathbf{v}\cdot\mathbf{n})_{|\partial\Omega} = 0$ *:*
$\exists \mathbf{w} \in H_0(\mathbf{rot}\,;\Omega) \cap H(\operatorname{div};\Omega)$, $\operatorname{div}\mathbf{w} = 0$ *dans* Ω, *tel que* $\mathbf{v} = \mathbf{rot}\,\mathbf{w}$ *dans* Ω.

2.3.2 Constructions des problèmes variationnels mixtes

Pour des données volumiques $\mathbf{f}_H$ et $\mathbf{f}_E$ dans $L^2(\Omega)^3$, on cherche $\mathbf{H}(t)$ et $\mathbf{E}(t)$ dans $H(\mathbf{rot}\,;\Omega)$. On a vu au §2.3.1 qu'on peut contrôler la trace tangentielle – mais pas la trace normale – des éléments de $H(\mathbf{rot}\,;\Omega)$. La condition aux limites sur $\mathbf{E}$, qui met en jeu sa trace tangentielle, pourra donc être considérée comme une *condition aux limites essentielle*, c'est-à-dire qu'elle sera incorporée dans l'espace des solutions et fonctions-tests. En d'autres termes, on va choisir $V_H = H(\mathbf{rot}\,;\Omega)$ pour le problème (2.100) d'inconnue le champ magnétique $\mathbf{H}$, et $V_E = H_0(\mathbf{rot}\,;\Omega)$ pour le problème (2.101) d'inconnue le champ électrique $\mathbf{E}$. Il nous reste maintenant à exprimer ces problèmes sous forme mixte.

Commençons par remarquer qu'il n'y a pas de multiplicateur dans les systèmes quasi-statiques (2.100) et (2.101) ! Nous allons donc en introduire un dans chaque système, purement "artificiel", c'est-à-dire nul ! Notons-les λ_E pour $\mathbf{E}$ et λ_H pour $\mathbf{H}$, respectivement appelés "pression électrique artificielle" et "pression magnétique artificielle".

Considérons le système (2.101) en $\mathbf{E}$, et prenons *formellement* le rotationnel de la première équation, auquel on ajoute $\varepsilon\nabla\lambda_E(t)$ – supposé nul – à gauche du signe $=$, pour aboutir à

$$\begin{cases} \text{trouver } \mathbf{E}(t) \text{ tel que} & \\ \mathbf{rot}\,\mathbf{rot}\,\mathbf{E}(t) + \varepsilon\nabla\lambda_E(t) = \mathbf{rot}\,\mathbf{f}_E(t) & \text{dans } \Omega, \\ \operatorname{div}\varepsilon\mathbf{E}(t) = g_E(t) & \text{dans } \Omega, \\ \mathbf{E}(t) \times \mathbf{n} = 0 & \text{sur } \partial\Omega. \end{cases} \tag{2.106}$$

Pour le système (2.100) en $\mathbf{H}$, on raisonne de même pour introduire la pression magnétique artificielle λ_H. On ajoute aussi explicitement la seconde condition aux limites, à savoir $\mathbf{rot}\,\mathbf{H}(t) \times \mathbf{n} = \mathbf{f}_H(t) \times \mathbf{n}$ sur $\partial\Omega$. On aboutit cette fois à :

$$\left\{\begin{array}{ll} \text{trouver } \mathbf{H}(t) \text{ tel que} & \\ \mathbf{rot}\,\mathbf{rot}\,\mathbf{H}(t) + \mu\nabla\lambda_H(t) = \mathbf{rot}\,\mathbf{f}_H(t) & \text{dans } \Omega, \\ \operatorname{div}\mu\mathbf{H}(t) = 0 & \text{dans } \Omega, \\ \mu\mathbf{H}(t)\cdot\mathbf{n} = 0 & \text{sur } \partial\Omega, \\ \mathbf{rot}\,\mathbf{H}(t) \times \mathbf{n} = \mathbf{f}_H(t) \times \mathbf{n} & \text{sur } \partial\Omega. \end{array}\right. \tag{2.107}$$

Remarque 2.31 *Concernant ces nouveaux systèmes quasi-statiques (2.106) et (2.107), on devra d'une part vérifier lors de leur résolution que les pressions artificielles λ_E et λ_H sont bien nulles. D'autre part, comme ils ont été construits formellement, il faudra également s'assurer qu'ils sont bien équivalents à (2.100) et (2.101).*

Poursuivons par l'étude proprement dite du problème (2.107) en $\mathbf{H}(t)$: les contraintes sont $\operatorname{div}\mu\mathbf{H}(t) = 0$ dans Ω, et $\mu\mathbf{H}(t)\cdot\mathbf{n}_{|\partial\Omega} = 0$. A l'aide de l'espace $L^2_0(\Omega)$ défini précédemment en (2.6), on peut établir un premier résultat.

Proposition 2.32 (Contraintes sur H) *Soit $\mathbf{v} \in L^2(\Omega)^3$. Alors*

$$\operatorname{div}\mathbf{v} = 0 \text{ dans } \Omega,\ \mathbf{v}\cdot\mathbf{n}_{|\partial\Omega} = 0 \quad \Longleftrightarrow \int_\Omega \mathbf{v}\cdot\nabla q\, d\Omega = 0,\ \forall q \in H^1(\Omega) \cap L^2_0(\Omega).$$

Remarque 2.33 *A droite, seul le gradient d'éléments de $H^1(\Omega)$ apparaît : les constantes sont donc superflues. D'où l'idée de se placer dans le sous-espace vectoriel formé des éléments à moyenne nulle (ce qui n'est pas le seul choix possible !).*

Démonstration : $\Longrightarrow$] Comme $\operatorname{div}\mathbf{v} = 0$, on a $\mathbf{v} \in H(\operatorname{div};\Omega)$. En intégrant par parties (cf. (2.102)) avec $\mathbf{v}\cdot\mathbf{n}_{|\partial\Omega} = 0$, on en déduit que pour tout $q \in H^1(\Omega)$:

$$\int_\Omega \mathbf{v}\cdot\nabla q\, d\Omega = 0.$$

$\Longleftarrow$] Vérifions tout d'abord que $\operatorname{div}\mathbf{v} = 0$ au sens des distributions. Soit donc $\varphi \in \mathcal{D}(\Omega)$:

$$\langle \operatorname{div}\mathbf{v}, \varphi\rangle = -\langle \mathbf{v}, \nabla\varphi\rangle = -\int_\Omega \mathbf{v}\cdot\nabla\varphi\, d\Omega = 0.$$

La première égalité est la définition de la dérivation au sens des distributions. La deuxième égalité provient de l'identification de $L^2(\Omega)$ à un sous-espace de $\mathcal{D}'(\Omega)$. La troisième égalité découle quant à elle de l'inclusion $\mathcal{D}(\Omega) \subset H^1(\Omega)$. Plus précisément, rien n'assure que $\varphi \in L^2_0(\Omega)$; néanmoins, si on pose $c_\varphi = \frac{1}{\operatorname{mes}\Omega}\int_\Omega \varphi\, d\Omega$, alors $\varphi - c_\varphi \in L^2_0(\Omega)$, et $\nabla(\varphi - c_\varphi) = \nabla\varphi$.
On en conclut que $\operatorname{div}\mathbf{v} = 0$ dans $\mathcal{D}'(\Omega)$: ainsi $\mathbf{v} \in H(\operatorname{div};\Omega)$. On peut ensuite utiliser l'intégration par parties (2.102), pour trouver :

$$0 = \int_\Omega \mathbf{v}\cdot\nabla(q - c_q)\, d\Omega = \int_\Omega \mathbf{v}\cdot\nabla q\, d\Omega = \langle \mathbf{v}\cdot\mathbf{n}_{|\partial\Omega}, q\rangle_{H^{-1/2}(\partial\Omega),H^{1/2}(\partial\Omega)},\ \forall q \in H^1(\Omega),$$

avec $c_q = \frac{1}{\text{mes}\Omega} \int_\Omega q \, d\Omega$.
Comme l'application trace γ_0 est surjective de $H^1(\Omega)$ dans $H^{1/2}(\partial\Omega)$, on a bien $\mathbf{v}\cdot\mathbf{n}_{|\partial\Omega} = 0$ dans $H^{-1/2}(\partial\Omega)$ comme annoncé. ∎

Pour le problème (2.107), les contraintes portent sur $\mu\mathbf{H}(t)$. On va donc choisir $M_H = H^1(\Omega) \cap L^2_0(\Omega)$, $c_H(\mathbf{v}, q) = \int_\Omega \mu\mathbf{v} \cdot \nabla q \, d\Omega$ et $\ell_{M_H}(q) = 0$.

Il reste à construire les formes $a_H(\cdot,\cdot)$ et $\ell_{V_H}(\cdot)$. Formellement, on note que $\mathbf{rot}\,\mathbf{rot}\,\mathbf{H}(t) + \mu\nabla\lambda_H(t) = \mathbf{rot}\,\mathbf{f}_H(t)$ dans Ω, additionné de la condition aux limites $\mathbf{rot}\,\mathbf{H}(t) \times \mathbf{n} = \mathbf{f}_H(t) \times \mathbf{n}$ sur $\partial\Omega$, entraine en particulier

$$\int_\Omega \mathbf{rot}\,\mathbf{H} \cdot \mathbf{rot}\,\mathbf{v} \, d\Omega + \int_\Omega \mu\nabla\lambda_H \cdot \mathbf{v} \, d\Omega = \int_\Omega \mathbf{f}_H \cdot \mathbf{rot}\,\mathbf{v} \, d\Omega, \quad \forall \mathbf{v} \in V_H, \quad (2.108)$$

par l'intégration par parties (2.104).
Ainsi, on peut considérer $a_H(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{rot}\,\mathbf{u} \cdot \mathbf{rot}\,\mathbf{v} \, d\Omega$ et $\ell_{V_H}(\mathbf{v}) = \int_\Omega \mathbf{f}_H \cdot \mathbf{rot}\,\mathbf{v} \, d\Omega$.

Pour finir, vérifions que l'on a bien une formulation sous la forme d'un problème mixte, comme en (2.7)–(2.8). Dans le cas présent, on doit résoudre

$$\begin{cases} \text{trouver } (\mathbf{H}, \lambda_H) \in H(\mathbf{rot}\,;\Omega) \times (H^1(\Omega) \cap L^2_0(\Omega)) \text{ tel que} \\ \displaystyle\int_\Omega \mathbf{rot}\,\mathbf{H} \cdot \mathbf{rot}\,\mathbf{v} d\Omega + \int_\Omega \mu\mathbf{v} \cdot \nabla\lambda_H d\Omega \\ \qquad\qquad = \displaystyle\int_\Omega \mathbf{f}_H \cdot \mathbf{rot}\,\mathbf{v} d\Omega \qquad \forall \mathbf{v} \in H(\mathbf{rot}\,;\Omega), \qquad (2.109) \\ \displaystyle\int_\Omega \mu\mathbf{H} \cdot \nabla q \, d\Omega = 0 \quad \forall q \in (H^1(\Omega) \cap L^2_0(\Omega)). \qquad (2.110) \end{cases}$$

Rappelons que nous avons utilisé deux hypothèses non justifiées : nous avons raisonné formellement pour passer du système quasi-statique (2.100) à (2.107), et donc pour passer de (2.100) à (2.109)–(2.110), et d'autre part nous avons supposé que $\nabla\lambda_H = 0$! Ceci étant rappelé, pourquoi la pression magnétique artificielle λ_H est-elle nulle ?

Notons que si $q \in H^1(\Omega)$, alors $\nabla q \in H(\mathbf{rot}\,;\Omega)$. En effet, comme $q \in H^1(\Omega)$, on sait que $\nabla q \in L^2(\Omega)^3$. Par ailleurs, on a la suite d'égalités, pour $\mathbf{z} \in \mathcal{D}(\Omega)^3$:

$$\langle \mathbf{rot}\,(\nabla q), \mathbf{z}\rangle = \langle \nabla q, \mathbf{rot}\,\mathbf{z}\rangle = -\langle q, \mathrm{div}(\mathbf{rot}\,\mathbf{z})\rangle = 0,$$

puisque $\mathrm{div}\,\mathbf{rot}\,\mathbf{z} = 0$. On en déduit que $\mathbf{rot}\,(\nabla q) = 0$ dans $\mathcal{D}'(\Omega)^3$, et donc que $\nabla q \in H(\mathbf{rot}\,;\Omega)$.

Cette observation – élémentaire – est d'une importance fondamentale. En effet, on en déduit que $\mathbf{v}_H = \nabla\lambda_H$ peut être employée comme fonction-test dans (2.109),

d'où il suit $\int_\Omega \mu|\nabla\lambda_H|^2\,d\Omega = 0$, puisque $\mathbf{rot}\,\mathbf{v}_H = 0$. Sachant que $\mu \geq \mu_\star > 0$ presque partout, on en conclut que $\int_\Omega \mu_\star|\nabla\lambda_H|^2\,d\Omega \leq 0$, et donc que $\nabla\lambda_H = 0$. La pression magnétique artificielle est donc constante dans Ω (supposé connexe). Qui plus est, λ_H est à valeur moyenne nulle sur Ω, et ainsi [5] $\lambda_H = 0$.

Evoquons plus brièvement la contrepartie électrique...

Proposition 2.34 (Contrainte sur E) *Soient* $\mathbf{v} \in L^2(\Omega)^3$ *et* $g \in H^{-1}(\Omega)$. *Alors*

$$\operatorname{div}\mathbf{v} = g \text{ dans } \Omega \quad \Longleftrightarrow \int_\Omega \mathbf{v}\cdot\nabla q\,d\Omega = -\langle g, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)},\ \forall q \in H_0^1(\Omega).$$

Démonstration : Laissée au lecteur. ∎

La contrainte portant sur $\varepsilon\mathbf{E}$, on choisit cette fois : $M_E = H_0^1(\Omega)$, $c_E(\mathbf{v},q) = \int_\Omega \varepsilon\mathbf{v}\cdot\nabla q\,d\Omega$ et $\ell_{M_E}(q) = -\langle g_E, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}$. Quant aux formes $a_E(\cdot,\cdot)$ et $\ell_{V_E}(\cdot)$, l'égalité $\mathbf{rot}\,\mathbf{rot}\,\mathbf{E} + \varepsilon\nabla\lambda_E = \mathbf{rot}\,\mathbf{f}_E$ dans Ω entraine par l'intégration par parties (2.105)

$$\int_\Omega \mathbf{rot}\,\mathbf{E}\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega + \int_\Omega \varepsilon\nabla\lambda_E\cdot\mathbf{v}\,d\Omega = \int_\Omega \mathbf{f}_E\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega, \quad \forall\mathbf{v}\in V_E, \tag{2.111}$$

et donc $a_E(\mathbf{u},\mathbf{v}) = \int_\Omega \mathbf{rot}\,\mathbf{u}\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega$ et $\ell_{V_E}(\mathbf{v}) = \int_\Omega \mathbf{f}_E\cdot\mathbf{rot}\,\mathbf{v}\,d\Omega$. Le problème mixte en $\mathbf{E}$ s'écrit alors

$$\begin{cases} \text{trouver } (\mathbf{E},\lambda_E)\in H_0(\mathbf{rot}\,;\Omega)\times H_0^1(\Omega) \text{ tel que} & \\ \displaystyle\int_\Omega \mathbf{rot}\,\mathbf{E}\cdot\mathbf{rot}\,\mathbf{v}d\Omega + \int_\Omega \varepsilon\mathbf{v}\cdot\nabla\lambda_E d\Omega & \\ \qquad\qquad = \displaystyle\int_\Omega \mathbf{f}_E\cdot\mathbf{rot}\,\mathbf{v}d\Omega & \forall\mathbf{v}\in H_0(\mathbf{rot}\,;\Omega), \quad (2.112)\\ \displaystyle\int_\Omega \varepsilon\mathbf{E}\cdot\nabla q\,d\Omega = -\langle g_E, q\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} & \forall q\in H_0^1(\Omega). \quad (2.113) \end{cases}$$

Encore une fois, il faut que la "pression électrique artificielle" λ_E disparaisse. On utilise une technique semblable à celle du cas magnétique.

Proposition 2.35 (Séquence exacte) *Soit* $q \in H_0^1(\Omega)$, *alors* $\nabla q \in H_0(\mathbf{rot}\,;\Omega)$.

Démonstration : Comme dans le cas magnétique, on constate tout d'abord que $\nabla q \in H(\mathbf{rot}\,;\Omega)$. Pourquoi ∇q appartient-il à $H_0(\mathbf{rot}\,;\Omega)$?

5. On rappelle que, d'après l'inégalité de Poincaré-Wirtinger, la semi-norme $|\cdot|_1$ définie par $|q|_1 = \left(\int_\Omega |\nabla q|^2\,d\Omega\right)^{1/2}$ est une norme sur $H^1(\Omega)\cap L_0^2(\Omega)$, équivalente à $\|\cdot\|_{H^1(\Omega)}$. Comme $\infty > \mu^\star \geq \mu \geq \mu_\star > 0$ presque partout, on remarque que $q \mapsto \left(\int_\Omega \mu|\nabla q|^2\,d\Omega\right)^{1/2}$ définit également une norme équivalente à $\|\cdot\|_{H^1(\Omega)}$. Ce résultat d'équivalence des normes est plus fort que celui que l'on vient d'utiliser.

Une réponse intrinsèque est la suivante : puisque $q \in H_0^1(\Omega) = \overline{\mathcal{D}(\Omega)}^{\,H^1(\Omega)}$, il existe $(q_m)_m$ une suite d'éléments de $\mathcal{D}(\Omega)$ telle que $q_m \to q$ dans $H^1(\Omega)$. En particulier, $(\nabla q_m)_m$ tend vers ∇q dans $L^2(\Omega)^3$. Par ailleurs, on a $\mathbf{rot}\,(\nabla q_m) = 0$ pour tout m, et bien sûr $\mathbf{rot}\,(\nabla q) = 0$, d'où l'on déduit que $(\nabla q_m)_m$ tend en fait vers ∇q dans $H(\mathbf{rot}\,;\Omega)$. Or, $(\nabla q_m)_m$ est une suite d'éléments de $\mathcal{D}(\Omega)^3$ et, par définition de $H_0(\mathbf{rot}\,;\Omega) = \overline{(\mathcal{D}(\Omega)^3)}^{\,H(\mathbf{rot}\,;\Omega)}$ (cf. définition 2.28), il ressort que l'on a $\nabla q \in H_0(\mathbf{rot}\,;\Omega)$. ∎

On peut démontrer, de façon tout à similaire, un second résultat de séquence exacte.

Proposition 2.36 (Séquence exacte) *Si* $\mathbf{v} \in H_0(\mathbf{rot}\,;\Omega)$, *alors* $\mathbf{rot}\,\mathbf{v}$ *appartient à* $H_0(\mathrm{div};\Omega)$.

Démonstration : Laissée au lecteur. ∎

Soit donc la fonction-test $\mathbf{v}_E = \nabla\lambda_E$ dans (2.112) : on arrive à $\int_\Omega \varepsilon|\nabla\lambda_E|^2\,d\Omega = 0$. Sachant que $\varepsilon \geq \varepsilon_\star > 0$ presque partout, on en conclut que $\nabla\lambda_E = 0$, et la pression électrique artificielle est constante dans Ω. S'annulant sur la frontière $\partial\Omega$, elle est nulle. [6].

En conclusion, nous avons montré que si $(\mathbf{E},\mathbf{H})$ est solution du modèle quasi-statique, alors $(\mathbf{E},0)$ et $(\mathbf{H},0)$ résolvent respectivement les problèmes mixtes (2.112)–(2.113) et (2.109)–(2.110). Qu'en est-il de la réciproque ?

2.3.3 Résolution des problèmes variationnels mixtes

Examinons maintenant le caractère bien posé des problèmes mixtes, ainsi que le retour au modèle quasi-statique d'origine. On va essentiellement traiter le cas électrique, le cas magnétique étant résolu par des considérations tout à fait similaires.

Commençons par la coercivité sur le noyau, qui est le point délicat. Soit donc C_E associée à $c_E(\cdot,\cdot)$ par (2.13).
D'après la proposition 2.34, on sait que

$$\mathrm{Ker}\,C_E = \{\mathbf{v} \in H_0(\mathbf{rot}\,;\Omega) \mid \mathrm{div}\,\varepsilon\mathbf{v} = 0 \text{ dans } \Omega\}\,.$$

Pour établir que $a_E(\cdot,\cdot)$ est coercive sur $\mathrm{Ker}\,C_E$, on va utiliser le résultat plus général suivant. Soit

$$X_E = \left\{\mathbf{v} \in H_0(\mathbf{rot}\,;\Omega) \;\middle|\; \mathrm{div}\,\varepsilon\mathbf{v} \in L^2(\Omega)\right\}, \tag{2.114}$$

un espace fonctionnel qui nous sera fort utile pour la suite, que l'on munit de la norme induite :

6. Dans $H_0^1(\Omega)$, $|\cdot|_1$ est une norme équivalente à $\|\cdot\|_{H^1(\Omega)}$, d'après l'inégalité de Poincaré. Comme $\infty > \varepsilon^\star \geq \varepsilon \geq \varepsilon_\star > 0$ presque partout, il en est de même pour $q \mapsto \left(\int_\Omega \varepsilon|\nabla q|^2\,d\Omega\right)^{1/2}$.

$$\|\mathbf{v}\|_\varepsilon = \left(\|\mathbf{v}\|^2_{L^2(\Omega)^3} + \|\mathbf{rot}\,\mathbf{v}\|^2_{L^2(\Omega)^3} + \|\operatorname{div}\varepsilon\mathbf{v}\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Lorsque ε est globalement constante sur Ω, on emploiera la terminologie :

$$X_E^{cst} = H_0(\mathbf{rot}\,;\Omega) \cap H(\operatorname{div};\Omega).$$

Le résultat suivant est établi dans [50, 31]. Nous en proposons une démonstration différente ci-dessous, qui repose explicitement sur le fait que ε est constante par morceaux.

Théorème 2.37 (Weber – cas électrique) *L'injection canonique de X_E dans $L^2(\Omega)^3$ est compacte.*

Remarque 2.38 *En d'autres termes, de toute suite bornée de X_E, on peut extraire une sous-suite qui converge dans $L^2(\Omega)^3$ (voir définition 1.8). Ce résultat est également à rapprocher du théorème de Rellich, à savoir l'injection compacte de $H^1(\Omega)$ dans $L^2(\Omega)$. Dans le cas présent néanmoins, il est impératif d'imposer une condition aux limites homogène (tangentielle dans la définition de X_E).*

Démonstration : Soit $(\Omega_i)_{1\le i\le I}$ une partition de Ω ($\bar\Omega = \cup_{i=1,I}\bar\Omega_i$, $\Omega_i \cap \Omega_{i'} = \emptyset$ si $i \neq i'$), telle que $\varepsilon_{|\Omega_i}$ soit constante pour tout i. Considérons maintenant $\mathbf{v}$ un élément quelconque de X_E. D'après [20], on peut décomposer $\mathbf{v}$ en

$$\mathbf{v} = \mathbf{w} + \nabla\phi_0, \tag{2.115}$$

avec

$$\begin{cases} \mathbf{w} \in X_E, \text{ tel que } \mathbf{w}_{|\Omega_i} \in H^1(\Omega_i)^3,\ 1 \le i \le I, \\ \phi_0 \in H^1_0(\Omega), \text{ tel que } \operatorname{div}\varepsilon\nabla\phi_0 \in L^2(\Omega). \end{cases}$$

De plus, cette décomposition est *continue*, à savoir qu'il existe une constante $C_{el} > 0$ indépendante de $\mathbf{v}$ telle que

$$\|\mathbf{w}\|_\varepsilon + \sum_{i=1,I} \|\mathbf{w}_{|\Omega_i}\|_{H^1(\Omega_i)^3} + \|\phi_0\|_{H^1(\Omega)} + \|\operatorname{div}\varepsilon\nabla\phi_0\|_{L^2(\Omega)} \le C_{el}\,\|\mathbf{v}\|_\varepsilon.$$

A l'aide de la théorie usuelle [39] des espaces de Sobolev $H^s(\Omega)$, avec $s \in [0,1]$, on peut vérifier que $\mathbf{w} \in \cap_{s<1/2} H^s(\Omega)^3$.
Par ailleurs, grâce au caractère constant par morceaux de ε, on peut réaliser une étude du comportement de ϕ_0, considéré comme solution d'un problème du type : trouver $\phi_0 \in H^1_0(\Omega)$ tel que $\operatorname{div}\varepsilon\nabla\phi_0 = f$ dans Ω, avec $f \in L^2(\Omega)$. On peut alors démontrer (cf. [20]) qu'il existe $\sigma_\varepsilon > 0$ tel que $\phi_0 \in \cap_{s<1+\sigma_\varepsilon} H^s(\Omega)$. Puisque la dérivation fait perdre un ordre de régularité, on en déduit que $\nabla\phi_0 \in \cap_{s<\sigma_\varepsilon} H^s(\Omega)^3$.
Si on note $\sigma'_\varepsilon = \min(1/2, \sigma_\varepsilon) > 0$, on a prouvé que $\mathbf{v} \in \cap_{s<\sigma'_\varepsilon} H^s(\Omega)^3$; de plus l'injection de X_E dans $\cap_{s<\sigma'_\varepsilon} H^s(\Omega)^3$ est continue. Or,
– d'une part, pour $s > 0$ et Ω borné, les espaces de Sobolev $H^s(\Omega)$ s'injectent de façon compacte dans $L^2(\Omega)$;
– d'autre part, la composition d'une application continue et d'une application compacte est elle-même compacte.

En écrivant $i_{X_E \to L^2(\Omega)^3} = i_{H^{\sigma'_\varepsilon/2}(\Omega)^3 \to L^2(\Omega)^3} \circ i_{X_E \to H^{\sigma'_\varepsilon/2}(\Omega)^3}$, le résultat suit. ∎

Remarque 2.39 *Que se passe-t-il si ε est globalement constante sur Ω ? En pratique, cette situation recouvre les cas où le volume Ω est occupé par de l'air, ou bien si on a fait le vide dans celui-ci... On peut alors écrire, pour tout élément $\mathbf{v}$ de X_E^{cst}, la décomposition (2.115), avec dans ce cas $\mathbf{w} \in X_E^{cst} \cap H^1(\Omega)^3$ et $\phi_0 \in H_0^1(\Omega)$ tel que $\Delta\phi_0 \in L^2(\Omega)$. D'après les résultats de régularité sur la solution du Laplacien avec condition aux limites de Dirichlet homogène (cf. [15]), on sait qu'il existe $\sigma_D > 1/2$ tel que $\phi_0 \in \cap_{s<1+\sigma_D} H^s(\Omega)$, alors que $\phi_0 \in H^{1+\sigma_D}(\Omega)$ n'est pas garanti (σ_D est l'exposant limite de régularité). Ainsi, on peut écrire que $X_E^{cst} \subset \cap_{s<\sigma_D} H^s(\Omega)^3$, avec injection continue.*

A l'aide de ce résultat, on peut établir un résultat d'équivalence de normes dans X_E.

Proposition 2.40 (Inégalité de Weber – cas électrique) *Il existe une constante C_W, dépendant seulement de Ω telle que :*

$$\int_\Omega |\mathbf{v}|^2 \, d\Omega \leq C_W \int_\Omega \left(|\mathbf{rot}\, \mathbf{v}|^2 + (\operatorname{div} \varepsilon \mathbf{v})^2 \right) d\Omega, \ \forall \mathbf{v} \in X_E. \tag{2.116}$$

Remarque 2.41 *Cette inégalité est à rapprocher de l'inégalité de Poincaré dans $H_0^1(\Omega)$.*

Démonstration : On raisonne par l'absurde, comme pour la démonstration de l'inégalité de Poincaré-Friedrichs [15]. Soit donc $(\mathbf{y}_k)_{k\geq 1}$ une suite d'éléments de X_E tels que

$$\|\mathbf{y}_k\|_{L^2(\Omega)^3} = 1, \ \|\mathbf{rot}\, \mathbf{y}_k\|^2_{L^2(\Omega)^3} + \|\operatorname{div} \varepsilon \mathbf{y}_k\|^2_{L^2(\Omega)} \leq \frac{1}{k}, \qquad \forall k \geq 1.$$

La suite $(\mathbf{y}_k)_{k\geq 1}$ est bornée dans X_E, puisque $\|\mathbf{y}_k\|_\varepsilon \leq \sqrt{2}$ pour tout $k \geq 1$. D'après le théorème de Weber, on peut en extraire une sous-suite $(\mathbf{y}_{k'})_{k'\geq 1}$ qui converge dans $L^2(\Omega)^3$, vers une limite notée $\mathbf{y}$: $\|\mathbf{y}_{k'} - \mathbf{y}\|_{L^2(\Omega)^3} \to 0$ et en particulier $\|\mathbf{y}\|_{L^2(\Omega)^3} = 1$.
Vérifions pour commencer que $\operatorname{div} \varepsilon \mathbf{y} = 0$ au sens des distributions. Soit $\varphi \in \mathcal{D}(\Omega)$:

$$\begin{aligned}\langle \operatorname{div} \varepsilon \mathbf{y}, \varphi \rangle &= -\langle \varepsilon \mathbf{y}, \nabla \varphi \rangle = -\int_\Omega \varepsilon \mathbf{y} \cdot \nabla \varphi \, d\Omega = -\int_\Omega \varepsilon \lim_{k'\to\infty} \mathbf{y}_{k'} \cdot \nabla \varphi \, d\Omega \\ &= -\lim_{k'\to\infty} \int_\Omega \varepsilon \mathbf{y}_{k'} \cdot \nabla \varphi \, d\Omega = \lim_{k'\to\infty} \int_\Omega (\operatorname{div} \varepsilon \mathbf{y}_{k'}) \varphi \, d\Omega = 0.\end{aligned}$$

Pour $f \in L^2(\Omega)$, on a identifié $\langle f, \varphi \rangle$ à $\int_\Omega f\varphi \, d\Omega$, puis on a utilisé la formule d'intégration par parties (2.102).
En particulier, $\operatorname{div} \varepsilon \mathbf{y}$ $(= 0)$ appartient à $L^2(\Omega)$ et on a : $\|\operatorname{div} \varepsilon \mathbf{y}_{k'} - \operatorname{div} \varepsilon \mathbf{y}\|_{L^2(\Omega)} \to 0$.
De la même façon, on vérifie que $\mathbf{rot}\, \mathbf{y} = 0$ au sens des distributions : ainsi, $\mathbf{y}$ appartient à $H(\mathbf{rot}\,; \Omega)$ et $\|\mathbf{y}_{k'} - \mathbf{y}\|_{H(\mathbf{rot}\,;\Omega)} \to 0$. Qui plus est, tous les termes de la suite $(\mathbf{y}_{k'})_{k'\geq 1}$ sont par hypothèse à trace tagentielle nulle. D'après la continuité de l'application trace tangentielle γ_T de $H(\mathbf{rot}\,; \Omega)$ dans $H^{-1/2}(\partial\Omega)^3$, on en déduit que $\mathbf{y} \times \mathbf{n}_{|\partial\Omega} = 0$.
Récapitulons, la limite $\mathbf{y}$ est telle que : $\mathbf{y} \in L^2(\Omega)^3$, $\|\mathbf{y}\|_{L^2(\Omega)^3} = 1$, $\|\operatorname{div} \varepsilon \mathbf{y}\|_{L^2(\Omega)} = 0$, $\|\mathbf{rot}\, \mathbf{y}\|_{L^2(\Omega)^3} = 0$ et $\mathbf{y} \times \mathbf{n}_{|\partial\Omega} = 0$. Montrons que ceci conduit à une contradiction. D'après la proposition 2.30 (ii), il existe q appartenant à $H_0^1(\Omega)$ tel que $\mathbf{y} = \nabla q$ dans Ω : d'après les propriétés de $\mathbf{y}$, q vérifie en plus $\|\nabla q\|_{L^2(\Omega)^3} = 1$ et $\operatorname{div} \varepsilon \nabla q = \operatorname{div} \varepsilon \mathbf{y} = 0$ dans Ω. Mais on a par intégration par parties :

$$\varepsilon_\star = \varepsilon_\star \|\nabla q\|^2_{L^2(\Omega)^3} \leq \int_\Omega \varepsilon|\nabla q|^2 d\Omega = \int_\Omega (\varepsilon\nabla q)\cdot\nabla q\, d\Omega = -\int_\Omega (\operatorname{div}\varepsilon\nabla q) q\, d\Omega = 0,$$

ce qui est une contradiction, puisque $\varepsilon_\star > 0$. ∎

Ainsi, d'après ce qui précède, on peut munir X_E d'une norme équivalente à $\|\cdot\|_\varepsilon$, à savoir

$$|\mathbf{v}|_\varepsilon = \Big(\|\mathbf{rot}\,\mathbf{v}\|^2_{L^2(\Omega)^3} + \|\operatorname{div}\varepsilon\mathbf{v}\|^2_{L^2(\Omega)}\Big)^{1/2}.$$

Or, $\operatorname{Ker} C_E$ étant un sous-espace (fermé) de X_E, on déduit de ce qui précède la coercivité de $a_E(\cdot,\cdot)$ sur $\operatorname{Ker} C_E$ puisque, pour tout élément $\mathbf{v}$ de ce noyau, on a justement $a_E(\mathbf{v},\mathbf{v}) = (\mathbf{rot}\,\mathbf{v},\mathbf{rot}\,\mathbf{v})_{L^2(\Omega)^3} = |\mathbf{v}|^2_\varepsilon$.

La condition inf–sup est quant à elle beaucoup plus simple à obtenir !
Pour cela, on va ré-utiliser la démarche permettant d'établir la nullité de la pression électrique artificielle λ_E. Pour $q \in H^1_0(\Omega)$, on sait que $\mathbf{v}_q = \nabla q \in H_0(\mathbf{rot}\,;\Omega)$ d'après la proposition de séquence exacte 2.35. Comme à la proposition 2.4, on écrit :

$$\sup_{\mathbf{v}\in H_0(\mathbf{rot}\,;\Omega)} \frac{c_E(\mathbf{v},q)}{\|\mathbf{v}\|_{H(\mathbf{rot}\,;\Omega)}|q|_1} \geq \frac{c_E(\mathbf{v}_q,q)}{\|\nabla q\|_{H(\mathbf{rot}\,;\Omega)}|q|_1} = \frac{\int_\Omega \varepsilon|\nabla q|^2\, d\Omega}{|q|^2_1} \geq \varepsilon_\star > 0.$$

Selon le théorème 2.1, la formulation variationnelle mixte en $\mathbf{E}$ est bien posée.

Faisons le point : quelle est la situation pour le problème quasi-statique en $\mathbf{E}$?
(i) Nous avons prouvé que toute solution du problème (2.101) vérifie les équations de la formulation variationnelle mixte en $\mathbf{E}$ (2.112)–(2.113), avec – automatiquement – $\lambda_E = 0$.
(ii) Nous venons de voir que cette formulation variationnelle mixte est bien posée.
Pour en conclure finalement que le problème de départ (2.101) est bien posé, il nous reste à prouver que la partie $\mathbf{E}$ de la formulation variationnelle mixte (2.112)–(2.113) vérifie bien les trois équations formant ce problème de départ. C'est cette dernière étape que nous formalisons ci-dessous. En particulier, nous ne "repassons" pas par le système *formel* (2.106), mais nous retrouvons *directement* (2.101).

Théorème 2.42 *Soient $\mathbf{f}_E \in L^2(\Omega)^3$ telle que $\operatorname{div}\mathbf{f}_E = 0$ dans Ω et $\mathbf{f}\cdot\mathbf{n}_{|\partial\Omega} = 0$, et $g_E \in H^{-1}(\Omega)$. Alors, le problème quasi-statique électrique (2.101) est bien posé dans $L^2(\Omega)^3$.*

Remarque 2.43 *Pour qu'un problème soit bien posé, il faut préciser l'espace fonctionnel dans lequel on cherche sa solution. C'est pourquoi, nous précisons que la solution $\mathbf{E}$ de (2.101) doit être cherchée dans $L^2(\Omega)^3$.*

Démonstration : D'après les étapes (i) et (ii) mentionnées ci-dessus, il reste à passer de (2.112)–(2.113) à (2.101). D'après la proposition 2.34, il est clair que l'on a bien $\operatorname{div}\varepsilon\mathbf{E} = g$ dans Ω. Et,

d'autre part, $\mathbf{E} \times \mathbf{n}_{|\partial\Omega} = 0$, puisque $\mathbf{E}$ appartient à $H_0(\mathbf{rot}\,;\Omega)$. Il nous reste à vérifier que $\mathbf{rot}\,\mathbf{E} = \mathbf{f}_E$ dans Ω. Pour cela, nous allons avoir besoin des propriétés satisfaites par la donnée $\mathbf{f}_E$ [7].
Formons $\mathbf{w} = \mathbf{rot}\,\mathbf{E} - \mathbf{f}_E$: par hypothèse, on a $\operatorname{div}\mathbf{w} = 0$ dans Ω et en particulier $\mathbf{w} \in H(\operatorname{div};\Omega)$. Par ailleurs, comme $\mathbf{E} \in H_0(\mathbf{rot}\,;\Omega)$, on sait que $\mathbf{rot}\,\mathbf{E} \in H_0(\operatorname{div};\Omega)$ d'après la deuxième propriété de séquence exacte (cf. proposition 2.36). Comme par ailleurs tout élément de $H_0(\operatorname{div};\Omega)$ est à trace normale nulle sur $\partial\Omega$ (d'après le théorème 2.27), on a donc $\mathbf{w} \cdot \mathbf{n}_{|\partial\Omega} = 0$.
Etudions ensuite le rotationnel de $\mathbf{w}$, au sens des distributions. Soit $\mathbf{z} \in \mathcal{D}(\Omega)^3$:

$$\langle \mathbf{rot}\,\mathbf{w}, \mathbf{z}\rangle = \langle \mathbf{w}, \mathbf{rot}\,\mathbf{z}\rangle = \int_\Omega \mathbf{w}\cdot\mathbf{rot}\,\mathbf{z}\,d\Omega = \int_\Omega (\mathbf{rot}\,\mathbf{E} - \mathbf{f}_E)\cdot\mathbf{rot}\,\mathbf{z}\,d\Omega = 0,$$

où la dernière égalité provient du fait que, $\mathbf{z}$ appartenant à $H_0(\mathbf{rot}\,;\Omega)$, elle peut être employée comme fonction-test dans (2.112). On en conclut que $\mathbf{rot}\,\mathbf{w} = 0$ dans Ω, et ainsi $\mathbf{w} \in H(\mathbf{rot}\,;\Omega)$. Récapitulons :

$$\mathbf{w} \in H_0(\operatorname{div};\Omega) \cap H(\mathbf{rot}\,;\Omega),\ \operatorname{div}\mathbf{w} = 0 \text{ et } \mathbf{rot}\,\mathbf{w} = 0 \text{ dans } \Omega.$$

Montrons maintenant que $\mathbf{w} = 0$. D'après la proposition 2.30 (i), il existe $q \in H^1(\Omega)$ tel que $\mathbf{w} = \nabla q$ dans Ω. Par construction, q est tel que :

$$q \in H^1(\Omega),\ \Delta q = \operatorname{div}\mathbf{w} = 0 \text{ dans } \Omega,\ \frac{\partial q}{\partial n}_{|\partial\Omega} = \mathbf{w}\cdot\mathbf{n}_{|\partial\Omega} = 0.$$

Ainsi, q est solution du problème de Laplace avec second membre nul, et condition aux limites de Neumann homogène : q est donc constant sur Ω, et son gradient est nul. On retrouve alors $\mathbf{w} = \nabla q = 0$ comme annoncé, c'est-à-dire que $\mathbf{rot}\,\mathbf{E} = \mathbf{f}_E$ dans Ω comme attendu. ∎

Ceci conclut l'étude du problème quasi-statique électrique.

Reprenons-en les grandes lignes, pour les transposer au cas magnétique.
Vérifions la coercivité sur le noyau de C_H, associée à $c_H(\cdot,\cdot)$ (cf. proposition 2.32) :

$$\operatorname{Ker} C_H = \left\{ \mathbf{v} \in H(\mathbf{rot}\,;\Omega) \,\middle|\, \operatorname{div}\mu\mathbf{v} = 0 \text{ dans } \Omega \text{ et } \mu\mathbf{v}\cdot\mathbf{n}_{|\partial\Omega} = 0 \right\}.$$

On introduit

$$X_H = \left\{ \mathbf{v} \in H(\mathbf{rot}\,;\Omega) \,\middle|\, \operatorname{div}\mu\mathbf{v} \in L^2(\Omega) \text{ et } \mu\mathbf{v}\cdot\mathbf{n}_{|\partial\Omega} = 0 \right\},$$

muni de la norme induite $\|\cdot\|_\mu$. Dans le cas où μ est globalement constante sur Ω, on utilisera la notation

$$X_H^{cst} = H(\mathbf{rot}\,;\Omega) \cap H_0(\operatorname{div};\Omega).$$

On peut établir une seconde inégalité de Weber dans X_H. Avant cela, on démontre un second résultat d'injection compacte, toujours dû à Weber [50, 31].

7. D'une part, c'est uniquement maintenant que nous utilisons les propriétés satisfaites par $\mathbf{f}_E$; d'autre part, rappelons que si $\mathbf{f}_E$ ne vérifie pas la propriété de divergence nulle ou celle de trace normale nulle, alors on ne peut pas avoir $\mathbf{rot}\,\mathbf{E} = \mathbf{f}_E$ dans Ω, avec $\mathbf{E} \in H_0(\mathbf{rot}\,;\Omega)$, d'après la deuxième propriété de séquence exacte pour la condition aux limites.

Théorème 2.44 (Weber – cas magnétique) *L'injection canonique de X_H dans $L^2(\Omega)^3$ est compacte.*

Démonstration (esquissée) : On raisonne comme pour la démonstration du théorème 2.37, en utilisant cette fois une décomposition continue [20] de tout élément $\mathbf{v}$ de X_H de la forme (2.115), avec une partie $\mathbf{w} \in X_H$ "régulière par morceaux", et un gradient d'un champ scalaire $\phi_0 \in H^1(\Omega)$, solution d'un problème du type $\operatorname{div} \mu \nabla \phi_0 = f$ dans Ω ($f \in L^2(\Omega)$) avec condition aux limites homogène de type Neumann. La régularité globale de chaque terme permet de conclure que $\mathbf{v} \in \cap_{s<\sigma'_\mu} H^s(\Omega)^3$, pour $\sigma'_\mu > 0$ indépendant de $\mathbf{v}$, et que de plus l'injection de X_H dans $\cap_{s<\sigma'_\mu} H^s(\Omega)^3$ est continue. La conclusion suit, par composition d'une application continue par une application compacte. ∎

A partir de là, la deuxième inégalité de Weber peut être démontrée.

Proposition 2.45 (Inégalité de Weber – cas magnétique) *Il existe une constante C'_W, dépendant seulement de Ω telle que :*

$$\int_\Omega |\mathbf{v}|^2 \, d\Omega \leq C'_W \int_\Omega \left(|\mathbf{rot}\, \mathbf{v}|^2 + (\operatorname{div} \mu \mathbf{v})^2 \right) d\Omega, \ \forall \mathbf{v} \in X_H. \tag{2.117}$$

Démonstration : Laissée en exercice (raisonner par l'absurde). ∎

Remarque 2.46 *(Cas $\varepsilon = \mu = 1$) Les deux inégalités de Weber sont valables dans les sous-espaces X_E^{cst} et X_H^{cst} de $H(\mathbf{rot}\,;\Omega) \cap H(\mathrm{div};\Omega)$, pour lesquels on annule soit la trace tangentielle, soit la trace normale sur la frontière. Si on choisit d'annuler l'une sur un bout de la frontière, et l'autre sur le reste, un résultat similaire peut-être établi. Nous renvoyons à [25] pour plus de détails, ainsi qu'à des généralisations au cas de traces non-nulles mais appartenant à $L^2(\partial\Omega)$, composante par composante. Enfin, notons que l'on peut facilement prouver que $X_E^{cst} \cap X_H^{cst} = H_0^1(\Omega)^3$ (cf. [3]). Bref, contrôler le rotationnel et la divergence revient à contrôler le gradient, à la condition expresse d'annuler toutes les traces des champs, tangentielle et normale.*

On conclut à la coercivité de $a_H(\cdot,\cdot)$ sur $\operatorname{Ker} C_H$.

La condition inf–sup s'obtient de façon élémentaire, comme dans le cas électrique, et la formulation variationnelle mixte (2.109)–(2.110) est bien posée.

Enfin, pour revenir de cette formulation mixte au problème quasi-statique magnétique (2.100), on démontre le résultat ci-dessous, encore une fois sans "repasser" par le système *formel* (2.107).

Théorème 2.47 *Soit $\mathbf{f}_H \in L^2(\Omega)^3$ telle que $\operatorname{div} \mathbf{f}_H = 0$ dans Ω. Alors, le problème quasi-statique magnétique (2.100) est bien posé dans $L^2(\Omega)^3$.*

Démonstration : Laissée en exercice. ∎

Remarque 2.48 *Si l'on revient au modèle quasi-statique initial (2.100-2.101), on se souvient que le second membre* $\mathbf{f}_H$ *est égal à* $\mathbf{J}$*, la densité de courant (qui est une donnée) : ainsi, la norme* $\|\mathbf{H}\|_\mu$ *dépend continûment de* $\|\mathbf{J}\|_{L^2(\Omega)^3}$*. Par contre, si* g_E *est bien une donnée (*$g_E = \rho$*), on a* $\mathbf{f}_E = -\mu\partial_t\mathbf{H}$*, ce qui n'est pas une donnée, mais la solution du problème quasi-statique magnétique. Toutefois, puisque* $\mathbf{H}$ *est solution de (2.100) avec pour donnée* $\mathbf{J}$*, alors* $\partial_t\mathbf{H}$ *est solution de (2.100) avec pour donnée* $\partial_t\mathbf{J}$ *! Donc* $\|\partial_t\mathbf{H}\|_\mu$ *dépend continûment de* $\|\partial_t\mathbf{J}\|_{L^2(\Omega)^3}$*, ce qui nous permet finalement de conclure que* $\|\mathbf{E}\|_\varepsilon$ *dépend continûment de* $\|\partial_t\mathbf{J}\|_{L^2(\Omega)^3}$ *et* $\|\rho\|_{H^{-1}(\Omega)}$*.*

2.3.4 Discrétisation

On suppose dans cette sous-section que les données satisfont aux hypothèses garantissant l'existence de solutions aux problèmes quasi-statiques électrique et magnétique. Voir pour cela les théorèmes 2.42 et 2.47, ainsi que la remarque 2.48. On suppose en outre que le domaine Ω est à frontière polyédrique.

Elément fini de Raviart-Thomas-Nédélec

Une des clefs pour la résolution des problèmes électrique ou magnétique est l'utilisation des propriétés de séquence exacte (voir la proposition 2.35) :

$$q \in M_E = H_0^1(\Omega) \Longrightarrow \nabla q \in V_E = H_0(\mathbf{rot}\,;\Omega),$$
$$q \in M_H = H^1(\Omega) \cap L_0^2(\Omega) \Longrightarrow \nabla q \in V_H = H(\mathbf{rot}\,;\Omega).$$

Au niveau discret, une idée "naturelle" (mais non triviale à réaliser en pratique !) est de reproduire ces propriétés...
Choisissons pour approcher les multiplicateurs de M_H, une discrétisation par éléments finis de Lagrange P^1. Soit donc $(\mathcal{T}_h)_h$ une suite de maillages [8] de Ω, composés de tétraèdres K :

$$M_h = \{q_h \in \mathcal{C}^0(\bar{\Omega}) \text{ tel que } q_{h|K} \in P^1(K),\ \forall K \in \mathcal{T}_h\},$$

et on choisit $M_h \cap L_0^2(\Omega)$ comme espace d'approximation ; et, pour ceux de M_E,

$$M_h^0 = \{q_h \in M_h \text{ tel que } q_{h|\partial\Omega} = 0\}.$$

Par construction, pour tout $q_h \in M_h$, on a en particulier $q_h \in H^1(\Omega)$. On déduit de la propriété de séquence exacte de la proposition 2.35 que $\mathbf{v}_h = \nabla q_h \in V_H$. Et, de plus, pour tout tétraèdre K de $\mathcal{T}_h$, $\mathbf{v}_{h|K}$ est un vecteur constant de $\mathbb{R}^3$ (on écrit

8. Les maillages $(\mathcal{T}_h)_h$ répondent aux critères usuels (cf. [15]). On suppose en particulier que $\bar{\Omega} = \cup_{K\in\mathcal{T}_h} K$, et que si on appelle h_K le rayon de la plus petite sphère circonscrite à K pour tout tétraèdre, $h = \max_K h_K$ dénote le pas de la triangulation.

$\mathbf{v}_{h|K} \in P^0(K)^3$).
Du point de vue de la construction d'un espace d'approximation, c'est un début... Malheureusement, on a automatiquement $\mathbf{rot}\,\mathbf{v}_h = 0$, puisque $\mathbf{v}_h$ est un gradient, et ceci resterait valable quel que soit l'ordre de l'élément fini de Lagrange définissant M_h.
Essayons donc d'adjoindre à ces approximations constantes par tétraèdre, des approximations à rotationnel constant par tétraèdre. En toute généralité, c'est un problème élémentaire, si l'on se souvient que, pour tout vecteur $\mathbf{b} \in \mathbb{R}^3$, on a la propriété :

$$\mathbf{rot}\,(\mathbf{b} \times \mathbf{x}) = 2\mathbf{b}\ !$$

Ainsi, un "bon" candidat à l'approximation par éléments finis conformes dans $H(\mathbf{rot}\,;\Omega)$ est du type :

$$\mathbf{v}_{h|K}(\mathbf{x}) = \mathbf{a}_K + \mathbf{b}_K \times \mathbf{x},\ \text{avec}\ \mathbf{a}_K, \mathbf{b}_K \in \mathbb{R}^3,\ \forall K \in \mathcal{T}_h.$$

De cette façon, on englobe les gradients d'éléments finis de Lagrange P^1, auxquels on a ajouté des champs discrets à rotationnels non-nuls. Ceci permet de transposer la propriété de séquence exacte *au niveau discret*.
La question est : ceci permet-il de définir un élément fini ? La réponse est oui, et elle a été apportée par Raviart-Thomas en 2D [44], et par Nédélec en 3D [41]. Rigoureusement, il nous reste à définir les degrés de liberté : la solution consiste à choisir des degrés de liberté de type moment, de support les *arêtes* de la frontière du tétraèdre.

Fixons le tétraèdre K, et introduisons

$$\mathcal{R}_1 = \{\mathbf{v} \in (P^1(K))^3 \text{ tel que } \mathbf{v}(\mathbf{x}) = \mathbf{a} + \mathbf{b} \times \mathbf{x},\ \mathbf{a}, \mathbf{b} \in \mathbb{R}^3\},$$

un espace de polynômes définis sur K, et de dimension 6. On désigne par A_K l'ensemble de ses six arêtes a et par $\mathbf{t}$ un vecteur unitaire dans la direction de a (voir la figure 2.3).

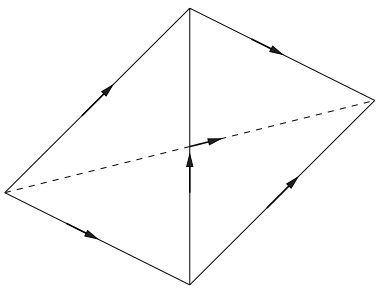


Figure 2.3. Un tétraèdre et ses six arêtes

On va tout d'abord montrer le résultat suivant.

Proposition 2.49 *Un champ de vecteurs* $\mathbf{v}$ *de* $\mathcal{R}_1$ *est entièrement déterminé par les six moments*

$$M_a(\mathbf{v}) = \int_a \mathbf{v}\cdot\mathbf{t}\, dl, \ \textit{pour } a \in A_K. \tag{2.118}$$

De plus, la quantité $\mathbf{v}\times\mathbf{n}_{|f}$ *s'annule sur une face* f *de* K *si, et seulement si, ses moments (2.118) s'annulent pour chaque arête* a *de la face* f*.*

Remarque 2.50 *Si on note* Σ *l'ensemble des six moments* $(M_a)_{a\in A_K}$*, on démontre ici l'unisolvance du triplet* $(K,\Sigma,\mathcal{R}_1)$*.*

Démonstration : Commençons par la seconde partie. Fixons f une face de K. Soit $\mathbf{v}\in\mathcal{R}_1$ de la forme $\mathbf{v}(\mathbf{x}) = \mathbf{a}+\mathbf{b}\times\mathbf{x}$ tel que $M_a(\mathbf{v}) = 0$ pour toutes les arêtes a de f. Supposons que f soit incluse dans le plan $\{\mathbf{x} : x_3 = 0\}$. Ainsi, la composante tangentielle $\mathbf{v}_T$ de $\mathbf{v}$ s'écrit sur la face f

$$\mathbf{v}_T(x_1,x_2) = \begin{pmatrix} v_1(x_1,x_2,0)\\ v_2(x_1,x_2,0)\end{pmatrix} = \begin{pmatrix} a_1 - b_3x_2\\ a_2+b_3x_1\end{pmatrix}.$$

Sous l'hypothèse $M_a(\mathbf{v}) = 0$ pour toute arête a de f, et à l'aide d'une formule d'intégration par parties sur la face, on trouve que

$$\int_f \mathrm{rot}\mathbf{v}_T\, ds = 0,$$

où le rotationnel bidimensionnel scalaire est défini par $\mathrm{rot}\mathbf{v}_T = \partial_1 v_2 - \partial_2 v_1$. Or, $\mathrm{rot}\mathbf{v}_T = 2b_3$, d'où $b_3 = 0$ et donc $\mathbf{v}_T$ est constant. Comme $M_a(\mathbf{v}) = 0$ pour chaque arête de f, on obtient que $\mathbf{v}_T = 0$ sur f. Ceci prouve que $\mathbf{v}_T = 0$ sur f si, et seulement si, $M_a(\mathbf{v}) = 0$ pour chaque arête a de f, puisque la réciproque est immédiate. Nous aurions également pu raisonner avec la quantité $\mathbf{v}\times\mathbf{n}_{|f}$, qui s'annule bien sûr si, et seulement si, $\mathbf{v}_T$ s'annule sur f.

Comme le nombre de moments $(M_a)_{a\in A_K}$ est égal à la dimension de $\mathcal{R}_1$, démontrer que tout élément $\mathbf{v}$ de $\mathcal{R}_1$ est complètement déterminé par les moments revient à établir que $M_a(\mathbf{v}) = 0$ pour tout $a\in A_K$ implique que $\mathbf{v} = 0$. Supposons donc que les moments (2.118) de $\mathbf{v}$ s'annulent pour chaque arête a de K. On sait alors, d'après ce que l'on vient de voir, que $\mathbf{v}\times\mathbf{n}$ s'annule sur chaque face de K, et grâce cette fois à la formule d'intégration par parties (2.104), on a

$$\int_K \mathbf{rot}\,\mathbf{v}\, d\Omega = 0.$$

Comme $\mathbf{rot}\,\mathbf{v} = 2\mathbf{b}$, on en déduit $\mathbf{b} = 0$ et donc $\mathbf{v}$ est constant sur K. Puisque $0 = \mathbf{v}\times\mathbf{n}$ sur *chaque face* de K, on en déduit que $\mathbf{v} = 0$. ∎

A partir du *même* maillage $\mathcal{T}_h$ que celui utilisé pour définir l'espace des multiplicateurs discrets M_h, on définit alors l'espace discret V_h avec

$$V_h = \{\mathbf{v}\in H(\mathbf{rot}\,;\Omega) \text{ tel que } \mathbf{v}_{|K}\in\mathcal{R}_1(K),\ \forall K\in\mathcal{T}_h\}.$$

En effet, d'après la proposition sur la trace tangentielle 2.25, on peut vérifier que toute fonction $\mathbf{v}$ définie sur Ω, et régulière sur chaque tétraèdre de $\mathcal{T}_h$, appartient à l'espace $H(\mathbf{rot}\,;\Omega)$ si, et seulement si, $\mathbf{v}\times\mathbf{n}$ est continue au travers de l'ensemble des faces communes à deux tétraèdres. Ainsi, en utilisant la proposition 2.49, on

peut choisir les degrés de liberté d'une fonction $\mathbf{v} \in V_h$ comme étant les moments (2.118) sur l'ensemble des arêtes de la triangulation $\mathcal{T}_h$ – on écrit $\{a \,|\, a \in \mathcal{T}_h\}$ – : il y a autant de degrés de liberté que d'arêtes dans la triangulation. Ceci permet d'approcher la solution du problème magnétique. Pour le problème électrique, on choisit

$$V_h^0 = V_h \cap H_0(\mathbf{rot}\,;\Omega) = \{\mathbf{v} \in V_h \text{ tel que } \mathbf{v} \times \mathbf{n}_{|\partial\Omega} = 0\}.$$

Toujours d'après la proposition précédente, on établit facilement qu'une fonction $\mathbf{v}$ de V_h appartiendra à V_h^0 (trace tangentielle nulle sur la frontière) si, et seulement si, ses moments (2.118) s'annulent sur chaque arête a de la frontière $\partial\Omega$. Une fonction de V_h^0 a donc pour degrés de liberté les moments (2.118) sur l'ensemble des arêtes *internes* de la triangulation $\mathcal{T}_h$ – on écrit $\{a \,|\, a \in \mathcal{T}_h \setminus \partial\Omega\}$ – : il y a autant de degrés de liberté que d'arêtes *internes* dans la triangulation.

Nous pouvons maintenant écrire les problèmes discrétisés. Pour le champ magnétique, il s'écrit

$$\begin{cases} \text{trouver } (\mathbf{H}_h, \lambda_H^h) \in V_h \times (M_h \cap L_0^2(\Omega)) \text{ tel que} \\ \displaystyle\int_\Omega \mathbf{rot}\,\mathbf{H}_h \cdot \mathbf{rot}\,\mathbf{v}_h \, d\Omega + \int_\Omega \mu \mathbf{v}_h \cdot \nabla \lambda_H^h \, d\Omega \\ \qquad\qquad = \displaystyle\int_\Omega \mathbf{f}_H \cdot \mathbf{rot}\,\mathbf{v}_h \, d\Omega \qquad \forall \mathbf{v}_h \in V_h, \qquad (2.119) \\ \displaystyle\int_\Omega \mu \mathbf{H}_h \cdot \nabla q_h \, d\Omega = 0 \quad \forall q_h \in (M_h \cap L_0^2(\Omega)). \qquad (2.120) \end{cases}$$

Comme on a préservé la propriété de séquence exacte au niveau discret, on peut utiliser $\mathbf{v}_h = \nabla \lambda_H^h$ comme fonction-test dans (2.119), pour trouver :

$$\int_\Omega \mu |\nabla \lambda_H^h|^2 \, d\Omega = 0.$$

Sachant que $\mu \geq \mu_\star > 0$ presque partout, le multiplicateur discret λ_H^h est constant. Comme il est de plus à moyenne nulle : $\lambda_H^h = 0$. C'est donc une excellente approximation de la pression magnétique artificielle λ_H, nulle elle aussi !
Pour le champ électrique, le problème discrétisé s'écrit cette fois

$$\begin{cases} \text{trouver } (\mathbf{E}_h, \lambda_E^h) \in V_h^0 \times M_h^0 \text{ tel que} \\ \displaystyle\int_\Omega \mathbf{rot}\,\mathbf{E}_h \cdot \mathbf{rot}\,\mathbf{v}_h \, d\Omega + \int_\Omega \varepsilon \mathbf{v}_h \cdot \nabla \lambda_E^h \, d\Omega = \int_\Omega \mathbf{f}_E \cdot \mathbf{rot}\,\mathbf{v}_h \, d\Omega \ \forall \mathbf{v}_h \in V_h^0, \\ \displaystyle\int_\Omega \varepsilon \mathbf{E}_h \cdot \nabla q_h \, d\Omega = -\langle g_E, q_h \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \quad \forall q_h \in M_h^0. \end{cases}$$

Grâce une nouvelle fois à la propriété de séquence exacte au niveau discret, on vérifie que $\lambda_E^h = 0$, ce qui constitue également une excellente approximation de la pression électrique artificielle $\lambda_E = 0$.

Résultats d'interpolation

Pour pouvoir définir un opérateur d'interpolation π_h à valeurs dans V_h, il faut que les moments $(M_a(\mathbf{v}))_{a\in\mathcal{T}_h}$ (cf. (2.118)) aient un sens.
Considérons d'abord l'opérateur d'interpolation local π_K, défini sur un tétraèdre K : d'après [3], on peut prouver que les moments $(M_a(\mathbf{v}))_{a\in A_K}$ ont un sens dès lors que $\mathbf{v} \in X_p(K)$, avec $p > 2$, où

$$X_p(K) = \{\mathbf{v} \in L^p(K)^3 \text{ tel que } \mathbf{rot}\,\mathbf{v} \in L^p(K)^3,\ \mathbf{v}\times\mathbf{n}_{|\partial K} \in L^p(\partial K)^3\}.$$

Par conséquent, $\pi_K\mathbf{v}$ existe pour tout $\mathbf{v} \in X_p(K)$: c'est par définition l'élément de $\mathcal{R}_1$ qui a pour six moments $(M_a(\mathbf{v}))_{a\in A_K}$.
Par extension, l'opérateur d'interpolation global π_h est à valeurs dans V_h, et tel que $\pi_h\mathbf{v}_{|K} = \pi_K\mathbf{v}$, pour tout $K \in \mathcal{T}_h$ (sous réserve que tous les $\pi_K\mathbf{v}$ existent...).

Remarque 2.51 *Classiquement, on peut démontrer que si* $\mathbf{v} \in H^s(K)^3$*, avec* $s > 1/2$*, il existe* $p_s > 2$ *(qui dépend de s) tel que d'une part* $\mathbf{v} \in L^{p_s}(K)^3$*, et d'autre part* $\mathbf{v}\times\mathbf{n}_{|\partial K} \in L^{p_s}(\partial K)^3$*.*

On en déduit que si $\mathbf{v} \in H^s(K)^3$ et $\mathbf{rot}\,\mathbf{v} \in H^s(K)^3$, avec $s > 1/2$, alors $\mathbf{v} \in X_p(K)$ et ainsi $\pi_K\mathbf{v}$ existe. Par ailleurs, on peut établir un résultat d'approximabilité (cf. [16, 12]), qui repose sur le même type de régularité.

Proposition 2.52 *Soit* $s \in]1/2, 1]$.
Pour tout $\mathbf{v} \in H^s(K)^3$ *tel que* $\mathbf{rot}\,\mathbf{v} \in H^s(K)^3$, $\pi_K\mathbf{v}$ *existe et on a*

$$\|\mathbf{v} - \pi_K\mathbf{v}\|_{H(\mathbf{rot};K)} \le C_s\, h_K^s \left(\|\mathbf{v}\|_{H^s(K)^3} + \|\mathbf{rot}\,\mathbf{v}\|_{H^s(K)^3}\right).$$

Pour tout $\mathbf{v} \in H^s(\Omega)^3$ *tel que* $\mathbf{rot}\,\mathbf{v} \in H^s(\Omega)^3$, $\pi_h\mathbf{v}$ *existe et on a*

$$\|\mathbf{v} - \pi_h\mathbf{v}\|_{H(\mathbf{rot};\Omega)} \le C_s\, h^s \left(\|\mathbf{v}\|_{H^s(\Omega)^3} + \|\mathbf{rot}\,\mathbf{v}\|_{H^s(\Omega)^3}\right).$$

Ci-dessus, $C_s > 0$ *est indépendante de* $\mathbf{v}$ *(et de K).*

Rappelons également le résultat technique suivant, dû à Nédélec [41].

Proposition 2.53 *Soit* $q \in H^1(\Omega)$ *et* $\mathbf{v} = \nabla q$*. Alors, si* $\pi_h\mathbf{v}$ *est bien défini, il existe* $q_h \in M_h$ *tel que* $\pi_h\mathbf{v} = \nabla q_h$*. Si de plus* $q \in H^1_0(\Omega)$*, alors* $q_h \in M_h^0$*.*

Comment garantir que l'opérateur d'interpolation π_h ait un sens, lorsqu'on l'applique aux solutions du problème quasi-statique (2.100-2.101) ?

Concentrons-nous sur le champ électrique $\mathbf{E}$, et plaçons-nous dans le cas où ε est globalement constante sur Ω. Lorsque *de plus* $g_E \in L^2(\Omega)$, $\mathbf{E} \in X_E^{cst}$ et la remarque 2.39 nous permet d'affirmer que $\mathbf{E} \in \cap_{s<\sigma_D} H^s(\Omega)^3$, avec $\sigma_D > 1/2$. Il reste maintenant à étudier la régularité de $\mathbf{rot}\,\mathbf{E}$. Tout d'abord, d'après la propriété de

séquence exacte de la proposition 2.36, on a $\mathbf{rot}\,\mathbf{E} \in H_0(\mathrm{div};\Omega)$. Qui plus est, on sait que $\mathbf{rot}\,\mathbf{E} = \mathbf{f}_E$ (cf. la première équation de (2.101)). Sous l'hypothèse *supplémentaire* que $\mathbf{f}_E \in H(\mathbf{rot}\,;\Omega)$, on en déduit que $\mathbf{rot}\,\mathbf{E} \in H_0(\mathrm{div};\Omega) \cap H(\mathbf{rot}\,;\Omega)$, c'est-à-dire l'espace fonctionnel X_H^{cst}. Si on reprend la démonstration du théorème 2.44 et la remarque 2.39 transposée au cas magnétique, on en déduit que $X_H^{cst} \subset \cap_{s<\sigma_N} H^s(\Omega)^3$, avec $\sigma_N > 1/2$ l'exposant limite de régularité du problème de Laplace avec condition aux limites homogène de Neumann.

On pourrait raisonner selon les mêmes lignes pour $\mathbf{H}$, sous des hypothèses similaires (μ constante dans Ω et $\mathbf{f}_H \in H_0(\mathbf{rot}\,;\Omega)$).

Proposition 2.54 *Supposons que ε soit constante sur Ω et $\mathbf{f}_E \in H(\mathbf{rot}\,;\Omega)$, $g_E \in L^2(\Omega)$. Alors*

$$\mathbf{E} \in \cap_{s<\sigma_D} H^s(\Omega)^3, \quad \mathbf{rot}\,\mathbf{E} \in \cap_{s<\sigma_N} H^s(\Omega)^3.$$

Supposons que μ soit constante sur Ω et $\mathbf{f}_H \in H_0(\mathbf{rot}\,;\Omega)$. Alors

$$\mathbf{H} \in \cap_{s<\sigma_N} H^s(\Omega)^3, \quad \mathbf{rot}\,\mathbf{H} \in \cap_{s<\sigma_D} H^s(\Omega)^3.$$

En conclusion, nous aboutissons à un résultat d'approximabilité des solutions du problème quasi-statique (2.100-2.101). Posons $\sigma = \min(\sigma_D, \sigma_N) > 1/2$.

Théorème 2.55 *Supposons que ε et μ soient constantes sur Ω, et $\mathbf{f}_E \in H(\mathbf{rot}\,;\Omega)$, $g_E \in L^2(\Omega)$, $\mathbf{f}_H \in H_0(\mathbf{rot}\,;\Omega)$. Alors*

$$\|\mathbf{E} - \pi_h\mathbf{E}\|_{H(\mathbf{rot}\,;\Omega)} \le C_s\, h^s \left(\|\mathbf{E}\|_{H^s(\Omega)^3} + \|\mathbf{rot}\,\mathbf{E}\|_{H^s(\Omega)^3}\right),$$
$$\|\mathbf{H} - \pi_h\mathbf{H}\|_{H(\mathbf{rot}\,;\Omega)} \le C'_s\, h^s \left(\|\mathbf{H}\|_{H^s(\Omega)^3} + \|\mathbf{rot}\,\mathbf{H}\|_{H^s(\Omega)^3}\right),$$

pour tout $s < \sigma$, avec $C_s, C'_s > 0$ indépendantes de $(\mathbf{E}, \mathbf{H})$.

Uniformité de la coercivité et vitesse de convergence

On a vu que dans le cas des équations de Stokes, la condition (2.67) de coercivité uniforme sur le noyau discret est automatiquement satisfaite, puisque la forme bilinéaire $a(\cdot,\cdot)$ est coercive sur l'espace discret global. La difficulté réside dans l'établissement de la condition inf–sup discrète uniforme (2.68).
Dans le cas de l'électromagnétisme quasi-statique, la situation est inverse ! En effet, la condition inf–sup discrète uniforme est simple à établir : son obtention suit pas à pas celle du problème continu, puisqu'on a conservé la propriété de séquence exacte au niveau discret (ce qui constitue une propriété remarquable de la paire d'éléments finis choisie). Par contre, l'établissement de la coercivité uniforme est plus délicat.
Nous nous concentrons dans ce qui suit sur le problème électrique : la forme bilinéaire $a_E(\mathbf{v}_h, \mathbf{w}_h) = (\mathbf{rot}\,\mathbf{v}_h, \mathbf{rot}\,\mathbf{w}_h)_{L^2(\Omega)^3}$ est définie ici sur $V_h^0 \times V_h^0$, où V_h^0 est muni du produit scalaire de $H(\mathbf{rot}\,;\Omega)$, et on note K_h le noyau

$$K_h = \{\mathbf{v}_h \in V_h^0 \text{ tel que } \int_\Omega \varepsilon \mathbf{v}_h \cdot \nabla q_h \, d\Omega = 0, \ \forall q_h \in M_h^0\}.$$

Proposition 2.56 *La propriété (2.67) de coercivité uniforme de la forme $a_E(\cdot,\cdot)$ est vérifiée sur le noyau K_h.*

Note : $a_E(\mathbf{v}_h, \mathbf{v}_h) = \|\mathbf{rot}\, \mathbf{v}_h\|^2_{L^2(\Omega)^3}$; lorsque $\mathbf{v}_h$ est un élément de K_h, il s'agit donc de prouver que l'on peut contrôler la norme $\|\mathbf{v}_h\|_{L^2(\Omega)^3}$ par $\|\mathbf{rot}\, \mathbf{v}_h\|_{L^2(\Omega)^3}$ pour établir la coercivité.

Démonstration : Soit $\mathbf{v}_h \in K_h$ non-nul.
On lui associe l'unique solution de problème variationnel

$$\begin{cases} \text{trouver } \varphi \in H_0^1(\Omega) \text{ tel que} \\ \displaystyle\int_\Omega \nabla\varphi \cdot \nabla\psi \, d\Omega = \int_\Omega \mathbf{v}_h \cdot \nabla\psi \, d\Omega, \ \forall \psi \in H_0^1(\Omega). \end{cases}$$

Posons $\mathbf{w} = \mathbf{v}_h - \nabla\varphi$. Par construction, on a les propriétés suivantes pour $\mathbf{w} \in L^2(\Omega)^3$:

$$\mathbf{rot}\, \mathbf{w} = \mathbf{rot}\, \mathbf{v}_h \text{ et } \operatorname{div} \mathbf{w} = 0 \text{ dans } \Omega, \quad \mathbf{w} \times \mathbf{n} = 0 \text{ sur } \partial\Omega.$$

Ainsi $\mathbf{w} \in X_E^{cst}$ et donc $\mathbf{w} \in H^{1/2+t}(\Omega)^3$ pour un certain $t > 0$, d'après la remarque 2.39. Qui plus est, comme $\mathbf{v}_h$ est un champ discret, son rotationnel est borné : $\mathbf{rot}\, \mathbf{w} = \mathbf{rot}\, \mathbf{v}_h \in L^\infty(\Omega)^3$. D'après la remarque 2.51, $\pi_h \mathbf{w}$ existe. Comme par définition $\pi_h \mathbf{v}_h = \mathbf{v}_h$, on en déduit par différence que $\pi_h(\nabla\varphi)$ existe aussi. D'après la proposition 2.53, il existe $\varphi_h \in M_h^0$ tel que

$$\mathbf{v}_h = \pi_h \mathbf{w} + \nabla\varphi_h.$$

Or, $\mathbf{v}_h$ est un élément du noyau K_h, on en déduit par orthogonalité que

$$\begin{aligned} &\varepsilon_\star \|\mathbf{v}_h\|^2_{L^2(\Omega)^3} \leq \int_\Omega \varepsilon |\mathbf{v}_h|^2 \, d\Omega = \int_\Omega \varepsilon \mathbf{v}_h \cdot \pi_h \mathbf{w} \, d\Omega \leq \varepsilon^\star \|\mathbf{v}_h\|_{L^2(\Omega)^3} \|\pi_h \mathbf{w}\|_{L^2(\Omega)^3}, \\ &\text{d'où : } \|\mathbf{v}_h\|_{L^2(\Omega)^3} \leq \frac{\varepsilon^\star}{\varepsilon_\star} \|\pi_h \mathbf{w}\|_{L^2(\Omega)^3}. \end{aligned} \tag{2.121}$$

A partir de là, nous allons estimer $\|\pi_h \mathbf{w}\|_{L^2(\Omega)^3}$.
$\mathbf{rot}\, \mathbf{w} = \mathbf{rot}\, \mathbf{v}_h$ est constant par tétraèdre, d'où en particulier $\mathbf{rot}\, \mathbf{w} \in H^s(K)^3$ pour tout $s > 0$ et tout K. D'après la proposition 2.52, on peut écrire, pour un certain $s \in]1/2, 1]$ qui sera précisé plus bas :

$$\begin{aligned} \|\mathbf{w} - \pi_K \mathbf{w}\|_{L^2(K)^3} &\leq C_s \, h_K^s \left(\|\mathbf{w}\|_{H^s(K)^3} + \|\mathbf{rot}\, \mathbf{w}\|_{H^s(K)^3} \right) \\ &\leq C_s \, h_K^s \left(\|\mathbf{w}\|_{H^s(K)^3} + \|\mathbf{rot}\, \mathbf{v}_h\|_{H^s(K)^3} \right) \\ &\leq C_s \, h_K^s \left(\|\mathbf{w}\|_{H^s(K)^3} + \|\mathbf{rot}\, \mathbf{v}_h\|_{L^2(K)^3} \right). \end{aligned}$$

Pour la dernière (in)égalité, nous avons utilisé le fait que les champs discrets de V_h^0 étant à rotationnel constant par tétraèdre, on a $\|\mathbf{rot}\, \mathbf{v}_h\|_{H^s(K)^3} = \|\mathbf{rot}\, \mathbf{v}_h\|_{L^2(K)^3}$. En sommant les inégalités précédentes sur tous les tétraèdres, on en déduit que :

$$\|\mathbf{w} - \pi_h \mathbf{w}\|_{L^2(\Omega)^3} \leq \sqrt{2} C_s \, h^s \left(\|\mathbf{w}\|_{H^s(\Omega)^3} + \|\mathbf{rot}\, \mathbf{v}_h\|_{L^2(\Omega)^3} \right).$$

A partir de maintenant, on suppose que $h \leq 1$, ce qui est loisible puisque h est un paramètre destiné à tendre vers 0. A l'aide de l'inégalité triangulaire, on estime $\|\pi_h \mathbf{w}\|_{L^2(\Omega)^3}$, pour trouver :

$$\|\pi_h \mathbf{w}\|_{L^2(\Omega)^3} \leq (1+\sqrt{2}C_s)\,\left(\|\mathbf{w}\|_{L^2(\Omega)^3} + \|\mathbf{w}\|_{H^s(\Omega)^3} + \|\mathbf{rot}\,\mathbf{v}_h\|_{L^2(\Omega)^3}\right).$$

Comme $\|\mathbf{w}\|_{L^2(\Omega)^3} \leq \|\mathbf{w}\|_{H^s(\Omega)^3}$ ($H^s(\Omega) \subset L^2(\Omega)^3$), la norme L^2 est absorbée par la norme H^s. Ensuite, on se souvient que l'espace X_E^{cst} s'injecte continûment dans $H^s(\Omega)^3$, pour $s > 1/2$ convenable, et donc on peut majorer la norme $\|\mathbf{w}\|_{H^s(\Omega)^3}$ par $C\,(\|\mathbf{rot}\,\mathbf{w}\|^2_{L^2(\Omega)^3} + \|\operatorname{div}\mathbf{w}\|^2_{L^2(\Omega)})^{1/2}$, avec une constante $C > 0$ de continuité indépendante de $\mathbf{w}$. Comme par construction $\operatorname{div}\mathbf{w} = 0$ et $\mathbf{rot}\,\mathbf{w} = \mathbf{rot}\,\mathbf{v}_h$ dans Ω, il reste finalement :

$$\|\pi_h \mathbf{w}\|_{L^2(\Omega)^3} \leq C_s' \,\|\mathbf{rot}\,\mathbf{v}_h\|_{L^2(\Omega)^3},$$

avec $C_s' > 0$ indépendante de $\mathbf{v}_h$. A l'aide de (2.121), on en conclut que

$$\|\mathbf{v}_h\|_{L^2(\Omega)^3} \leq C_s' \,\frac{\varepsilon^\star}{\varepsilon_\star}\,\|\mathbf{rot}\,\mathbf{v}_h\|_{L^2(\Omega)^3},$$

ce qui est le résultat attendu. ∎

La théorie s'applique donc (cf. le théorème 2.12). Pour estimer l'erreur d'approximation $\inf_{\mathbf{v}_h \in V_h} \|\mathbf{E} - \mathbf{v}_h\|_{H(\mathbf{rot}\,;\Omega)}$, on utilise le résultat d'approximabilité du théorème 2.55. Comme en outre la pression électrique artificielle λ_E et son approximation λ_E^h s'annulent toutes les deux, la théorie se simplifie puisque, bien sûr, $\|\lambda_E - \lambda_E^h\|_{H^1(\Omega)} = 0$, et $\inf_{\mu_h \in M_h^0} \|\lambda_E - \mu_h\|_{H^1(\Omega)} = 0$. En conclusion, on a le résultat d'approximation suivant (pour la dépendance explicite par rapport aux données $\mathbf{J}$ et ρ, voir la remarque 2.48).

Théorème 2.57 *Supposons que ε soit constante sur Ω et $\mathbf{f}_E \in H(\mathbf{rot}\,;\Omega)$, $g_E \in L^2(\Omega)$. Alors*

$$\|\mathbf{E} - \mathbf{E}_h\|_{H(\mathbf{rot}\,;\Omega)} \leq C_s\, h^s \left(\|\mathbf{f}_E\|_{H(\mathbf{rot}\,;\Omega)} + \|g_E\|_{L^2(\Omega)}\right),$$

pour tout $s < \sigma$, avec $C_s > 0$ indépendante de $\mathbf{E}$.

Pour le problème magnétique, la même démarche s'applique, pour aboutir au second résultat ci-dessous.

Théorème 2.58 *Supposons que μ soit constante sur Ω et $\mathbf{f}_H \in H_0(\mathbf{rot}\,;\Omega)$. Alors*

$$\|\mathbf{H} - \mathbf{H}_h\|_{H(\mathbf{rot}\,;\Omega)} \leq C_s'\, h^s\, \|\mathbf{f}_H\|_{H(\mathbf{rot}\,;\Omega)},$$

pour tout $s < \sigma$, avec $C_s' > 0$ indépendante de $\mathbf{H}$.

2.4 Illustrations numériques

Afin d'appréhender les problèmes de mise en œuvre de la résolution des problèmes mixtes, nous allons considérer dans un premier temps le cas emblématique de la résolution du problème de Stokes puis celui de l'approximation du problème de Maxwell quasi–statique, mettant en œuvre des éléments finis d'arêtes s'appuyant sur des degrés de liberté moins usuels.

2.4.1 Résolution du problème de Stokes

Dans la suite nous nous intéressons au problème de Stokes posé dans un ouvert borné Ω de $\mathbb{R}^2$:

$$\begin{cases} -\nu \Delta \mathbf{u} + \nabla p = \mathbf{f} \text{ dans } \Omega, \\ \operatorname{div} \mathbf{u} = 0 \quad \text{dans } \Omega, \\ \mathbf{u} = \mathbf{g} \quad \text{sur } \partial\Omega. \end{cases}$$

On suppose que $\mathbf{f} \in L^2(\Omega)^2$ et $\mathbf{g} \in H^{1/2}(\partial\Omega)^2$ et on introduit la formulation variationnelle :

$$\begin{cases} \text{trouver } (\mathbf{u}, p) \in H^1(\Omega)^2 \times L^2_0(\Omega) \text{ tel que} \\ \nu \displaystyle\int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, d\Omega - \int_\Omega p \operatorname{div} \mathbf{v} \, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\Omega, \quad \forall \mathbf{v} \in H^1_0(\Omega)^2, \\ \displaystyle\int_\Omega q \operatorname{div} \mathbf{u} \, d\Omega = 0, \quad \forall q \in L^2_0(\Omega) \\ \mathbf{u} = \mathbf{g} \text{ sur } \partial\Omega. \end{cases}$$

En vertu des résultats de la section 2.1.2, ce problème admet une unique solution. Soient V_h et M_h deux espaces d'approximation interne de $H^1(\Omega)^2$ et $L^2_0(\Omega)$ de dimensions finies respectives n_h et m_h. On note $V_{h0} = \{\mathbf{v}_h \in V_h, \mathbf{v}_h = \mathbf{0} \text{ sur } \partial\Omega\}$ de dimension n_{h0}. Par la suite, $(\mathbf{w}_I)_{I=1,n_h}$ désigne une base de l'espace V_h et $(\tau_k)_{k=1,m_h}$ une base de M_h. On désigne par $\mathcal{I}$ l'ensemble des indices $I \in 1, ..., n_h$ correspondant aux fonctions $\mathbf{w}_I$ qui s'annulent sur $\partial\Omega$ et $\mathcal{D}$ le complémentaire de $\mathcal{I}$ dans $\{1, ..., n_h\}$. Notons que $V_{h0} = \text{vect}_{I \in \mathcal{I}} \mathbf{w}_I$. On définit également l'ensemble $\mathcal{K} = \{1, ..., m_h\}$. On considère le problème discrétisé :

$$\begin{cases} \text{trouver } (\mathbf{u}_h, p_h) \in V_h \times M_h \text{ tels que} \\ \nu \displaystyle\int_\Omega \nabla \mathbf{u}_h : \nabla \mathbf{v}_h \, d\Omega - \int_\Omega p_h \operatorname{div} \mathbf{v}_h \, d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \, d\Omega, \quad \forall \mathbf{v}_h \in V_{h0}, \\ \displaystyle\int_\Omega q_h \operatorname{div} \mathbf{u}_h \, d\Omega = 0, \quad \forall q_h \in M_h, \\ \mathbf{u}_h = \mathbf{g}_h \text{ sur } \partial\Omega, \end{cases}$$

où $\mathbf{g}_h$ est une projection de $\mathbf{g}$ sur l'espace V_h (par exemple l'interpolé $\Pi_h \mathbf{g}$ dès lors que $\mathbf{g}$ est continu sur $\partial\Omega$).
Le problème discrétisé précédent s'écrit sous la forme matricielle :

$$\begin{bmatrix} \nu\mathbb{K} & -\mathbb{D}^t \\ -\mathbb{D} & 0 \end{bmatrix} \begin{pmatrix} \overrightarrow{U} \\ \overrightarrow{P} \end{pmatrix} = \begin{pmatrix} \overrightarrow{F} \\ 0 \end{pmatrix} \text{ et } U_I = (g_h)_I \, \forall I \in \mathcal{D} \tag{2.122}$$

avec $\overrightarrow{U}$ le vecteur des composantes de $\mathbf{u}_h$ dans la base $(\mathbf{w}_I)_{I=1,n_h}$, $\overrightarrow{P}$ le vecteur des composantes de p_h dans la base $(\tau_k)_{k=1,m_h}$ et $\forall I \in \mathcal{I},\ J = 1, ..., n_h,\ k = 1, ..., m_h$:

$$\mathbb{K}_{IJ} = \int_\Omega \nabla \mathbf{w}_I \colon \nabla \mathbf{w}_J \, d\Omega, \; \mathbb{D}_{kJ} = \int_\Omega \tau_k \, \operatorname{div} \mathbf{w}_J \, d\Omega, \; F_I = \int_\Omega \mathbf{f} \cdot \mathbf{w}_I d\Omega.$$

Attention, la matrice $\mathbb{K}$ n'est pas une matrice carrée (n_{h0} lignes et n_h colonnes). Nous allons modifier le système afin de prendre en compte la condition de Dirichlet sur la vitesse. On introduit le vecteur $\overrightarrow{G}$ de composantes :

$$G_I = \begin{cases} 0 & \text{si } I \in \mathcal{I} \\ (g_h)_I & \text{si } I \in \mathcal{D} \end{cases}$$

Le système (2.122) est équivalent au système :

$$\begin{bmatrix} \nu\mathbb{K}_{\mathcal{I}\mathcal{I}} & 0 & -\mathbb{D}^t_{\mathcal{K}\mathcal{I}} \\ 0 & \mathbb{I}_{\mathcal{D}\mathcal{D}} & 0 \\ -\mathbb{D}_{\mathcal{K}\mathcal{I}} & 0 & 0 \end{bmatrix} \begin{pmatrix} \overrightarrow{U}_{\mathcal{I}} \\ \overrightarrow{U}_{\mathcal{D}} \\ \overrightarrow{P} \end{pmatrix} = \begin{pmatrix} (\overrightarrow{F} - \mathbb{K}\overrightarrow{G})_{\mathcal{I}} \\ \overrightarrow{G}_{\mathcal{D}} \\ \mathbb{D}\overrightarrow{G} \end{pmatrix}, \tag{2.123}$$

où $\mathbb{K}_{\mathcal{I}\mathcal{I}}$ et $\mathbb{D}_{\mathcal{K}\mathcal{I}}$ sont les matrices extraites de $\mathbb{K}$ et $\mathbb{D}$ à partir des ensembles d'indice $\mathcal{I}$ et $\mathcal{K}$ et $\mathbb{I}_{\mathcal{D}\mathcal{D}}$ la matrice identité restreinte aux indices $\mathcal{D}$.
En pratique, on construit la matrice $\mathbb{K}$ sans se préoccuper des indices correspondants aux degrés de liberté de Dirichlet. C'est donc une matrice ayant n_h lignes. Pour se ramener au système précédent, il suffit d'annuler tous les coefficients des lignes et des colonnes d'indices appartenant à l'ensemble $\mathcal{D}$ hormis les termes diagonaux (pseudo-élimination). Il n'est pas simple de fabriquer un espace d'approximation de $L^2_0(\Omega)$ dans lequel les fonctions sont à moyenne nulle. La pression étant définie à une constante près dans les équations de Stokes, il suffit de fixer cette constante en annulant un des degrés de liberté de la pression.

Mise en œuvre

L'essentiel du travail de mise en œuvre consiste d'une part, à calculer les différentes matrices du système linéaire sans prendre en compte la condition de Dirichlet et, d'autre part, à procéder à la pseudo-élimination des conditions de Dirichlet et au calcul du second membre. Nous nous sommes intéressés principalement aux cas des approximations $P^1_{\text{bul}} - P^1$, $P^2 - P^0$, $P^2 - P^1$ et $P^1 - P^0$; cette dernière approximation n'étant pas adaptée (verrouillage numérique) ! Pour stocker les matrices, nous avons choisi par commodité (ce n'est pas le choix optimal), de ranger les différents coefficients des matrices suivant l'ordre suivant :

$$(\overrightarrow{U}^l_x, \overrightarrow{U}^l_y, \overrightarrow{U}^b_x, \overrightarrow{U}^b_y, \overrightarrow{P})$$

où $\overrightarrow{U}^l_x$ et $\overrightarrow{U}^l_y$ représentent respectivement les vecteurs de déplacement suivant x et y associés aux degrés de liberté de Lagrange, $\overrightarrow{U}^b_x$ et $\overrightarrow{U}^b_y$ représentent respectivement les vecteurs de déplacement suivant x et y associés aux degrés de liberté

bulle lorsqu'il y en a, et $\overrightarrow{P}$ représente le vecteur associé à l'approximation de la pression. Nous donnons ci-après le code Matlab permettant de calculer les matrices du système de Stokes pour différents choix d'approximation. La fonction Matlab **EFStokes.m** calcule à partir des données de maillages (Su,Tu,Sp,Tp), du choix des approximations (apu,app) et de la fonction Matlab **fS.m** définissant la fonction **f**, les matrices $\mathbb{K}$, $\mathbb{D}$ ainsi que les matrices de masse $\mathbb{M}$ associées à chaque approximation et le vecteur B correspondant au vecteur $\overrightarrow{F}$ du système linéaire. Cette fonction repose sur la fonction Matlab **EFelmStokes.m** qui réalise les calculs au niveau de chaque élément.

```
function [K,D,Mu,Mp,B]=EFStokes(apu,Su,Tu,app,Sp,Tp,fS)
nt=size(Tu,1);nsu=size(Su,1);nu=2*nsu;np=nt;          %choix approximation
if(strcmp(apu,'P1_bulle')==1) nu=2*nsu+2*nt;end
if(strcmp(app,'P1')==1) np=size(Sp,1); end
K=sparse(nu,nu);Mu=sparse(nu,nu);D=sparse(np,nu);Mp=sparse(np,np);
B=zeros(nu,1);
for t=1:nt,                                          %boucle sur les éléments
   [Kt,Dt,Mut,Mpt,Bt]=EFelmStokes(Su(Tu(t,:),:),apu,app,fS);
   I=[Tu(t,:) nsu+Tu(t,:)];J=[t];                                %assemblage
   if(strcmp(apu,'P1_bulle')==1)
       I=[Tu(t,:) 2*nsu+t nsu+Tu(t,:) 2*nsu+nt+t];end,
   if(strcmp(app,'P1')==1) J=[Tp(t,1:3)];end
   K(I,I)=K(I,I)+Kt;D(J,I)=D(J,I)+Dt;
   Mu(I,I)=Mu(I,I)+Mut;Mp(J,J)=Mp(J,J)+Mpt;
   B(I)=B(I)+Bt;
 end

function [K,D,Mu,Mp,Bu]=EFelmStokes(S,apu,app,fS)
os=sqrt(15);s3=1./3.;nbq=7;                          %Formule de quadrature
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;pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];
n=0;m=0;
S21=S(2,:)-S(1,:);S31=S(3,:)-S(1,:);
delta=S21(1)*S31(2)-S21(2)*S31(1);
Jflmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;                %transfo affine
switch apu
          case 'P1' n=3;
          case 'P2' n=6;
          case 'P1_bulle' n=4;
end,
switch app
          case 'P0' m=1;
          case 'P1' m=3;
end,
K=zeros(n,n);Mu=zeros(n,n);D=zeros(m,2*n);Mp=zeros(m,m);
Bu=zeros(2*n,1);
for k=1:nbq,                                     %boucle points de quadrature
      x=pts_quadT(k,1);y=pts_quadT(k,2);
      switch apu                                 %calcul des fonctions de base
          case 'P1' w=[1-x-y x y];gw=[-1 1 0;-1 0 1];
          case 'P2' w=[(1-x-y)*(1-2*x-2*y) x*(2*x-1) y*(2*y-1) ...
                       4*x*(1-x-y) 4*x*y 4*y*(1-x-y)];
                    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)];
        case 'P1_bulle'
             tb=60*(1-x-y)*x*y;
            dxtb=60*(1-2*x-y)*y;dytb=60*(1-2*y-x)*x;
            w=[1-x-y-tb/3  x-tb/3  y-tb/3  tb];
            gw=[-1-dxtb/3  1-dxtb/3  -dxtb/3  dxtb;
                -1-dytb/3  -dytb/3  1-dytb/3  dytb];
    end,
    switch app
        case 'P0' t=[1];
        case 'P1' t=[1-x-y  x  y];
    end,
    pk=pds_quadT(k)*abs(delta);          %calculs des matrices élémentaires
    jg=Jflmt*gw;
    K=K+pk*jg'*jg;Mu=Mu+pk*w'*w;
    D=D+pk*t'*[jg(1,:)  jg(2,:)];Mp=Mp+pk*t'*t;
    F=fS(S(1:3,:)'*[1-x-y;  x;  y]);                    %calcul second membre
    Bu=Bu+[F(1)*pk*w';F(2)*pk*w'];
end,
K=[K  0*K;0*K  K];                                      %assemblage vectoriel
Mu=[Mu  0*Mu;0*Mu  Mu];
```

A partir de cette fonction et des outils de maillage déjà utilisés dans d'autres exemples, il est facile d'écrire un script réalisant la résolution du problème de Stokes. Nous donnons les scripts Matlab correspondant à l'approximation $P^1_{\text{bul}}-P^1$ et P^2-P^0 dans le cas où la donnée de Dirichlet sur le bord est $\mathbf{g}=0$. Il est facile de déduire les scripts correspondant à l'approximation P^1-P^0 et P^2-P^1.

Script de l'approximation $P^1_{\text{bul}}-P^1$

```
v=1;n=30;
[S,T,BR,RT]=triangle_rectangle([0  1  0  1],n,n);    %maillage P1 du carré
[K,D,Mu,Mp,Bu]=EFStokes('P1_bulle',S,T,'P1',S,T,@fStokes);  %calculs EF
A=[v*K  -D';-D  0*Mp];                                %assemblage du système
ns=size(S,1);;nt=size(T,1);nu=size(K,1);np=size(Mp,1);
B=[Bu;zeros(np,1)];
ND=noeud_bords(S,T,BR,[1  2  3  4]);                  %condition limite u=0
ID=find(ND==1);DA=diag(A);                           %par pseudo-élimination
A(ID,:)=0;A(ns+ID,:)=0;A(:,ID)=0;
A(:,ns+ID)=0;B(ID)=0;B(ns+ID)=0;
for i=1:size(ID,1),
    A(ID(i),ID(i))=DA(ID(i));
    A(ns+ID(i),ns+ID(i))=DA(ns+ID(i));
end
i=nu+1;A(i,:)=0;A(:,i)=0;A(i,i)=1;B(i)=0;            %élimination pression
X=A\B;                                                %résolution
U=[X(1:ns),X(ns+1:2*ns)];P=X(nu+1:end);
```

Script de l'approximation P^2-P^0

```
v=1;n=30;
[S,T,BR,RT]=triangle_rectangle([0  1  0  1],n,n);    %maillage P1 du carré
[S2,T2,BR2,RT2]=maillageP2(S,T,BR,RT);                %maillage P2
[T2,S2]=renume(T2,S2);                               %renumérotation
[K,D,Mu,Mp,Bu]=EFStokes('P2',S2,T2,'P0',S,T,@fStokes);  %calculs EF
A=[v*K  -D';-D  0*Mp];                                %assemblage
```

```
ns=size(S2,1);; nt=size(T2,1); nu=size(K,1); np=size(Mp,1);
B=[Bu; zeros(np,1)];
ND=noeud_bords(S2,T2,BR2,[1 2 3 4]);                    %conditions limite u=0
ID=find(ND==1);DA=diag(A);                              %par pseudo-élimination
A(ID,:)=0;A(ns+ID,:)=0;A(:,ID)=0;
A(:,ns+ID)=0;B(ID)=0;B(ns+ID)=0;
for i=1:size(ID,1),
    A(ID(i),ID(i))=DA(ID(i));
    A(ns+ID(i),ns+ID(i))=DA(ns+ID(i));
end
i=nu+1;A(i,:)=0;A(:,i)=0;A(i,i)=1;B(i)=0;               %élimination pression
X=A\B;                                                  %résolution
U=[X(1:ns),X(ns+1:2*ns)];P=X(nu+1:end);
```

Notons que ces deux scripts sont très voisins. Cela résulte du choix de rangement des inconnues du système qui "masque" le type des approximations.

Etude numérique

Nous avons testé la résolution du problème de Stokes sur le carré unité $\Omega =]0,1[\times]0,1[$, en choisissant la donnée :

$$\mathbf{f}(x,y) = \begin{pmatrix} \sin ay \left\{ (ab - 2\nu a^2)\cos ax + \nu a^2 \right\} \\ \sin ax \left\{ (ab + 2\nu a^2)\cos ax + \nu a^2 \right\} \end{pmatrix}$$

conduisant à la solution :

$$\mathbf{u}(x,y) = \begin{pmatrix} (1-\cos(ax))\sin(ay) \\ (\cos(ay)-1)\sin(ax) \end{pmatrix} \text{ et } p(x,y) = b\sin(ax)\sin(ay).$$

C'est une solution des équations de Stokes et en choisissant $a = 2k\pi$, le vecteur vitesse $\mathbf{u}$ s'annule sur le bord $\partial\Omega$. Dans les expériences numériques, nous avons choisi $a = 2\pi$ et $b = 1$ (cf figure 2.4).

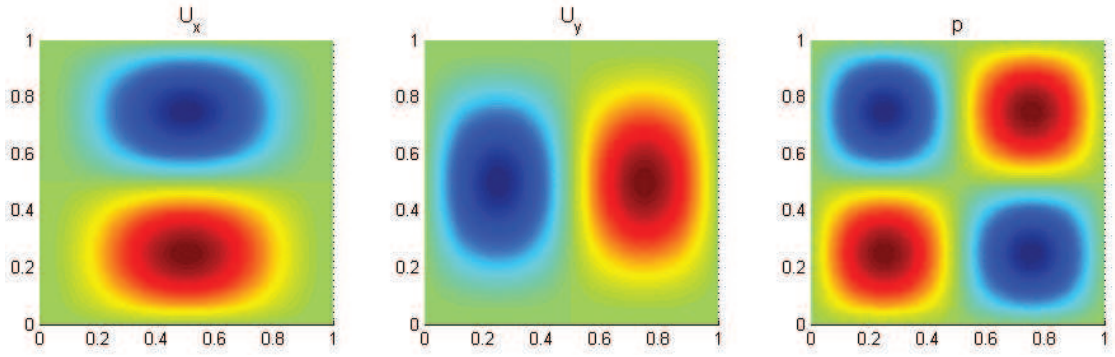


Figure 2.4. Champ de vitesse et pression exacts

Sur les figures 2.5, 2.6, 2.7 et 2.8 nous représentons les champs de vitesses et pression ainsi que les champs d'erreurs pour les approximations $P^1_{\text{bul}} - P^1$, $P^2 - P^0$,

$P^2 - P^1$ et $P^1 - P^0$ sur un même maillage (le carré découpé 30×30).
Dans tous les cas, on note que le champ de vitesses est toujours mieux approché

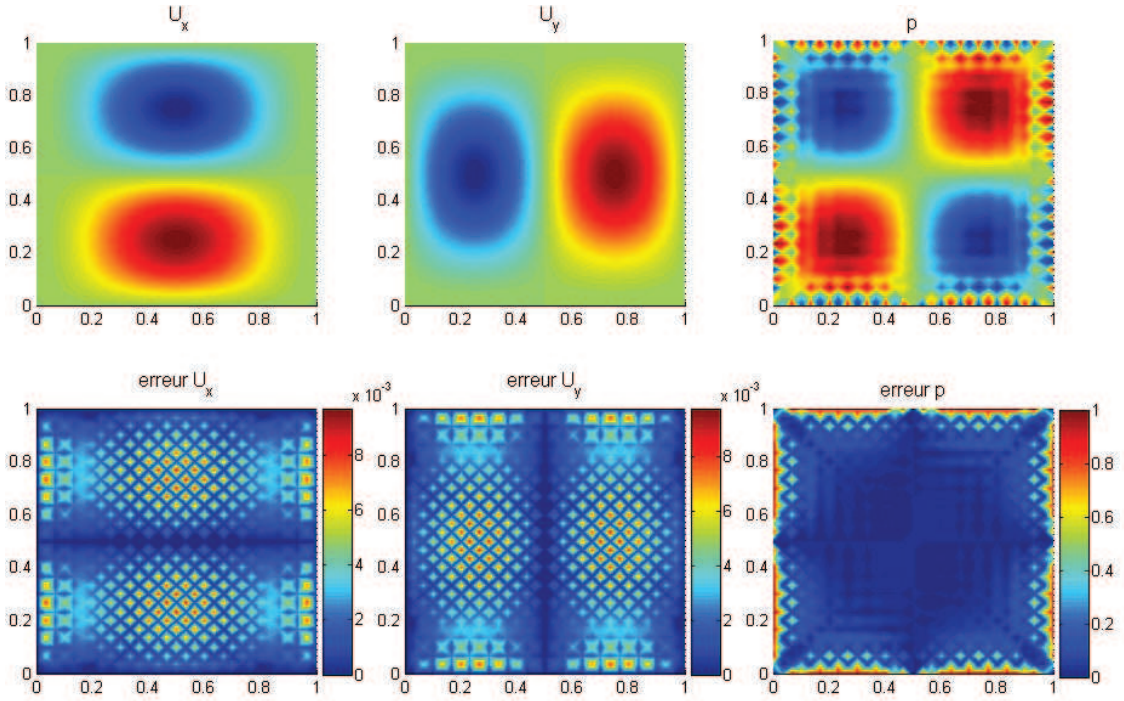


Figure 2.5. Solution et erreur obtenues avec l'approximation $P^1_{\text{bul}} - P^1$

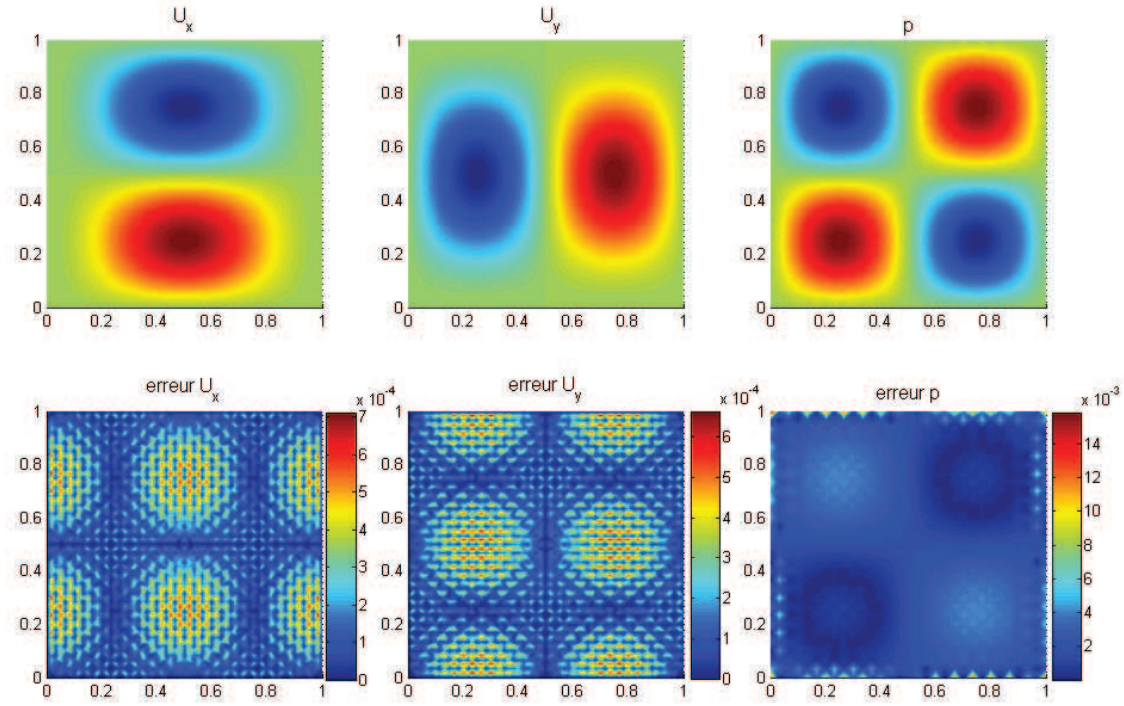


Figure 2.6. Solution et erreur obtenues avec l'approximation $P^2 - P^0$

que le champ de pression. La meilleure approximation est bien évidemment obtenue avec les approximations $P^2 - P^0$ et $P^2 - P^1$ car ce sont des approximations d'ordre 2. Les approximations $P^1 - P^0$ et $P^1_{\text{bul}} - P^1$ conduisent à une précision sur le champ de vitesse compatible avec l'ordre 1. L'approximation du champ de pression est, quant à elle, de beaucoup moins bonne qualité, voire fausse dans le cas $P^1 - P^0$! Dans ce dernier cas, on sait d'après la théorie que cette approximation conduit à un verrouillage numérique. Remarquons que le champ de pression est relativement bien approché par les approximations $P^2 - P^0$ et $P^2 - P^1$, erreur maximum de l'ordre de 10^{-2} avec des effets de concentration de l'erreur au voisinage des bords. Dans le cas de l'approximation $P^1_{\text{bul}} - P^1$, le champ de pression obtenu est fortement perturbé par des effets de bord. Au vu de cette erreur, il n'est pas clair que l'approximation $P^1_{\text{bul}} - P^1$ soit correcte. Afin de se persuader de la pertinence de ces résultats nous avons mené une étude de convergence vis-à-vis du paramètre de discrétisation.

Sur la figure 2.9 nous donnons les courbes d'erreurs en norme L^2 en échelle loga-

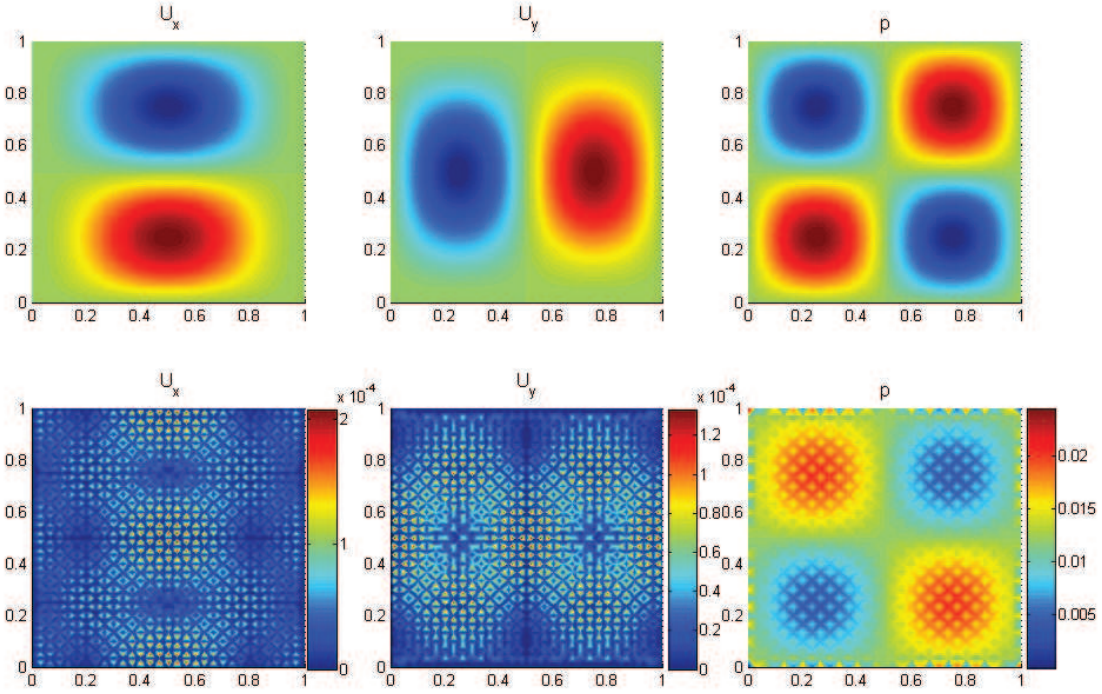


Figure 2.7. Solution et erreur obtenues avec l'approximation $P^2 - P^1$

rithmique obtenues pour la pression (trait pointillé) et la vitesse (trait plein) et ce pour les approximations $P^1 - P^0$ et $P^1_{\text{bul}} - P^1$. On observe que l'approximation de la pression est divergente pour l'approximation $P^1 - P^0$. Ce qui n'est pas surprenant car cette approximation n'est pas convergente théoriquement. Rappelons que dans ce cas, il y a trop de contraintes par rapport au nombre de degré de liberté et le système linéaire n'est pas inversible. Matlab arrive toutefois à l'inverser à cause

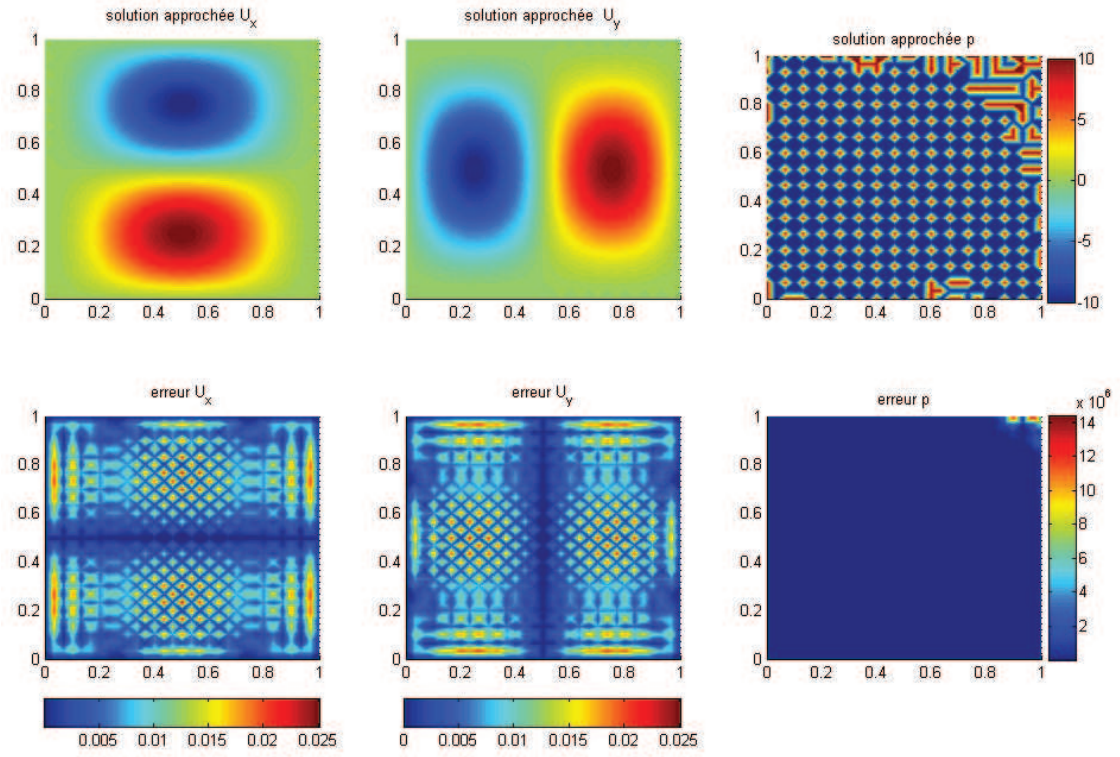


Figure 2.8. Solution et erreur obtenues avec l'approximation $P^1 - P^0$

des erreurs d'arrondis mais indique des conditionnements très élevés (typiquement 10^{18}). Observons néanmoins que l'approximation de la vitesse reste convergente, en h^2, ordre compatible avec une approximation par éléments finis d'ordre 1.

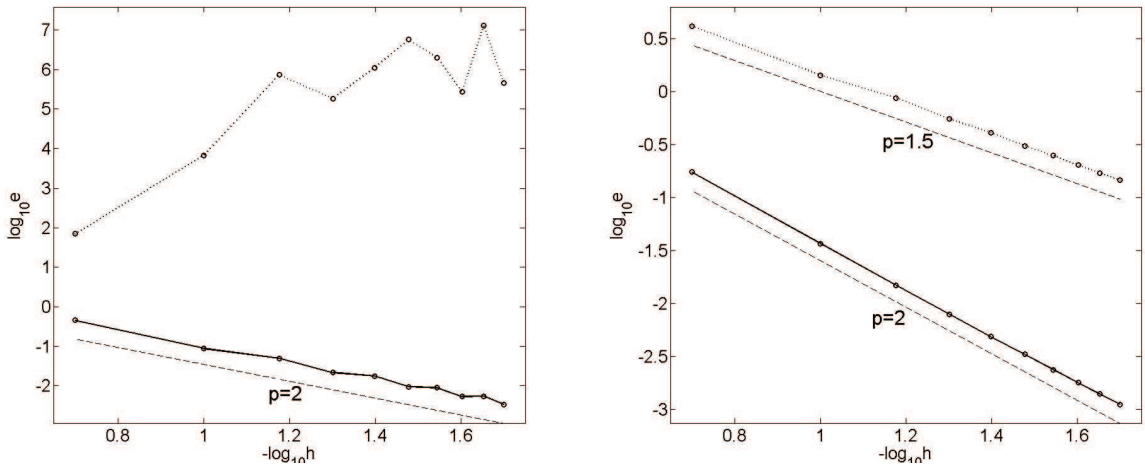


Figure 2.9. Erreurs en norme L^2 obtenues avec les approximations $P^1 - P^0$ (à gauche) et $P^1_{\text{bul}} - P^1$ (à droite), l'erreur sur la pression est en trait pointillé et l'erreur sur la vitesse en trait plein

Sur la figure 2.10 nous représentons les courbes d'erreurs en norme L^2 pour des approximations P^2 de la vitesse et respectivement P^0 et P^1 pour la pression. On

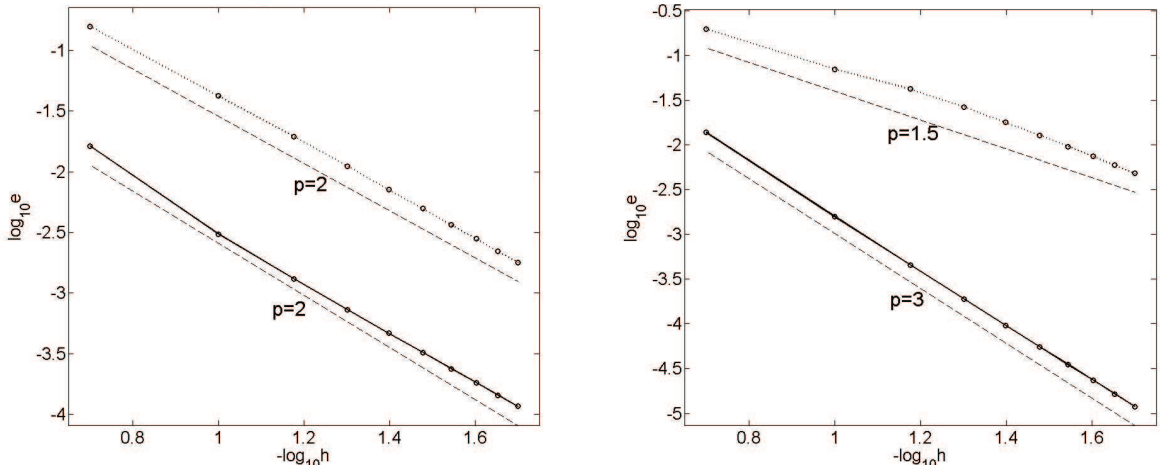


Figure 2.10. Erreurs en norme L^2 obtenues avec les approximations $P^2 - P^0$ (à gauche) et $P^2 - P^1$ (à droite), l'erreur sur la pression est en pointillé et celle sur la vitesse en trait plein

obtient des ordres de convergence d'au moins 2 pour l'approximation de la vitesse et d'au moins 1.5 pour l'approximation de la pression. Il est intéressant d'observer qu'une discrétisation d'ordre 0 pour la pression fournit une meilleure approximation de la pression (convergence à l'ordre 2) qu'une approximation P^1 qui conduit à une approximation seulement d'ordre 1.5 de la pression. Par contre, l'approximation P^0 de la pression ne fournit qu'une précision d'ordre 2 de la vitesse alors qu'une approximation P^1 de la pression conduit à une précision d'ordre 3 de la vitesse. En conclusion, si on souhaite utiliser une approximation du même ordre pour la vitesse et la pression il est plus intéressant d'utiliser une discrétisation $P^2 - P^0$ et si on désire obtenir une très bonne précision sur la vitesse il est préférable d'utiliser la discrétisation $P^2 - P^1$ sachant par ailleurs que cette dernière requiert un petit peu moins de degrés de liberté.

La cavité entraînée

Afin de terminer cette expérimentation numérique, nous avons considéré l'exemple de la cavité entraînée. Il s'agit d'un problème de Stokes homogène où on impose

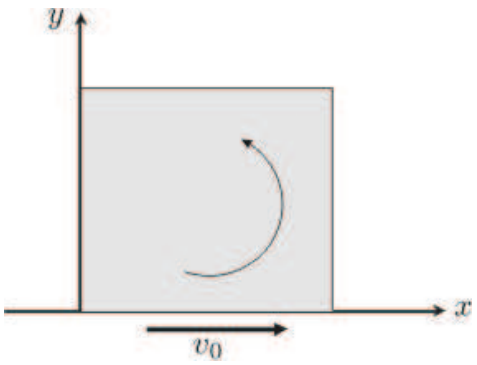


Figure 2.11. Cavité entrainée

une vitesse d'entraînement sur un des côtés de la cavité $\Omega =]0,1[\times]0,1[$:

$$\begin{cases} -\nu\Delta\,\mathbf{u} + \nabla p = \mathbf{f} \text{ dans } \Omega, \\ \operatorname{div}\mathbf{u} = 0 \qquad\qquad \text{dans } \Omega, \\ \mathbf{u} = (v_0, 0) \qquad \text{sur } \Sigma = \{0\}\times]0,1[, \\ \mathbf{u} = \mathbf{0} \qquad\qquad \text{sur } \partial\Omega \setminus \Sigma \end{cases}$$

Le script Matlab suivant réalise la résolution de ce problème à l'aide d'une ap-

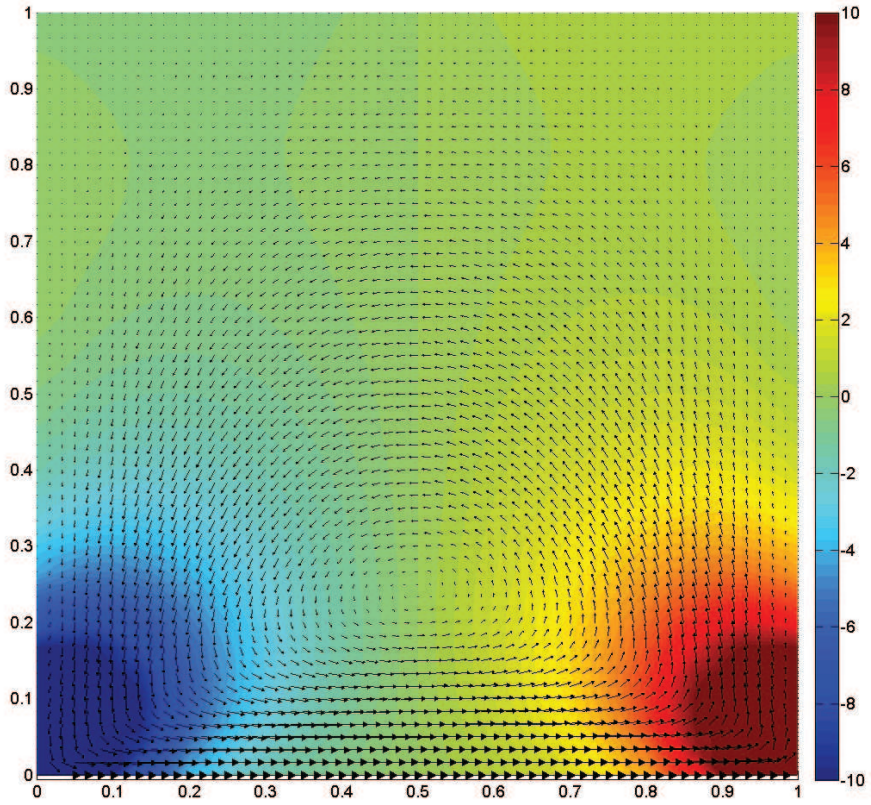


Figure 2.12. Solution $P^2 - P^0$ du problème de la cavité entraînée

proximation $P^2 - P^0$. Il s'agit d'une adaptation des scripts précédents permettant de prendre en compte une condition de Dirichlet non homogène.

```
v=1;v0=1;n=30;
[S1,T1,BR1,RT1]=triangle_rectangle([0 1 0 1],n,n); %maillage P1 du carré
[S2,T2,BR2,RT2]=maillageP2(S1,T1,BR1,RT1);                %maillage P2
[T2,S2]=renume(T2,S2);
[K,D,Mu,Mp,Bu]=EFStokes('P2',S2,T2,'P0',S1,T1,@fStokes);      %calculs EF
A=[v*K -D';-D 0*Mp];                          %assemblage du système de Stokes
ns=size(S2,1);nt=size(T2,1);nu=size(K,1);np=size(Mp,1);l=size(A,1);
ND1=noeud_bords(S2,T2,BR2,[1]);ID1=find(ND1==1);                %relèvement
c=v0*(ND1==1);G=[c;zeros(ns,1)];B=g*[-v*K*G;D*G];
ND=noeud_bords(S2,T2,BR2,[1 2 3 4]);ID=find(ND==1);  %élimination vitesse
DA=diag(A);A(ID,:)=0;A(ns+ID,:)=0;A(:,ID)=0;A(:,ns+ID)=0;u=0
B(ID)=0;B(ns+ID)=0;B(ID1)=DA(ID1).*c(ID1);
for i=1:size(ID,1),
    A(ID(i),ID(i))=DA(ID(i));A(ns+ID(i),ns+ID(i))=DA(ns+ID(i));
```

```
end
i=nu+1;A(i,:)=0;A(:,i)=0;A(i,i)=1;B(i)=0;          %élimination pression
X=A\B;                                              %résolution
U=[X(1:ns),X(ns+1:2*ns)];P=X(nu+1:end);P=P-mean(P);
```

Nous représentons sur la figure 2.12 le champ de vitesse (flèches) et de pression (isovaleurs) obtenus à l'aide de ce script. Notons que la solution présente une singularité aux points $(0,0)$ et $(1,0)$ car la donnée de Dirichlet y est discontinue !

2.4.2 Résolution des équations de Maxwell

On s'intéresse maintenant à la résolution numérique des équations de Maxwell. Cette illustration numérique va permettre d'appréhender de façon concrète une classe d'éléments finis moins standard : les éléments finis d'arêtes auxquels sont attachés des degrés de liberté qui ne sont plus nodaux mais de type moment. Nous allons nous intéresser à la formulation en champ électrique posée dans un domaine borné Ω de $\mathbb{R}^2$ lorsque $g_E = 0$:

$$\begin{cases} \operatorname{rot} \mathbf{E}(t) = f_E(t) & \text{dans } \Omega, \\ \operatorname{div} \varepsilon \mathbf{E}(t) = 0 & \text{dans } \Omega, \\ \mathbf{E}(t) \times \mathbf{n} = 0 & \text{sur } \partial\Omega, \end{cases}$$

dont une formulation mixte est :

$$\begin{cases} \text{trouver } (\mathbf{E}, \lambda_E) \in H_0(\operatorname{rot};\Omega) \times H_0^1(\Omega) \text{ tel que} \\ \displaystyle\int_\Omega \operatorname{rot}\mathbf{E}\, \operatorname{rot}\mathbf{v}\, d\Omega + \int_\Omega \varepsilon \mathbf{v} \cdot \nabla\lambda_E\, d\Omega = \int_\Omega f_E\, \operatorname{rot}\mathbf{v}\, d\Omega \quad \forall \mathbf{v} \in H_0(\operatorname{rot};\Omega), \\ \displaystyle\int_\Omega \varepsilon \mathbf{E} \cdot \nabla q\, d\Omega = 0 \quad \forall q \in H_0^1(\Omega). \end{cases}$$

Il s'agit maintenant de construire des espaces d'approximation conformes de $H_0(\operatorname{rot};\Omega)$ et de $H_0^1(\Omega)$. Pour ce faire, on considère un maillage conforme $\mathcal{T}_h$ de Ω constitué des triangles $(T^\ell)_{\ell=1,L}$ de sommets $(S_I)_{I=1,N}$. On supposera par la suite pour simplifier que Ω est un polygone et que $\bigcup_{\ell=1,L} T^\ell = \overline{\Omega}$. On introduit les espaces :

$$\begin{aligned} V_h^0 &= \{\mathbf{v}_h \in H_0(\operatorname{rot};\Omega) \text{ tel que } \mathbf{v}_{h|T^\ell} \in \mathcal{R}^1(T^\ell)\ \forall \ell = 1, L\} \\ &\text{où } \mathcal{R}^1 = \{\mathbf{v} = \mathbf{a} + \gamma \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix},\ \mathbf{a} \in \mathbb{R}^2 \text{ et } \gamma \in \mathbb{R}\} \\ M_h^0 &= \{v_h \in H_0^1(\Omega) \text{ tel que } v_{h|T^\ell} \in P^1(T^\ell)\ \forall \ell = 1, L\}. \end{aligned}$$

Les espaces V_h^0 et M_h^0 constituent respectivement des approximations conformes de $H_0(\mathbf{rot}\,;\Omega)$ et $H_0^1(\Omega)$. On considère le problème variationnel discret :

$$\begin{cases} \text{trouver } (\mathbf{E}_h, \lambda_h) \in V_h^0 \times M_h^0 \text{ tel que} \\ \displaystyle\int_\Omega \operatorname{rot} \mathbf{E}_h \operatorname{rot} \mathbf{v}_h \, d\Omega + \int_\Omega \mathbf{v} \cdot \nabla \lambda_h \, d\Omega = \int_\Omega f_E \operatorname{rot} \mathbf{v}_h \, d\Omega \ \forall \mathbf{v}_h \in V_h^0, \quad (2.124) \\ \displaystyle\int_\Omega \mathbf{E}_h \cdot \nabla q = 0 \ \forall q \in M_h^0 \quad (2.125) \end{cases}$$

L'espace M_h^0 est l'espace classique associé aux éléments finis de Lagrange P^1 à trace nulle. Dans la suite, nous notons $(w_i)_{i=1,N}$ les fonctions de bases globales P^1 associées au maillage. Nous devons maintenant construire de façon explicite une classe d'éléments finis associée à l'espace V_h^0. Nous allons utiliser la classe des éléments finis 2D d'arête de Raviart-Thomas-Nédélec d'ordre 1 (version 2D de la classe d'éléments finis présentée auparavant en 3D). Signalons que l'on trouve une présentation assez exhaustive des éléments finis d'arêtes dans la référence [52].

Elément fini 2D de Raviart-Thomas-Nédélec (RTN)

Sur un triangle T^ℓ de sommets $(S_1^\ell, S_2^\ell, S_3^\ell)$ on considère les 3 degrés de liberté de type moment $i = 1, 2, 3$ (avec la convention $S_4^\ell = S_1^\ell$) :

$$M_i^\ell(\mathbf{v}) = \int_{A_i^\ell} \mathbf{v} \cdot \mathbf{t}_i^\ell \, dA \quad \text{avec } A_i^\ell =]S_i^\ell, S_{i+1}^\ell[\text{ et } \mathbf{t}_i^\ell = \frac{\overrightarrow{S_i^\ell S_{i+1}^\ell}}{\text{mes} A_i^\ell}.$$

En notant $(\lambda_i^\ell)_{i=1,2,3}$ les coordonnées barycentriques par rapport aux sommets du triangle T^ℓ, on introduit les fonctions à valeurs vectorielles

$$\boldsymbol{\tau}_i^\ell = \lambda_i^\ell \nabla \lambda_{i+1}^\ell - \lambda_{i+1}^\ell \nabla \lambda_i^\ell \ \text{ pour } i = 1, 2, 3.$$

On peut vérifier que $\boldsymbol{\tau}_i^\ell \in \mathcal{R}^1(T^\ell)$ et comme $\lambda_i^\ell + \lambda_{i+1}^\ell = 1$ sur A_i^ℓ, il est facile de montrer que

$$\boldsymbol{\tau}_{i|A_i^\ell}^\ell \cdot \mathbf{t}_i^\ell = \frac{1}{\text{mes} A_i^\ell} \text{ et } \boldsymbol{\tau}_{i|A_j^\ell}^\ell \cdot \mathbf{t}_j^\ell = 0 \ \ \forall j \neq i.$$

Ce calcul prouve que

$$M_i^\ell(\boldsymbol{\tau}_j^\ell) = \delta_{ij} \ \forall i, j = 1, 2, 3$$

et que pour tout $\mathbf{v} \in \mathcal{R}^1(T^\ell)$ on a la décomposition :

$$\mathbf{v} = \sum_{i=1,2,3} M_i^\ell(\mathbf{v}) \, \boldsymbol{\tau}_i^\ell.$$

Il est assez facile d'expliciter les fonctions de base $\boldsymbol{\tau}_i^\ell$ en coordonnées cartésiennes. En effet, comme $\boldsymbol{\tau}_i^\ell \cdot \mathbf{t}_j^\ell = 0 \ \ \forall j \neq i$, on a $\boldsymbol{\tau}_i^\ell(S_{i+2}^\ell) = 0$ et donc :

$$\boldsymbol{\tau}_i^\ell(x, y) = \Delta_i \begin{pmatrix} -(y - y_{i+2}^\ell) \\ (x - x_{i+2}^\ell) \end{pmatrix}.$$

On détermine la constante Δ_i en utilisant le fait que $\text{mes}A_i^\ell \ \boldsymbol{\tau}_{i|A_i^\ell}^\ell \cdot \mathbf{t}_i^\ell = 1$:

$$\boldsymbol{\tau}_i^\ell(x,y) = \frac{1}{\det\left(\overrightarrow{S_{i+1}^\ell S_{i+2}^\ell}, \overrightarrow{S_{i+1}^\ell S_i^\ell}\right)} \begin{pmatrix} -(y - y_{i+2}^\ell) \\ (x - x_{i+2}^\ell) \end{pmatrix}.$$

On a donc ainsi construit les fonctions de base locales $(\boldsymbol{\tau}_i^\ell)_{i=1,2,3}$ associées aux degrés de liberté locaux $(M_i^\ell)_{i=1,2,3}$. Il s'agit maintenant de construire les degrés de liberté globaux et les fonctions de base globales associées. L'idée naturelle consiste à réunir en un seul degré de liberté global les deux degrés de liberté définis de part et d'autre d'une arête commune à deux triangles. Il faut prendre quelques précautions car le vecteur tangent sur une arête est défini à un signe près (orientation du vecteur tangent). Il faut fixer sur chaque arête l'orientation du vecteur tangent, par exemple en décidant que le vecteur tangent est orienté suivant la numérotation globale croissante des sommets du maillage.

Plus précisément, considérons une arête interne A_k commune aux triangles T^p

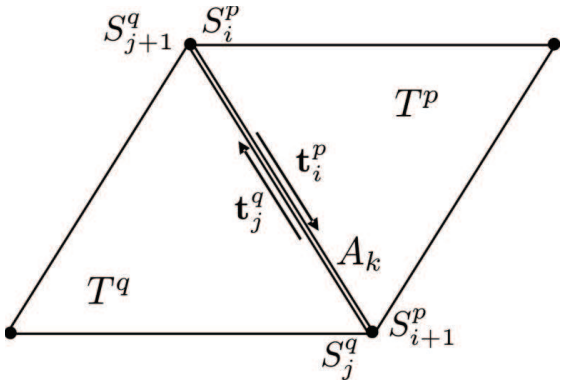


Figure 2.13. Elément fini d'arête 2D

et T^q. L'arête A_k correspond au segment $]S_i^p, S_{i+1}^p[$ du triangle T^p et au segment $]S_j^q, S_{j+1}^q[$ du triangle T_q . Si lg désigne l'application qui passe de la numérotation locale à la numérotation globale des sommets, on note :

$$\begin{aligned} k_1 &= \min(lg(p,i), lg(p,i+1)) = \min(lg(q,j,), lg(q,j+1)) \\ k_2 &= \max(lg(p,i), lg(p,i+1)) = \max(lg(q,j), lg(q,j+1)) \end{aligned}$$

et on choisit le vecteur tangent suivant sur l'arête A_k :

$$\mathbf{t}_k = \frac{\overrightarrow{S_{k_2} S_{k_1}}}{\text{mes}A_k}.$$

Par la suite, on pose :

$$s_p = \begin{cases} 1 \quad \text{si } k_1 = lg(p,i) \\ -1 \text{ si } k_1 \neq lg(p,i) \end{cases} \text{ et } s_q = \begin{cases} 1 \quad \text{si } k_1 = lg(q,j) \\ -1 \text{ si } k_1 \neq lg(q,j) \end{cases}.$$

On définit le degré de liberté global sur l'arête k :

$$M_k(\mathbf{v}) = \int_{A_k} \mathbf{v} \cdot \mathbf{t}_k \, dA = s_p M_i^p(\mathbf{v}) = s_q M_j^q(\mathbf{v})$$

et la fonction de base globale $\mathbf{r}_k$ associée au degré de liberté global M_k :

$$\mathbf{r}_k = \begin{cases} s_p \boldsymbol{\tau}_i^p \ \text{ sur } T_p \\ s_q \boldsymbol{\tau}_j^q \ \text{ sur } T_q \end{cases}.$$

En notant $\mathcal{K}$ l'ensemble des indices des arêtes internes du maillage, $(\mathbf{r}_k)_{k\in\mathcal{K}}$ constitue une base de l'espace V_h^0 associée aux degrés de liberté globaux $(M_k)_{k\in\mathcal{K}}$. On a pour tout $\mathbf{v} \in V_h^0$:

$$\mathbf{v} = \sum_{k\in\mathcal{K}} M_k(\mathbf{v}) \mathbf{r}_k.$$

On note $\mathcal{I}$ l'ensemble des indices des sommets qui ne sont pas situés sur la frontière $\partial\Omega$, de telle sorte que $(w_I)_{I\in\mathcal{I}}$ constitue une base de M_h^0. Nous sommes maintenant en mesure d'écrire le système linéaire associé à la formulation variationnelle discrète (2.124)-(2.125) :

$$\begin{bmatrix} \mathbb{R} & \mathbb{D} \\ \mathbb{D}^t & 0 \end{bmatrix} \begin{pmatrix} \overrightarrow{E} \\ \overrightarrow{\lambda} \end{pmatrix} = \begin{pmatrix} \overrightarrow{F} \\ \overrightarrow{0} \end{pmatrix}, \tag{2.126}$$

avec

$$\mathbf{E}_h = \sum_{k\in\mathcal{K}} E_k \, \mathbf{r}_k, \ \lambda_h = \sum_{I\in\mathcal{I}} \lambda_I \, w_I,$$
$$\mathbb{R}_{kl} = \int_\Omega \text{rot}\, \mathbf{r}_k \ \text{rot}\, \mathbf{r}_l \, d\Omega, \ \mathbb{D}_{kj} = \int_\Omega \varepsilon \mathbf{r}_k \cdot \nabla w_J \, d\Omega, \ k, \ell \in \mathcal{K} \text{ et } J \in \mathcal{I}$$
$$F_k = \int_\Omega \text{rot}\, \mathbf{r}_k \, f_E \, d\Omega.$$

Mise en œuvre

L'implémentation des éléments finis d'arêtes requiert une indexation des arêtes du maillage. Il est assez facile de construire un tableau (L×3) indiquant pour chaque arête d'un triangle son numéro global dans la liste de toutes les arêtes du maillage. Le script Matlab suivant construit ainsi la liste globale A des arêtes ($A(a,1)$ et $A(a,2)$ contenant les numéros des extrémités de l'arête a rangés dans l'ordre croissant) ainsi que le tableau de numérotation locale des arêtes NA ($NA(t,a)$ indiquant le numéro global de la $a^{ème}$ arête du triangle t, son signe fournissant son orientation par rapport à l'orientation de référence). Les tableaux A et NA sont redondants, mais suivant les situations il est plus commode de disposer de l'un ou de l'autre.

```
function [A,NA]=aretes(S,T)
ns=size(S,1);nt=size(T,1);na=0;
G=sparse(ns,ns);NA=zeros(nt,3);                    % matrice des arêtes
for t=1:nt,                                        % boucle sur les triangles
    for a=1:3,                                     % boucle sur les arêtes
        ap=a+1;if(ap==4) ap=1;end,
        if(T(t,a)<T(t,ap)) I=T(t,a);J=T(t,ap);s=1;
        else I=T(t,ap);J=T(t,a);s=-1;
        end
        I=min(T(t,a),T(t,ap));J=max(T(t,a),T(t,ap));      % orientation
        if(G(I,J)==0)                                  % nouvelle arête
            na=na+1;G(I,J)=na;
        end,
        NA(t,a)=s*G(I,J);                  % numérotation locale de l'arête
    end,
end,
A=zeros(na,2);[r,c,v]=find(G);                     % liste globale des arêtes
A(abs(v),1)=r;A(abs(v),2)=c;
```

On a également besoin de repérer les arêtes du bord sur lequel on impose que la composante tangentielle du champ électrique est nulle. Le script suivant construit à partir de la numérotation des arêtes et des références des arêtes de bord, le tableau AB indiquant si une arête appartient aux parties de la frontière considérées.

```
function [AB]=aretes_bords(A,NA,B,RB)
na=size(A,1);;nt=size(NA,1);AB=zeros(na,1)
for t=1:nt,
    for a=1:3;
        if(ismember(B(t,a),RB))                    %arête a sur un bord
           AB(abs(NA(t,a)))=1;
        end,
    end,
end,
```

Une fois ces outils construits, nous sommes en mesure de construire la matrice et le vecteur constituant le système linéaire (2.126). Le script Matlab suivant réalise à partir d'un maillage P^1 et de la donnée fonction f_E, le calcul des matrices $\mathbb{R}$, $\mathbb{D}$ et du vecteur $\overrightarrow{F}$. On supppose ici pour simplifier que ε est constant égal à 1. Ces matrices sont évaluées sans tenir compte des conditions essentielles, on utilise par la suite une technique de pseudo-élimination afin de prendre en compte ces conditions. Afin de projeter le champ approché $\mathbf{E}_h$ dans l'espace M_h (utile pour réaliser des représentations graphiques et des calculs d'erreurs), ce script calcule également les matrices de masse $\mathbb{M}$ et $\mathbb{N}$ définies par :

$$\mathbb{M}_{IJ} = \int_\Omega w_I\, w_J\, d\Omega,\ \mathbb{N}_{Il} = \int_\Omega w_i\, \mathbf{r}_l\, d\Omega,\quad I,J = 1,N,\ l = 1,L.$$

Ce script fonctionne encore suivant un principe d'assemblage de calculs élémentaires effectués par la fonction **EFelmMaxwell_hrot_P1.m**.

```
function [R,D,F,M,N]=EFMaxwell_hrot_P1(S,T,NAr,na,fE)
nt=size(T,1);ns=size(S,1);
```

```
R=sparse(na,na);D=sparse(na,ns);F=zeros(na,1);
M=sparse(ns,ns);N=sparse(2*ns,na);
for t=1:nt,                                          %boucle sur les éléments
    [Rt,Dt,Ft,Mt,Nt]=EFelmMaxwell_hrot_p1(S(T(t,:),:),sign(NAr(t,:)),fE);
    I=T(t,:);K=abs(NAr(t,:));
    R(K,K)=R(K,K)+Rt;D(K,I)=D(K,I)+Dt;
    M(I,I)=M(I,I)+Mt;N([I I+ns],K)=N([I I+ns],K)+Nt;
    F(K)=F(K)+Ft;
 end

function [R,D,F,M,N]=EFelmMaxwell_hrot_p1(S,sA,fE)
nbq=7;os=sqrt(15);s3=1./3.;              %Formule de quadrature à 7 points
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;pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];
S21=S(2,:)-S(1,:);S31=S(3,:)-S(1,:);     %calcul des matrices élémentaires
delta=S21(1)*S31(2)-S21(2)*S31(1);aire=abs(delta)/2;
Jflmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;              %transfo affine!
D=zeros(3,3);F=zeros(3,1);N=zeros(6,3);M=zeros(3,3);
rotr=2*sA/delta;R=aire*rotr'*rotr;          %int rot rk.rot rl (constant)
for k=1:nbq,                              %int rk.grad(wj) et int_T fE rot(rk)
      x=pts_quadT(k,1);y=pts_quadT(k,2);
      pk=pds_quadT(k)*abs(delta);
      P=S(1:3,:)'*[1-x-y; x; y];
      w=[1-x-y x y];gw=[-1 1 0;-1 0 1]; jg=Jflmt*gw;
      x=P(1);y=P(2);
      r=[sA;sA].*[S(3,2)-y S(1,2)-y  S(2,2)-y;
                  x-S(3,1) x-S(1,1)  x-S(2,1)]/delta;
      D=D+pk*r'*jg;F=F+pk*fE(P)*rotr';
      M=M+pk*w'*w;N=N+[pk*w'*r(1,:);pk*w'*r(2,:)];
end,
```

Notons que le calcul du terme rot rot est réalisé dans le plan physique, sans utiliser de calcul par quadrature numérique. Ce qui est possible ici car les fonctions de base $\mathbf{r}_k$ sont de degré 1, et beaucoup plus simple car le changement de variable dans l'opérateur rot n'est pas très sympathique ! La résolution par éléments finis d'arêtes du problème de Maxwell est finalement réalisée à l'aide du script suivant :

```
n=10;
[S,T,BR,RT]=triangle_rectangle([0 1 0 1],n,n);     %maillage P1 du carré
[Ar,NAr]=aretes(S,T);                              %indexation des arêtes
ns=size(S,1);na=size(Ar,1);
[R,D,F,M,N]=EFMaxwell_hrot_P1(S,T,NAr,na,@fE);                %calculs EF
A=[R D;D' sparse(ns,ns)];B=[F;zeros(ns,1)];                   %assemblage
[AB]=aretes_bords(Ar,NAr,BR,[1 2 3 4]);          %condition limite Exn=0
ID=find(AB==1);DA=diag(A);                       %par pseudo élimination
A(ID,:)=0;A(:,ID)=0;B(ID)=0;
for i=1:size(ID,1),A(ID(i),ID(i))=DA(ID(i));end
ND=noeud_bords(S,T,BR,[1 2 3 4]);                %condition limite L=0
ID=na+find(ND==1);                               %par pseudo élimination
A(ID,:)=0;A(:,ID)=0;B(ID)=0;
for i=1:size(ID,1),A(ID(i),ID(i))=DA(ID(i));end
X=A\B;                                                        %résolution
E=X(1:na);L=X(na+1:end);
Ex_P1=M\(N(1:ns,:)*E);Ey_P1=M\(N(ns+1:end,:)*E);              %projection P1
```

Expérimentation numérique

On valide le script précédent en choisissant la solution :

$$\mathbf{E} = \begin{pmatrix} -\cos(ax)\sin(ay) \\ \sin(ax)\cos(ay) \end{pmatrix}$$

qui est à divergence nulle. En prenant $a = k\pi$, le champ $\mathbf{E}$ vérifie la condition aux limites $\mathbf{E} \times \mathbf{n}_{|\partial\Omega} = 0$. Il suffit donc de choisir la donnée :

$$f_E = \mathrm{rot}\,\mathbf{E} = 2a\cos(ax)\cos(ay).$$

Nous indiquons sur les figures 2.14 et 2.15, le champ électrique obtenu à l'aide de l'approximation précédente avec un pas de maillage $h = 0.05$ ($n = 50$). Sur ces figures, sont également représentées le champ exact ainsi que l'écart entre le champ approché et le champ exact. La figure 2.14 correspond au cas $a = \pi$ tandis

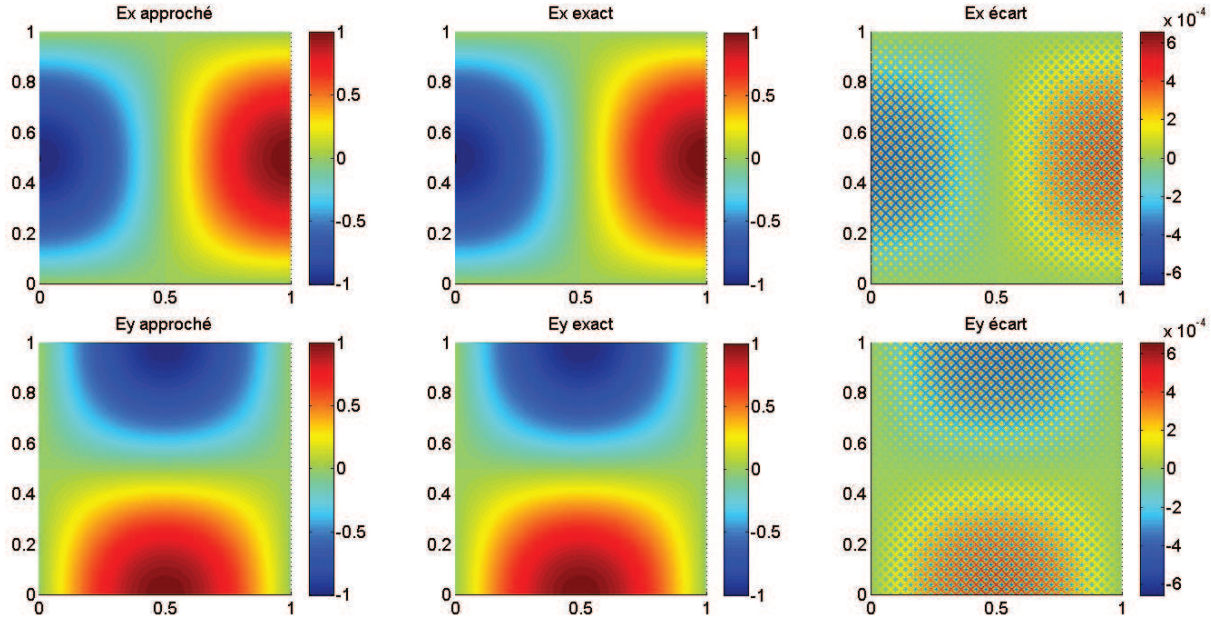


Figure 2.14. Champs électriques approché et exact, écart entre les deux champs ($a = \pi$)

que la figure 2.15 correspond au cas $a = 2\pi$. Signalons que les éléments finis de Raviart-Nédélec ne conduisent pas à des champs continus, seule la composante tangentielle du champ électrique le long des arêtes est continue. C'est pourquoi on utilise une projection $\hat{\mathbf{E}}_h$ sur l'espace $M_h \times M_h$ de ces champs approchés en utilisant la relation :

$$\int_\Omega \hat{\mathbf{E}}_h w_I \, d\Omega = \int_\Omega \mathbf{E}_h w_I \, d\Omega, \quad \forall I = 1, N$$

soit sous forme matricielle :

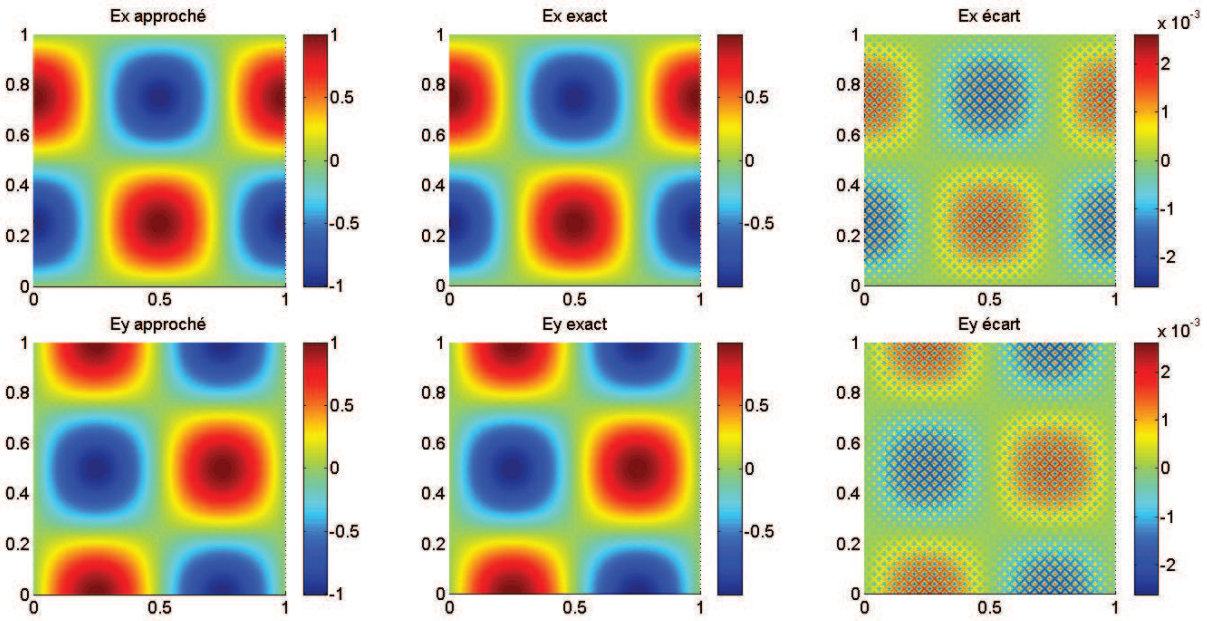


Figure 2.15. Champs électriques approché et exact, écart entre les deux champs ($a = 2\pi$)

$$\mathbb{M}\vec{\vec{E}} = \mathbb{N}\vec{E}.$$

L'approximation est de bonne qualité, les écarts étant de l'ordre de 10^{-3} pour un champ électrique de l'ordre de l'unité. Nous n'avons pas représenté le multiplicateur λ_h car ce dernier est très petit, de l'ordre de 10^{-13}. Rappelons qu'il est théoriquement nul dans le cas de la formulation continue. Cette propriété n'est pas exactement vérifiée dans le cas discret, à cause des erreurs d'arrondis. Bien qu'insignifiant, il n'est pas possible de supprimer le multiplicateur de la formulation discrète car dans ce cas cette dernière n'est plus inversible !

Nous avons également réalisé une étude expérimentale de convergence. Nous re-

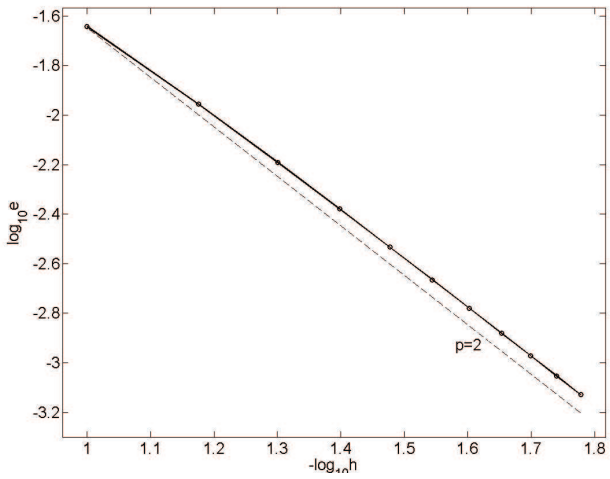


Figure 2.16. Convergence en norme L^2 de l'approximation mixte RTN-P^1 de l'équation de Maxwell quasi-statique

présentons sur la figure 2.16 l'erreur en norme L^2 en fonction du pas de maillage en échelle logarithmique. On observe une convergence à l'ordre 2 ; ce qui est tout à fait compatible avec la théorie.

Elimination du multiplicateur

Comme nous l'avons déjà indiqué, le multiplicateur ne peut pas être éliminé de la formulation en considérant par exemple la nouvelle formulation :

$$\left\{ \begin{array}{l} \text{trouver } \mathbf{E}_h \in V_h^0 \text{ tel que} \\ \displaystyle\int_\Omega \text{rot}\, \mathbf{E}_h \ \text{rot}\, \mathbf{v}_h \, d\Omega = \int_\Omega f_E \ \text{rot}\, \mathbf{v}_h \, d\Omega \ \forall \mathbf{v}_h \in V_h^0, \end{array} \right. \tag{2.127}$$

car alors plus rien ne garantit la nullité de la divergence du champ !
Néanmoins, il est possible dans certaines situations de considérer une telle formulation non mixte. C'est ainsi le cas pour les équations de Maxwell en régime harmonique (dépendance en $e^{-i\omega t}$ avec $\omega \neq 0$) et $\mu = cte$:

$$\left\{ \begin{array}{ll} \mathbf{rot}\ \text{rot}\, \mathbf{E} - \omega^2 \mu\varepsilon \mathbf{E} = -i\omega\mu \mathbf{J} & \text{dans } \Omega, \\ \text{div}(\varepsilon \mathbf{E}) = 0 & \text{dans } \Omega, \\ \mathbf{E} \times \mathbf{n} = 0 & \text{sur } \partial\Omega. \end{array} \right.$$

Comme $\text{div}\, \mathbf{J} = 0$, en prenant la divergence de la première équation, on retrouve, car $\omega >\neq 0$, que $\text{div}(\varepsilon \mathbf{E}) = 0$ dans Ω. Il n'est donc plus nécessaire de l'imposer !
On peut ainsi considérer la formulation variationnelle discrète :

$$\left\{ \begin{array}{l} \text{trouver } \mathbf{E}_h \in V_h^0 \text{ tel que} \\ \displaystyle\int_\Omega \text{rot}\, \mathbf{E}_h \ \text{rot}\, \mathbf{v}_h \, d\Omega - \omega^2 \mu \int_\Omega \varepsilon \mathbf{E}_h \cdot \mathbf{v}_h \, d\Omega \\ \qquad\qquad = -i\omega\mu \displaystyle\int_\Omega \mathbf{J} \cdot \mathbf{v}_h \, d\Omega \end{array} \right. \forall \mathbf{v}_h \in V_h^0. \tag{2.128}$$

A partir des scripts précédents il est facile d'écrire ceux correspondant à cette formulation. Nous montrons sur la figure 2.17, le champ obtenu à l'aide de cette approximation dans le cas où $\omega = 1$ et où le courant $\mathbf{J}$ est choisi de telle sorte que la solution demeure la même ($a = 2\pi,\ n = 50$). On note que le niveau de l'écart est du même ordre (10^{-3}) que celui observé avec la formulation mixte.

Utilisation d'éléments finis conforme H^1

On peut se demander si une approximation du champ électrique par éléments finis de Lagrange peut donner un résultat correct. Nous indiquons sur la figure 2.18 le champ obtenu (solution de (2.124)-(2.125)) à l'aide d'une approximation P^1 des composantes du champ électrique et du multiplicateur. Comme

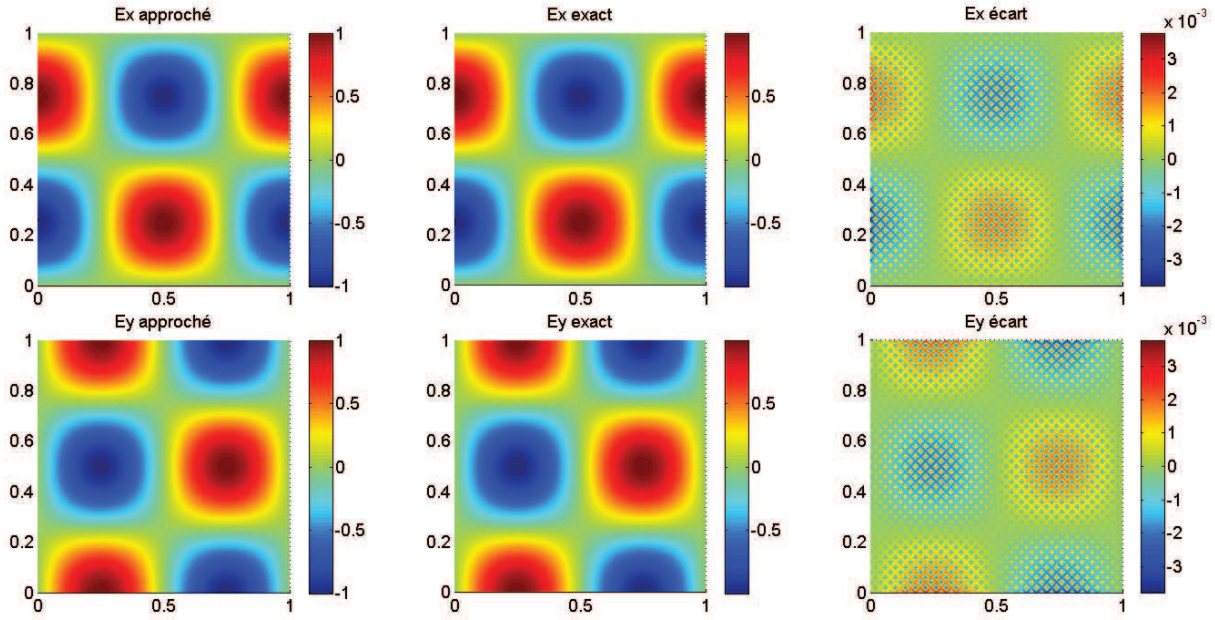


Figure 2.17. Champ électrique approché, exact et écart en régime harmonique

on pouvait s'y attendre cette approximation ne marche pas, la forme bilinéaire $(\mathbf{u}, \mathbf{v}) \mapsto (\operatorname{rot} \mathbf{u}, \operatorname{rot} \mathbf{v})_{L^2(\Omega)}$ n'étant pas coercive sur $H^1(\Omega)^2$.

En fait, on peut mener la réflexion un peu plus loin en introduisant un problème dit

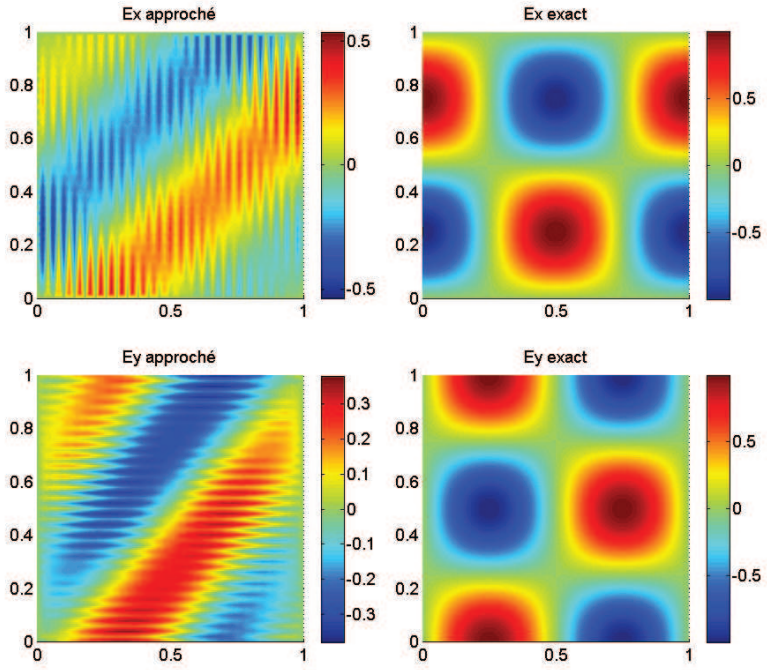


Figure 2.18. Champ électrique approché et exact avec une approximation P^1-P^1 ($a = 2\pi$)

régularisé qui consiste à ajouter à l'équation le terme de régularisation $\alpha\varepsilon\nabla \operatorname{div} \varepsilon\mathbf{E}$ ($\alpha > 0$). Comme $\operatorname{div} \varepsilon\mathbf{E} = 0$, on n'a en fait rien rajouté ! On considère donc le problème régularisé suivant :

$$\begin{cases} \mathbf{rot}\ \mathrm{rot}\,\mathbf{E} - \alpha\varepsilon\nabla\,\mathrm{div}\,\varepsilon\mathbf{E} = \mathbf{rot}\,\mathbf{f}_E & \text{dans } \Omega, \\ \mathbf{E}\times\mathbf{n} = 0 & \text{sur } \partial\Omega, \\ \mathrm{div}(\varepsilon\mathbf{E}) = 0 & \text{sur } \partial\Omega. \end{cases} \tag{2.129}$$

On conserve la condition de divergence nulle sur le bord afin d'assurer la réciproque. En effet, en prenant la divergence de l'équation en volume, on obtient :

$$\begin{cases} \mathrm{div}(\varepsilon\nabla\,\mathrm{div}\,\varepsilon\mathbf{E}) = 0 & \text{dans } \Omega, \\ \mathrm{div}(\varepsilon\mathbf{E}) = 0 & \text{sur } \partial\Omega \end{cases}$$

dont l'unique solution dans $H^1(\Omega)$ est $\mathrm{div}\,\varepsilon\mathbf{E} = 0$.
La formulation variationnelle dans X_E (cf. (2.114)) du problème (2.129) est :

$$\begin{cases} \text{trouver } \mathbf{E}\in X_E \text{ tel que} \\ \displaystyle\int_\Omega \mathrm{rot}\,\mathbf{E}\ \mathrm{rot}\,\mathbf{v}\,d\Omega + \alpha\int_\Omega \mathrm{div}(\varepsilon\mathbf{E})\,\mathrm{div}(\varepsilon\mathbf{v})\,d\Omega = \int_\Omega f_E\ \mathrm{rot}\,\mathbf{v}\,d\Omega\ \forall\mathbf{v}\in X_E. \end{cases}$$

Ce problème est bien posé dans l'espace W. Si ε est suffisamment régulier (continu), on peut donc en faire une approximation par éléments finis P^1. Sur l'espace W_h défini par :

$$W_h = \{\mathbf{v}_h \in X_E \text{ tel que } \mathbf{v}_{h|T^\ell} \in P^1(T^\ell)^2\ \forall \ell = 1, L\},$$

on introduit la formulation variationnelle discrète :

$$\begin{cases} \text{trouver } \mathbf{E}_h\in W_h \text{ tel que} \\ \displaystyle\int_\Omega \mathrm{rot}\,\mathbf{E}_h\ \mathrm{rot}\,\mathbf{v}_h\,d\Omega + \alpha\int_\Omega \mathrm{div}(\varepsilon\mathbf{E}_h)\,\mathrm{div}(\varepsilon\mathbf{v}_h)\,d\Omega = \int_\Omega f_E\ \mathrm{rot}\,\mathbf{v}_h\,d\Omega\ \forall\mathbf{v}_h\in W_h. \end{cases}$$

Nous donnons les scripts Matlab permettant de calculer la solution associée à cette formulation :

```
n=50;alpha=0.01;
[S,T,BR,RT]=triangule_rectangle([0 1 0 1],n,n);     %maillage P1 du carré
ns=size(S,1);
[R,Q,F]=EFMaxwell_P1_P1(S,T,@fE);                          %calculs EF
A=R+alpha*Q;B=F;
ND=noeud_bords(S,T,BR,[1 3]);          %condition limite Ex=0 sur 1 et 3
ID=find(ND==1);dA=diag(A);
A(ID,:)=0;A(:,ID)=0;B(ID)=0;
for i=1:size(ID,1),A(ID(i),ID(i))=dA(ID(i));end
ND=noeud_bords(S,T,BR,[2 4]);          %condition limite Ey=0 sur 2 et 4
ID=ns+find(ND==1);
A(ID,:)=0;A(:,ID)=0;B(ID)=0;
for i=1:size(ID,1),A(ID(i),ID(i))=dA(ID(i));end
X=A\B;                                                     %résolution
Ex_P1=X(1:ns);Ey_P1=X(ns+1:2*ns);
```

```
function [R,Q,F]=EFMaxwell_P1_P1(S,T,fE)
nt=size(T,1); ns=size(S,1);
R=sparse(2*ns,2*ns);Q=sparse(2*ns,2*ns);F=zeros(2*ns,1);
for t=1:nt,                                          %boucle sur les éléments
    [Rt,Qt,Ft]=EFelmMaxwell_P1_p1(S(T(t,:),:),fE);
    I=T(t,:);K=[I ns+I] ;
    R(K,K)=R(K,K)+Rt;
    Q(K,K)=Q(K,K)+Qt;
    F(K)=F(K)+Ft;
end

function [R,Q,F]=EFelmMaxwell_P1_p1(S,fE)
nbq=7;os=sqrt(15);s3=1./3.;                          %formule de quadrature
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;pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];
S21=S(2,:)-S(1,:);S31=S(3,:)-S(1,:);S23=S(2,:)-S(3,:);
delta=S21(1)*S31(2)-S21(2)*S31(1);aire=abs(delta)/2;
Jflmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;          %transfo affine!
rotr=[S23(1) S31(1) -S21(1) S23(2) S31(2) -S21(2)]/delta;      % rot(wi)
divr=[S23(2) S31(2) -S21(2) -S23(1) -S31(1) S21(1)]/delta;     % div(wi)
R=aire*rotr'*rotr;Q=aire*divr'*divr;
F=zeros(6,1);
for k=1:nbq,                              %calcul par quadrature de int fE rot(wi)
    x=pts_quadT(k,1);y=pts_quadT(k,2);
    pk=pds_quadT(k)*abs(delta);
    P=S(1:3,:)'*[1-x-y; x; y];
    F=F+pk*fE(P)*rotr';
end,
```

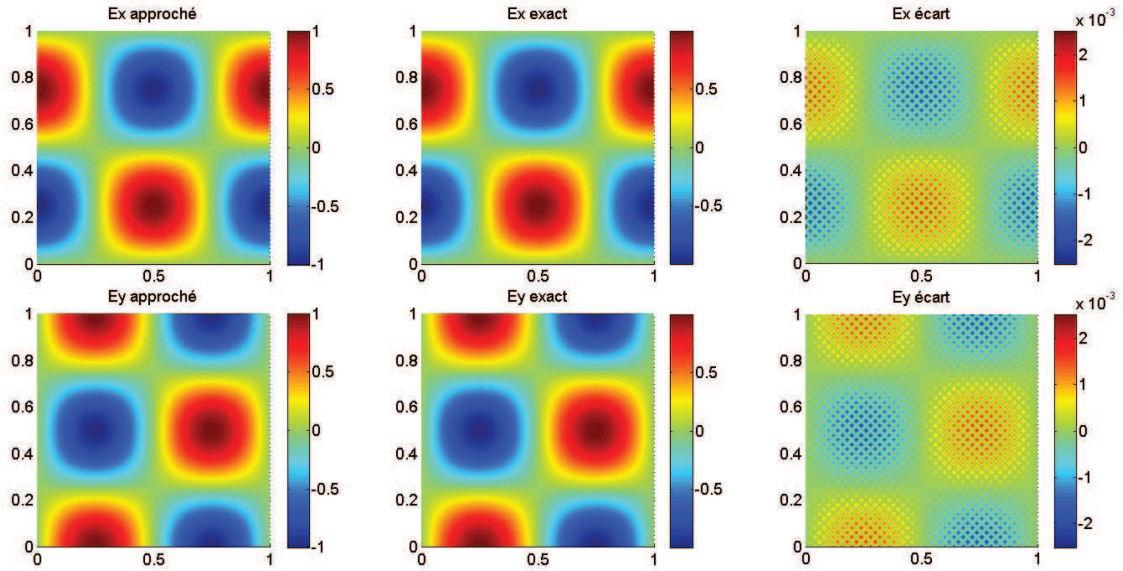


Figure 2.19. Champ électrique approché, exact et écart pour le problème de Maxwell régularisé ($a = 2\pi$)

La figure 2.19 montre le champ électrique obtenu à l'aide de la formulation régularisée ($\alpha = 0.01$), toujours comparable à la même solution exacte ($a = 2\pi$). L'ordre

de l'écart avec la solution exacte est encore de l'ordre de 10^{-3}. Nous représentons également, pour différentes valeurs du paramètre de régularisation, l'évolution de l'erreur en norme L^2 en fonction du pas de maillage (cf. figure 2.20). On observe une convergence à l'ordre 2 pour toutes les valeurs du paramètre de régularisation α. On note que certains choix du paramètre de régularisation se révèlent meilleurs, par exemple $\alpha = 0.1$ dans notre cas.

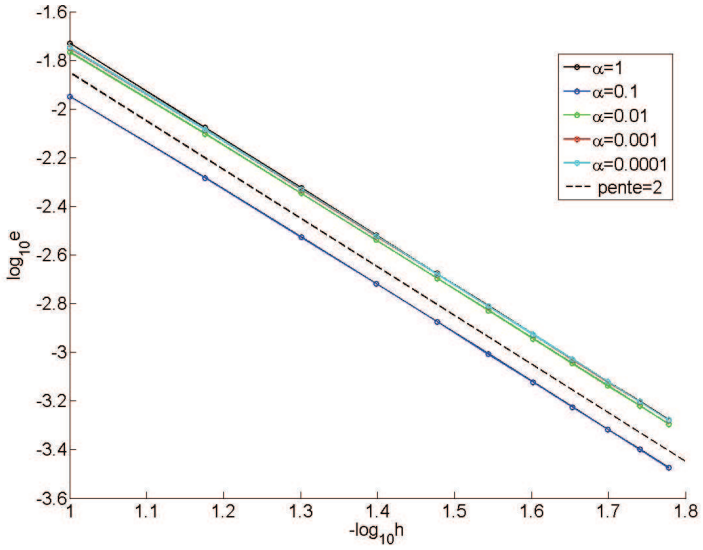


Figure 2.20. Convergence en norme L^2 de l'approximation P^1 du problème de Maxwell régularisé

Signalons une petite finesse qui a des conséquences importantes. Le problème (2.4.2) est également bien posé dans l'espace $W = \mathbf{v} \in H_0^1(\Omega)^2,\ \mathbf{E} \times \mathbf{n}_{|\partial\Omega}$ (voir [19]). Lorsque l'ouvert Ω est convexe, on montre que $W = X_E$; égalité qui n'est plus vraie dans le cas d'ouverts non convexes, par exemple des ouverts avec des coins réentrants. Dans cette situation, il existe des fonctions de X_E qui n'appartiennent pas dans $H^1(\Omega)$, fonctions que l'on qualifie de fonctions singulières. A titre d'illustration, nous donnons sur la figure 2.21, les solutions obtenues, d'une part, avec l'approximation RTN-P^1 de la formulation mixte et d'autre part, avec l'approximation P^1 de la formulation régularisée. Ces solutions ont été obtenues avec la donnée $f_E(x, y) = xy$. On observe qu'au voisinage du coin réentrant les solutions ne sont qualitativement pas les mêmes. D'une certaine façon, l'approximation P^1 du problème régularisée "étale" un comportement légèrement singulier. Certains auteurs ont proposé diverses techniques pour résoudre cette difficulté ; par exemple en ajoutant à l'espace d'approximation des fonctions ayant le comportement singulier adéquat au voisinage des coins réentrants (voir [4, 32]). Ces techniques sont relativement faciles à mettre en œuvre pour des géométries bidimensionnelles (nombre fini de fonctions singulières) mais ardues pour des géométries tridimensionnelles (en particulier pour les arêtes où il y a un continuum de fonctions singulières). Bien

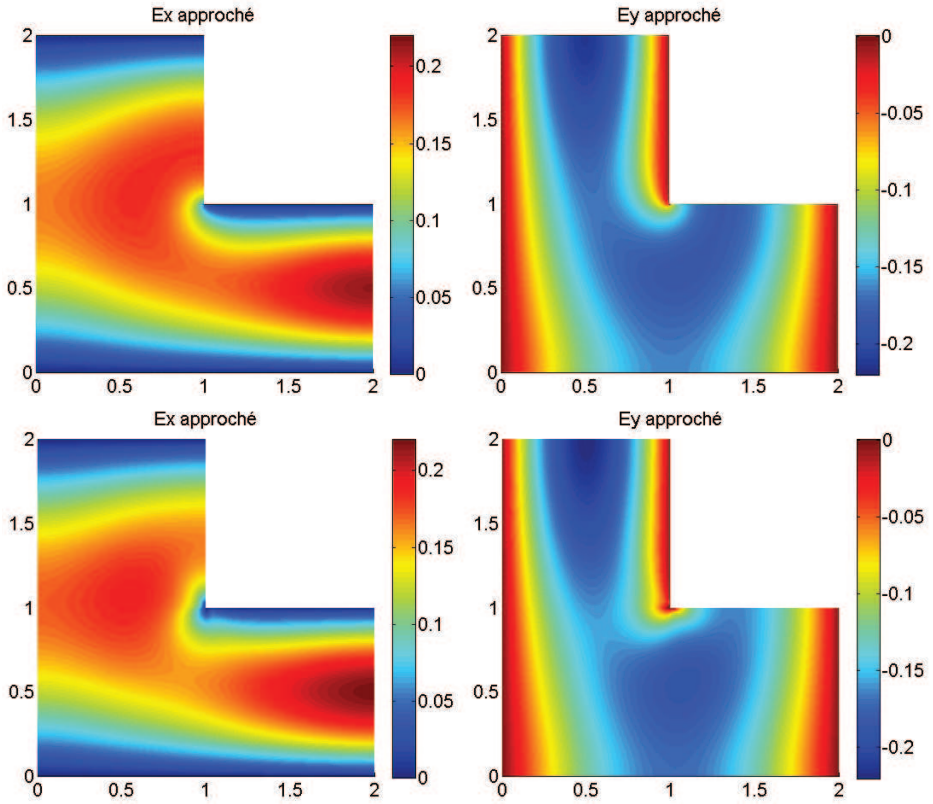


Figure 2.21. Comparaison entre l'approximation P^1 de la formulation régularisée (en bas) et l'approximation RTN-P^1 de la formulation mixte (en haut) dans le cas d'un coin réentrant

évidemment les approximations par éléments finis d'arête conformes dans $H(\mathbf{rot})$ ne présentent pas ces difficultés. Elles sont néanmoins plus délicates à mettre en œuvre et il est plus difficile de monter en ordre.

3

Etude et approximation de l'équation de la chaleur

Ce chapitre et le suivant sont consacrés à l'étude des équations d'évolution (ou instationnaires) dont l'opérateur aux dérivées partielles spatiales est elliptique. Citons deux exemples fondamentaux de telles équations. L'*équation de la chaleur* :

$$\frac{\partial u}{\partial t} - \Delta u = 0$$

caractéristique de la classe des équations paraboliques du second ordre :

$$\frac{\partial u}{\partial t} - Pu = 0$$

avec P un opérateur elliptique [23, 15], à comparer à l'équation des ondes :

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = 0$$

dont l'étude fait l'objet du chapitre 4.

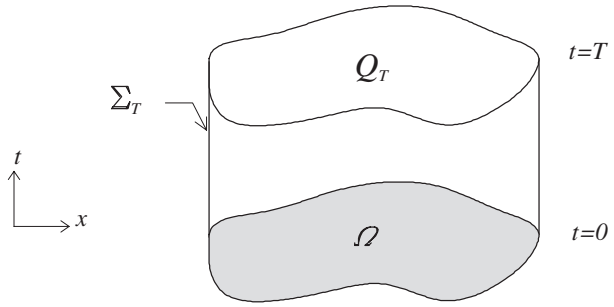
Nous consacrons la première partie de ce chapitre à l'étude théorique de l'équation de la chaleur : formulation variationnelle, existence de solution, principe du maximum, propriété dissipative. Ces résultats reposent essentiellement sur les propriétés spectrales du Laplacien, étudiées au chapitre 1, ainsi que sur des techniques d'estimations d'énergie.

La seconde partie de ce chapitre est consacrée à l'approximation de l'équation de la chaleur. Comme le temps ne joue pas du tout le même rôle que les variables d'espace, les discrétisations en espace et en temps sont a priori découplées. Ainsi, en espace, on utilise soit des éléments finis [15], soit des différences finies [13, 33] pour discrétiser le laplacien. Ces deux approches conduisent alors à des systèmes différentiels en temps. Nous étudions ensuite l'approximation en temps de ces systèmes différentiels fondée sur des schémas aux différences finies classiques [33] pour

les équations différentielles. Nous terminons cette étude en donnant des résultats de convergence en espace-temps et quelques illustrations numériques.

Dans tout le chapitre, nous traitons le problème modèle suivant, où Ω est un ouvert borné de $\mathbb{R}^n$, de frontière $\partial\Omega$ supposée "suffisamment régulière" et $T > 0$:

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = f(x,t), & (x,t) \in Q_T = \Omega\times]0,T[, \\ u(x,t) = 0, & (x,t) \in \Sigma_T = \partial\Omega\times]0,T[, \\ u(x,0) = u_0(x), & x \in \Omega. \end{cases} \tag{3.1}$$



La première équation de (3.1) est l'équation de la chaleur avec une source volumique $f(x,t)$, la seconde est une condition de Dirichlet homogène et traduit une condition de température nulle sur la frontière et la dernière équation est une condition initiale du problème. On étudie donc l'évolution de la température dans Ω, notée u, entre l'*instant initial* $t = 0$ et l'*instant final* $t = T$.

Le traitement d'une condition de Neumann-Fourier $\dfrac{\partial u}{\partial n} + \lambda u = 0$ sur Σ_T, avec $\lambda \geq 0$, serait similaire.

3.1 Théorie variationnelle de l'équation de la chaleur

Afin de donner une formulation variationnelle du problème (3.1), il est nécessaire de préciser les espaces de fonctions en (x,t) que l'on va utiliser.

3.1.1 Espaces de fonctions à valeurs fonctions

La variable temps t ne jouant pas le même rôle que la variable d'espace x, on introduit la notation $u(t)$ pour désigner la fonction $x \mapsto u(x,t)$ à t fixé. Par conséquent, la fonction $(x,t) \mapsto u(x,t)$ s'interprète comme une fonction à valeurs dans un espace de fonctions :

$$\begin{array}{rccc} u & : & [0,T] & \to & V(\Omega) \\ & & t & \mapsto & u(t) \end{array}$$

où $V(\Omega)$ est un espace de fonctions définies sur Ω.

Cette interprétation conduit à l'introduction des espaces suivants :

$\mathcal{C}^0(0,T;V(\Omega))$: ensemble des fonctions continues sur $[0,T]$ à valeurs dans $V(\Omega)$.

$L^2(0,T;V(\Omega))$: ensemble des fonctions mesurables et de carré intégrable sur $]0,T[$ à valeurs dans $V(\Omega)$.

Lorsque $V(\Omega)$ est un espace de Banach, muni de la norme $\|\cdot\|_{V(\Omega)}$, $\mathcal{C}^0(0,T;V(\Omega))$ est également un espace de Banach muni de la norme :

$$\|u\|_{\mathcal{C}^0(0,T;V)} = \sup_{t\in[0,T]} \|u(t)\|_{V(\Omega)}.$$

Si $V(\Omega)$ est un espace de Hilbert, muni du produit scalaire $(\cdot,\cdot)_V$, $L^2(0,T;V(\Omega))$ est également un espace de Hilbert muni du produit scalaire :

$$(u,v)_{L^2(0,T;V)} = \int_0^T (u(t),v(t))_V dt$$

dont la norme associée est :

$$\|u\|_{L^2(0,T;V)} = \left(\int_0^T \|u(t)\|_V^2 dt\right)^{1/2}.$$

Evidemment, on peut réaliser l'identification :

$$L^2(0,T;L^2(\Omega)) = L^2(]0,T[\times\Omega).$$

3.1.2 Formulation variationnelle de l'équation de la chaleur

Pour clarifier les calculs, on laisse ici la dépendance explicite de u par rapport à x et t, ainsi que celle des fonctions-test (de $H_0^1(\Omega)$) par rapport à x. On procède de même pour les éléments d'intégration : $d\Omega$ est (temporairement) remplacé par dx. Supposons que la fonction $u(x,t)$ soit solution de (3.1) et suffisamment régulière, c'est-à-dire $u \in \mathcal{C}^1(0,T;L^2(\Omega))$ et $u \in L^2(0,T;H^1(\Delta;\Omega))$, avec $H^1(\Delta;\Omega) = \{v \in H^1(\Omega)$ tel que $\Delta v \in L^2(\Omega)\}$. La régularité de la donnée f sera précisée plus loin (cf. (3.3)). Multiplions la première équation de (3.1) par une fonction test $v \in H_0^1(\Omega)$ et intégrons sur Ω :

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)\, v(x)\, dx - \int_\Omega \Delta u(x,t)\, v(x)\, dx = \int_\Omega f(x,t)\, v(x)\, dx.$$

En appliquant une formule de Green (cf. [15]), on obtient :

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)\, v(x)\, dx + \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\, dx = \int_\Omega f(x,t)\, v(x)\, dx.$$

Dès que $u \in \mathcal{C}^1(0,T;L^2(\Omega))$ on a :

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)\, v(x)\, dx = \frac{d}{dt}\left(\int_\Omega u(x,t)\, v(x)\, dx\right)$$

car

$$t \mapsto \int_\Omega u(x,t)\, v(x) dx$$

est une fonction dérivable.

Si on suppose simplement que $u \in \mathcal{C}^0(0,T;L^2(\Omega))$ alors la fonction :

$$t \mapsto \int_\Omega u(x,t)\, v(x)\, dx$$

est seulement continue sur $[0,T]$ et par conséquent,

$$\frac{d}{dt}\left(\int_\Omega u(x,t)\, v(x)\, dx\right)$$

est une distribution sur $]0,T[$ définie par, pour tout $\psi \in \mathcal{D}(]0,T[)$:

$$\begin{aligned}\left\langle \frac{d}{dt}\left(\int_\Omega u(x,t)v(x)\, dx\right); \psi \right\rangle &= -\left\langle \left(\int_\Omega u(x,t)v(x)\, dx\right); \frac{d\psi}{dt} \right\rangle \\ &= -\int_0^T \int_\Omega u(x,t)v(x)\frac{d\psi}{dt}(t)\, dxdt = -\int_{Q_T} u(x,t)v(x)\frac{d\psi}{dt}(t)\, dxdt.\end{aligned}$$

On introduit alors la formulation variationnelle (dite faible) du problème (3.1) :

$$\begin{cases} \text{trouver} \quad u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega)) \text{ tel que} \\ \dfrac{d}{dt}\left(\displaystyle\int_\Omega u(x,t)\, v(x)\, dx\right) + \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\, dx = \int_\Omega f(x,t)\, v(x)\, dx, \\ \qquad\qquad\qquad\qquad\qquad\qquad \forall v \in H_0^1(\Omega), \text{ p.p. } t \in]0,T[\\ u(x,0) = u_0(x) \quad \text{ sur } \quad \Omega. \end{cases} \tag{3.2}$$

La formulation (3.2) requiert une régularité minimale sur les données. Nous nous placerons par la suite dans le cas :

$$f \in L^2(0,T;L^2(\Omega)) \quad \text{et} \quad u_0 \in L^2(\Omega). \tag{3.3}$$

On peut démontrer plus généralement que si $u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ est solution de (3.2) alors u vérifie (3.1) au sens des distributions sur Q_T. Par ailleurs, pour une fonction-test $v \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$, on doit remplacer l'égalité apparaissant dans (3.2) par l'égalité suivante, valable pour presque tout $t \in]0,T[$:

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)\, v(x,t)\, dx + \int_\Omega \nabla u(x,t) \cdot \nabla v(x,t)\, dx = \int_\Omega f(x,t)\, v(x,t)\, dx. \qquad (3.4)$$

Remarque 3.1 *La condition initiale de la formulation (3.2) a bien un sens sous les hypothèses (3.3) car la solution u est cherchée dans l'espace $\mathcal{C}^0\left(0,T,L^2(\Omega)\right)$. Notons que cette formulation n'est pas symétrique en espace et en temps.*

Il y a évidemment équivalence entre la formulation forte et la formulation faible au sens suivant :

Proposition 3.2 *Toute fonction $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Delta;\Omega))$ est solution du problème (3.1) si et seulement si elle vérifie la formulation faible (3.2).*

Démonstration : l'implication a déjà été établie. Réciproquement, il suffit de réintégrer par parties. ∎

Remarque 3.3 *La formulation variationnelle (3.2) se généralise sous la forme :*

$$\begin{cases} \text{trouver } u \in \mathcal{C}^0(0,T;H) \cap L^2(0,T;V) \text{ tel que :} \\ \dfrac{d}{dt}(u(t),v)_H + a(u(t),v) = \ell(t,v) \qquad \forall v \in V,\ p.p.\ t > 0 \\ u(0) = u_0 \end{cases}$$

où H est un espace de Hilbert muni du produit scalaire $(\cdot,\cdot)_H$, V un sous-espace de Hilbert dense dans H tel que l'injection de V dans H soit compacte, $a(\cdot,\cdot)$ une forme bilinéaire continue sur V telle qu'il existe $\nu \in \mathbb{R}$ rendant $v \mapsto a(v,v)+\nu\|v\|_H^2$ coercive sur V, $\ell(t,\cdot)$ une forme linéaire continue sur V, et enfin $u_0 \in H$.

3.1.3 Existence d'une solution

Il existe plusieurs façons de démontrer l'existence d'une solution au problème (3.2). Celles fondées sur la théorie de Hille-Yoshida ou la théorie des semi-groupes (voir [9]) ainsi que celles fondées sur une théorie variationnelle (voir [39, 23]) sont assez abstraites. C'est pourquoi, nous allons utiliser une technique, basée sur la décomposition spectrale de l'opérateur Δ, qui est simple, et de surcroît très explicite.

Rappelons que, d'après la théorie spectrale (voir le théorème 1.13) le problème de Dirichlet homogène admet pour éléments propres $(\lambda_i, v_i)_{i\geq 1}$ avec :

– $0 < \lambda_1 \leq \lambda_2 \cdots \leq \lambda_i \leq \cdots \to +\infty$

– $(v_i)_i$ base hilbertienne de $L^2(\Omega)$, et base orthogonale de $H_0^1(\Omega)$.

$$\int_\Omega \nabla v_i \cdot \nabla v \, d\Omega = \lambda_i \int_\Omega v_i \, v \, d\Omega \qquad \forall v \in H_0^1(\Omega), \qquad \forall i \geq 1 \tag{3.5}$$

La base de fonctions propres $(v_i)_i$ nous permet d'expliciter la solution du problème variationnel (3.2).

Proposition 3.4 *Si u est solution du problème (3.2) alors on a :*

$$u(t) = \sum_{i\geq 1} \left\{ (u_0, v_i)_{L^2(\Omega)} e^{-\lambda_i t} + \int_0^t (f(s), v_i)_{L^2(\Omega)} e^{-\lambda_i (t-s)} ds \right\} v_i \qquad \forall t \in [0,T], \tag{3.6}$$

la série étant convergente dans $L^2(\Omega)$ pour presque tout t.

Démonstration : Comme $u \in \mathcal{C}^0\left(0,T; L^2(\Omega)\right)$ et $(v_i)_i$ est une base de $L^2(\Omega)$, pour tout $t \in [0,T]$, on a la décomposition modale

$$u(t) = \sum_{i\geq 1} (u(t), v_i)_{L^2(\Omega)} v_i.$$

En outre, $u(t) \in H_0^1(\Omega)$ pour presque tout t, on a donc :

$$\int_\Omega \nabla u(t) \cdot \nabla v_i \, d\Omega = \lambda_i (u(t), v_i)_{L^2(\Omega)} \quad \forall i \geq 1,$$

d'où on déduit de (3.2) (en choisissant $v = v_i$ comme fonction-test) que la fonction continue $\alpha_i(t) = (u(t), v_i)_{L^2(\Omega)}$ est solution de l'équation différentielle :

$$\begin{cases} \dfrac{d}{dt}\alpha_i(t) + \lambda_i \alpha_i(t) = (f(t), v_i)_{L^2(\Omega)} \text{ p.p. } t \in]0,T[\\ \alpha_i(0) = (u_0, v_i)_{L^2(\Omega)} \end{cases} \tag{3.7}$$

dont l'unique solution est donnée par :

$$\alpha_i(t) = (u_0, v_i)_{L^2(\Omega)} e^{-\lambda_i t} + \int_0^t (f(s), v_i)_{L^2(\Omega)} e^{-\lambda_i (t-s)} ds. \qquad \blacksquare$$

Remarque 3.5 *La technique de décomposition spectrale permet également de préciser la dépendance (continue) de la solution vis-à-vis des données u_0 et f :*

$$\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} e^{-\lambda_1 t} + \int_0^t \|f(s)\|_{L^2(\Omega)} e^{-\lambda_1 (t-s)} ds, \quad \forall t \in [0,T]$$

En effet, on vérifie par des majorations élémentaires que :

$$\| \sum_{i\geq 1} (u_0, v_i)_{L^2(\Omega)} e^{-\lambda_i t} v_i \|_{L^2(\Omega)} = \left[\sum_{i\geq 1} (u_0, v_i)^2_{L^2(\Omega)} e^{-2\lambda_i t} \right]^{1/2} \leq \|u_0\|_{L^2(\Omega)} e^{-\lambda_1 t},$$

$$\begin{aligned}\| \sum_{i\geq 1}(f(s),v_i)_{L^2(\Omega)} e^{-\lambda_i(t-s)} v_i\|_{L^2(\Omega)} &= \left[\sum_{i\geq 1}(f(s),v_i)^2_{L^2(\Omega)} e^{-2\lambda_i(t-s)}\right]^{1/2} \\ &\leq \|f(s)\|_{L^2(\Omega)} e^{-\lambda_1(t-s)}.\end{aligned}$$

Nous présentons à la section suivante, la démonstration de ce résultat par une technique fondée sur des estimations d'énergie.

Grâce à cette représentation en série de la solution u, nous allons maintenant établir le résultat d'existence :

Théorème 3.6 (existence) *Si $f \in L^2\left(0,T;L^2(\Omega)\right)$ et $u_0 \in L^2(\Omega)$, alors le problème variationnel (3.2) admet une unique solution $u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$.*

Démonstration : Dans cette démonstration, on note $(\cdot,\cdot)$ le produit scalaire dans $L^2(\Omega)$. Pour établir l'existence d'une solution au problème (3.2), il suffit de vérifier que la série (3.6) converge dans $\mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$, dès lors que $u_0 \in L^2(\Omega)$ et $f \in L^2(0,T;L^2(\Omega))$.
Posons, pour $m \geq 1$,

$$V_m = \underset{i=1,m}{\text{Vect}}\ (v_i)$$

l'espace vectoriel de dimension m engendré par les fonctions propres $(v_i)_{i=1,m}$ et remplaçons alors le problème continu (3.2) par le problème approché :

$$\left\{\begin{array}{l} \text{trouver } u_m : [0,T] \mapsto u_m(t) \in V_m \text{ tel que} \\ \dfrac{d}{dt}\left(\displaystyle\int_\Omega u_m(t) v\, d\Omega\right) + \displaystyle\int_\Omega \nabla u_m(t)\cdot\nabla v\, d\Omega = \int_\Omega f(t) v\, d\Omega \qquad \forall v \in V_m, \\ u_m(0) = u_{0,m} = \displaystyle\sum_{i=1,m} (u_0,v_i)\, v_i. \end{array}\right. \tag{3.8}$$

Il est clair que le problème (3.8), équivalent à un système différentiel linéaire de dimension finie, admet une unique solution $u_m \in \mathcal{C}^0\left(0,T;V_m\right)$ donnée par :

$$u_m(t) = \sum_{i=1}^{m}\left\{(u_0,v_i)\, e^{-\lambda_i t} + \int_0^t (f(s),v_i)\, e^{-\lambda_i(t-s)} ds\right\} v_i, \tag{3.9}$$

c'est-à-dire la somme partielle d'ordre m de la série (3.6).

Nous allons maintenant démontrer que la suite $(u_m)_m$ est une suite de Cauchy à la fois dans les espaces $\mathcal{C}^0\left(0,T;L^2(\Omega)\right)$ et $L^2\left(0,T;H_0^1(\Omega)\right)$.
Soient m,p deux entiers tels que $p > m \geq 1$. Comme $(v_i)_i$ est une base orthonormale de $L^2(\Omega)$, on a :

$$\|u_p(t) - u_m(t)\|_{L^2(\Omega)} = \left[\sum_{i=m+1}^{p}\left((u_0,v_i)\, e^{-\lambda_i t} + \int_0^t (f(s),v_i)\, e^{-\lambda_i(t-s)} ds\right)^2\right]^{1/2},$$

d'où on déduit que (inégalité triangulaire) :

$$\begin{aligned}\|u_p(t) - u_m(t)\|_{L^2(\Omega)} \quad &\leq \left[\sum_{i=m+1}^{p}(u_0,v_i)^2\, e^{-2\lambda_i t}\right]^{1/2} \\ &+ \left[\sum_{i=m+1}^{p}\left(\int_0^t (f(s),v_i)\, e^{-\lambda_i(t-s)} ds\right)^2\right]^{1/2}.\end{aligned} \tag{3.10}$$

D'après l'inégalité de Cauchy-Schwarz et le fait que $\lambda_i > 0$ on a :

$$\begin{aligned}\left(\int_0^t (f(s), v_i)\, e^{-\lambda_i (t-s)} ds\right)^2 &\leq \left(\int_0^t (f(s), v_i)^2\, ds\right)\left(\int_0^t e^{-2\lambda_i (t-s)} ds\right) \\ &\leq \frac{1}{2\lambda_i}\int_0^t (f(s), v_i)^2\, ds\end{aligned} \tag{3.11}$$

qui injecté dans (3.10) conduit à :

$$\begin{aligned}\sup_{t\in[0,T]} \|u_p(t) - u_m(t)\|_{L^2(\Omega)} &\leq \left[\sum_{i=m+1}^{p} (u_0, v_i)^2\right]^{1/2} \\ &\quad + \left[\frac{1}{2\lambda_1}\sum_{i=m+1}^{p}\int_0^T (f(s), v_i)^2\, ds\right]^{1/2}.\end{aligned} \tag{3.12}$$

Par ailleurs, en vertu de (3.5) on a :

$$\begin{aligned}\int_\Omega |\nabla u_p(t) - \nabla u_m(t)|^2\, d\Omega &= \sum_{i=m+1}^{p} \lambda_i \left(\int_\Omega (u_p(t) - u_m(t))\, v_i\, d\Omega\right)^2 \\ \text{(d'après (3.9))} \qquad &= \sum_{i=m+1}^{p} \lambda_i \left((u_0, v_i)\, e^{-\lambda_i t} + \int_0^t (f(s), v_i)\, e^{-\lambda_i (t-s)} ds\right)^2,\end{aligned}$$

qui conduit à la majoration suivante :

$$\|\nabla u_p(t) - \nabla u_m(t)\|^2_{L^2(\Omega)^n} \leq 2\sum_{i=m+1}^{p} \lambda_i \left[(u_0, v_i)^2\, e^{-2\lambda_i t} + \left(\int_0^t (f(s), v_i)\, e^{-\lambda_i (t-s)} ds\right)^2\right].$$

Comme d'une part :

$$\lambda_i \int_0^T e^{-2\lambda_i t} dt = \frac{1}{2}\left(1 - e^{-2\lambda_i T}\right) < \frac{1}{2}$$

et d'autre part (d'après (3.11)) :

$$\lambda_i \int_0^T \left(\int_0^t (f(s), v_i)\, e^{-\lambda_i (t-s)} ds\right)^2 dt \leq \frac{T}{2}\int_0^T (f(t), v_i)^2\, dt,$$

on déduit finalement que :

$$\int_0^T \|\nabla u_m(t) - \nabla u_p(t)\|^2_{L^2(\Omega)^n}\, dt \leq \sum_{i=m+1}^{p} \left[(u_0, v_i)^2 + T\int_0^T (f(t), v_i)^2\, dt\right]. \tag{3.13}$$

Si $u_0 \in L^2(\Omega)$ alors $\|u_0\|^2_{L^2(\Omega)} = \sum_{i\geq 1} (u_0, v_i)^2 < +\infty$, ce qui prouve que :

$$\lim_{m,p\to+\infty} \left[\sum_{i=m+1}^{p} (u_0, v_i)^2\right] = 0.$$

Si $f \in L^2\left(0, T; L^2(\Omega)\right)$ alors $\sum_{i\geq 1}\int_0^T (f(t), v_i)^2\, dt < +\infty$, qui prouve cette fois que :

$$\lim_{m,p\to+\infty}\left[\sum_{i=m+1}^{p}\int_0^T (f(t),v_i)^2\,dt\right]=0.$$

Par conséquent, les inégalités (3.12) et (3.13) montrent respectivement que :

$$\begin{cases}\lim\limits_{m,p\to+\infty}\|u_p-u_m\|_{\mathcal{C}^0\left(0,T;L^2(\Omega)\right)} &=0,\\ \lim\limits_{m,p\to+\infty}\|u_p-u_m\|_{L^2\left(0,T;H_0^1(\Omega)\right)} &=0.\end{cases}$$

Rappelons en effet qu'en vertu de l'inégalité de Poincaré, $v\mapsto\|\nabla v\|_{L^2(\Omega)^n}$ définit une norme équivalente à la norme $H^1(\Omega)$ sur $H_0^1(\Omega)$.
Ceci prouve que $(u_m)_m$ est une suite de Cauchy dans $\mathcal{C}^0\left(0,T;L^2(\Omega)\right)$ et dans $L^2\left(0,T;H_0^1(\Omega)\right)$. Puisque ces deux espaces sont complets, la suite $(u_m)_m$ converge dans chacun de ces espaces. Par ailleurs, comme les injections canoniques $L^2\left(0,T;H_0^1(\Omega)\right)\subset L^2\left(0,T;L^2(\Omega)\right)$ et $\mathcal{C}^0\left(0,T;L^2(\Omega)\right)\subset L^2\left(0,T;L^2(\Omega)\right)$ sont continues, la limite de $(u_m)_m$ est la même dans les deux espaces. On a donc prouvé que :

$$u_m\underset{m\to+\infty}{\longrightarrow}u \text{ dans } L^2\left(0,T;H_0^1(\Omega)\right)\cap\mathcal{C}^0\left(0,T;L^2(\Omega)\right)\cdot \tag{3.14}$$

Il reste à démontrer que u est solution du problème (3.2). D'après (3.8), on a $\forall m\geq l\geq 1$, $\forall v\in V_l,\ \forall\psi\in\mathcal{D}(]0,T[)$:

$$-\int_0^T(u_m(t),v)\,\frac{d\psi}{dt}(t)dt+\int_0^T\left(\int_\Omega\nabla u_m(t)\cdot\nabla v\,d\Omega\right)\psi(t)dt=\int_0^T(f(t),v)\psi(t)dt.$$

Passons à la limite, lorsque $m\to+\infty$. Compte-tenu de (3.14), on a, $\forall v\in V_l$:

$$-\int_0^T(u(t),v)\frac{d\psi}{dt}(t)dt+\int_0^T\left(\int_\Omega\nabla u(t)\cdot\nabla v\,d\Omega\right)\psi(t)dt=\int_0^T(f(t),v)\psi(t)dt. \tag{3.15}$$

Comme $\oplus_{l\geq1}V_l$ est dense dans $H_0^1(\Omega)$ car $(v_i)_i$ est une base de $H_0^1(\Omega)$, la relation (3.15) a lieu pour tout $v\in H_0^1(\Omega)$, ce qui montre que u vérifie (3.2) au sens des distributions sur $]0,T[$.

En outre d'après (3.14), on a en particulier :

$$u_m(0)\to u(0)\quad\text{ dans } L^2(\Omega),$$

et comme :

$$u_m(0)=\sum_{i=1,m}(u_0,v_i)\,v_i\underset{m\to+\infty}{\longrightarrow}u_0\quad\text{dans } L^2(\Omega),$$

on en déduit que $u(0)=u_0$ sur Ω, au sens des fonctions de $L^2(\Omega)$ (i.e. presque partout). Ceci montre que u est bien la solution de (3.2).
Enfin, en vertu de la proposition 3.4 il est clair que la solution du problème (3.2) est unique. ■

Remarque 3.7 *On peut généraliser sans difficulté cette démonstration au problème abstrait de la remarque 3.3.*

3.2 Propriétés de l'équation de la chaleur

Dans cette section, nous allons commencer par établir des estimations d'énergie qui nous permettront de démontrer, par la suite, le caractère dissipatif de l'équation de la chaleur ainsi qu'un principe du maximum.

3.2.1 Estimations d'énergie et caractère dissipatif

Au cours de cette sous-section, nous noterons toujours u la solution du problème (3.2). D'après le théorème 3.6, on sait que $u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$. Par conséquent, pour presque tout t, on peut choisir dans la formulation variationnelle (3.4) $v(x,t) = u(x,t)$ et intégrer sur l'intervalle $]0,t[$. On obtient ainsi l'égalité :

$$\int_0^t \left(\int_\Omega \frac{\partial u}{\partial t}(s) u(s)\, d\Omega \right) ds + \int_0^t \left(\int_\Omega \nabla u(s) \cdot \nabla u(s)\, d\Omega \right) ds = \int_0^t \left(\int_\Omega f(s) u(s)\, d\Omega \right) ds,$$

soit encore :

$$\frac{1}{2} \int_0^t \frac{d}{dt} \left(\int_\Omega u^2(s)\, d\Omega \right) ds + \int_0^t \left(\int_\Omega |\nabla u(s)|^2\, d\Omega \right) ds = \int_0^t \left(\int_\Omega f(s) u(s)\, d\Omega \right) ds,$$

qui conduit à la *première égalité d'énergie* :

Lemme 3.8 (première égalité d'énergie) *On a l'égalité*

$$\frac{1}{2}\|u(t)\|^2_{L^2(\Omega)} + \int_0^t \|\nabla u(s)\|^2_{L^2(\Omega)^n} ds = \int_0^t (f(s), u(s))_{L^2(\Omega)}\, ds + \frac{1}{2}\|u_0\|^2_{L^2(\Omega)}. \tag{3.16}$$

Cette égalité permet de démontrer facilement l'unicité de la solution du problème (3.2).

Proposition 3.9 *Le problème (3.2) admet au plus une solution.*

Démonstration : Soient u_1 et u_2 deux solutions de (3.2). Posons $w = u_1 - u_2$. Alors il est clair, compte-tenu de la linéarité de (3.2), que w vérifie (d'après (3.16)) :

$$\frac{1}{2}\|w(t)\|^2_{L^2(\Omega)} + \int_0^t \|\nabla w(s)\|^2_{L^2(\Omega)^n} ds = 0,$$

qui montre que $w(x,t) \equiv 0$ p.p $x \in \Omega$, $t \in]0,T[$. ∎

L'égalité (3.16) montre que l'équation de la chaleur est *dissipative*, c'est-à-dire que s'il n'y a aucune source de chaleur ($f \equiv 0$) alors la température décroit au cours du temps.

Proposition 3.10 *Soit u la solution du problème (3.2) avec $f \equiv 0$. Soit $t_1 \geq 0$, si $u(t_1) \not\equiv 0$ alors :*

$$\|u(t_1)\|_{L^2(\Omega)} > \|u(t_2)\|_{L^2(\Omega)} \qquad \forall t_2 > t_1 \geq 0.$$

Démonstration : Soient $0 \leq t_1 < t_2 \leq T$**,** d'après (3.16) on a :

$$\frac{1}{2}\|u(t_2)\|^2_{L^2(\Omega)} = \frac{1}{2}\|u(t_1)\|^2_{L^2(\Omega)} - \int_{t_1}^{t_2} \|\nabla u(s)\|^2_{L^2(\Omega)^n} ds.$$

Si $\nabla u(t) \equiv 0 \ \forall t \in [t_1, t_2]$ alors comme $u(t) \in H^1_0(\Omega)$, on aurait $u(t) = cte = 0$, $\forall t \in [t_1, t_2]$ donc $u(t_1) \equiv 0$. Ce qui prouve le résultat. ∎

Remarque 3.11 *Cet effet dissipatif de l'équation de la chaleur est à comparer à l'effet conservatif de l'équation des ondes (voir §4.3.1).*

En fait, on a un résultat plus précis de décroissance exponentielle :

Proposition 3.12 *Soit u la solution de (3.2) avec $f \equiv 0$. Alors il existe une constante $\alpha > 0$, indépendante de t, telle que :*

$$\|u(t)\|^2_{L^2(\Omega)} \leq \|u_0\|^2_{L^2(\Omega)} e^{-\alpha t}.$$

Démonstration : par dérivation (au sens des distributions dans $\mathcal{D}'(]0,T[)$) en temps de l'égalité (3.16), on obtient, en posant $\varphi(t) = \|u(t)\|^2_{L^2(\Omega)}$:

$$\varphi \in L^2(]0,T[) \text{ et } \frac{1}{2}\varphi'(t) + \|\nabla u(t)\|^2_{L^2(\Omega)^n} = 0.$$

Cette égalité montre que $\varphi' \in L^2(]0,T[)$ car $u \in L^2(0,T; H^1_0(\Omega))$ et donc $\|\nabla u\|_{L^2(\Omega)^n} \in L^2(]0,T[)$ et, par conséquent, que $\varphi \in H^1(]0,T[)$.
Par ailleurs, en vertu de l'inégalité de Poincaré dans Ω, on a :

$$\|\nabla u(t)\|^2_{L^2(\Omega)^n} \geq \frac{1}{C_p}\varphi(t),$$

qui conduit à l'estimation :

$$\varphi'(t) + \alpha\varphi(t) \leq 0 \quad p.p\ t \in]0,T[\quad \text{avec } \alpha = \frac{2}{C_p}.$$

Posons $h(t) = e^{\alpha t}\varphi(t) \in H^1(]0,T[)$. De l'inégalité précédente, on déduit que :

$$e^{-\alpha t}h'(t) \leq 0 \quad p.p\ t \in]0,T[,$$

qui conduit à l'inégalité :

$$\int_0^s h'(t)dt \leq 0 \quad \forall s \in [0,T].$$

Comme $h \in H^1(]0,T[)$ on en déduit que :

$$h(s) \leq h(0) \quad \forall s \in [0,T],$$

ce qui achève la démonstration. ∎

3.2.2 Dépendance continue des solutions

Nous allons montrer dans cette sous-section comment l'estimation d'énergie (3.16) permet de démontrer que la solution de l'équation de la chaleur dépend continûment des données f et u_0. Ce résultat s'appuie sur les inégalités suivantes dites de Gronwall.

Lemme 3.13 (Inégalités de Gronwall) *Soient $\varphi(t)$ une fonction continue positive définie sur $[0,T]$, $m(t)$ une fonction positive de $L^1(]0,T[)$, $\gamma \in]0,1[$, et $C \geq 0$. Alors si φ vérifie :*

$$\varphi(t) \leq C + \int_0^t m(s)\varphi(s)^\gamma ds \qquad \forall t \in [0,T],$$

on a :

$$\varphi(t) \leq \left(C^{1-\gamma} + (1-\gamma) \int_0^t m(s) ds \right)^{\frac{1}{1-\gamma}} \qquad \forall t \in [0,T]. \tag{3.17}$$

Dans le cas où $\gamma = 1$, on a :

$$\varphi(t) \leq C\, e^{\int_0^t m(s)ds} \qquad \forall t \in [0,T]. \tag{3.18}$$

Démonstration : Posons $G(t) = C + \displaystyle\int_0^t m(s)\varphi(s)^\gamma ds$; on a $G'(t) = m(t)\varphi(t)^\gamma$. D'après l'hypothèse, on sait que : $0 \leq \varphi(t) \leq G(t)$, soit $0 \leq \varphi(t)^\gamma \leq G(t)^\gamma$, d'où on tire que :

$$G(t)^{-\gamma} G'(t) \leq m(t).$$

En intégrant entre 0 et t, il vient :

$$\frac{1}{1-\gamma}(G(t)^{1-\gamma} - G(0)^{1-\gamma}) \leq \int_0^t m(s)ds,$$

mais comme $1-\gamma > 0$ et $G(0) = C$, on en déduit que :

$$G(t)^{1-\gamma} \leq C^{1-\gamma} + (1-\gamma) \int_0^t m(s)ds,$$

qui donne le résultat du lemme, compte-tenu du fait que la fonction $x \mapsto x^{1/1-\gamma}$ est croissante lorsque $1-\gamma > 0$. L'autre résultat s'obtient de façon similaire. ∎

Proposition 3.14 (continuité des solutions) *La solution $u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$ du problème (3.2) dépend continûment des données $u_0 \in L^2(\Omega)$ et $f \in L^2(0,T;L^2(\Omega))$. Plus précisément on a les estimations suivantes :*

$$\begin{aligned} \|u\|_{\mathcal{C}^0(0,T;L^2(\Omega))} &\leq \|u_0\|_{L^2(\Omega)} + \sqrt{T}\|f\|_{L^2(0,T;L^2(\Omega))}, \\ \|u\|_{L^2(0,T;H_0^1(\Omega))} &\leq \|u_0\|_{L^2(\Omega)} + \sqrt{T}\|f\|_{L^2(0,T;L^2(\Omega))}. \end{aligned}$$

Démonstration : Dans cette démonstration, on note $\|\cdot\|$ la norme de $L^2(\Omega)$ ou de $L^2(\Omega)^n$. D'après l'égalité d'énergie (3.16), on a en vertu de l'inégalité de Cauchy-Schwarz, $\forall t \in [0,T]$:

$$\frac{1}{2}\|u(t)\|^2 + \int_0^t \|\nabla u(s)\|^2 ds \leq \frac{1}{2}\|u_0\|^2 + \int_0^t \|f(s)\|\,\|u(s)\| ds, \tag{3.19}$$

d'où on déduit, $\forall t \in [0,T]$:

$$\|u(t)\|^2 \leq \|u_0\|^2 + 2\int_0^t \|f(s)\|\,\|u(s)\| ds.$$

En appliquant le lemme 3.13 avec :

$$m(t) = 2\|f(t)\|,\ C = \|u_0\|^2,\ \varphi(t) = \|u(t)\|^2 \text{ et } \gamma = \frac{1}{2},$$

on obtient, $\forall t \in [0,T]$:

$$\|u(t)\|^2 \leq \left(\|u_0\| + \int_0^t \|f(s)\| ds\right)^2,$$

qui montre (car $t < T$) que :

$$\|u(t)\| \leq \|u_0\| + \int_0^T \|f(s)\| ds, \qquad \forall t \in [0,T]. \tag{3.20}$$

Par ailleurs, de l'inégalité (3.19) il vient :

$$\int_0^T \|\nabla u(s)\|^2 ds \leq \frac{1}{2}\|u_0\|^2 + \sup_{s\in[0,T]} \|u(s)\| \int_0^T \|f(s)\| ds,$$

qui compte-tenu de l'inégalité (3.20) conduit à :

$$\left(\int_0^T \|\nabla u(s)\|^2 ds\right)^{1/2} \leq \left(\frac{1}{2}\|u_0\|^2 + \|u_0\| \int_0^T \|f(s)\| ds + \left(\int_0^T \|f(s)\| ds\right)^2\right)^{1/2},$$

soit finalement :

$$\left(\int_0^T \|\nabla u(s)\|^2 ds\right)^{1/2} \leq \|u_0\| + \int_0^T \|f(s)\| ds. \tag{3.21}$$

Les estimations (3.20) et (3.21) conduisent aux estimations souhaitées car :

$$\int_0^T \|f(s)\| ds \leq \sqrt{T}\left(\int_0^T \|f(s)\|^2 ds\right)^{1/2} \qquad (\text{d'après Cauchy} - \text{Schwarz}).$$

Ce qui démontre la proposition. ∎

Remarque 3.15 *Si l'on suppose plus généralement que $f \in L^1(0,T;H^1(\Omega))$, alors d'après (3.20) et (3.21), on a :*

$$\begin{aligned} \|u\|_{\mathcal{C}^0(0,T;L^2(\Omega))} &\leq \|u_0\|_{L^2(\Omega)} + \|f\|_{L^1(0,T;L^2(\Omega))}, \\ \|u\|_{L^2(0,T;H_0^1(\Omega))} &\leq \|u_0\|_{L^2(\Omega)} + \|f\|_{L^1(0,T;L^2(\Omega))}, \end{aligned}$$

c'est-à-dire une dépendance continue par rapport à la norme L^1 de f.

Remarque 3.16 *Si l'on suppose que $u_0 \in H^1(\Omega)$, alors on peut établir la* seconde égalité d'énergie *:*

$$\frac{1}{2}\|\nabla u(t)\|^2_{L^2(\Omega)^n} + \int_0^t \|\frac{\partial}{\partial t}u(s)\|^2_{L^2(\Omega)} ds = \frac{1}{2}\|\nabla u_0\|^2_{L^2(\Omega)^n} + \int_0^t (\frac{\partial}{\partial t}u(s), f(s))_{L^2(\Omega)} ds$$

(en multipliant par la fonction test $v = \dfrac{\partial u}{\partial t}(t)$ la première équation du problème (3.1)). A l'aide de cette égalité, on peut alors montrer, par des procédés similaires aux précédents, que la solution est plus régulière :

$$u \in \mathcal{C}^0\left(0, T; H_0^1(\Omega)\right) \quad \text{et} \quad \frac{\partial u}{\partial t} \in L^2\left(0, T; L^2(\Omega)\right),$$

avec dépendance continue vis-à-vis des données f et u_0.

3.2.3 Principe du maximum

De même que pour le problème de Dirichlet, l'équation de la chaleur vérifie un principe de positivité et un principe du maximum.

Proposition 3.17 (principe de positivité) *Soient $f \in L^2(0,T;L^2(\Omega))$ et $u_0 \in L^2(\Omega)$ tels que $u_0 \geq 0$ sur Ω et $f \geq 0$ sur Q_T. Alors la solution u du problème (3.2) vérifie $u \geq 0$ sur Q_T.*

Démonstration : On utilise la même technique que dans la démonstration du principe de positivité pour l'équation de Laplace [15]. Notons $u^+ = \sup(u, 0)$, $u^- = \inf(u, 0)$,

$$Q^+ = \{(x,t) \in Q_T \text{ tels que } u(x,t) \geq 0\} = \text{Supp } u^+$$
$$Q^- = \{(x,t) \in Q_T \text{ tels que } u(x,t) \leq 0\} = \text{Supp } u^-$$

et enfin χ^+ $\left(\text{resp. } \chi^-\right)$ la fonction caractéristique de Q^+ $\left(\text{resp. } Q^-\right)$.
Comme $u \in L^2(0,T;H_0^1(\Omega))$, $u^+(t)$ et $u^-(t)$ appartiennent à $H_0^1(\Omega)$ pour presque tout t et

$$\frac{\partial}{\partial t}u^\mp = \chi^\mp \frac{\partial u}{\partial t}, \quad \nabla u^\mp = \chi^\mp \nabla u.$$

On peut donc prendre $v = u^-$ dans (3.4), pour trouver :

$$\int_\Omega \frac{\partial u}{\partial t}(t) u^-(t)\, d\Omega + \int_\Omega \nabla u(t) \cdot \nabla u^-(t)\, d\Omega = \int_\Omega f(t) u^-(t)\, d\Omega.$$

Or on a :

$$\int_\Omega \frac{\partial u}{\partial t}(t) u^-(t)\, d\Omega = \int_\Omega \frac{\partial u^-}{\partial t}(t) u^-(t)\, d\Omega = \frac{1}{2}\frac{d}{dt}\left(\int_\Omega (u^-(t))^2\, d\Omega\right),$$

où la dernière égalité est à prendre au sens des distributions sur $]0,T[$.
Par ailleurs :

$$\int_\Omega \nabla u(t) \cdot \nabla u^-(t)\, d\Omega = \int_\Omega |\nabla u^-(t)|^2\, d\Omega.$$

D'où on déduit que

$$\frac{1}{2}\frac{d}{dt}\left(\int_\Omega \left(u^-(t)\right)^2\, d\Omega\right) = -\int_\Omega |\nabla u^-(t)|^2\, d\Omega + \int_\Omega f(t)u^-(t)\, d\Omega,$$

qui compte-tenu de l'hypothèse $f \geq 0$, prouve que :

$$\int_\Omega \left(u^-(t)\right)^2\, d\Omega \leq \int_\Omega \left(u_0^-\right)^2\, d\Omega.$$

Ceci implique que $u^-(t) \equiv 0,\ \forall t \in]0,T[$ car $u_0^- \equiv 0$ ($u_0 \geq 0$ par hypothèse) et par conséquent $u \geq 0$ sur Q_T. ∎

On a également le principe du maximum suivant :

Proposition 3.18 (principe du maximum) *Soit u la solution de (3.2) avec $f \equiv 0$. On a alors :*

$$\min\left(0, \inf_\Omega u_0\right) \leq u(x,t) \leq \max\left(0, \sup_\Omega u_0\right) \qquad p.p\ (x,t) \in Q_T.$$

Démonstration : On pose $K = \mathrm{Max}\,(0, \sup_\Omega u_0)$ et $w = K - u$. Notons que, pour presque tout t, on a $w(t) \in H^1(\Omega)$, mais que par contre $w(t) \notin H_0^1(\Omega)$. Ceci étant, $w^-(t) \in H_0^1(\Omega)$ car :

$$w^-{}_{|\partial\Omega}(t) = 0 \quad \text{ sachant que } K \geq 0 \text{ et } u \in H_0^1(\Omega).$$

En choisissant $v = w^-$ dans (3.4) et compte-tenu du fait que :

$$\frac{\partial}{\partial t}K = 0\ ,\ \nabla K = 0\ ,\ \text{et } f \equiv 0,$$

on obtient :

$$\int_\Omega \frac{\partial w}{\partial t}(t)\, w^-(t)\, d\Omega + \int_\Omega \nabla w(t) \cdot \nabla w^-(t)\, d\Omega = 0,$$

soit encore :

$$\frac{1}{2}\,\frac{d}{dt}\left(\int_\Omega (w^-(t))^2\, d\Omega\right) + \int_\Omega |\nabla w^-(t)|^2\, d\Omega = 0.$$

En intégrant sur l'intervalle de temps $[0,s]$ et en remarquant que $w^-(0) \equiv 0$ on déduit que $w^-(s) \equiv 0\ \forall s$. On a donc prouvé que $u \leq K$ sur Q_T. L'autre inégalité se démontre de façon similaire. ∎

Remarque 3.19 *Ce résultat a pour conséquence la* stabilité L^∞ *de l'équation de la chaleur lorsque $f \equiv 0$: si $u_0 \in L^\infty(\Omega)$, alors $\forall t \in [0,T]$, $u(t) \in L^\infty(\Omega)$ et*

$$\|u(t)\|_{L^\infty(\Omega)} \leq \|u_0\|_{L^\infty(\Omega)}\ .$$

3.2.4 Caractère régularisant

Terminons cette étude par la présentation de quelques propriétés très particulières de l'équation de la chaleur : effet régularisant, non réversibilité et propagation à vitesse infinie. Nous renvoyons à [9] pour la démonstration de ces résultats.

Proposition 3.20 *Soit u la solution du problème (3.2) avec $f \equiv 0$. Alors, $\forall \varepsilon > 0$ on a $u \in \mathcal{C}^\infty(\overline{\Omega} \times [\varepsilon, T])$.*

Cette propriété est liée à l'*effet régularisant* du noyau de l'équation de la chaleur. En effet, la solution élémentaire dans $\mathbb{R}^n$ de l'équation de la chaleur est donnée par :

$$E(x,t) = \frac{H(t)}{\left(\sqrt{4\pi t}\right)^n} e^{-\frac{|x|^2}{4t}},$$

où H est la fonction de Heavyside définie par

$$H(t) = \begin{cases} 1 \text{ si } t > 0, \\ 0 \text{ si } t < 0. \end{cases}.$$

Comme E est localement intégrable et $\mathcal{C}^\infty$ sauf au point 0, la solution u de l'équation de la chaleur homogène dans $\mathbb{R}^n$ est donnée par :

$$u(x,t) = E * (u_0 \otimes \delta) = \frac{1}{\left(\sqrt{4\pi t}\right)^n} \int_{\mathbb{R}^n} e^{-\frac{|x-y|^2}{4t}} u_0(y)\, dy$$

et est $\mathcal{C}^\infty$ en dehors du support de la distribution $u_0 \otimes \delta$, i.e. $\mathbb{R}^n \times \{0\}$ (voir [35]).

La proposition 3.20 exprime, par exemple, le fait que même si u_0 est discontinue la solution u est $\mathcal{C}^\infty$ dès que $t >0$.

Cet effet régularisant a une conséquence inattendue, à savoir que l'équation de la chaleur est *non réversible*, c'est-à-dire qu'il est impossible de retrouver la condition initiale u_0 à partir de la connaissance de $u(t)$ sur Ω à un instant donné $t > 0$.

Par ailleurs, cet effet régularisant est lié à la *propagation à vitesse infinie* de la chaleur. C'est-à-dire que :

$$\text{si } u_0 \neq 0,\ u_0 \geq 0 \text{ p.p et } f \equiv 0 \text{ alors } u(x,t) > 0\ \forall x \in \Omega,\ \forall t > 0.$$

En particulier, même si u_0 est à support compact dans Ω, $u(t)$, $t > 0$ est à support non-compact dans Ω !

Les propriétés d'effet régularisant, de non-réversibilité et de propagation à vitesse infinie sont des propriétés très spécifiques à l'équation de la chaleur. Evidemment, l'équation des ondes présente des comportements tout-à-fait différents, comme on le verra au chapitre suivant. Ces différences de comportement réapparaissent lors de la discrétisation en temps et conduisent à des schémas numériques de nature différente.

3.3 Discrétisation

3.3.1 Semi-discrétisation en espace

La méthode des éléments finis étant bien adaptée à la discrétisation des opérateurs elliptiques, il est naturel de l'utiliser pour discrétiser spatialement l'équation de la chaleur. Par ailleurs, on dispose d'une formulation variationnelle en espace de cette équation (cf. l'équation (3.2)).

Considérons donc $\mathcal{T}_h$ une triangulation admissible [1] du domaine Ω, composée d'éléments finis de Lagrange d'ordre k : $(K_\ell, \Sigma_\ell, P_\ell)_{\ell=1,L}$.

On note :

– $(M_I)_{I=1,N}$ les nœuds du maillage.
– $(w_I)_{I=1,N}$ les fonctions de base globales, respectivement attachées aux nœuds M_I, qui vérifient :

$$w_I = 0 \qquad \text{sur } \partial\Omega.$$

– $V_h = \underset{I=1,N}{\text{Vect}}\,(w_I)$ l'espace vectoriel engendré par les fonctions de base globales.

Enfin, on suppose que, outre la propriété d'approximabilité (1.54) toutes les hypothèses sur le maillage sont satisfaites de telle sorte que les propriétés suivantes soient vérifiées :

$$V_h \subset \mathcal{C}^0\left(\overline{\Omega}\right) \qquad \text{et} \qquad V_h \subset H_0^1(\Omega), \tag{3.22}$$

$$\|v - \Pi_h v\|_{H^1(\Omega)} \leq C h^k \mid v \mid_{k+1,\Omega} \qquad \forall v \in H^{k+1}(\Omega), \tag{3.23}$$

où Π_h est l'opérateur d'interpolation habituel (cf. [15]).

Problème variationnel approché

L'approximation dans l'espace V_h, du problème (3.1) conduit à la formulation variationnelle semi-discrète suivante :

$$\begin{cases} \text{trouver } u_h \in \mathcal{C}^0(0,T;V_h) \text{ tel que :} \\ \dfrac{d}{dt}\displaystyle\int_\Omega u_h(t)\, v_h \, d\Omega + \int_\Omega \nabla u_h(t) \cdot \nabla v_h \, d\Omega = \int_\Omega f(t)\, v_h \, d\Omega \qquad \begin{matrix} \forall v_h \in V_h \\ \forall t \in]0,T[\end{matrix} \\ u_h(0) = u_{h,0} \end{cases} \tag{3.24}$$

où $u_{h,0}$ désigne une approximation dans V_h de la fonction u_0 (par exemple $u_{h,0} = \Pi_h u_0$ si u_0 est continue sur $\overline{\Omega}$). En tout état de cause, on suppose que

$$\lim_{h\to 0} \|u_{h,0} - u_0\|_{L^2(\Omega)} = 0. \tag{3.25}$$

1. On supposera, par souci de simplification, que Ω est un ouvert polyédrique borné de $\mathbb{R}^n$.

Interprétation matricielle

Le problème discrétisé (3.24) est un système différentiel d'ordre N. En effet, notons $\overrightarrow{U}(t)$ le vecteur de composantes $U_I(t) = u_h(M_I, t),\ I = 1, N,\ t \in [0, T[$. Par construction :

$$u_h(t) = \sum_{J=1,N} u_h(M_J, t) w_J = \sum_{J=1,N} U_J(t)\, w_J. \tag{3.26}$$

En substituant cette expression dans (3.24) et en prenant $v_h = w_I$ on obtient :

$$\frac{d}{dt} \sum_{J=1,N} \left(\int_\Omega w_J\, w_I\, d\Omega \right) U_J(t) + \sum_{J=1,N} \left(\int_\Omega \nabla w_J \cdot \nabla w_I\, d\Omega \right) U_J(t) = \int_\Omega f(t) w_I\, d\Omega.$$

En notant :

– $\mathbb{K}$ la matrice symétrique de $\mathbb{R}^{N\times N}$ définie par $\mathbb{K}_{IJ} = \displaystyle\int_\Omega \nabla w_I \cdot \nabla w_J\, d\Omega$,

– $\mathbb{M}$ la matrice symétrique de $\mathbb{R}^{N\times N}$ définie par $\mathbb{M}_{IJ} = \displaystyle\int_\Omega w_I\, w_J\, d\Omega$,

– $\overrightarrow{F}(t)$ le vecteur de $\mathbb{R}^N$ défini par $F_I(t) = \displaystyle\int_\Omega f(t)\, w_I\, d\Omega$,

la formulation (3.24) est équivalente au système différentiel :

$$\begin{cases} \text{trouver } \overrightarrow{U} \in \mathcal{C}^0(0, T; \mathbb{R}^N) \text{ tel que} \\ \dfrac{d}{dt}\mathbb{M}\overrightarrow{U}(t) + \mathbb{K}\overrightarrow{U}(t) = \overrightarrow{F}(t) \quad \forall t \in]0, T[, \\ \overrightarrow{U}(0) = \overrightarrow{U}_0. \end{cases} \tag{3.27}$$

où $\overrightarrow{U}_0$ est le vecteur de composantes $u_{h,0}(M_I),\ I = 1, N$.

Existence d'une solution au problème semi-discrétisé

En vertu du lemme 1.23 et du théorème 1.25, le problème aux valeurs propres :

$$\mathbb{K}\overrightarrow{V} = \lambda \mathbb{M}\overrightarrow{V} \tag{3.28}$$

admet N valeurs propres $(\lambda_{m,h})_{m=1,N}$ avec $\lambda_{m,h} > 0$ et une base de vecteurs propres $(\overrightarrow{V}_m)_{m=1,N}$ orthonormale pour le produit scalaire[2] $(\mathbb{M}\cdot|\cdot)$ sur $\mathbb{R}^N$. Qui plus est, si on définit des fonctions $(v_{m,h})_{m=1,N}$ de V_h par :

$$v_{m,h} = \sum_{I=1,N} V_m^I w_I, \tag{3.29}$$

2. Rappel : on note $(\cdot|\cdot)$ le produit scalaire usuel de $\mathbb{R}^N$.

où $V_m^I = v_{m,h}(M_I)$ est la $I^{ème}$ composante du vecteur $\overrightarrow{V}_m$, alors celles-ci forment une base de V_h, orthonormale dans $L^2(\Omega)$, et elles sont aussi les fonctions propres discrètes du Laplacien avec condition aux limites de Dirichlet homogène (solutions dans $H_0^1(\Omega)$).

On peut donc décomposer $\overrightarrow{U}(t)$ sur la base de vecteurs propres : on écrit

$$\overrightarrow{U}(t) = \sum_{m=1,N} \alpha_{m,h}(t)\overrightarrow{V}_m, \text{ avec } \alpha_{m,h}(t) = (\mathbb{M}\overrightarrow{U}(t)|\overrightarrow{V}_m). \tag{3.30}$$

Revenons au système différentiel (3.27) et effectuons le produit scalaire usuel avec $\overrightarrow{V}_\ell,\ \forall \ell = 1, N$:

$$\begin{cases} \dfrac{d}{dt}\displaystyle\sum_{m=1,N} \alpha_{m,h}(t)(\mathbb{M}\overrightarrow{V}_m|\overrightarrow{V}_\ell) + \sum_{m=1,N}^{N} \alpha_{m,h}(t)(\mathbb{K}\overrightarrow{V}_m|\overrightarrow{V}_\ell) = (\overrightarrow{F}(t)|\overrightarrow{V}_\ell), \\ \alpha_{m,h}(0) = (\mathbb{M}\overrightarrow{U}_0|\overrightarrow{V}_m) \qquad \forall m = 1, N, \end{cases}$$

qui compte-tenu de (3.28) et des propriétés d'orthogonalité de la base de vecteurs propres conduit aux équations différentielles, $\forall m = 1, N$:

$$\begin{cases} \dfrac{d}{dt}\alpha_{m,h}(t) + \lambda_{m,h}\alpha_{m,h}(t) = (\overrightarrow{F}(t)|\overrightarrow{V}_m), \\ \\ \alpha_{m,h}(0) = (\mathbb{M}\overrightarrow{U}_0|\overrightarrow{V}_m). \end{cases} \tag{3.31}$$

Ces équations admettent pour solution

$$\alpha_{m,h}(t) = \alpha_{m,h}(0)e^{-\lambda_{m,h}t} + \int_0^t (\overrightarrow{F}(s)|\overrightarrow{V}_m)e^{-\lambda_{m,h}(t-s)}ds, \quad m = 1, N. \tag{3.32}$$

On en conclut finalement :

Proposition 3.21 *Le problème semi-discrétisé (3.24) admet une unique solution $u_h \in \mathcal{C}^0(0, T; V_h)$ de la forme :*

$$u_h(t) = \sum_{m=1,N} \alpha_{m,h}(t)v_{m,h} \tag{3.33}$$

où $\alpha_{m,h}(t)$ est donnée par (3.32) et $v_{m,h}$ par (3.29).

Démonstration : A l'aide de (3.26), puis (3.30) considérée composante par composante, et enfin (3.29), on écrit la suite d'égalités :

$$\begin{aligned} u_h(t) &\overset{(3.26)}{=} \sum_{I=1,N} U_I(t)\, w_I \\ &\overset{(3.30)}{=} \sum_{I=1,N} \left(\sum_{m=1,N} \alpha_{m,h}(t)V_m \right)^I w_I = \sum_{I=1,N} \sum_{m=1,N} \alpha_{m,h}(t)V_m^I\, w_I \\ &= \sum_{m=1,N} \alpha_{m,h}(t) \sum_{I=1,N} V_m^I\, w_I \overset{(3.29)}{=} \sum_{m=1,N} \alpha_{m,h}(t)v_{m,h}. \end{aligned}$$

■

Discrétisation par différences finies

Mentionnons rapidement dans cette sous-section à quel type de système conduit une discrétisation du laplacien par différences finies (voir [13, 33]).

Notons cette fois $(M_I)_{I=1,M}$ les points de la grille d'approximation et Δ_h l'opérateur aux différences finies approchant Δ. On approche alors l'équation de la chaleur (3.1) par le problème :

$$\begin{cases} \dfrac{d}{dt}\underline{u}_h^I(t) - (\Delta_h \underline{u}_h)^I(t) = f(M_I,t) & M_I \notin \partial\Omega \\ \underline{u}_h^I(t) = 0 & M_I \in \partial\Omega \\ \underline{u}_h^I(0) = u_0(M_I) & M_I \in \overline{\Omega} \end{cases} \tag{3.34}$$

$\underline{u}_h^I(t)$ désignant l'approximation au temps t et au point M_I de la solution u.

Notons $\overrightarrow{\underline{U}}(t)$ le vecteur de composantes $\underline{U}_I(t) = \underline{u}_h^I(t)$, $I = 1, N$ où N désigne le nombre de points qui ne se trouvent pas sur la frontière (on a supposé que les points M_I situés sur la frontière $\partial\Omega$ sont numérotés de $N+1$ à M).

Après élimination des conditions de Dirichlet, les équations (3.34) conduisent au système différentiel d'ordre N suivant :

$$\begin{cases} \text{trouver } \overrightarrow{\underline{U}} \in \mathcal{C}^0(0,T;\mathbb{R}^N) \text{ tel que :} \\ \dfrac{d}{dt}\overrightarrow{\underline{U}}(t) + \mathbb{D}\overrightarrow{\underline{U}}(t) = \overrightarrow{\underline{F}}(t) \qquad \forall t \in]0,T[\\ \overrightarrow{\underline{U}}(0) = \overrightarrow{\underline{U}}_0 \end{cases} \tag{3.35}$$

où $\mathbb{D}$ désigne la matrice de différences finies du Laplacien après prise en compte de la condition de Dirichlet, $\overrightarrow{\underline{F}}(t)$ le vecteur de composantes $\underline{F}^I = F(M_I,t)$ et $\overrightarrow{\underline{U}}_0$ le vecteur de composantes $u_0(M_I)$, pour $I = 1, N$.

Remarque 3.22 *Comme on considère les valeurs ponctuelles $u_0(M_I)$ et $f(M_I,t)$, il faut supposer que $u_0 \in \mathcal{C}^0\left(\overline{\Omega}\right)$ et $f \in \mathcal{C}^0\left(\overline{\Omega} \times [0,T]\right)$.*

Dans [13, 33], il est établi que l'approximation par différences finies du problème de Dirichlet conduit à une matrice $\mathbb{D}$ symétrique définie-positive. Par conséquent, on démontre, de façon analogue à la proposition 3.21, qu'il existe une unique solution $\overrightarrow{\underline{U}} \in \mathcal{C}^0\left(0,T;\mathbb{R}^N\right)$ au système différentiel (3.35).

Le système différentiel (3.35) a une structure différente du système différentiel (3.27), issu de la discrétisation par éléments finis : en effet, on a le terme $\frac{d}{dt}\overrightarrow{\underline{U}}(t)$ au lieu du terme $\frac{d}{dt}\mathbb{M}\overrightarrow{U}(t)$. Classiquement les différences finies reposent sur des estimations ponctuelles (d'où l'absence de matrice de masse) alors que les éléments finis reposent sur des estimations en normes liées à des espaces de Sobolev (d'où la matrice de masse $\mathbb{M}$). Toutefois, si on utilise la méthode de "lumping" (ou

condensation de masse, voir §4.5.1) de la matrice de masse, la matrice $\mathbb{M}$ devient diagonale et le système (3.27) retrouve une structure équivalente au système (3.35), à un facteur en h^n près, où n désigne ici la dimension de l'espace, $\Omega \subset \mathbb{R}^n$.

3.3.2 Discrétisation totale

Que ce soit par éléments finis ou différences finies, la semi-discrétisation en espace conduit à des systèmes différentiels du premier ordre que nous allons discrétiser maintenant par des techniques classiques de différences finies.

Afin de résoudre numériquement le système différentiel (3.27) ou (3.35), nous utilisons des schémas aux différences finies pour approcher le terme dérivé $\partial_t u$.

Par la suite, nous mènerons l'analyse de ces schémas sur le cas du système (3.27) (plus général que (3.35)) tout en gardant à l'esprit que la matrice $\mathbb{M}$ peut toujours être choisie de type diagonal dans le cas de l'approximation par éléments finis.

Dans toute la suite, $(t_k)_{k=0,K}$ désigne une suite d'instants croissant de $t_0 = 0$ à $t_K = T$. Afin de simplifier la présentation, on suppose que $t_k = k\Delta t$ et on note $\overrightarrow{U}^k \in \mathbb{R}^N$ le vecteur approché à l'instant t_k du vecteur $\overrightarrow{U}(t_k)$. En d'autres termes, la $I^{ème}$ composante U_I^k du vecteur $\overrightarrow{U}^k \in \mathbb{R}^N$ représente une approximation de la solution u du problème (3.1) à l'instant t_k et au nœud M_I :

$$U_I^k \simeq U_I(t_k) \simeq u(M_I, t_k).$$

Dans cette sous-section, nous allons présenter deux types de schémas, qui illustrent bien les diverses propriétés et difficultés que l'on peut rencontrer lors de l'étape de discrétisation en temps.

Schémas à un pas de temps décentrés vers l'avant

Afin d'obtenir des schémas explicites en temps, l'idée la plus naturelle pour approcher la dérivation en temps consiste à utiliser la différence finie :

$$D_{\Delta t}^+ \overrightarrow{U}^k = \frac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^k}{\Delta t} \approx \frac{d\overrightarrow{U}}{dt}(t_k).$$

Ceci conduit à introduire le schéma *explicite* suivant pour approcher (3.27) :

$$\begin{cases} \dfrac{\mathbb{M}(\overrightarrow{U}^{k+1} - \overrightarrow{U}^k)}{\Delta t} + \mathbb{K}\overrightarrow{U}^k = \overrightarrow{F}(t_k) \qquad k = 0, K-1, \\ \overrightarrow{U}^0 = \overrightarrow{U}_0. \end{cases} \tag{3.36}$$

Bien qu'il soit *a priori* nécessaire d'inverser la matrice $\mathbb{M}$, ce schéma est qualifié d'explicite car cette "inversion" (dans la pratique une factorisation de Cholesky $\mathbb{L}\mathbb{L}^t$ sachant que $\mathbb{M}$ est symétrique définie-positive) une fois effectuée au début du processus, conduit au schéma réellement explicite :

$$\overrightarrow{U}^{k+1} = \overrightarrow{U}^k + \Delta t \mathbb{M}^{-1} \left(\overrightarrow{F}(t_k) - \mathbb{K}\overrightarrow{U}^k \right).$$

Ci-dessus, on ne calcule pas $\mathbb{M}^{-1}$, mais on résout le système linéaire

$$\mathbb{M}\overrightarrow{X}_k = \overrightarrow{F}(t_k) - \mathbb{K}\overrightarrow{U}^k,$$

par un algorithme de descente-remontée à l'aide de la factorisation de Cholesky. Il reste à effectuer la mise à jour $\overrightarrow{U}^{k+1} = \overrightarrow{U}^k + \Delta t \overrightarrow{X}_k$. Si la matrice $\mathbb{M}$ est diagonale, l'inversion est fictive, ce qui est notamment le cas lorsque l'on utilise la méthode de "lumping".

Evidemment, on peut rendre *implicite* le schéma (3.36) en se plaçant à l'instant t_{k+1} :

$$\begin{cases} \dfrac{\mathbb{M}(\overrightarrow{U}^{k+1} - \overrightarrow{U}^k)}{\Delta t} + \mathbb{K}\overrightarrow{U}^{k+1} = \overrightarrow{F}(t_{k+1}) \qquad k = 0, K-1, \\ \overrightarrow{U}^0 = \overrightarrow{U}_0, \end{cases} \tag{3.37}$$

qui s'ecrit également :

$$(\mathbb{M} + \Delta t \mathbb{K})\overrightarrow{U}^{k+1} = \Delta t \overrightarrow{F}(t_{k+1}) + \mathbb{M}\overrightarrow{U}^k.$$

Ici, au début du processus, il faut "inverser" la matrice $(\mathbb{M} + \Delta t \mathbb{K})$, ce qui reste faisable à l'aide d'une factorisation de Cholesky, puisque $\mathbb{K}$ est symétrique et positive.

Les schémas (3.36) et (3.37) peuvent être "moyennés" donnant ainsi naissance à la classe de schémas, $\theta \in [0, 1]$:

$$\begin{cases} \dfrac{\mathbb{M}(\overrightarrow{U}^{k+1} - \overrightarrow{U}^k)}{\Delta t} + \theta \mathbb{K}\overrightarrow{U}^{k+1} + (1-\theta)\mathbb{K}\overrightarrow{U}^k \\ \qquad\qquad = \theta \overrightarrow{F}(t_{k+1}) + (1-\theta)\overrightarrow{F}(t_k), \quad k = 0, K-1, \\ \overrightarrow{U}^0 = \overrightarrow{U}_0, \end{cases} \tag{3.38}$$

que l'on appelle la classe des *θ-schémas*.

Le schéma (3.36) correspond à $\theta = 0$ et le schéma (3.37) à $\theta = 1$.

Dans la pratique, on utilise la forme :

$$(\mathbb{M}+\theta\Delta t\mathbb{K})\overrightarrow{U}^{k+1} = (\mathbb{M}-\Delta t(1-\theta)\mathbb{K})\overrightarrow{U}^{k} + \Delta t\theta\overrightarrow{F}(t_{k+1}) + \Delta t(1-\theta)\overrightarrow{F}(t_k). \quad (3.39)$$

On rencontre les dénominations usuelles suivantes :
– $\theta = 0$: schéma d'*Euler* explicite,
– $\theta = 1$: schéma d'*Euler* implicite ou Euler rétrograde,
– $\theta = 1/2$: schéma de *Crank-Nikolson*.

Comme précédemment, il faut "inverser" la matrice $(\mathbb{M} + \theta\Delta t\mathbb{K})$ à l'aide d'une factorisation de Cholesky. Les schémas considérés jusqu'à présent sont des *schémas à deux niveaux de temps* (indices ^{k} et $^{k+1}$).

Un schéma à deux pas de temps centré

L'utilisation d'une approximation par différences finies centrées :

$$\frac{d\overrightarrow{U}^{k}}{dt} \simeq \frac{\overrightarrow{U}(t_{k+1}) - \overrightarrow{U}(t_{k-1})}{2\Delta t}$$

conduit à un schéma à trois niveaux de temps, dit schéma de *Richardson* :

$$\begin{cases} \dfrac{\mathbb{M}(\overrightarrow{U}^{k+1} - \overrightarrow{U}^{k-1})}{2\Delta t} + \mathbb{K}\overrightarrow{U}^{k} = \overrightarrow{F}(t_k) \qquad k = 1, K-1, \\ \overrightarrow{U}^{0} = \overrightarrow{U}_0, \\ \overrightarrow{U}^{1} = \overrightarrow{U}^{0} + \Delta t\mathbb{M}^{-1}(\overrightarrow{F}(0) - \mathbb{K}\overrightarrow{U}^{0},) \end{cases} \quad (3.40)$$

qui est explicite.

Notons, puisqu'on utilise un *schéma à trois niveaux de temps*, qu'il est nécessaire de connaître $\overrightarrow{U}^{1}$. Ici, on utilise le schéma d'Euler explicite (3.36) pour définir $\overrightarrow{U}^{1}$.

Les θ-schémas et celui de Richardson sont les plus simples que l'on puisse imaginer. Il en existe de plus sophistiqués : méthodes de Runge-Kutta, méthodes multipas, prédicteur-correcteur (voir [22, 48] par exemple), mais ceux que nous avons introduits sont largement suffisants à notre propos.

3.4 Convergence temporelle du schéma

Dans l'analyse d'erreur qui va suivre, nous nous intéressons à la convergence en temps des schémas. En d'autres termes, nous considérons que la discrétisation en espace est fixée et nous nous demandons dans quelle mesure :

$$\overrightarrow{U}^{k} \underset{\Delta t\to 0}{\longrightarrow} \overrightarrow{U}(t_k) \qquad h \text{ fixé.}$$

Nous traiterons les propriétés de convergence en espace au §3.5.

3.4.1 Consistance

Rappelons que $\overrightarrow{U}^k$ est un vecteur de $\mathbb{R}^N$ et que la matrice de masse $\mathbb{M}$ est symétrique définie-positive (dans le cas des différences finies $\mathbb{M} = \mathbb{I}$). On munit donc $\mathbb{R}^N$ du produit scalaire $(\mathbb{M}\cdot|\cdot)$ et de la norme $|\cdot|_{\mathbb{M}} = (\mathbb{M}\cdot|\cdot)^{1/2}$. Ce choix correspond au fait que, si $\overrightarrow{U}$ désigne la solution du système différentiel (3.27) :

$$(\mathbb{M}\overrightarrow{U}(t)|\overrightarrow{U}(t)) = \int_\Omega u_h(t)^2\, d\Omega = \|u_h(t)\|^2_{L^2(\Omega)}. \tag{3.41}$$

Regroupons sous forme abstraite les schémas (3.38) et (3.40) :

$$S(\overrightarrow{U}^{k-1}, \overrightarrow{U}^k, \overrightarrow{U}^{k+1}) = 0, \qquad \text{avec } S(\overrightarrow{U}, \overrightarrow{V}, \overrightarrow{W}) = \frac{\overrightarrow{W}}{\Delta t} - H(\overrightarrow{U}, \overrightarrow{V}, \Delta t). \tag{3.42}$$

Remarque 3.23 *Dans le cas des schémas à deux niveaux de temps, H ne dépend pas de $\overrightarrow{U}$ et nous écrivons le schéma sous la forme :*

$$S(\overrightarrow{U}^k, \overrightarrow{U}^{k+1}) = 0, \qquad \textit{avec } S(\overrightarrow{U}, \overrightarrow{V}) = \frac{\overrightarrow{V}}{\Delta t} - \mathbb{G}(\Delta t)\overrightarrow{U} - \Phi(\overrightarrow{F}_k, \overrightarrow{F}_{k+1}), \tag{3.43}$$

avec $\mathbb{G}$ une matrice de $\mathbb{R}^{N\times N}$ indépendante de k et $\overrightarrow{F}_k = \overrightarrow{F}(t_k)$.

La notion de consistance de ce schéma est similaire à celle rencontrée lors de l'approximation des équations hyperboliques par différences finies (cf. [33]). Seule la norme dans laquelle est exprimée l'erreur de consistance varie au sens où la matrice de masse est présente dans la définition de la norme $|\cdot|_{\mathbb{M}}$.

Définition 3.24 (consistance en temps) *On dit que le schéma S est consistant (resp. consistant à l'ordre p) avec le système différentiel (3.27) si (resp. $\exists C_{cons} > 0$ telle que) :*

$$\max_{k=1,K} \left| S(\overrightarrow{U}(t_{k-1}), \overrightarrow{U}(t_k), \overrightarrow{U}(t_{k+1})) \right|_{\mathbb{M}} \underset{\Delta t\to 0}{\longrightarrow} 0 \quad (resp.\ \leq C_{cons}(\Delta t)^p). \tag{3.44}$$

Ci-dessus, la norme $|\cdot|_{\mathbb{M}}$ est associée à la matrice de masse $\mathbb{M}$, ce qui revient à une évaluation en norme $\|\cdot\|_{L^2(\Omega)}$ des schémas, d'après (3.41). On a le résultat suivant pour le θ-schéma :

Proposition 3.25 *Le θ-schéma est consistant à l'ordre 1 et il est consistant à l'ordre 2 si et seulement si $\theta = \frac{1}{2}$.*

Démonstration : On suppose que la fonction $t \mapsto \overrightarrow{F}(t)$ est suffisamment régulière ($\mathcal{C}^4$) sur $[0, T]$. Par conséquent, d'après (3.27), la fonction $t \mapsto \overrightarrow{U}(t)$ est régulière. Nous adoptons la notation $\overrightarrow{U}^{(i)}(t) = \dfrac{d^i}{dt^i}\overrightarrow{U}(t)$ dans ce qui suit.

En vertu des formules de Taylor on a :

$$\overrightarrow{U}(t_{k+1}) = \overrightarrow{U}(t_k) + \Delta t\, \overrightarrow{U}^{(1)}(t_k) + \frac{\Delta t^2}{2}\overrightarrow{U}^{(2)}(t_k) + \frac{\Delta t^3}{6}\overrightarrow{U}^{(3)}(t_k) + O(\Delta t^4),$$
$$\overrightarrow{U}(t_{k-1}) = \overrightarrow{U}(t_k) - \Delta t\, \overrightarrow{U}^{(1)}(t_k) + \frac{\Delta t^2}{2}\overrightarrow{U}^{(2)}(t_k) - \frac{\Delta t^3}{6}\overrightarrow{U}^{(3)}(t_k) + O(\Delta t^4).$$

En remplaçant dans l'expression (3.43) du θ-schéma, on obtient :

$$\begin{aligned}
&\mathbb{M}S(\overrightarrow{U}(t_k), \overrightarrow{U}(t_{k+1})) = \\
&\quad \mathbb{M}\left\{\overrightarrow{U}^{(1)}(t_k) + \frac{\Delta t}{2}\overrightarrow{U}^{(2)}(t_k) + \frac{\Delta t^2}{6}\overrightarrow{U}^{(3)}(t_k) + O(\Delta t^3)\right\} \\
&\quad +\theta\mathbb{K}\left\{\overrightarrow{U}(t_k) + \Delta t\overrightarrow{U}^{(1)}(t_k) + \frac{\Delta t^2}{2}\overrightarrow{U}^{(2)}(t_k) + \frac{\Delta t^3}{6}\overrightarrow{U}^{(3)}(t_k) + O(\Delta t^4)\right\} + (1-\theta)\mathbb{K}\overrightarrow{U}(t_k) \\
&\quad -\theta\left\{\overrightarrow{F}(t_k) + \Delta t\overrightarrow{F}^{(1)}(t_k) + \frac{\Delta t^2}{2}\overrightarrow{F}^{(2)}(t_k) + \frac{\Delta t^3}{6}\overrightarrow{F}^{(3)}(t_k) + O(\Delta t^4)\right\} - (1-\theta)\overrightarrow{F}(t_k),
\end{aligned}$$

qui s'écrit encore :

$$\begin{aligned}
&\mathbb{M}S(\overrightarrow{U}(t_k), \overrightarrow{U}(t_{k+1})) = \\
&\qquad \mathbb{M}\frac{d}{dt}\overrightarrow{U}(t_k) + \mathbb{K}\overrightarrow{U}(t_k) - \overrightarrow{F}(t_k) \qquad (= 0 \text{ car } \overrightarrow{U} \text{ est solution de (3.27)}) \\
&\qquad +\Delta t\left\{\frac{1}{2}\mathbb{M}\overrightarrow{U}^{(2)}(t_k) + \theta\mathbb{K}\overrightarrow{U}^{(1)}(t_k) - \theta\overrightarrow{F}^{(1)}(t_k)\right\} \qquad (3.45) \\
&\qquad +\Delta t^2\left\{\frac{1}{6}\mathbb{M}\overrightarrow{U}^{(3)}(t_k) + \frac{\theta}{2}\mathbb{K}\overrightarrow{U}^{(2)}(t_k) - \frac{\theta}{2}\overrightarrow{F}^{(2)}(t_k)\right\} + O(\Delta t^3).
\end{aligned}$$

Ce qui donne le résultat car lorsque $\theta = \frac{1}{2}$, le terme en Δt devient

$$\frac{\Delta t}{2}\left\{\mathbb{M}\frac{d^2}{dt^2}\overrightarrow{U}(t_k) + \mathbb{K}\frac{d}{dt}\overrightarrow{U}(t_k) - \frac{d}{dt}\overrightarrow{F}(t_k)\right\} = 0 \quad \text{(dériver (3.27) par rapport à } t\text{)}.$$

Le θ-schéma ne peut pas être d'ordre 3 car le terme suivant n'est pas automatiquement nul (pour toutes fonctions $\overrightarrow{U}$ et $\overrightarrow{F}$ satisfaisant à (3.27)). ■

Remarque 3.26 *On remarque que dans l'expression (3.44), la constante C_{cons} va dépendre de h : c'est le cas pour l'analyse du θ-schéma, par l'intermédiaire des matrices $\mathbb{M}$ et $\mathbb{K}$, qui dépendent de la discrétisation, et donc de h. Il n'est généralement pas facile d'obtenir cette constante de manière explicite et donc en particulier de connaître son comportement en fonction de h.*

A titre d'exercice, nous laissons le soin au lecteur de vérifier que le schéma de Richardson est d'ordre 2.

3.4.2 Stabilité et convergence

Le point clef de la convergence d'un schéma est sa stabilité. Nous considérons ici des schémas à deux niveaux de temps écrits sous la forme (3.43) et on introduit l'application :

$$E \; : \quad \begin{array}{ccc} \mathbb{R}^N & \longrightarrow & \mathbb{R}^N \\ \overrightarrow{X} & \mapsto & \Delta t \mathbb{G}(\Delta t)\overrightarrow{X}. \end{array}$$

Le schéma (3.43) – multiplié par Δt – s'écrit alors :

$$\overrightarrow{U}^{k+1} = E(\overrightarrow{U}^k) + \Delta t \Phi(\overrightarrow{F}_k, \overrightarrow{F}_{k+1}). \tag{3.46}$$

Le θ-schéma s'inscrit dans ce cadre avec :

$$\begin{array}{l} E_\theta(\overrightarrow{X}) = (\mathbb{M} + \theta \Delta t \mathbb{K})^{-1}(\mathbb{M} - (1-\theta)\Delta t \mathbb{K})\overrightarrow{X}, \\ \Phi(\overrightarrow{X}, \overrightarrow{Y}) = (\mathbb{M} + \theta \Delta t \mathbb{K})^{-1}((1-\theta)\overrightarrow{X} + \theta \overrightarrow{Y}). \end{array} \tag{3.47}$$

On introduit maintenant la définition de stabilité suivante :

Définition 3.27 (stabilité) *Le schéma (3.43) est stable si il existe une constante C_{stab}, indépendante de k et Δt telle que :*

$$| \, E^k \overrightarrow{X} \, |_{\mathbb{M}} \leq C_{stab} \, | \, \overrightarrow{X} \, |_{\mathbb{M}} \qquad \forall \overrightarrow{X} \in \mathbb{R}^N, \; \forall k = 0, K. \tag{3.48}$$

L'utilisation de la norme $| \cdot |_{\mathbb{M}}$ correspond à la stabilité en norme $\| \cdot \|_{L^2(\Omega)}$ des schémas (cf. (3.41)). La condition (3.48) exprime la continuité uniforme vis-à-vis de Δt des opérateurs $(E^k)_{k=0,K}$:

$$\|E^k\|_{\mathbb{M}} = \sup_{\overrightarrow{X} \in \mathbb{R}^N \setminus \{0\}} \frac{| \, E^k \overrightarrow{X} \, |_{\mathbb{M}}}{| \, \overrightarrow{X} \, |_{\mathbb{M}}} \leq C_{stab} \qquad \forall k = 0, K.$$

Cette définition de la stabilité, nous conduit au résultat classique de convergence :

Théorème 3.28 (convergence) *Si le schéma (3.43) est consistant à l'ordre p avec l'équation (3.27) et stable, alors il existe une constante C, indépendante de k et Δt, telle que :*

$$| \, \overrightarrow{U}^k - \overrightarrow{U}(t_k) \, |_{\mathbb{M}} \leq C(\Delta t)^p \quad \forall k = 0, K. \tag{3.49}$$

Ce théorème établit la convergence du schéma (3.43) pour la norme $\mathcal{C}^0(0, T; \mathbb{R}^N)$.

Démonstration : En vertu de la consistance du schéma, on a :

$$\overrightarrow{U}(t_{k+1}) = E\overrightarrow{U}(t_k) + \Delta t \Phi(\overrightarrow{F}(t_k), \overrightarrow{F}(t_{k+1})) + (\Delta t)^{p+1} R_k \;\; \text{avec} \;\; | \, R_k \, |_{\mathbb{M}} \leq C_{cons}, \;\; \forall k,$$

qui par différence avec (3.46), nous donne

$$\overrightarrow{U}^{k+1} - \overrightarrow{U}(t_{k+1}) = E(\overrightarrow{U}^k - \overrightarrow{U}(t_k)) + (\Delta t)^{p+1} R_k,$$

ce qui conduit, par récurrence, à l'estimation :

$$\begin{aligned}\left|\overrightarrow{U}^{k+1} - \overrightarrow{U}(t_{k+1})\right|_{\mathbb{M}} &\leq \left|E^k(\overrightarrow{U}^0 - \overrightarrow{U}(0))\right|_{\mathbb{M}} + (\Delta t)^{p+1} \sum_{\ell=1,k} \mid E^{k-\ell} R_\ell \mid_{\mathbb{M}},\\ &\leq \|E^k\|_{\mathbb{M}} \left|\overrightarrow{U}^0 - \overrightarrow{U}(0)\right|_{\mathbb{M}} + (\Delta t)^{p+1} \sum_{\ell=1,k} \|E^{k-\ell}\|_{\mathbb{M}} \mid R_\ell \mid_{\mathbb{M}},\\ &\leq C_{stab}\left(\left|\overrightarrow{U}^0 - \overrightarrow{U}(0)\right|_{\mathbb{M}} + k(\Delta t)^{p+1} C_{cons}\right).\end{aligned}$$

Comme $k \leq K = \frac{T}{\Delta t}$ et $\overrightarrow{U}^0 = \overrightarrow{U}(0)$ on déduit finalement que :

$$\left|\overrightarrow{U}^{k+1} - \overrightarrow{U}(t_{k+1})\right|_{\mathbb{M}} \leq C(\Delta t)^p, \quad \text{avec } C = C_{stab} C_{cons}\, T,$$

ce qui achève la démonstration. ∎

Stabilité du θ-schéma

Nous allons maintenant étudier la stabilité du θ-schéma. Pour estimer la norme de l'opérateur E_θ, on utilise à nouveau les couples vecteur propres – valeurs propres $(\overrightarrow{V}_m, \lambda_{m,h})_{m=1,N}$ du problème $\mathbb{K}\overrightarrow{V} = \lambda \mathbb{M}\overrightarrow{V}$, cf. §3.3.1.

Proposition 3.29 (stabilité du θ-schéma) *Si $\theta \geq \frac{1}{2}$ le θ-schéma est inconditionnellement stable. Si $\theta \in [0, \frac{1}{2}[$ le θ-schéma est stable si*

$$\lambda_{m,h}\Delta t \leq \frac{2}{1-2\theta} \qquad \forall m = 1, N. \tag{3.50}$$

Démonstration : Décomposons un vecteur quelconque $\overrightarrow{X}$ sur la base $(\overrightarrow{V}_m)_{m=1,N}$:

$$\overrightarrow{X} = \sum_{m=1,N} x_m \overrightarrow{V}_m \qquad \text{avec } \; x_m = (\mathbb{M}\overrightarrow{X} | \overrightarrow{V}_m) \text{ et } |\overrightarrow{X}|^2_{\mathbb{M}} = \sum_{m=1,N} x_m^2.$$

On a :

$$\begin{aligned}E_\theta \overrightarrow{X} &= (\mathbb{M} + \theta\Delta t\mathbb{K})^{-1}(\mathbb{M} - (1-\theta)\Delta t\mathbb{K}) \sum_{m=1,N} x_m \overrightarrow{V}_m,\\ &= (\mathbb{M} + \theta\Delta t\mathbb{K})^{-1}\left(\sum_{m=1,N} x_m(\mathbb{M}\overrightarrow{V}_m - (1-\theta)\Delta t\mathbb{K}\overrightarrow{V}_m)\right),\\ &= (\mathbb{M} + \theta\Delta t\mathbb{K})^{-1}\left(\sum_{m=1,N} x_m(1-(1-\theta)\Delta t\lambda_{m,h})\mathbb{M}\overrightarrow{V}_m\right).\end{aligned}$$

Remarquons ensuite que :

$$\left\{\begin{array}{c}(\mathbb{M} + \theta\Delta t\mathbb{K})\overrightarrow{V}_m = (1 + \theta\Delta t\lambda_{m,h})\mathbb{M}\overrightarrow{V}_m, \text{ soit encore}\\ (\mathbb{M} + \theta\Delta t\mathbb{K})^{-1}\mathbb{M}\overrightarrow{V}_m = \dfrac{1}{(1+\theta\Delta t\lambda_{m,h})}\overrightarrow{V}_m.\end{array}\right.$$

D'après ce qui précède :

$$E_\theta \overrightarrow{X} = \sum_{m=1,N}\left(\frac{1-(1-\theta)\Delta t\lambda_{m,h}}{1+\theta\Delta t\lambda_{m,h}}\right) x_m \overrightarrow{V}_m.$$

On en déduit que

$$E_\theta^k \overrightarrow{X} = \sum_{m=1,N} \left(\frac{1-(1-\theta)\Delta t\lambda_{m,h}}{1+\theta\Delta t\lambda_{m,h}} \right)^k x_m \overrightarrow{V}_m,$$

et par conséquent :

$$| E_\theta^k \overrightarrow{X} |_{\mathbb{M}}^2 = \sum_{m=1,N} \left(\frac{1-(1-\theta)\Delta t\lambda_{m,h}}{1+\theta\Delta t\lambda_{m,h}} \right)^{2k} x_m^2 \leq \max_{m=1,N} \left(\frac{1-(1-\theta)\Delta t\lambda_{m,h}}{1+\theta\Delta t\lambda_{m,h}} \right)^{2k} |\overrightarrow{X}|_{\mathbb{M}}^2,$$

d'où finalement l'estimation :

$$\frac{| E_\theta^k \overrightarrow{X} |_{\mathbb{M}}}{| \overrightarrow{X} |_{\mathbb{M}}} \leq \max_{m=1,N} \left| \frac{1-(1-\theta)\Delta t\lambda_{m,h}}{1+\theta\Delta t\lambda_{m,h}} \right|^k.$$

Pour que le θ-schéma soit stable, il suffit, d'après la définition 3.27, que

$$-1 \leq \frac{1-\lambda_{m,h}(1-\theta)\Delta t}{1+\lambda_{m,h}\theta\Delta t} \leq 1 \qquad \forall m = 1, N.$$

L'inégalité de droite est toujours vérifiée car

$$\frac{1-\lambda_{m,h}(1-\theta)\Delta t}{1+\lambda_{m,h}\theta\Delta t} = 1 - \frac{\lambda_{m,h}\Delta t}{1+\lambda_{m,h}\theta\Delta t} \qquad \text{avec } \theta \geq 0, \lambda_{m,h} \geq 0, \Delta t \geq 0.$$

L'inégalité de gauche est équivalente à

$$\lambda_{m,h}\Delta t(1-2\theta) \leq 2.$$

Cette condition est automatiquement satisfaite pour tout Δt si $\theta \geq \frac{1}{2}$.

Par contre si $\theta \in [0, \frac{1}{2}[$, Δt doit satisfaire à la condition de stabilité (3.50) comme annoncé. ∎

La proposition 3.29 permet de conclure sur la convergence en temps du θ-schéma. Sous les conditions de stabilité et à h fixé, on a :

$$|\overrightarrow{U}^k - \overrightarrow{U}(t_k)|_{\mathbb{M}} \leq C(\Delta t)^p \quad \text{avec} \begin{cases} p = 1 \ \text{ si } \theta \neq \frac{1}{2}, \\ p = 2 \ \text{ si } \theta = \frac{1}{2}. \end{cases}$$

Remarque 3.30 *La condition (3.50) montre que lorsque le θ-schéma est instable, il ne l'est que sur les composantes de haute fréquence, c'est-à-dire $\lambda_{m,h}$ grand. Par ailleurs, on peut réduire la condition de stabilité (3.50) à :*

$$\left(\max_{m=1,N} \lambda_{m,h} \right) \Delta t \leq \frac{2}{1-2\theta}.$$

Cette remarque a une contrepartie pratique. En effet, il est facile et assez rapide de calculer cette valeur propre (méthode de la puissance présentée au §1.3.3) fournissant ainsi un critère de choix du pas de temps exploitable dans la pratique. Notons, par ailleurs, que les plus grandes valeurs propres du laplacien discrétisé se

comportent comme h^{-2} (voir [13] pour les différences finies). Par conséquent, la condition de stabilité (3.50) devient :

$$\frac{\Delta t}{h^2} \leq C\frac{2}{1-2\theta},$$

ce qui montre que le pas de temps doit être choisi d'autant plus petit que le pas d'espace est petit et ce dans un rapport $\frac{1}{4}$ si l'on diminue le pas d'espace par 2.

Remarque 3.31 *Lorsque $\theta > \frac{1}{2}$, on a, $\lambda_{1,h}$ désignant la plus petite valeur propre :*

$$\left(\frac{1-\lambda_{m,h}(1-\theta)\Delta t}{1+\lambda_{m,h}\theta\Delta t}\right)^2 \leq \left(\frac{1-\lambda_{1,h}(1-\theta)\Delta t}{1+\lambda_{1,h}\theta\Delta t}\right)^2 \quad \text{pour } \Delta t \text{ petit.}$$

Ceci montre que :

$$|E_\theta^k \overrightarrow{X}|_{\mathbb{M}}^2 \leq (1-\lambda_{1,h}\Delta t)^{2k}|\overrightarrow{X}|_{\mathbb{M}}^2 \leq e^{-2\lambda_{1,h}k\Delta t}|\overrightarrow{X}|_{\mathbb{M}}^2.$$

Puisqu'on a $\lambda_{1,h} \geq \lambda_1$ (λ_1 la plus petite valeur propre du problème de Dirichlet homogène) d'après le lemme 1.26, on arrive finalement à :

$$|E_\theta^k \overrightarrow{X}|_{\mathbb{M}} \leq e^{-\lambda_1 k\Delta t}|\overrightarrow{X}|_{\mathbb{M}} \qquad \forall \overrightarrow{X} \in \mathbb{R}^N.$$

La stabilité est donc de type exponentiel et le θ-schéma est asymptotiquement stable. Cette propriété est reliée à la propriété de décroissance exponentielle des solutions de l'équation de la chaleur (proposition 3.12). C'est pourquoi il est souvent préférable d'utiliser le θ-schéma avec $\theta > \frac{1}{2}$.

Exemple monodimensionnel

Les résultats précédents se retrouvent de façon explicite dans le cas monodimensionnel. Reprenons le formalisme développé dans [33], et considérons le schéma d'Euler explicite [3] de type différences finies :

$$\begin{cases} \dfrac{U_I^{k+1}-U_I^k}{\Delta t} - \dfrac{U_{I+1}^k + U_{I-1}^k - 2U_I^k}{\Delta x^2} = 0 & \forall I, \\ U_I^0 = u_0(I\Delta x) & \forall I, \end{cases}$$

discrétisant l'équation de la chaleur monodimensionnelle :

$$\begin{cases} \dfrac{\partial}{\partial t}u(x,t) - \dfrac{\partial^2}{\partial x^2}u(x,t) = 0 & x \in \mathbb{R},\ t > 0, \\ u(0,x) = u_0(x) & x \in \mathbb{R}. \end{cases}$$

3. on choisit, ici, la notation $h = \Delta x$.

Menons l'analyse de Fourier de ce schéma. Supposons que l'on ait, pour $\xi \in \mathbb{R}$ donné :

$$U_J^k = e^{i(J\Delta x\xi)} \quad \forall J \in \mathbb{Z}.$$

Avec ce choix, le schéma conduit à la relation :

$$U_J^{k+1} = g(\xi, \Delta t, \Delta x)U_J^k$$

avec le *coefficient d'amplification* :

$$g(\xi, \Delta t, \Delta x) = 1 + 2\frac{\Delta t}{\Delta x^2}(\cos(\Delta x\xi) - 1).$$

En appliquant le critère de *stabilité de Von Neumann*, on obtient :

$$|g(\xi, \Delta t, \Delta x)| = \left|1 - 4\frac{\Delta t}{\Delta x^2}\sin^2\left(\frac{\xi\Delta x}{2}\right)\right| \leq 1,$$

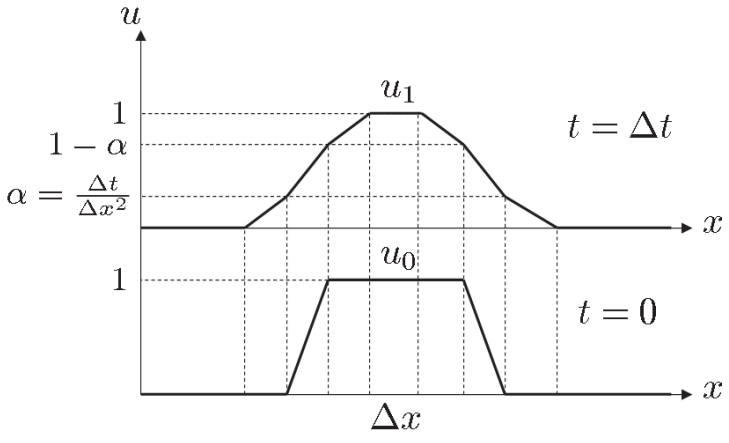
qui donne la condition :

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2\sin^2\left(\frac{\xi\Delta x}{2}\right)},$$

d'où la condition de stabilité :

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}.$$

Par ailleurs, on remarque sur la figure ci-dessous, que ce schéma explicite "propage" à vitesse finie, ce qui n'est pas le cas des schémas implicites. Par conséquent, il se révèle mal adapté à la résolution de l'équation de la chaleur.



Instabilité du schéma de Richardson

Il n'est pas question, ici, de développer la théorie générale de la stabilité des schémas à trois niveaux de temps. Illustrons-la seulement sur l'exemple du schéma de Richardson.

En décomposant sur la base spectrale $(\overrightarrow{V}_m)_m$:

$$\overrightarrow{U}^k = \sum_{m=1,N} \alpha_m^k \overrightarrow{V}_m,$$

le schéma de Richardson s'écrit dans le cas où $\overrightarrow{F} = 0$ (par souci de simplification) :

$$\left\{ \begin{array}{l} \mathbb{M}\overrightarrow{U}^{k+1} = \mathbb{M}\overrightarrow{U}^{k-1} - 2\Delta t \mathbb{K}\overrightarrow{U}^k \\ \\ \mathbb{M}\overrightarrow{U}^1 = \mathbb{M}\overrightarrow{U}^0 - \Delta t \mathbb{K}\overrightarrow{U}^0 \end{array} \right. , \text{ ou } \left\{ \begin{array}{l} \alpha_m^{k+1} = \alpha_m^{k-1} - 2\Delta t \lambda_{m,h} \alpha_m^k \\ \\ \alpha_m^1 = \alpha_m^0 - \Delta t \lambda_{m,h} \alpha_m^0 \end{array} \right. \quad m = 1, N, \tag{3.51}$$

soit encore, sous forme matricielle :

$$\begin{pmatrix} \alpha_m^{k+1} \\ \alpha_m^k \end{pmatrix} = \begin{bmatrix} -2\lambda_{m,h}\Delta t & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \alpha_m^k \\ \alpha_m^{k-1} \end{pmatrix}.$$

Les valeurs propres de la matrice de transfert sont :

$$\gamma^{\pm} = -\lambda_{m,h}\Delta t \pm \sqrt{1 + (\lambda_{m,h}\Delta t)^2}.$$

On a toujours $|\gamma^-| > 1$ et $|\gamma^+| < 1$. En effectuant le changement d'inconnue suivant :

$$\left\{ \begin{array}{l} v_k = \gamma^+ \alpha_m^k + \alpha_m^{k-1}, \\ \tilde{v}_k = \gamma^- \alpha_m^k + \alpha_m^{k-1}, \end{array} \right.$$

on aboutit à

$$v_k = \gamma^+ v_{k-1} \quad \text{et} \quad \tilde{v}_k = \gamma^- \tilde{v}_{k-1},$$

qui conduit par récurrence aux expressions :

$$\left\{ \begin{array}{l} v_k = (\gamma^+)^{k-1}(\gamma^+ \alpha_m^1 + \alpha_m^0), \\ \tilde{v}_k = (\gamma^-)^{k-1}(\gamma^- \alpha_m^1 + \alpha_m^0). \end{array} \right.$$

Finalement, on peut écrire :

$$\alpha_m^k = \frac{1}{\gamma^+ - \gamma^-} \left\{ \left((\gamma^+)^k - (\gamma^-)^k \right) \alpha_m^1 - \gamma^+\gamma^- \left((\gamma^+)^{k-1} - (\gamma^-)^{k-1} \right) \alpha_m^0 \right\}.$$

Néanmoins, puisque $|\gamma^-| > 1$ on en déduit que le schéma de Richardson n'est jamais stable, puisque la condition (3.48) n'est pas valable : il est *inconditionnellement instable*. C'est pourquoi, ce schéma d'ordre 2 est inutilisable dans la pratique.

3.5 Résultats de convergence

Dans cette section, nous allons indiquer les principaux résultats de convergence en temps et en espace pour une approximation par éléments finis en espace et par différences finies en temps (θ-méthode en temps).

Nous commençons par donner une estimation de l'écart entre la solution exacte u du problème (3.1) et la solution u_h du problème semi-discrétisé (3.24).

3.5.1 Convergence du problème semi-discrétisé en espace

On utilise les éléments finis de Lagrange d'ordre k pour la discrétisation en espace. On introduit l'opérateur de projection elliptique P_h de $V = H_0^1(\Omega)$ sur V_h (approximation interne) caractérisé par :

$$a(P_h v - v, v_h) = 0 \qquad \forall v_h \in V_h, \tag{3.52}$$

où

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \; d\Omega. \tag{3.53}$$

L'existence et l'unicité de $P_h v$ est garantie par le théorème de Lax-Milgram appliqué dans l'espace V_h. Attention, il ne faut pas confondre l'opérateur de projection P_h avec l'opérateur d'interpolation Π_h habituel (cf. [15]). On peut en fait montrer que la projection elliptique est également une approximation dans V_h. En effet, nous avons le résultat abstrait suivant :

Lemme 3.32 *On a l'estimation*

$$\|v - P_h v\|_V \leq \frac{C_a}{\alpha} \inf_{v_h \in V_h} \|v - v_h\|_V, \qquad \forall v \in V, \tag{3.54}$$

où α est la constante de coercivité de a et C_a sa constante de continuité.

Démonstration : En utilisant la définition de la projection elliptique, il est facile de voir que

$$\begin{aligned} \alpha \|P_h v - v\|_V^2 \leq a(P_h v - v, P_h v - v) &= a(P_h v - v, v_h - v), \qquad \forall v_h \in V_h, \\ &\leq C_a \, \|P_h v - v\|_V \|v_h - v\|_V, \qquad \forall v_h \in V_h, \end{aligned}$$

d'où le résultat. ∎

Sous l'hypothèse d'approximabilité (1.54) de l'espace V par les $(V_h)_h$ (et d'après la remarque 1.37), on déduit de l'estimation (3.54) que $P_h v$ approche v dans V :

$$\|v - P_h v\|_V \underset{h \to 0}{\longrightarrow} 0 \qquad \forall v \in V. \tag{3.55}$$

Si la fonction v est plus régulière (ce qui est notre cas), plus précisément si $v \in H^2(\Omega)$, par définition de $P_h v$ et en vertu des estimations d'erreur d'interpolation pour les éléments finis de Lagrange, on a

$$\|v - P_h v\|_{H^1(\Omega)} \leq \|v - \Pi_h v\|_{H^1(\Omega)} \leq Ch|v|_{H^2(\Omega)}. \tag{3.56}$$

Enfin, si le problème adjoint est régulier (cf. [15]), une technique analogue à celle utilisée pour établir un résultat de convergence en norme $\|\cdot\|_{L^2(\Omega)}$, permet également de démontrer que (voir [48]) :

$$\|v - P_h v\|_{L^2(\Omega)} \leq Ch^2|v|_{H^2(\Omega)}. \tag{3.57}$$

On démontre alors le résultat :

Proposition 3.33 *Si la solution u du problème (3.1) appartient à $\mathcal{C}^1(0,T;H_0^1(\Omega))$ alors on a, $\forall t \in [0,T]$:*

$$\begin{aligned} \|u_h(t) - u(t)\|_{L^2(\Omega)} &\leq \|u_{h,0} - P_h u_0\|_{L^2(\Omega)} e^{-\lambda_1 t} + \|(I - P_h)u(t)\|_{L^2(\Omega)} \\ &\quad + \int_0^t \|(I - P_h)\frac{du}{dt}(s)\|_{L^2(\Omega)} e^{-\lambda_1(t-s)} ds, \end{aligned} \tag{3.58}$$

où I désigne l'opérateur identité dans $H_0^1(\Omega)$, et λ_1 est la plus petite valeur propre du problème de Dirichlet homogène.

La démonstration de ce résultat s'appuie sur une représentation spectrale discrète de la solution u_h et des estimations analogues à celles utilisées dans la démonstration d'existence de la solution u du problème continu (théorème 3.6). Afin de bien comprendre comment un même outil permet de prouver à la fois l'existence d'une solution et la convergence spatiale du problème semi-discrétisé (3.24) nous allons détailler celle-ci.

Démonstration de la proposition 3.33 : Par différence de (3.24) et (3.2) on obtient :

$$\frac{d}{dt}\int_\Omega (u_h(t) - u(t))\, v_h \, d\Omega + \int_\Omega \nabla\,(u_h(t) - u(t)) \cdot \nabla v_h \, d\Omega = 0 \quad \forall v_h \in V_h.$$

Compte-tenu de (3.52), on déduit de l'égalité précédente :

$$\frac{d}{dt}\int_\Omega \xi_h(t) v_h \, d\Omega + \int_\Omega \nabla \xi_h \cdot \nabla v_h \, d\Omega = \int_\Omega g_h \, v_h \, d\Omega \quad \forall v_h \in V_h, \tag{3.59}$$

où on a posé :

$$\xi_h(t) = u_h(t) - P_h u(t) \;\text{ et }\; g_h(t) = (I - P_h)\frac{d}{dt}u(t).$$

L'égalité dans (3.59) résulte du fait que

$$\frac{d}{dt}\int_\Omega (u(t) - P_h u(t))\, v_h \, d\Omega = \int_\Omega \left(\frac{d}{dt}u(t) - P_h \frac{d}{dt}u(t)\right) v_h \, d\Omega,$$

car $u \in \mathcal{C}^1(0,T;H_0^1(\Omega))$ et en particulier $\dfrac{d}{dt}u(t) \in H_0^1(\Omega)$; on peut donc lui appliquer P_h.
Les fonctions propres discrètes $(v_{m,h})_{m=1,N}$ (cf. §3.3.1), ordonnées par valeurs propres $(\lambda_{m,h})_{m=1,N}$ croissantes, vérifient :

$$\int_\Omega \nabla v_{m,h} \cdot \nabla v_h \, d\Omega = \lambda_{m,h} \int_\Omega v_{m,h} v_h \, d\Omega \quad \forall v_h \in V_h \quad \text{et} \quad \int_\Omega v_{m,h} v_{\ell,h} \, d\Omega = \delta_{m\ell} \quad \forall m, \ell = 1, N.$$

En décomposant $g_h(t)$ sur cette base de fonctions propres, on obtient (voir le calcul de la proposition 3.4) :

$$\xi_h(t) = \sum_{m=1}^{N} \left((\xi_h(0), v_{m,h})_{L^2(\Omega)} e^{-\lambda_{m,h} t} + \int_0^t (g_h(s), v_{m,h})_{L^2(\Omega)} e^{-\lambda_{m,h}(t-s)} ds \right) v_{m,h}.$$

En vertu des propriétés d'orthonormalité des fonctions propres $(v_{m,h})_m$ que l'on vient de rappeler, on a les estimations suivantes :

$$\begin{aligned} \| \textstyle\sum_{m=1}^N (\xi_h(0), v_{m,h})_{L^2(\Omega)} e^{-\lambda_{m,h} t} v_{m,h} \|_{L^2(\Omega)} &= \left[\textstyle\sum_{m=1}^N (\xi_h(0), v_{m,h})^2_{L^2(\Omega)} e^{-2\lambda_{m,h} t} \right]^{1/2} \\ &\leq \|\xi_h(0)\|_{L^2(\Omega)} e^{-\lambda_{1,h} t}, \end{aligned}$$

$$\begin{aligned} \left\| \sum_{m=1}^{N} (g_h(s), v_{m,h})_{L^2(\Omega)} e^{-\lambda_{m,h}(t-s)} v_{m,h} \right\|_{L^2(\Omega)} &= \left[\sum_{m=1}^{N} (g_h(s), v_{m,h})^2_{L^2(\Omega)} e^{-2\lambda_{m,h}(t-s)} \right]^{1/2} \\ &\leq \|g_h(s)\|_{L^2(\Omega)} e^{-\lambda_{1,h}(t-s)}. \end{aligned}$$

Comme d'après le lemme 1.26 on a $\lambda_{1,h} \geq \lambda_1$, on déduit l'estimation suivante :

$$\|\xi_h(t)\|_{L^2(\Omega)} \leq \|\xi_h(0)\|_{L^2(\Omega)} e^{-\lambda_1 t} + \int_0^t \|g_h(s)\|_{L^2(\Omega)} e^{-\lambda_1 (t-s)} ds,$$

qui, compte-tenu des définitions de ξ_h et g_h et de l'inégalité :

$$\|u_h(t) - u(t)\|_{L^2(\Omega)} \leq \|(I - P_h) u(t)\|_{L^2(\Omega)} + \|\xi_h(t)\|_{L^2(\Omega)},$$

conduit à l'estimation (3.58). ∎

Cette proposition fournit des résultats de convergence pour la norme de l'espace $\mathcal{C}^0(0, T; L^2(\Omega))$ du problème semi-discrétisé (3.24) dès lors que l'on suppose que $u_{h,0}$ converge vers u_0 en norme $L^2(\Omega)$. En outre, si la solution est régulière (c'est-à-dire si $u \in \mathcal{C}^1(0, T; H^2(\Omega))$, la convergence est d'ordre 2 en h.

Théorème 3.34 (convergence en espace) *Selon la régularité de la solution du problème (3.1), on a les résultats suivants :*
i) Si la solution u appartient à $\mathcal{C}^1(0, T; H_0^1(\Omega))$ alors :

$$\forall t \in [0, T[, \quad \lim_{h \to 0} \|u_h(t) - u(t)\|_{L^2(\Omega)} = 0. \tag{3.60}$$

ii) Si la solution u appartient à $\mathcal{C}^1(0, T; H_0^1(\Omega) \cap H^2(\Omega))$ et si :

$$\|u_{h,0} - u_0\|_{L^2(\Omega)} \leq C h^2, \tag{3.61}$$

alors il existe une constante C, indépendante de h telle que :

$$\forall t \in [0, T], \quad \|u_h(t) - u(t)\|_{L^2(\Omega)} \leq C h^2. \tag{3.62}$$

Démonstration : i) Il suffit de remarquer que

$$\forall v \in \mathcal{C}^0(0,T;H_0^1(\Omega)), \quad \lim_{h\to 0} \sup_{0\le t\le T} \|(I-P_h)v\|_{H_0^1(\Omega)} = 0.$$

En effet, la famille de fonctions $((I-P_h)v)_{h\ge 0}$ est uniformément équicontinue [4] de $[0,T]$ dans $H_0^1(\Omega)$ et on a, en vertu de (3.55) :

$$\forall t \in [0,T], \quad \lim_{h\to 0} \|(I-P_h)v(t)\|_{H_0^1(\Omega)} = 0.$$

Le théorème d'Ascoli[4] permet alors de conclure. D'après (3.25) et ce qui précède, on déduit facilement l'estimation (3.60).

ii) En raisonnant de la même façon qu'au i), compte-tenu, cette fois ci, de (3.61) et de l'estimation (3.57) on obtient (3.62). ∎

Remarque 3.35 *Lorsque $u_0 \in H^2(\Omega)$ avec $u_{h,0} = \Pi_h u_0$, en vertu de l'estimation d'interpolation pour les éléments finis de Lagrange, l'estimation (3.61) est vérifiée.*

3.5.2 Convergence globale

Notons u_h^k l'approximation par le θ-schéma de la fonction u_h à l'instant t_k :

$$u_h^k = \sum_{I=1,N} U_I^k w_I.$$

On doit estimer l'écart total :

$$\|u_h^k - u(t_k)\|_{L^2(\Omega)}, \qquad \forall k = 0, K.$$

Une idée naturelle consiste à découper cet écart en deux parties :

$$\|u_h^k - u(t_k)\|_{L^2(\Omega)} \le \|u_h^k - u_h(t_k)\|_{L^2(\Omega)} + \|u_h(t_k) - u(t_k)\|_{L^2(\Omega)}.$$

Le second terme est contrôlé (par $C_2 h^2$ par exemple) en vertu du théorème 3.34 et le premier terme par $C_1(\Delta t)^p$ ($p = 1$ ou 2 suivant θ) en vertu du théorème 3.28. On s'attend donc à une estimation d'erreur globale en $(\Delta t)^p + h^2$ lorsque la solution u est suffisamment régulière. Malheureusement, la constante C_1, égale à $C_1 = C_{stab} C_{cons} T$, intervenant dans l'estimation de l'erreur temporelle, dépend de h de façon non explicite ce qui ne permet pas de conclure directement. C'est pourquoi, il est nécessaire d'utiliser une démontration directe fondée sur des estimations d'énergie (cette démonstration est librement inspirée de [48]).

4. **Théorème d'Ascoli**. (voir par exemple [9]) Soit $\mathcal{H}$ un sous-ensemble borné de $\mathcal{C}^0([0,T])$. On suppose que $\mathcal{H}$ est uniformément équicontinu, c'est-à-dire que : $\forall \varepsilon > 0$, $\exists \alpha > 0$ tel que $\forall (s,t) \in [0,T]$, $|s-t| < \alpha \Rightarrow \sup_{f\in\mathcal{H}} |f(s) - f(t)| < \varepsilon$. Alors, $\mathcal{H}$ est relativement compact dans $\mathcal{C}^0([0,T])$.

Théorème 3.36 *On suppose que $u_0 \in H^2(\Omega)$. Si u est une solution suffisamment régulière*[5] *du problème (3.1) alors la solution u_h^k du problème discrétisé par éléments finis de Lagrange et le θ-schéma converge vers u au sens suivant :*

$$\sup_{k=0,K} \|u(t_k) - u_h^k\|_{L^2(\Omega)} \leq C\left((\Delta t)^p + h^2\right),$$

avec $p = 1$ si $\theta \neq \frac{1}{2}$ et $p = 2$ si $\theta = \frac{1}{2}$.

Démonstration : Dans un souci de simplification nous faisons la démonstration dans le cas implicite $\theta = 1$, le cas général se traitant de façon similaire.
On écrit :

$$u_h^k - u(t_k) = (u_h^k - P_h u(t_k)) + (P_h u(t_k) - u(t_k)) \underset{def}{=} \xi_k + \mu_k.$$

En vertu de (3.57), u étant supposé suffisamment régulier, on a l'estimation

$$\|\mu_k\|_{L^2(\Omega)} \leq Ch^2. \tag{3.63}$$

Il reste à estimer $\|\xi_k\|_{L^2(\Omega)}$. C'est l'objet du reste de la preuve. Pour commencer, u_h^k vérifie, avec $D_{\Delta t}^+ u_h^{k-1} = (u_h^k - u_h^{k-1})/\Delta t$:

$$\int_\Omega D_{\Delta t}^+ u_h^{k-1} v_h \, d\Omega + \int_\Omega \nabla u_h^k \cdot \nabla v_h \, d\Omega = \int_\Omega f(t_k) v_h \, d\Omega \quad \forall v_h \in V_h.$$

Par définition de l'opérateur P_h et puisque u est solution de (3.1) on a

$$\int_\Omega D_{\Delta t}^+ P_h u(t_{k-1}) v_h \, d\Omega + \int_\Omega \nabla P_h u(t_k) \cdot \nabla v_h \, d\Omega = \int_\Omega f(t_k) v_h \, d\Omega + \int_\Omega \omega_k v_h \, d\Omega \quad \forall v_h \in V_h$$

avec

$$\omega_k = D_{\Delta t}^+ P_h u(t_{k-1}) - \frac{du}{dt}(t_k).$$

De ce qui précède, on déduit

$$\int_\Omega D_{\Delta t}^+ \xi_{k-1} v_h \, d\Omega + \int_\Omega \nabla \xi_k \cdot \nabla v_h \, d\Omega = -\int_\Omega \omega_k \, v_h \, d\Omega \quad \forall v_h \in V_h.$$

Maintenant, prenons $v_h = \xi_k$ (ce qui est licite car u_h^k et $P_h u(t_k)$ appartiennent tous deux à l'espace V_h). On obtient alors l'estimation, sachant que $\|\nabla \xi_k\|_{L^2(\Omega)^n} \geq 0$,

$$\int_\Omega D_{\Delta t}^+ \xi_{k-1} \xi_k \, d\Omega \leq -\int_\Omega \omega_k \, \xi_k \, d\Omega, \text{ ou } \|\xi_k\|_{L^2(\Omega)}^2 \leq \int_\Omega (\xi_{k-1} - \Delta t \, \omega_k) \xi_k \, d\Omega$$

qui conduit, par application de Cauchy-Schwarz, à l'estimation fondamentale

$$\|\xi_k\|_{L^2(\Omega)} \leq \|\xi_{k-1}\|_{L^2(\Omega)} + \Delta t \|\omega_k\|_{L^2(\Omega)}.$$

Par récurrence, on obtient alors

$$\|\xi_k\|_{L^2(\Omega)} \leq \|\xi_0\|_{L^2(\Omega)} + \Delta t \sum_{i=1}^{k} \|\omega_i\|_{L^2(\Omega)}.$$

5. Précisément, $u \in L^2(0,T;H^2(\Omega)) \cap \mathcal{C}^m(0,T;L^2(\Omega))$ et $\dfrac{du}{dt} \in L^2(0,T;H^2(\Omega))$ où $m = 2$ si $\theta \neq 1/2$ et $m = 3$ si $\theta = 1/2$. Toutes ces hypothèses seront utilisées au cours de la preuve.

D'après (3.57) et l'hypothèse de régularité sur u_0 on a

$$\|\xi_0\|_{L^2(\Omega)} = \|u_{h,0} - P_h u_0\|_{L^2(\Omega)} \leq \|u_{h,0} - u_0\|_{L^2(\Omega)} + \|u_0 - P_h u_0\|_{L^2(\Omega)} \leq Ch^2.$$

Ecrivons :

$$\omega_k = ((P_h - I)D^+_{\Delta t} u(t_{k-1})) + \left(D^+_{\Delta t} u(t_{k-1}) - \frac{du}{dt}(t_k) \right) \underset{def}{=} \rho_k + \delta_k.$$

On veut donc majorer d'une part $\Delta t \sum_{i=1}^k \|\rho_i\|_{L^2(\Omega)}$, et d'autre part $\Delta t \sum_{i=1}^k \|\delta_i\|_{L^2(\Omega)}$. On utilise le résultat ci-dessous pour évaluer la norme $L^2(\Omega)$ de $\zeta = \displaystyle\int_t^{t'} v(s)\, ds$:

$$\begin{aligned}
&\|\zeta\|^2_{L^2(\Omega)} = \int_\Omega \left(\int_t^{t'} v(s)\, ds \right)^2 d\Omega\,; \\
&\text{or, } \left| \int_t^{t'} v(s)\, ds \right| = \left| \int_t^{t'} 1 \times v(s)\, ds \right| \leq \left[\int_t^{t'} 1\, ds \right]^{1/2} \left[\int_t^{t'} v(s)^2\, ds \right]^{1/2} \\
&\qquad\qquad = |t' - t|^{1/2} \left[\int_t^{t'} v(s)^2\, ds \right]^{1/2}; \\
&\Longrightarrow \|\zeta\|^2_{L^2(\Omega)} \leq \int_\Omega \left(|t' - t| \int_t^{t'} v(s)^2\, ds \right) d\Omega = |t' - t| \int_t^{t'} \|v(s)\|^2_{L^2(\Omega)}\, ds\,. \qquad (3.64)
\end{aligned}$$

Notons que

$$\rho_i = (P_h - I)\frac{1}{\Delta t} \int_{t_{i-1}}^{t_i} \frac{du}{dt}(s) ds = \int_{t_{i-1}}^{t_i} \frac{1}{\Delta t}(P_h - I)\frac{du}{dt}(s) ds.$$

Choisissons donc $\zeta = \rho_i$ dans (3.64) :

$$\|\rho_i\|^2_{L^2(\Omega)} \leq \frac{1}{\Delta t} \times \int_{t_{i-1}}^{t_i} \left\| (P_h - I)\frac{du}{dt}(s) \right\|^2_{L^2(\Omega)} ds\,.$$

A l'aide de (3.57), on en déduit

$$\|\rho_i\|^2_{L^2(\Omega)} \leq \frac{C^2 h^4}{\Delta t} \times \int_{t_{i-1}}^{t_i} \left| \frac{du}{dt}(s) \right|^2_{H^2(\Omega)} ds\,.$$

On utilise ensuite la majoration usuelle

$$\sum_{i=1}^k \|\rho_i\|_{L^2(\Omega)} \leq \left(\sum_{i=1}^k \|\rho_i\|^2_{L^2(\Omega)} \right)^{1/2} \left(\sum_{i=1}^k 1 \right)^{1/2} \leq \frac{\sqrt{T}}{(\Delta t)^{1/2}} \left(\sum_{i=1}^k \|\rho_i\|^2_{L^2(\Omega)} \right)^{1/2}.$$

(Pour la dernière inégalité, on se souvient que l'inégalité $k\Delta t \leq T$ est valable pour tout k).
On regroupe le tout, pour aboutir à

$$\Delta t \sum_{i=1}^k \|\rho_i\|_{L^2(\Omega)} \leq C\sqrt{T}\, h^2 \times \left(\int_0^{t_k} \left| \frac{du}{dt}(s) \right|^2_{H^2(\Omega)} ds \right)^{1/2}.$$

Pour la seconde somme (avec les δ_i), rappelons que

$$\delta_i = \frac{u(t_i) - u(t_i)}{\Delta t} - \frac{du}{dt}(t_i) = -\int_{t_{i-1}}^{t_i} \frac{1}{\Delta t}(s - t_{i-1}) \frac{d^2 u}{dt^2}(s) ds,$$

Pour $\zeta = \delta_i$ dans (3.64), on trouve cette fois

$$\begin{aligned}
\|\delta_i\|^2_{L^2(\Omega)} &\leq \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} (s-t_{i-1})^2 \left\|\frac{d^2u}{dt^2}(s)\right\|^2_{L^2(\Omega)} ds \\
&\leq \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} (s-t_{i-1})^2 \, ds \times \sup_{s\in[t_{i-1},t_i]} \left(\left\|\frac{d^2u}{dt^2}(s)\right\|^2_{L^2(\Omega)}\right) \\
&= \frac{(\Delta t)^2}{3} \times \left(\sup_{s\in[t_{i-1},t_i]} \left\|\frac{d^2u}{dt^2}(s)\right\|_{L^2(\Omega)}\right)^2 \\
&\leq \frac{(\Delta t)^2}{3} \times \left(\sup_{s\in[0,T]} \left\|\frac{d^2u}{dt^2}(s)\right\|_{L^2(\Omega)}\right)^2 .
\end{aligned}$$

L'estimation précédente est indépendante de i. On trouve alors, puisque $k\Delta t \leq T$:

$$\Delta t \sum_{i=1}^{k} \|\delta_i\|_{L^2(\Omega)} \leq k\frac{(\Delta t)^2}{\sqrt{3}} \times \sup_{s\in[0,T]} \left\|\frac{d^2u}{dt^2}(s)\right\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{3}} T\,\Delta t \times \sup_{s\in[0,T]} \left\|\frac{d^2u}{dt^2}(s)\right\|_{L^2(\Omega)} .$$

En réunissant toutes les estimations précédentes on aboutit finalement à

$$\|\xi_k\|_{L^2(\Omega)} \leq C(\Delta t + h^2), \tag{3.65}$$

qui allié à (3.63) démontre le théorème. ∎

Au cours de ce chapitre, nous n'avons traité que certains aspects de l'approximation de l'équation de la chaleur (consistance, stabilité et convergence). D'autres seraient intéressants à analyser. En particulier, la propriété[6] de *positivité d'un schéma* :

$$\inf_x u_h^k(x) \geq 0 \Rightarrow \inf_x u_h^{k+1}(x) \geq 0$$

est importante car l'équation de la chaleur la vérifie (voir la proposition 3.17). Ainsi, on peut montrer que le θ-schéma est positif si l'on utilise une approximation éléments finis P^1-Lagrange. En outre, la positivité du schéma interdit l'apparition d'oscillations, ce qui le rend bien meilleur.

Nous nous sommes intéressés au cas de l'équation de la chaleur, mais il est bien évident que les techniques d'approximation s'adaptent au cadre général de la remarque 3.3.

3.6 Illustrations numériques

Nous donnons dans cette section quelques illustrations numériques de la résolution par éléments finis de problèmes de diffusion. Nous nous intéressons dans un premier

6. La positivité du schéma est à relier à la propriété de monotonie des schémas utilisés pour discrétiser les équations hyperboliques (cf. [33]).

temps, à l'étude de la convergence numérique de la discrétisation de l'équation de la chaleur par éléments finis et le θ-schéma. Nous illustrons ensuite les effets de régularisation et de dissipation sur quelques exemples simples. Enfin, nous présentons le calcul du prix d'une option européenne à deux actifs basé sur la résolution d'une équation de diffusion à coefficients variables [47, 37, 2].

3.6.1 Convergence numérique du θ-schéma

Afin de réaliser une étude numérique de convergence, on se place sur le carré $\Omega =]0,1[\times]0,1[$ sur lequel on considère le problème modèle :

$$\left\{\begin{array}{ll} \dfrac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = 0 & (x,t) \in Q_T = \Omega\times]0,T[, \\ \dfrac{\partial u}{\partial n}(x,t) = 0 & (x,t) \in \Sigma_T = \partial\Omega\times]0,T[, \\ u(x,0) = u_0(x) & x \in \Omega. \end{array}\right. \tag{3.66}$$

La formulation variationnelle de ce problème est

$$\left\{\begin{array}{l} \text{trouver} \quad u \in \mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega)) \text{ tel que} \\ \dfrac{d}{dt}\left(\displaystyle\int_\Omega u(x,t)\,v(x)\,dx\right) + \displaystyle\int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,dx = 0, \quad \begin{array}{c}\forall v \in H^1(\Omega), \\ \forall t \in]0,T[,\end{array} \\ u(x,0) = u_0(x) \quad x \in \Omega. \end{array}\right. \tag{3.67}$$

Considérons une approximation par éléments finis de Lagrange ($(M_I)_{I=1,N}$ nœuds du maillage) et par le θ-schéma (Δt pas de temps constant, $K = E(T/\Delta t)$) d'inconnue $(u_h^k)_{k=0,K}$. La formulation variationnelle (3.67) conduit au schéma discret :

$$\left\{\begin{array}{l} \overrightarrow{U}^0 = \overrightarrow{U}_0 \\ (\mathbb{M} + \theta\Delta t\mathbb{K})\overrightarrow{U}^{k+1} = (\mathbb{M} + (1-\theta)\Delta t\mathbb{K})\overrightarrow{U}^k, \quad k = 0, K-1 \end{array}\right. \tag{3.68}$$

où $\overrightarrow{U}_0$ est défini par $U_0^I = u_0(M_I)$, $1 \le I \le N$, et $\overrightarrow{U}^k$ est le vecteur de composantes $(u_I^k)_{I=1,N}$ représentant une approximation de la solution $u(M_I, t_k)$ au nœud M_I à l'instant $t_k = k\Delta t$ et $\mathbb{M}$, $\mathbb{K}$ les matrices de masse et de rigidité usuelles. Lorsque les coefficients du problème de diffusion sont indépendants du temps et que le pas de temps est constant, ce qui est le cas présent, il est possible et même fortement recommandé de pré-factoriser la matrice $\mathbb{M}+\theta\Delta t\mathbb{K}$. Ici, cette matrice étant définie-positive, on utilise la factorisation de Cholesky :

$$\mathbb{M} + \theta\Delta t\mathbb{K} = \mathbb{L}\mathbb{L}^t,$$

ce qui conduit à la forme plus efficace de l'algorithme (résolution de systèmes triangulaires par descente-remontée) :

$$\begin{cases} \overrightarrow{U}^0 = \overrightarrow{U}_0 \\ \mathbb{L}\overrightarrow{V}^k = (\mathbb{M} + (1-\theta)\Delta t\mathbb{K})\overrightarrow{U}^k \\ \mathbb{L}^t\overrightarrow{U}^{k+1} = \overrightarrow{V}^k \end{cases}, \quad k = 0, K-1. \tag{3.69}$$

Ce schéma reste implicite (résolution d'un système linéaire) mais le coût de la résolution du système triangulaire, proportionnel à $N^{3/2}$ (si $\Omega \subset \mathbb{R}^n$, en $N^{2-1/n}$), reste raisonnable par rapport au coût de calcul du produit matrice×vecteur intervenant dans l'algorithme, proportionnel lui à N. Lorsque $\theta = 0$, le schéma, bien que qualifié d'explicite, requiert également la résolution d'un système linéaire ("inversion" de la matrice de masse $\mathbb{M}$). En utilisant une formule de quadrature adaptée pour le calcul des coefficients de la matrice $\mathbb{M}$ ou la technique de condensation de masse (voir §4.5.1), on peut se ramener à une matrice de masse diagonale, restituant ainsi le caractère explicite du schéma pour $\theta = 0$.

Remarque 3.37 *Lorque le nombre de pas de temps est très grand, il peut être plus intéressant de calculer l'inverse de la matrice du système une fois pour toute !*

En choisissant la donnée initiale $(m, n \geq 0)$:

$$u_0(x_1, x_2) = \cos(m\pi x_1)\cos(n\pi x_2),$$

la solution du problème (3.66) est donnée par :

$$u(x_1, x_2, t) = \cos(m\pi x_1)\cos(n\pi x_2)e^{-\lambda_{mn}t} \text{ avec } \lambda_{mn} = (m^2+n^2)\pi^2.$$

Dans l'étude numérique qui suit, nous avons choisi $m = n = 1$; la solution $u(x,t)$ correspondante est représentée sur la figure 3.1 à l'instant $t = 0.1$.
Le code Matlab suivant permet de construire les nappes d'erreur représentant $\sup_{k=0,K} \|u_h^k - u(t_k)\|_{L^2(\Omega)}$ en fonction du paramètre de finesse du maillage h et du pas de temps Δt.

```
theta=0.5;tf=0.1;
m=1;n=1;lmn=(m^2+n^2)*pi^2;
at=10.^(-4:-0.2:-7);q=5:3:25;
[h,dt]=meshgrid(1./q,at);
er_sup_l2=0*h;ndt=size(h,1);ndh=size(h,2);

for i=1:ndh,                                    %variation du pas de maillage
  [S,T,BR,RT]=triangule_rectangle([0 1 0 1],q(i),q(i),1);   %maillage P1
  [S2,T2,BR2,RT2]=maillageP2(S,T,BR,RT);                    %maillage P2
  [T2,S2]=renume(T2,S2);                                    %renumerotation
  S=S2;T=T2;BR=BR2;RT=RT2;
  [K,M]=calcul_EF_2D(S,T,RT);                               %matrices EF
  Uex=cos(m*pi*S(:,1)).*cos(n*pi*S(:,2));                   %donnee initiale
  for j=1:ndt,                                              %variation de dt
    t=0;d=dt(j,i);Un=Uex;
```

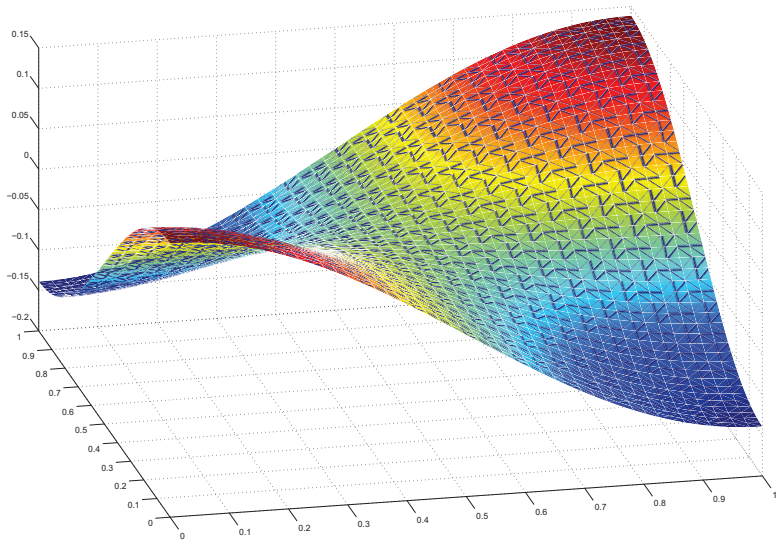


Figure 3.1. Solution exacte ($m = n = 1$) de l'équation de la chaleur (3.66) à $t = 0.1$

```
    I=(M+theta*d*K);E=(M-(1-theta)*d*K);          %assemblage des mat. EF
    L=chol(I);                                    %fact. de Cholesky
    Un=Uex;
    while(t<tf)                                   %boucle en temps du theta-schema
        R=E*Un;
        Unp1=L\(L'\R);
        t=t+d;Un=Unp1;
        ec=Un-exp(-lmn*t)*Uex;                    %calcul de l'erreur
        er_l2=sqrt(ec'*M*ec);
        er_sup_l2(j,i)=max(er_sup_l2(j,i),er_l2);
    end,
  end,
end,
surf(log10(h),log10(dt),log10(er_sup_l2));        %surface d'erreur
```

Nous donnons sur la figure 3.2, les nappes d'erreur (en échelle logarithmique) obtenues en utilisant une approximation par éléments finis P^2 et ce pour les valeurs $\theta = 0, \frac{1}{2}, 1$ du θ-schema. Ces résultats sont obtenus avec $T = 0.5$; temps relativement court qui permet de conserver une solution ne décroissant pas trop vite ($\lambda_{11} \approx 20$). Les bornes de h et Δt varient d'une situation à l'autre afin de se placer dans des plages significatives. Bien évidemment, lorque $\theta = 0$ il faut satisfaire une condition de stabilité (voir proposition 3.29). C'est la raison pour laquelle ne sont pas représentés les résultats pour les couples $(h, \Delta t)$ qui ne satisfont pas cette condition de stabilité. Pour $\theta \geq \frac{1}{2}$, le θ-schéma est inconditionnellement stable.
Les résultats obtenus sont en accord avec l'estimation théorique indiquée dans le théorème 3.36. En particulier, on observe que la précision suivant h est en h^3 ce qui est conforme avec l'ordre des éléments finis P^2 utilisés (à savoir, convergence en norme L^2 d'ordre h^{p+1} pour des éléments finis d'ordre p, si la solution est suf-

fisamment régulière), et que la précision suivant Δt est en Δt lorsque $\theta \neq \frac{1}{2}$ et en Δt^2 lorsque $\theta = \frac{1}{2}$.

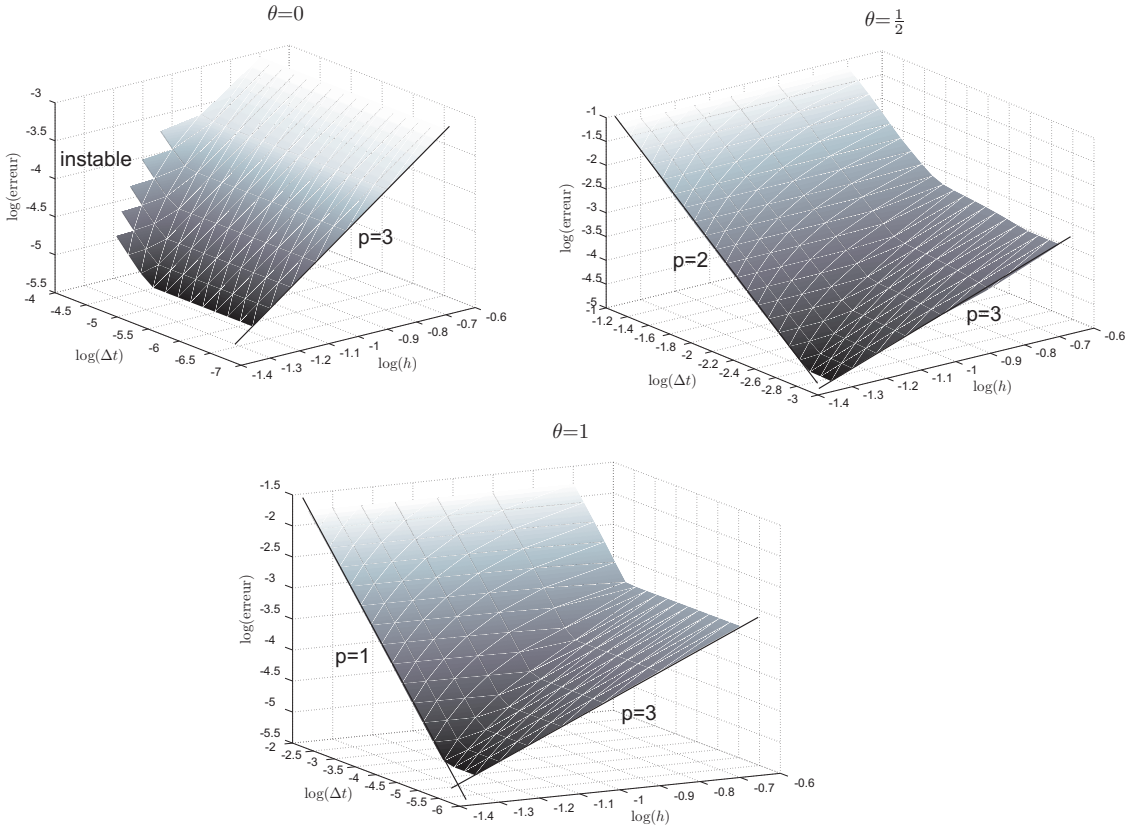


Figure 3.2. Surface d'erreur $\sup_{k=0,K} \|u_h^k - u(t_k)\|_{L^2(\Omega)}$ pour une approximation par éléments finis P^2, et $\theta = 0, \frac{1}{2}, 1$ (représentation logarithmique)

3.6.2 Effet dissipatif et régularisant

A l'aide de l'exemple suivant, nous allons mettre en évidence l'effet régularisant et dissipatif de l'équation de la chaleur. Cet effet est à mettre en regard du caractère non régularisant et conservatif des équations hyperboliques qui sont abordées au chapitre suivant. Nous considérons, à nouveau, l'équation de la chaleur (3.66) dans le domaine $\Omega =]0,1[\times]0,1[$, mais en choisissant comme donnée initiale une fonction discontinue :

$$u_0(x) = \begin{cases} 1 & \text{si } |x - \frac{1}{2}| < \frac{1}{10}, \\ 0 & \text{si } |x - \frac{1}{2}| > \frac{1}{10}. \end{cases}$$

En Matlab, le vecteur $\overrightarrow{U}_0$ (de composantes $U_0^I = u_0(M_I)$, $1 \leq I \leq N$) se construit à partir de la fonction u_0, par exemple, à l'aide des commandes suivantes :

```
W=(S(:,1)-0.5).*(S(:,1)-0.5)+(S(:,2)-0.5).*(S(:,2)-0.5);
U0=sqrt(W)<=0.1;
```

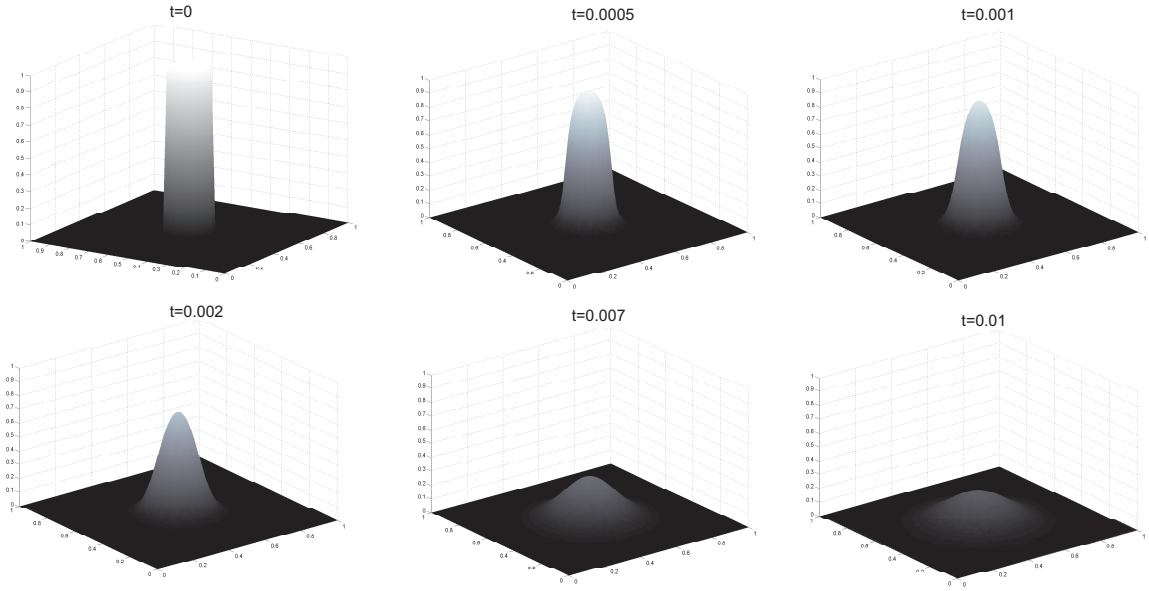


Figure 3.3. Répartition de la température à différents instants

Pour cette donnée initiale, nous donnons sur la figure 3.3 la solution obtenue à différents instants à l'aide d'une approximation P^2, $\theta = 1$, $h = 0.025$ et $\Delta t = 0.0005$. On observe d'une part, que la solution devient régulière, en fait instantanément C^∞ d'après la théorie (effet régularisant) et d'autre part, que la solution décroît et tend rapidement vers 0 (effet dissipatif). La figure 3.4 montre l'évolution de la température au cours du temps au point $(0.5, 0.5)$ (trait plein) et au point $(0, 0)$ (trait pointillé). Ces courbes mettent en évidence de façon plus quantitative ces mêmes phénomènes. Notons que d'après le principe du maximum, la solution doit

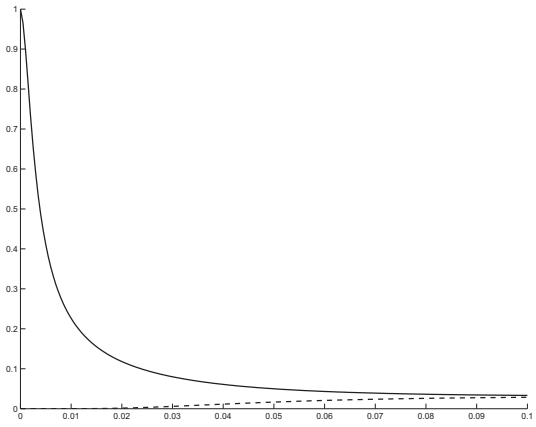


Figure 3.4. Evolution de la température au point $(0.5, 0.5)$ (trait plein) et au point $(0, 0)$ (trait pointillé)

rester positive. Ce que l'on observe numériquement, bien que le schéma ne soit pas positif car, contrairement à l'approximation P^1, l'approximation par élément fini P^2 ne conserve pas la positivité (fonctions de bases non positives) !

3.6.3 Calcul d'une option européenne

Dans le contexte des marchés financiers, il est courant de manipuler des produits financiers appelés options. Il s'agit de contrats d'assurance d'une durée T (maturité) émis par un organisme financier et visant à garantir à un client le prix K (appelé *strike*) d'un actif a (actions, matières premières,...) d'une fluctuation aléatoire de sa valeur x (un cours sur un marché). Lorsqu'il s'agit d'un droit de vente on parle de *put* et on parle de *call* dès lors qu'il s'agit d'un droit d'achat. Suivant les droits de vente du titulaire du contrat, on distingue plusieurs types d'options. La plus simple, l'*option européenne*, stipule que le droit de vente ne peut s'exercer qu'à la fin du contrat (maturité). Lorsqu'à tout instant $t < T$, on peut exercer son droit à vendre au prix K on parle d'*option américaine*. Il existe bien évidemment bien d'autre types de règles. Dans la suite, nous nous plaçons dans le cadre d'une option européenne. On définit alors la "récompense" (*payoff*) $R(t) = (x(t) - K)^+$ qui traduit le fait que si le cours est supérieur au prix garanti K, le détenteur de l'actif a intérêt à vendre sans faire jouer son contrat. Dans le cas contraire, le détenteur du contrat fera jouer son option et vendra donc au prix K à maturité. La problématique consiste donc à déterminer, dans un contexte fluctuant, le prix $p(x,t)$ d'un tel contrat à tout instant t en fonction de la valeur x de l'actif sous-jacent à l'option. Dans le cadre du modèle de Black-Scholes à risque neutre où la fluctuation aléatoire suit un processus stochastique particulier (*brownien*), on montre (voir [37]) que le prix $p(x,t)$ est donné par l'équation de diffusion rétrograde :

$$\frac{\partial p}{\partial t}(x,t) + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 p}{\partial x^2}(x,t) + r\, x \frac{\partial p}{\partial x}(x,t) - r\, p(x,t) = 0, \quad x \in \mathbb{R}_*^+,\ 0 < t < T,$$

où σ^2 est la variance de l'actif a (*volatilité*) et r le taux d'intérêt. A l'échéance T du contrat, le prix est

$$p(x,T) = (x - K)^+.$$

Ce modèle se généralise à plusieurs actifs. Par exemple, pour un modèle à deux actifs, on considère le problème ($x = (x_1, x_2) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+$) :

$$\begin{cases} \dfrac{\partial p}{\partial t}(x,t) + \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{2} \Xi_{ij} x_i x_j \frac{\partial^2 p}{\partial x_i \partial x_j}(x,t) \\ \qquad\qquad + r \displaystyle\sum_{i=1} x_i \frac{\partial p}{\partial x_i}(x,t) - r\, p(x,t) = 0 \quad x \in (\mathbb{R}_*^+)^2,\ 0 < t < T, \\ p(x,T) = (x_1 + x_2 - K)^+ \ x \in (\mathbb{R}_*^+)^2, \end{cases} \tag{3.70}$$

où Ξ est la matrice de covariance associée aux actifs (a_1, a_2) que l'on suppose, pour simplifier, indépendante de x et t. En effectuant le changement $t \to -t$ (renversement du temps), on aboutit à une équation de la chaleur à coefficients variables :

$$\begin{cases} \dfrac{\partial p}{\partial t}(x,t) - \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{2} \Xi_{ij} x_i x_j \dfrac{\partial^2 p}{\partial x_i \partial x_j}(x,t) \\ \qquad\qquad -r \displaystyle\sum_{i=1} x_i \dfrac{\partial p}{\partial x_i}(x,t) + r\, p(x,t) = 0 \quad x \in (\mathbb{R}_*^+)^2,\ 0 < t < T, \\ p(x,0) = (x_1 + x_2 - K)^+ \quad x \in (\mathbb{R}_*^+)^2. \end{cases} \tag{3.71}$$

Cette équation dégénère en $x_1 = 0$ ou $x_2 = 0$ car alors, des termes d'ordre 2 disparaissent. A cause de cette dégénérescence il n'est pas nécessaire [7], et en fait pas possible, de préciser la condition aux limites en $x_1 = 0$ ou $x_2 = 0$! L'étude théorique de cette équation présente des difficultés qui sortent du cadre de cet ouvrage. En utilisant des approches plus générales (voir [39]) on peut montrer qu'il existe une unique solution $p \in L^2(0,T;V) \cap C^0(0,T;L^2((\mathbb{R}_*^+)^2))$ avec

$$V = \left\{ v \in L^2((\mathbb{R}_*^+)^2), x_i \partial_i v \in L^2((\mathbb{R}_*^+)^2), i = 1,\ 2 \right\}.$$

L'espace V muni de la norme

$$\|v\|_V^2 = \|v\|^2_{L^2((\mathbb{R}_*^+)^2)} + \|x_1 \partial_1 v\|^2_{L^2((\mathbb{R}_*^+)^2)} + \|x_2 \partial_2 v\|^2_{L^2((\mathbb{R}_*^+)^2)}$$

est un espace de Hilbert, dense dans $L^2((\mathbb{R}_*^+)^2)$.
Notons que comme la solution $p \in C^0(0,T;L^2((\mathbb{R}_*^+)^2))$, $p(x,t)$ tend vers 0 lorsque $|x|$ tend vers $+\infty$ (les fonctions de L^2 tendent vers 0 à l'infini !). Dans la perspective de la résolution numérique de ce problème, il convient de borner l'espace ; ce qui est raisonnable en pratique car x_1 ou x_2 n'ont pas de raison de tendre vers l'infini. On supposera donc, dorénavant, que le problème est posé sur le carré : $\Omega_L =]0, L[\times]0, L[$, sur lequel on définit l'espace fonctionnel V_L (avec des notations évidentes), et compte tenu de l'observation précédente, on impose la condition aux limites suivante sur la frontière artificielle $\Gamma_{art} = \{x \in \partial\Omega_L \,|\, x_1 = L \text{ ou } x_2 = L\}$:

$$\frac{\partial}{\partial n} p(x,t) = 0 \text{ sur } \Gamma_{art}.$$

On peut également imposer une condition de Dirichlet homogène $p = 0$ sur Γ_{art}.
Finalement, on cherche à résoudre numériquement le problème

7. Pour s'en convaincre, il suffit de construire la formulation variationnelle (3.73) ci-après par intégration par parties. La présence des facteurs x_1 ou x_2 annule la contribution sur la partie de la frontière incluse dans $\{x \,|\, x_1 = 0\}$ ou $\{x \,|\, x_2 = 0\}$.

$$\begin{cases} \dfrac{\partial p}{\partial t}(x,t) - \mathrm{div}(A(x)\nabla p(x,t)) \\ \qquad\qquad + W(x)\cdot\nabla p(x,t) + r\,p(x,t) = 0 \quad x\in\Omega_L,\ 0<t<T, \\ p(x,0) = (x_1+x_2-K)^+ \quad x\in\Omega_L, \\ \dfrac{\partial}{\partial n}p(x,t) = 0 \quad x\in\Gamma_{art},\ 0<t<T, \end{cases} \tag{3.72}$$

où on a posé

$$A(x) = \frac{1}{2}\begin{bmatrix} \Xi_{11}x_1^2 & \Xi_{12}x_1x_2 \\ \Xi_{21}x_1x_2 & \Xi_{22}x_2^2 \end{bmatrix} \text{ et } W(x) = \begin{pmatrix} (\Xi_{11}+\frac{1}{2}\Xi_{21}-r)x_1 \\ (\Xi_{22}+\frac{1}{2}\Xi_{12}-r)x_2 \end{pmatrix}.$$

La formulation variationnelle de ce problème, obtenue après intégration par parties, est

$$\begin{cases} \text{trouver } \ p\in L^2(0,T;V_L)\cap C^0(0,T;L^2(\Omega_L)) \text{ tel que} \\ \dfrac{d}{dt}\displaystyle\int_{\Omega_L} p(x,t)\,q(x)\,dx + \int_{\Omega_L} A(x)\nabla p(x,t)\cdot\nabla q(x)\,dx \\ + \displaystyle\int_{\Omega_L} W(x)\cdot\nabla p(x,t)\,q(x)dx + r\int_{\Omega_L} p(x,t)\,q(x)\,dx = 0, \quad \begin{matrix}\forall q\in V_L,\\ \forall t\in]0,T[,\end{matrix} \\ p(x,0) = (x_1+x_2-K)^+ \quad x\in\Omega_L. \end{cases} \tag{3.73}$$

Comme l'espace $H^1(\Omega_L)$ est inclus dans l'espace V_L, les approximations par éléments finis de Lagrange sont conformes dans l'espace V_L. On peut donc appliquer à la formulation variationnelle (3.73), une approximation par éléments finis de Lagrange. Soit $(w_I)_{I=1,N}$ la famille des fonctions de base attachées à un maillage du domaine Ω_L et $V_h = \mathrm{vect}(w_I)_{I=1,N}$ l'espace d'approximation interne ($V_h\subset V_L$). On note $p_h(t)\in V_h$ l'approximation au temps t de la solution et $\overrightarrow{P}(t)$ le vecteur des composantes de $p_h(t)$ dans la base $(w_I)_{I=1,N}$. En considérant la formulation discrète dans V_h associée à la formulation continue (3.73), on obtient le système d'équations suivant

$$\begin{cases} \mathbb{M}\dfrac{d}{dt}\overrightarrow{P}(t) + (\mathbb{K}+\mathbb{B}+r\mathbb{M})\overrightarrow{P}(t) = 0 \quad 0<t<T, \\ \overrightarrow{P}(0) = \overrightarrow{Q}_0 \end{cases} \tag{3.74}$$

avec, pour $I,\ J = 1, N$:

$$\mathbb{M}_{IJ} = \int_{\Omega_L} w_J\,w_I\,dx,\ \mathbb{K}_{IJ} = \int_{\Omega_L} A(x)\nabla w_J\cdot\nabla w_I\,dx,\ \mathbb{B}_{IJ} = \int_{\Omega_L}(W(x)\cdot\nabla w_J)w_I\,dx$$

et $\overrightarrow{Q}_0$ le vecteur des composantes de l'interpolé de la fonction $(x_1+x_2-K)^+$.

On pose par la suite

$$\mathbb{D} = \mathbb{K} + \mathbb{B} + r\mathbb{M}.$$

En utilisant, par exemple, le schéma d'Euler implicite ($\theta = 1$) avec un pas de temps constant ($t_k = k\Delta t,\ 0 \leq k \leq K$), on est conduit au schéma itératif suivant, où $\overrightarrow{P}^k$ est l'approximation du vecteur $\overrightarrow{P}(t_k)$:

$$\begin{cases} \overrightarrow{P}^0 = \overrightarrow{Q}_0 \\ (\mathbb{M} + \Delta t\mathbb{D})P^{k+1} = \mathbb{M}P^k, \quad k = 0, K-1. \end{cases} \tag{3.75}$$

L'implémentation Matlab de ce schéma est similaire à celle déjà présentée auparavant (pré-factorisation de la matrice et boucle en temps). La seule différence réside dans le calcul des matrices éléments finis à coefficients variables $\mathbb{K}$ et $\mathbb{B}$ dont nous donnons ci-après le code Matlab qui est une adaptation de code **calcul_EF_2D.m** :

```
function [K,B,M]=calcul_EF_2D_var(Corneu,Numtri,Reftri,fK,fM,fB)
os=sqrt(15);s3=1./3.;                                  %formule de quadrature
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;pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];q=size(Numtri,2);
nt=size(Numtri,1);ns=size(Corneu,1);nbq=length(pds_quadT);
K=sparse(ns,ns);M=sparse(ns,ns);B=sparse(ns,ns);
A=[1 0;0 1];W=[0;0];c=1;
for t=1:nt,                                          %boucle sur les éléments
  S=[Corneu(Numtri(t,1),:);Corneu(Numtri(t,2),:);Corneu(Numtri(t,3),:)];
  S21=S(2,:)-S(1,:);S31=S(3,:)-S(1,:);
  delta=S21(1)*S31(2)-S21(2)*S31(1);
  Jflmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;          %transfo affine
  Mt=zeros(q,q);Kt=zeros(q,q);Bt=zeros(q,q);
  for k=1:nbq,                          %boucle sur les points de quadrature
    x=pts_quadT(k,1);y=pts_quadT(k,2);
    if(q==3)                                  %calculs des fonctions de base P1
      w=[1-x-y x y];gw=[-1 1 0;-1 0 1];
    else                                      %calculs des fonctions de base P2
      w=[(1-x-y)*(1-2*x-2*y) x*(2*x-1) y*(2*y-1)
          4*x*(1-x-y) 4*x*y 4*y*(1-x-y)];
      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=S'*[1-x-y; x; y];
    if(nargin>5) W=fB(P(1),P(2),Reftri(t));end;       %calcul des matrices
    if(nargin>4) c=fM(P(1),P(2),Reftri(t));end,              %élémentaires
    if(nargin>3) A=fK(P(1),P(2),Reftri(t));end;
    pk=pds_quadT(k)*abs(delta);jg=Jflmt*gw;
    Mt=Mt+c*pk*w'*w;Kt=Kt+pk*jg'*A*jg;Bt=Bt+pk*w'*W'*jg;
  end,
  In=Numtri(t,:);                                  %assemblage de K, M et B
  K(In,In)=K(In,In)+Kt; M(In,In)=M(In,In)+Mt;B(In,In)=B(In,In)+Bt;
end,
```

Le programme Matlab suivant réalise l'implémentation du schéma itératif (3.75) :

```
global r Xi
r=0.05;Xi=[0.04 -0.024; -0.024 0.04];SK=25;          %donnee du probleme
n=50;a=50;
[S,T,BR,RT]=triangle_rectangle([0 a 0 a],n,n,1);    %maillage P1 du carre
[K,B,M]=calcul_EF_2D_var(S,T,RT,@fA,@fUn,@fW);          %matrices EF
D=K+B+r*M;
t=0,tf=2;dt=0.01;nt=ceil(tf/dt)+1;                    %initialisation
p=zeros(size(S,1),1);k=1;
p(:,1)=max(SK-S(:,1)-S(:,2),0);                       %donnee initiale
[L,U]=lu(M+dt*D);                                     %factorisation LU
while(t<tf)                                           %boucle en temps
      p(:,k+1)=U\(L\(M*p(:,k)));
      t=t+dt;k=k+1;
end,
```

où les coefficients variables $A(x)$ et $W(x)$ sont définis dans les fonctions Matlab :

```
function [A]=fA(x,y,rf)
global Xi; A=0.5*[Xi(1,1)*x*x Xi(1,2)*x*y;Xi(2,1)*x*y Xi(2,2)*y*y];
function [W]=fW(x,y,rf)
global r Xi; W=[(Xi(1,1)+0.5*Xi(2,1)-r)*x;(Xi(2,2)+0.5*Xi(1,2)-r)*y];
function [f]=fUn(x,y,rf)
f=1.;
```

Nous donnons sur la figure 3.5, la solution obtenue avec le code précédent et les paramètres suivants :

$$\Xi = \begin{bmatrix} 0.04 & -0.024 \\ -0.024 & 0.04 \end{bmatrix} \text{ et } r = 0.05,$$

soit une volatilité de 20% et un taux d'intérêt de 5% par an. La simulation a été réalisée sur 2 ans avec un pas de temps de l'ordre de 3 jours. Afin de mieux

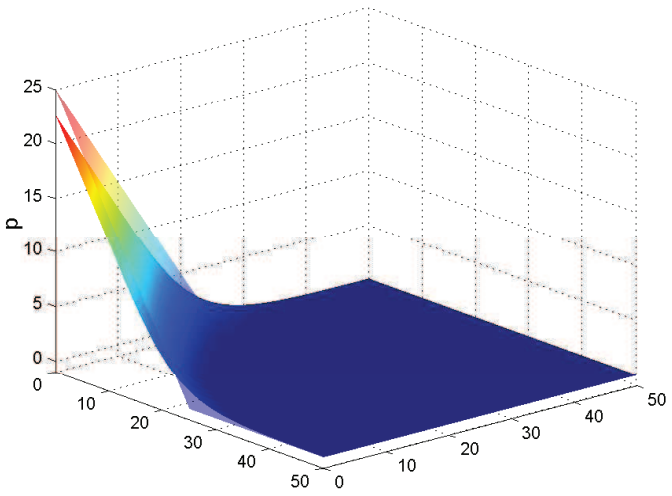


Figure 3.5. Isosurface du prix p à l'instant 0 (nappe transparente) et à l'instant $T = 2$

observer l'évolution, nous représentons sur la figure 3.6 les isosurfaces de $\log(p)$ en ces mêmes instants.

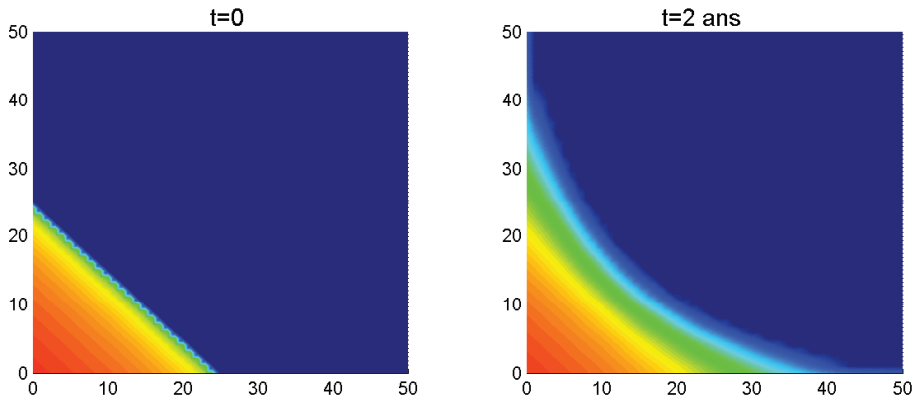


Figure 3.6. Isosurface de $\log(p)$ aux instants 0 et $T = 2$ ans

4

Étude et approximation de l'équation des ondes

Ce chapitre est consacré à l'étude de l'*équation des ondes* :

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = 0 \tag{4.1}$$

caractéristique des équations hyperboliques du second ordre :

$$\frac{\partial^2 u}{\partial t^2} - Pu = 0. \tag{4.2}$$

avec P un opérateur elliptique [23, 15]. L'équation des ondes en dimension 1, s'écrit également, en introduisant les fonctions intermédiaires $v = \dfrac{\partial u}{\partial t}$ et $w = c\dfrac{\partial u}{\partial x}$:

$$\frac{\partial}{\partial t}\begin{pmatrix} v \\ w \end{pmatrix} - \begin{bmatrix} 0 & c\partial_x \\ c\partial_x & 0 \end{bmatrix}\begin{pmatrix} v \\ w \end{pmatrix} = 0,$$

faisant ainsi apparaître un système hyperbolique du premier ordre. Par conséquent, l'étude de l'équation des ondes et de ses propriétés rejoint, dans une certaine mesure, celle menée dans [33].

Dans tout le chapitre, nous traiterons le problème modèle suivant, où Ω est un ouvert de $\mathbb{R}^n$, de frontière $\partial\Omega$ supposée "suffisamment régulière" et $T > 0$:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,t) - c^2 \Delta u(x,t) = f(x,t), & (x,t) \in Q_T = \Omega \times]0,T[, \\ u(x,t) = 0, & (x,t) \in \Sigma_T = \partial\Omega \times]0,T[, \\ u(x,0) = u_0(x), \ \dfrac{\partial u}{\partial t}(x,0) = u_1(x), & x \in \Omega, \end{cases} \tag{4.3}$$

où f est un terme source, u_0 et u_1 sont les conditions initiales et c désigne la vitesse de propagation dans le milieu qui est supposée constante dans tout ce chapitre. Les équations d'ondes, de la forme (4.2) interviennent dans plusieurs domaines de la physique, citons les plus courants :

– Les équations de Maxwell pour la modélisation des ondes électromagnétiques. Le champ u représente alors soit le champ électrique, soit le champ magnétique.
– Les équations de l'acoustique pour la propagation des ondes sonores dans un fluide. Le champ u représente alors le champ de pression ou le potentiel des vitesses.
– Les équations de l'élastodynamique pour la propagation d'ondes élastiques dans un solide. Le champ u représente alors le champ de déplacement des particules.
– Les équations de l'hydrodynamique pour la modélisation de la houle.

Remarque 4.1 *La plupart des résultats que nous donnons peuvent être étendus au cas d'un milieu hétérogène. L'équation gouvernant la propagation d'ondes devient alors*

$$\rho \frac{\partial^2 u}{\partial t^2} - div(\mu \nabla u) = f, \tag{4.4}$$

les coefficients $\rho(x)$ et $\mu(x)$ caractérisant le milieu de propagation. La vitesse dans le milieu est alors dépendante du point x et s'exprime sous la forme

$$c(x) = \sqrt{\frac{\mu(x)}{\rho(x)}}.$$

Le lecteur intéressé pourra retrouver les résultats plus complets sur l'équation d'ondes en milieu hétérogène dans [34]. Nous avons fait le choix ici de nous restreindre aux milieux homogènes (c constante) par souci de simplification de la présentation.

4.1 Le cas 1D : la formule de D'Alembert et ses conséquences

4.1.1 La formule de D'Alembert

On considère l'équation des ondes en dimension 1 avec une vitesse constante c et on s'intéresse au problème posé sur la droite réelle :

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,t) - c^2 \dfrac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x), \quad x \in \mathbb{R}. \end{cases} \tag{4.5}$$

On peut résoudre ce problème de façon explicite.

Théorème 4.2 *La solution du problème (4.5) est donnée par la formule de d'Alembert :*

$$u(x,t) = \frac{u_0(x+ct)+u_0(x-ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)ds \\ +\frac{1}{2c}\int_0^t ds \int_{|y-s|<c(t-s)} f(y,s)dy. \tag{4.6}$$

Démonstration : On se restreint au cas où $f = 0$, le cas $f \neq 0$ est un peu plus technique mais le principe de la démonstration est le même. L'idée est d'utiliser l'identité

$$\partial_t^2 - c^2\partial_x^2 = (\partial_t - c\partial_x)(\partial_t + c\partial_x),$$

c'est-à-dire de voir l'équation des ondes comme deux équations de transport se propageant en sens inverse, et d'introduire des variables qui permettent d'intégrer directement :

$$\partial_\xi \longleftrightarrow \partial_t + c\partial_x, \quad \partial_\eta \longleftrightarrow \partial_t - c\partial_x,$$

c'est-à-dire

$$\begin{cases} \xi = x - ct \\ \eta = x + ct \end{cases} \iff \begin{cases} x = \dfrac{\xi+\eta}{2} \\ t = \dfrac{\xi-\eta}{2c}, \end{cases}$$

et $U(\xi,\eta) = u(x,t)$. On a donc

$$\begin{cases} \partial_\xi U = \partial_x u \partial_\xi x + \partial_t u \partial_\xi t = \dfrac{1}{2c}(\partial_t + c\partial_x)u, \\ \partial_\eta U = \partial_x u \partial_\eta x + \partial_t u \partial_\eta t = -\dfrac{1}{2c}(\partial_t - c\partial_x)u, \end{cases}$$

et l'équation d'ondes se réécrit simplement

$$\partial^2_{\xi\eta} U(\xi,\eta) = 0. \tag{4.7}$$

Les conditions initiales en $t = 0$ deviennent des conditions en $\xi = \eta$:

$$\begin{aligned} &U(\xi,\xi) = u_0(\xi), \\ &\partial_\eta U(\xi,\xi) - \partial_\xi U(\xi,\xi) = \frac{u_1(\xi)}{c}. \end{aligned} \tag{4.8}$$

En intégrant (4.7) respectivement par rapport à ξ et η on obtient

$$\text{(i)}\ \ \partial_\eta U(\xi,\eta) = G(\eta),\ \text{(ii)}\ \ \partial_\xi U(\xi,\eta) = F(\xi),$$

et la deuxième condition initiale devient

$$G(\xi) - F(\xi) = \frac{u_1(\xi)}{c}.$$

Intégrons (i) par rapport à η, entre $\eta_0 = \xi$ et η :

$$U(\xi,\eta) - U(\xi,\xi) = \int_\xi^\eta G(s)ds,$$

ce qui conduit, grâce à la première condition initiale, à

$$U(\xi,\eta) - u_0(\xi) = \int_\xi^\eta G(s)ds.$$

De même on intègre (ii) par rapport à ξ, entre $\xi_0 = \eta$ et ξ :

$$U(\xi,\eta) - U(\eta,\eta) = -\int_{\xi}^{\eta} F(s)ds \Rightarrow U(\xi,\eta) - u_0(\eta) = -\int_{\xi}^{\eta} F(s)ds,$$

et en sommant, on obtient

$$2U(\xi,\eta) = u_0(\xi) + u_0(\eta) + \int_{\xi}^{\eta} \frac{u_1(s)}{c} ds,$$

ce qui donne la formule de D'Alembert. ■

Remarque 4.3 *On peut retrouver la formule de D'Alembert en passant par une transformée de Fourier en espace :*

$$\mathcal{F}(v)(\kappa) = \widehat{v}(\kappa) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} v(x) e^{-i\kappa x} dx. \tag{4.9}$$

On rappelle qu'on a alors :

$$v(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{v}(\kappa) e^{i\kappa x} d\kappa. \tag{4.10}$$

Il est alors facile de montrer que la transformée de Fourier de la solution de (4.5) a l'expression suivante, lorsque $f = 0$:

$$\widehat{u}(\kappa,t) = \widehat{u}_0(\kappa)\cos(c\kappa t) + \widehat{u}_1(\kappa)\frac{\sin(c\kappa t)}{c\kappa}. \tag{4.11}$$

La formule de D'Alembert s'obtient alors en calculant les transformées de Fourier inverses de ces distributions.

4.1.2 Propriétés qualitatives

Cône de dépendance, propagation à vitesse finie, conservation d'énergie

D'après la formule de D'Alembert, on remarque que pour calculer la solution au point $M = (x,t)$, on a besoin des conditions initiales et du second membre f seulement dans le cône de dépendance $D(M)$ délimité par les droites passant par M et de pentes $\pm 1/c$ (voir figure 4.1). On en déduit une propriété essentielle des ondes qui est leur *propagation à vitesse finie*. En effet, si on se donne f à support compact $K = [a,b]$ (par rapport à x) et les conditions intiales u_0 et u_1 dans le même support K, à un instant t, la solution $u(x,t)$ est à support $K_t = [a-ct, b+ct]$.

On définit l'énergie par

$$E(t) = \frac{1}{2}\int_{\mathbb{R}} \left(\left(\frac{\partial u}{\partial t}\right)^2 + \left(c\frac{\partial u}{\partial x}\right)^2 \right) dx. \tag{4.12}$$

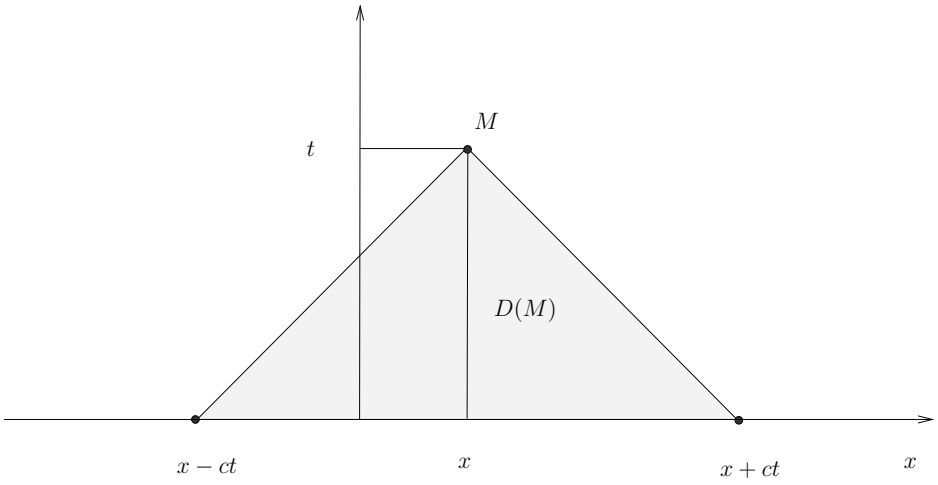


Figure 4.1. Cône de dépendance

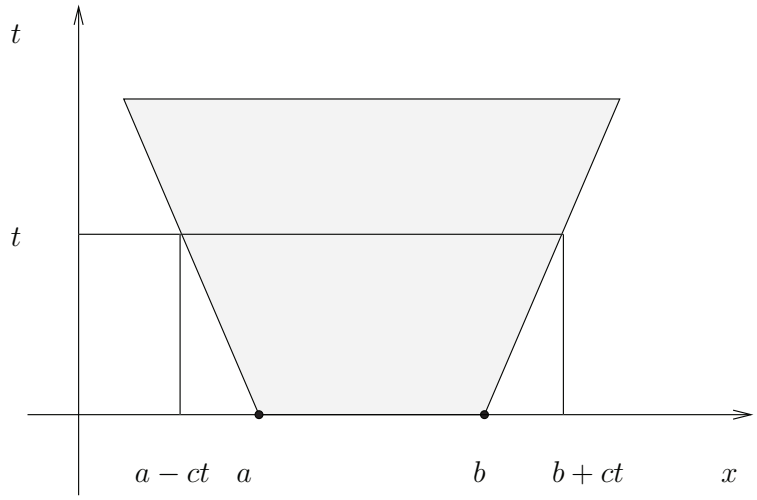


Figure 4.2. Propagation à vitesse finie

En absence de terme source ($f = 0$), on peut montrer qu'il y a conservation de l'énergie, i.e.

$$E(t) = E(0) = \frac{1}{2} \|u_1\|^2_{L^2(\mathbb{R})} + \frac{1}{2} \left\| c \frac{\partial u_0}{\partial x} \right\|^2_{L^2(\mathbb{R})}, \quad \forall t > 0.$$

On peut montrer ce résultat de deux façons. La première façon consiste à partir de l'équation des ondes, à la multiplier par $\partial_t u$ et à intégrer par parties (voir section 4.3.1). La deuxième démonstration se fait directement à partir de la formule de D'Alembert.

Ondes planes harmoniques

Ce sont des solutions particulières de l'équation des ondes, qui s'écrivent sous la forme :

$$u(x,t) = e^{i(\kappa x - \omega t)}, \tag{4.13}$$

où $\kappa \in \mathbb{R}$ est le nombre d'onde, ω la pulsation. On note $\lambda = 2\pi/\kappa$ la longueur d'onde (période en espace) et $T = 2\pi/\omega$ la période temporelle. La fonction u est une solution de l'équation des ondes si la *relation de dispersion* est vérifiée :

$$\omega^2 = c^2\kappa^2, \tag{4.14}$$

ce qui montre en particulier que $\lambda = 2\pi c/\omega = cT$. Ces solutions particulières jouent un rôle important : en utilisant la transformation de Fourier en espace (4.9) on montre que toute solution d'énergie finie de l'équation des ondes est une superposition d'ondes planes harmoniques. Plus précisément, la formule (4.11) permet de montrer que si u est une solution d'énergie finie, on peut l'écrire comme

$$u(x,t) = u^+(x,t) + u^-(x,t),$$

où u^+ est une onde se propageant vers la droite et u^- une onde se propageant vers la gauche, avec

$$u^+(x,t) = \int_{\mathbb{R}} a^+(\kappa) e^{i(\kappa x - \omega^+(\kappa)t)} d\kappa, \qquad \omega^+(\kappa) = c\kappa,$$
$$u^-(x,t) = \int_{\mathbb{R}} a^-(\kappa) e^{i(\kappa x - \omega^-(\kappa)t)} d\kappa, \qquad \omega^-(\kappa) = -c\kappa.$$

Les amplitudes $a^+(\kappa)$ et $a^-(\kappa)$ dépendent des conditions initiales :

$$a^+(\kappa) = \frac{1}{2}\widehat{u}_0(\kappa) + \frac{i}{2c\kappa}\widehat{u}_1(\kappa),$$
$$a^-(\kappa) = \frac{1}{2}\widehat{u}_0(\kappa) - \frac{i}{2c\kappa}\widehat{u}_1(\kappa).$$

Ce résultat s'obtient facilement à partir de l'expression en Fourier de la solution (4.11), qui montre que (voir également §4.3.3)

$$\widehat{u}(\kappa,t) = a^+(\kappa)e^{-ic\kappa t} + a^-(\kappa)e^{ic\kappa t}.$$

4.2 Théorie variationnelle de l'équation des ondes

4.2.1 Formulation variationnelle de l'équation des ondes

Comme pour l'équation de la chaleur, les variables temps t et espace x ne jouent pas le même rôle, ce qui conduit à considérer la fonction $t \mapsto u(t)$ à valeurs dans un espace $V(\Omega)$ de fonctions définies sur Ω. Nous renvoyons à la section 3.1.1 pour plus de détails.

Solutions fortes ou solutions classiques de l'équation des ondes

On admet que pour des données "assez régulières", le problème (4.3) admet une unique solution "très régulière", appelée *solution forte* ou *solution classique*. L'outil pour étudier les solutions fortes des équations d'évolution est le théorème de Hille-Yosida, que nous ne présenterons pas ici (le lecteur intéressé peut par exemple consulter [34]). Nous nous contentons d'énoncer le théorème suivant, application directe du théorème de Hille-Yosida :

Théorème 4.4 *Si on fait les hypothèses :*

$$\begin{cases} (u_0, u_1) \in H_0^1(\Delta, \Omega) \times H_0^1(\Omega), \\ f \qquad\quad \in \mathcal{C}^1(0, T; L^2(\Omega)), \end{cases}$$

le problème (4.3) admet une unique solution forte :

$$u \in \mathcal{C}^2(0, T; L^2(\Omega)) \cap \mathcal{C}^1(0, T; H_0^1(\Omega)) \cap \mathcal{C}^0(0, T; H_0^1(\Delta, \Omega)).$$

où

$$H_0^1(\Delta, \Omega) = \{u \in H_0^1(\Omega) \ / \ \Delta u \in L^2(\Omega)\}.$$

Solutions faibles de l'équation des ondes

Dans les applications réelles, les données sont habituellement moins régulières que ce qui a été supposé au théorème 4.4 et il en résulte une solution elle-même moins régulière, dite solution faible. La régularité de cette solution faible est en fait liée à des propriétés d'énergie (voir section 4.3.1). Pour affaiblir la régularité, on établit une formulation variationnelle de (4.3). Supposons que $u \in \mathcal{C}^2(0, T; L^2(\Omega)) \cap \mathcal{C}^1(0, T; H_0^1(\Omega)) \cap \mathcal{C}^0(0, T; H_0^1(\Delta, \Omega))$ est la solution forte de (4.3). Multiplions la première équation de (4.3) par une fonction test $v \in H_0^1(\Omega)$ et intégrons sur Ω :

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\, v(x)\, dx - \int_\Omega c^2 \Delta u(x,t)\, v(x)\, dx = \int_\Omega f(x,t)\, v(x)\, dx.$$

En appliquant une formule de Green (cf. [15]), on obtient :

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\, v(x)\, dx + \int_\Omega c^2 \nabla u(x,t) \cdot \nabla v(x)\, dx = \int_\Omega f(x,t)\, v(x)\, dx.$$

Dès que $u \in \mathcal{C}^2(0, T; L^2(\Omega))$ on a :

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\, v(x)\, dx = \frac{d^2}{dt^2} \left(\int_\Omega u(x,t)\, v(x)\, dx \right),$$

car :

$$t \mapsto \int_\Omega u(x,t)\, v(x) dx$$

est une fonction de $\mathcal{C}^2(0,T)$. Si on relaxe la régularité de la solution en temps par rapport à la solution forte, et qu'on suppose simplement que $u \in \mathcal{C}^1(0,T;L^2(\Omega))$ alors la fonction :

$$t \mapsto \int_\Omega u(x,t)\, v(x)\, dx,$$

est seulement $\mathcal{C}^1$ sur $[0,T]$ et par conséquent,

$$\frac{d^2}{dt^2}\left(\int_\Omega u(x,t)\, v(x)\, dx\right)$$

est une distribution sur $]0,T[$ définie par, pour tout $\psi \in \mathcal{D}(]0,T[)$:

$$\begin{aligned}\left\langle \frac{d^2}{dt^2}\left(\int_\Omega u(x,t)v(x)\, dx\right); \psi \right\rangle &= -\left\langle \frac{d}{dt}\left(\int_\Omega u(x,t)v(x)\, dx\right); \frac{d\psi}{dt} \right\rangle \\ &= \left\langle \left(\int_\Omega u(x,t)v(x)\, dx\right); \frac{d^2\psi}{dt^2} \right\rangle \\ &= \int_0^T \int_\Omega u(x,t)v(x)\frac{d^2\psi}{dt^2}(t)\, dt.\end{aligned}$$

On introduit alors la formulation variationnelle (dite faible) du problème (4.3) :

$$\begin{cases} \text{trouver} \quad u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega)) \text{ tel que} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega u(x,t)\, v(x)\, dx\right) + \displaystyle\int_\Omega c^2 \nabla u(x,t) \cdot \nabla v(x)\, dx \\ \qquad\qquad = \displaystyle\int_\Omega f(x,t)\, v(x)\, dx, \quad \forall v \in H_0^1(\Omega), \text{ p.p. } t \in]0,T[, \\ u(x,0) = u_0(x), \quad \dfrac{\partial u}{\partial t}(x,0) = u_1(x) \quad \text{dans} \quad \Omega. \end{cases} \tag{4.15}$$

L'un des intérêts de cette formulation est qu'elle garde un sens si on demande moins de régularité sur les données. Nous nous placerons par la suite dans le cas :

$$f \in L^1(0,T;L^2(\Omega)), \qquad u_0 \in H_0^1(\Omega) \quad \text{et} \quad u_1 \in L^2(\Omega). \tag{4.16}$$

Notons que pour satisfaire la deuxième condition initiale, c'est-à-dire pouvoir définir la vitesse à l'instant zéro comme une fonction de $L^2(\Omega)$, on demande à la solution d'être $\mathcal{C}^1$ en temps et non pas seulement $\mathcal{C}^0$ comme c'était le cas pour l'équation de la chaleur. On peut démontrer plus généralement que si $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$ est solution de (4.15) alors u vérifie (4.3) au sens des distributions sur Q_T.

Par ailleurs, pour une fonction test $v \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$, on doit remplacer la première egalité aparaissant dans (4.15) par l'égalité suivante, valable pour presque tout $t \in]0,T[$:

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)\, v(x,t)\, dx + \int_\Omega c^2 \nabla u(x,t) \cdot \nabla v(x,t)\, dx = \int_\Omega f(x,t)\, v(x,t)\, dx. \quad (4.17)$$

Remarque 4.5 *Par construction toute solution forte de (4.3) est a fortiori solution faible de (4.15).*

Remarque 4.6 *La formulation variationnelle (4.15) se généralise sous la forme*

$$\begin{cases} \text{trouver } u \in \mathcal{C}^1(0,T;H) \cap \mathcal{C}^0(0,T;V) \text{ tel que :} \\ \dfrac{d^2}{dt^2}(u(t),v)_H + a(u(t),v) = \ell(t,v) \quad \forall v \in V, \text{ p.p. } t > 0, \\ u(0) = u_0, \\ \dfrac{du}{dt}(0) = u_1. \end{cases}$$

où H est un espace de Hilbert muni du produit scalaire $(\cdot,\cdot)_H$, V un sous-espace de Hilbert dense dans H, $a(\cdot,\cdot)$ une forme bilinéaire continue sur V, $\ell(t,\cdot)$ une forme linéaire continue sur V, $u_0 \in V$ et enfin $u_1 \in H$.

4.2.2 Existence d'une solution

Comme pour l'équation de la chaleur, il existe plusieurs façons de démontrer l'existence d'une solution au problème (4.15), citons [9, 34] pour l'application de la théorie de Hille-Yosida aux problèmes d'évolution et [39, 23, 34] pour des techniques variationnelles. Ces techniques s'appliquent pour des domaines Ω de $\mathbb{R}^n$ non nécessairement bornés.
Nous faisons le choix ici d'utiliser la décomposition spectrale de l'opérateur Δ, technique qui a été suivie à la section 3.1.3, et nous ferons donc l'hypothèse dans cette section que Ω est un ouvert borné de $\mathbb{R}^n$.
La base hilbertienne de fonctions propres $(v_i)_i$ du laplacien (cf. section 3.1.3) nous permet ici encore d'expliciter la solution du problème variationnel (4.15).

Proposition 4.7 *Si u est solution du problème (4.15) alors on a :*

$$\begin{cases} u(t) = \displaystyle\sum_{i=1}^{+\infty} \alpha_i(t)\, v_i, \\ \alpha_i(t) = (u_0,v_i)_{L^2(\Omega)}\, \cos(c\sqrt{\lambda_i}\, t) + (u_1,v_i)_{L^2(\Omega)}\, \dfrac{\sin(c\sqrt{\lambda_i}\, t)}{c\sqrt{\lambda_i}} \\ \qquad + \displaystyle\int_0^t \frac{\sin(c\sqrt{\lambda_i}\,(t-s))}{c\sqrt{\lambda_i}}\, (f(s),v_i)_{L^2(\Omega)}\, ds, \end{cases} \quad (4.18)$$

la série étant convergente dans $L^2(\Omega)$ pour presque tout t.

Démonstration : elle découle de la décomposition modale de la solution sous la forme :

$$u(t) = \sum_{i\geq 1} (u(t), v_i)_{L^2(\Omega)} v_i \equiv \sum_{i\geq 1} \alpha_i(t) v_i \quad \text{avec} \quad \alpha_i(t) = (u(t), v_i)_{L^2(\Omega)}.$$

En injectant cette solution dans la formulation variationnelle (4.15) on aboutit au système suivant vérifié par les $\alpha_i(t)$:

$$\begin{cases} \dfrac{d^2\alpha_i}{dt^2}(t) + c^2\lambda_i\alpha_i(t) = (f(t), v_i)_{L^2(\Omega)}, \\ \alpha_i(0) = (u_0, v_i)_{L^2(\Omega)}, \\ \dfrac{d\alpha_i}{dt}(0) = (u_1, v_i)_{L^2(\Omega)}, \end{cases} \tag{4.19}$$

dont l'unique solution est donnée par la deuxième ligne de (4.18). ∎

Grâce à cette représentation en série de la solution u, nous allons maintenant établir le résultat d'existence :

Théorème 4.8 (existence) *Si on suppose que $f \in L^1\left(0,T;L^2(\Omega)\right)$, $u_0 \in H_0^1(\Omega)$ et $u_1 \in L^2(\Omega)$ alors le problème variationnel (4.15) admet une unique solution $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$.*

Démonstration : Nous reprenons exactement la même démarche et les mêmes notations que celle suivie pour l'équation de la chaleur (cf. théorème 3.6), en particulier on note $(\cdot,\cdot)$ le produit scalaire dans $L^2(\Omega)$. Pour établir l'existence d'une solution au problème (4.15), il suffit de vérifier que la série (4.18) converge dans $\mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$, dès lors que $u_0 \in H_0^1(\Omega)$, $u_1 \in L^2(\Omega)$ et $f \in L^1(0,T;L^2(\Omega))$.
Posons, pour $m \geq 1$,

$$V_m = \underset{i=1,m}{\text{Vect}} \,(v_i)\,,$$

l'espace vectoriel de dimension m engendré par les fonctions propres $(v_i)_{i=1,m}$ et remplaçons alors le problème continu (4.15) par le problème approché :

$$\begin{cases} \text{trouver } u_m : [0,T] \mapsto u_m(t) \in V_m \text{ solution du problème :} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega u_m(t) v \, d\Omega\right) + \int_\Omega c^2 \nabla u_m(t) \cdot \nabla v \, d\Omega = \int_\Omega f(t) v \, d\Omega \qquad \forall v \in V_m, \\ u_m(0) = u_{0,m} = \displaystyle\sum_{i=1,m} (u_0, v_i)_{L^2(\Omega)} \, v_i, \\ \dfrac{du_m}{dt}(0) = u_{1,m} = \displaystyle\sum_{i=1,m} (u_1, v_i)_{L^2(\Omega)} \, v_i. \end{cases} \tag{4.20}$$

Il est clair que le problème (4.20), équivalent à un système différentiel linéaire de dimension finie, admet une unique solution $u_m \in \mathcal{C}^1(0,T;V_m)$ donnée par :

$$u_m(t) = \sum_{i=1}^{m} \alpha_i(t) v_i, \tag{4.21}$$

où les α_i sont définis en (4.18), c'est-à-dire la somme partielle d'ordre m de la série (4.18).

– Nous allons maintenant démontrer que la suite $(u_m)_m$ est une suite de Cauchy à la fois dans les espaces $\mathcal{C}^1(0,T;L^2(\Omega))$ et $\mathcal{C}^0(0,T;H_0^1(\Omega))$. Les calculs étant analogues à ceux développés dans la démonstration du théorème 3.6, nous ne les détaillons pas ici et présentons juste les étapes.
Les hypothèses sur les données se traduisent par

– $u_1 \in L^2(\Omega)$

$$\int_\Omega u_1^2 \, d\Omega < +\infty \Longleftrightarrow \lim_{m,p\to+\infty} \left[\sum_{i=m+1}^{p} (u_1, v_i)^2 \right] = 0. \tag{4.22}$$

– $u_0 \in H_0^1(\Omega)$

$$\begin{aligned} &\text{(a)} \int_\Omega u_0^2 \, d\Omega < +\infty \iff \lim_{m,p\to+\infty} \left[\sum_{i=m+1}^{p} (u_0, v_i)^2 \right] = 0, \\ &\text{(b)} \int_\Omega |\nabla u_0|^2 \, d\Omega < +\infty \Longleftrightarrow \lim_{m,p\to+\infty} \left[\sum_{i=m+1}^{p} \lambda_i \, (u_0, v_i)^2 \right] = 0. \end{aligned} \tag{4.23}$$

– $f \in L^1\left(0,T;L^2(\Omega)\right)$

$$\int_0^T \|f(t)\|_{L^2(\Omega)} dt < +\infty \Longleftrightarrow \lim_{m,p\to+\infty} \left[\int_0^T \left(\sum_{i=m+1}^{p} (f(t), v_i)^2 \right)^{1/2} dt \right] = 0. \tag{4.24}$$

On veut montrer que la suite $(u_m)_m$ est une suite de Cauchy à la fois dans les espaces $\mathcal{C}^1\left(0,T;L^2(\Omega)\right)$ et $\mathcal{C}^0\left(0,T;H_0^1(\Omega)\right)$ ce qui revient à montrer

$$\begin{aligned} &\text{(i)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \|u_p(t) - u_m(t)\|_{L^2(\Omega)} = 0, \\ &\text{(ii)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \|(u_p - u_m)'(t)\|_{L^2(\Omega)} = 0, \\ &\text{(iii)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \|\nabla u_p(t) - \nabla u_m(t)\|_{L^2(\Omega)^n} = 0. \end{aligned} \tag{4.25}$$

Rappelons en effet qu'en vertu de l'inégalité de Poincaré, $\|\nabla v\|_{L^2(\Omega)}$ définit une norme équivalente à la norme $H^1(\Omega)$ sur $H_0^1(\Omega)$. Après avoir décomposé u_m et u_p sur la base de vecteurs propres, (4.25) s'exprime sous la forme :

$$\begin{aligned} &\text{(i)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \sum_{i=m+1}^{p} (\alpha_i(t))^2 = 0, \\ &\text{(ii)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \sum_{i=m+1}^{p} (\alpha_i'(t))^2 = 0, \\ &\text{(iii)} \quad \lim_{m,p\to+\infty} \sup_{t\in[0,T]} \sum_{i=m+1}^{p} \lambda_i (\alpha_i(t))^2 = 0. \end{aligned} \tag{4.26}$$

Pour démontrer (4.26)-(i), nous utilisons l'estimation $|\sin s| \leq |s|$ pour majorer α_i (dont l'expression est donnée en (4.18)) :

$$|\alpha_i(t)| \leq |(u_0, v_i)| + t|(u_1, v_i)| + \int_0^t (t-s)|(f(s), v_i)| ds,$$

dont on déduit aisément

$$(\alpha_i(t))^2 \leq 3 \left((u_0, v_i)^2 + T^2 (u_1, v_i)^2 + T^2 \left(\int_0^T |(f(s), v_i)| ds \right)^2 \right).$$

Par conséquent

$$\sum_{i=m+1}^{p} (\alpha_i(t))^2 \leq C(T) \left(\sum_{i=m+1}^{p} (u_0, v_i)^2 + \sum_{i=m+1}^{p} (u_1, v_i)^2 + \sum_{i=m+1}^{p} \left(\int_0^T |(f(s), v_i)| ds \right)^2 \right),$$

où $C(T)$ est une constante ne dépendant que de T. Nous estimons le dernier terme :

$$\begin{aligned}
\sum_{i=m+1}^{p} \left(\int_0^T |(f(s), v_i)| ds \right)^2 &= \sum_{i=m+1}^{p} \int_0^T \int_0^T |(f(s), v_i)| \, |(f(\tau), v_i)| \, ds d\tau \\
&= \int_0^T \int_0^T \sum_{i=m+1}^{p} |(f(s), v_i)| \, |(f(\tau), v_i)| \, ds d\tau \\
&\leq \int_0^T \int_0^T \left(\sum_{i=m+1}^{p} (f(s), v_i)^2 \right)^{1/2} \left(\sum_{i=m+1}^{p} (f(\tau), v_i)^2 \right)^{1/2} ds d\tau \\
&\leq \left(\int_0^T \left(\sum_{i=m+1}^{p} (f(s), v_i)^2 \right)^{1/2} ds \right)^2 .
\end{aligned}$$

On obtient finalement

$$\sum_{i=m+1}^{p} \left(\int_0^T |(f(s), v_i)| ds \right)^2 \leq \left(\int_0^T \left(\sum_{i=m+1}^{p} (f(s), v_i)^2 \right)^{1/2} ds \right)^2 . \tag{4.27}$$

En reportant dans l'estimation précédente, on a

$$\sum_{i=m+1}^{p} (\alpha_i(t))^2 \leq C(T) \left(\sum_{i=m+1}^{p} (u_0, v_i)^2 + \sum_{i=m+1}^{p} (u_1, v_i)^2 + \left(\int_0^T \left(\sum_{i=m+1}^{p} (f(s), v_i)^2 \right)^{1/2} ds \right)^2 \right).$$

On obtient donc (4.26)-(i) en utilisant les hypothèses sur les données (4.22), (4.23)-(a) et (4.24). Pour démontrer (4.26)- (ii), nous dérivons α_i par rapport au temps :

$$\alpha_i'(t) = -(u_0, v_i) c\sqrt{\lambda_i} \, \sin(c\sqrt{\lambda_i}\, t) + (u_1, v_i) \, \cos(c\sqrt{\lambda_i}\, t) + \int_0^t \cos(c\sqrt{\lambda_i}\, (t-s)) \, (f(s), v_i) \, ds,$$

d'où on déduit

$$\sum_{i=m+1}^{p} (\alpha_i'(t))^2 \leq C(T) \left(c^2 \sum_{i=m+1}^{p} \lambda_i (u_0, v_i)^2 + \sum_{i=m+1}^{p} (u_1, v_i)^2 + \sum_{i=m+1}^{p} \left(\int_0^T |(f(s), v_i)| ds \right)^2 \right),$$

où $C(T)$ est un nom générique pour une constante ne dépendant que de T (constante différente à chaque fois). L'estimation (4.27) et les hypothèses (4.22), (4.23)-(b) et (4.24) permettent de conclure.
Enfin, pour démontrer (4.26)-(iii), nous utilisons l'estimation

$$c\sqrt{\lambda_i} |\alpha_i(t)| \leq |(u_0, v_i)| c\sqrt{\lambda_i} + |(u_1, v_i)| + \int_0^t |(f(s), v_i)| ds,$$

qui implique

$$c^2 \lambda_i (\alpha_i(t))^2 \leq C(T) \left(c^2 \lambda_i (u_0, v_i)^2 + (u_1, v_i)^2 + \left(\int_0^T |(f(s), v_i)| ds \right)^2 \right).$$

On en déduit

$$c^2 \sum_{i=m+1}^{p} \lambda_i (\alpha_i(t))^2 \leq C(T) \left(c^2 \sum_{i=m+1}^{p} \lambda_i (u_0, v_i)^2 + \sum_{i=m+1}^{p} (u_1, v_i)^2 + \sum_{i=m+1}^{p} \left(\int_0^T |(f(s), v_i)| ds \right)^2 \right)$$

et de nouveau, l'estimation (4.27) ainsi que les hypothèses (4.22), (4.23)-(b) et (4.24), permettent de conclure.
En conclusion, on a montré (4.26), qui est équivalent à (4.25) et qui montre que la suite $(u_m)_m$ est une suite de Cauchy dans $\mathcal{C}^1(0,T;L^2(\Omega))$ et dans $\mathcal{C}^0(0,T;H_0^1(\Omega))$. Puisque ces deux espaces sont complets, la suite $(u_m)_m$ converge dans chacun de ces espaces. Par ailleurs, comme les injections canoniques $\mathcal{C}^0(0,T;H_0^1(\Omega)) \subset L^2(0,T;L^2(\Omega))$ et $\mathcal{C}^1(0,T;L^2(\Omega)) \subset L^2(0,T;L^2(\Omega))$ sont continues, la limite de $(u_m)_m$ est la même dans les deux espaces. On a donc prouvé que :

$$u_m \underset{m\to+\infty}{\longrightarrow} u \text{ dans } \mathcal{C}^0(0,T;H_0^1(\Omega)) \cap \mathcal{C}^1(0,T;L^2(\Omega)). \tag{4.28}$$

– Il reste à démontrer que u est solution du problème (4.15).
D'après (4.20), on a $\forall m \geq l \geq 1,\ \forall v \in V_l,\ \forall \psi \in \mathcal{D}(]0,T[)$:

$$\int_0^T (u_m(t), v) \frac{d^2\psi}{dt^2}(t) dt + \int_0^T \left(\int_\Omega c^2 \nabla u_m(t) \cdot \nabla v \, d\Omega \right) \psi(t) dt = \int_0^T (f(t), v) \psi(t) dt.$$

Passons à la limite, lorsque $m \to +\infty$. Compte-tenu de (4.28), on a, $\forall v \in V_l$:

$$\int_0^T (u(t), v) \frac{d^2\psi}{dt^2}(t) dt + \int_0^T \left(\int_\Omega c^2 \nabla u(t) \cdot \nabla v \, d\Omega \right) \psi(t) dt = \int_0^T (f(t), v) \psi(t) dt. \tag{4.29}$$

Comme $\cup_{l\geq 1} V_l$ est dense dans $H_0^1(\Omega)$ car $(v_i)_i$ est une base de $H_0^1(\Omega)$, la relation (4.29) a lieu pour tout $v \in H_0^1(\Omega)$, ce qui montre que u vérifie (4.15) au sens des distributions sur $]0,T[$.

En outre d'après (4.28), on a en particulier :

$$u_m(0) \to u(0) \quad \text{dans } H_0^1(\Omega),$$

et comme :

$$u_m(0) = \sum_{i=1,m} (u_0, v_i)\, v_i \underset{m\to+\infty}{\longrightarrow} u_0 \quad \text{dans } H_0^1(\Omega),$$

on en déduit que $u(0) = u_0$ sur Ω, au sens des fonctions de $H_0^1(\Omega)$. De même, on a

$$\frac{du_m}{dt}(0) \to \frac{du}{dt}(0) \quad \text{dans } L^2(\Omega),$$

et comme :

$$\frac{du_m}{dt}(0) = \sum_{i=1,m} (u_1, v_i)\, v_i \underset{m\to+\infty}{\longrightarrow} u_1 \quad \text{dans } L^2(\Omega),$$

on en déduit que $\dfrac{du}{dt}(0) = u_1$ dans $L^2(\Omega)$. Ceci montre que u est bien la solution de (4.15).
Enfin, en vertu de la proposition 4.7 il est clair que la solution du problème (4.15) est unique. ■

Remarque 4.9 *On peut généraliser sans difficulté cette démonstration au problème abstrait de la remarque 4.6.*

4.3 Propriétés de l'équation des ondes

4.3.1 Estimations d'énergie et estimations a priori

Estimations d'énergie et caractère conservatif

Au cours de cette sous-section, nous notons toujours u la solution du problème (4.15). D'après le théorème 4.8, on sait que $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$. Par conséquent, pour presque tout t, on peut choisir dans la formulation variationnelle (4.17) $v(x,t) = \partial_t u(x,t)$ et intégrer sur l'intervalle $]0,t[$. On obtient ainsi l'égalité

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(s)\frac{\partial u}{\partial t}(s)\,d\Omega + \int_\Omega c^2 \nabla u(s)\cdot\frac{\partial \nabla u}{\partial t}(s)\,d\Omega = \int_\Omega f(s)\frac{\partial u}{\partial t}(s)\,d\Omega,$$

soit encore :

$$\frac{1}{2}\frac{d}{dt}\left(\int_\Omega \left(\frac{\partial u}{\partial t}\right)^2(s)\,d\Omega\right) + \frac{1}{2}\frac{d}{dt}\left(\int_\Omega |c\nabla u(s)|^2\,d\Omega\right) = \int_\Omega f(s)\frac{\partial u}{\partial t}(s)\,d\Omega,$$

qui conduit à l'identité d'énergie :

Lemme 4.10 (identité d'énergie) *On a l'identité*

$$\frac{dE(t)}{dt} = (f(t), \frac{\partial u}{\partial t}(t))_{L^2(\Omega)}, \tag{4.30}$$

où $E(t)$ est l'énergie à l'instant t définie par

$$E(t) = \frac{1}{2}\|\frac{\partial u}{\partial t}(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|c\nabla u(t)\|^2_{L^2(\Omega)^n}. \tag{4.31}$$

Cette identité est équivalente à

$$E(t) = \int_0^t (f(s), \frac{\partial u}{\partial t}(s))_{L^2(\Omega)}\,ds + E_0, \tag{4.32}$$

où $E_0 = E(0)$ est l'énergie à l'instant 0 :

$$E_0 = \frac{1}{2}\|u_1\|^2_{L^2(\Omega)} + \frac{1}{2}\|c\nabla u_0\|^2_{L^2(\Omega)^n}. \tag{4.33}$$

Cette égalité permet de démontrer facilement l'unicité de la solution du problème (4.15).

Proposition 4.11 *Le problème (4.15) admet une unique solution.*

Démonstration : Soient u_1 et u_2 deux solutions de (4.15). Posons $w = u_1 - u_2$. Alors il est clair, compte-tenu de la linéarité de (4.15), que w vérifie (d'après (4.32)) :

$$\|\frac{\partial w}{\partial t}(t)\|^2_{L^2(\Omega)} + \|c\nabla w(t)\|^2_{L^2(\Omega)^n} = 0,$$

qui montre que $w(x,t) \equiv 0$ p.p $x \in \Omega$, $t \in]0,T[$. ∎

L'égalité (4.32) montre que l'équation des ondes est *conservative*, c'est-à-dire qu'en l'absence de source ($f \equiv 0$), l'énergie est conservée.

Proposition 4.12 *Soit u la solution du problème (4.15) avec $f \equiv 0$, alors :*

$$E(t) = E_0, \quad \forall t \geq 0.$$

Démonstration : immédiat d'après (4.32). ∎

Remarque 4.13 *Cet effet conservatif de l'équation des ondes est à comparer à l'effet dissipatif de l'équation de la chaleur (voir §3.2.1).*

Estimations a priori et résultats de régularité

Nous allons maintenant déduire de l'identité d'énergie (4.32) des estimations a priori sur la solution, c'est-à-dire des estimations de certaines normes de la solution u sans connaître son expression. Ces résultats montrent la dépendance continue de la solution par rapport aux données.

Proposition 4.14 (continuité des solutions) *La solution $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H^1_0(\Omega))$ du problème (4.15) dépend continûment des données $u_0 \in H^1_0(\Omega)$, $u_1 \in L^2(\Omega)$ et $f \in L^1(0,T;L^2(\Omega))$. Plus précisément on a les estimations suivantes, pour tout $t \geq 0$:*

$$\begin{aligned}
&\|\frac{du}{dt}(t)\|_{L^2(\Omega)} \leq (2E_0)^{1/2} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds \ , \\
&\|c\nabla u(t)\|_{L^2(\Omega)^n} \leq (2E_0)^{1/2} + \int_0^t \|f(s)\|_{L^2(\Omega)} ds \ , \\
&\|u(t)\|_{L^2(\Omega)} \leq \|u_0\|_{L^2(\Omega)} + t(2E_0)^{1/2} + \int_0^t (t-s)\|f(s)\|_{L^2(\Omega)} ds \ .
\end{aligned} \tag{4.34}$$

Démonstration : grâce à l'inégalité de Cauchy-Schwarz, nous avons

$$|(f(s), \frac{du}{dt}(s))_{L^2(\Omega)}| \ \leq \| \ \frac{du}{dt}(s) \ \|_{L^2(\Omega)} \ \| \ f(s) \ \|_{L^2(\Omega)},$$

soit encore, comme $\| \ \frac{du}{dt}(s) \ \|^2_{L^2(\Omega)} \leq \ 2E(s)$:

$$|(f(s), \frac{du}{dt}(s))_{L^2(\Omega)}| \ \leq \ \sqrt{2} \ E(s)^{\frac{1}{2}} \ \| \ f(s) \ \|_{L^2(\Omega)} \ .$$

En reportant dans (4.32), nous obtenons :

$$E(t) \ \leq \ E_0 + \sqrt{2}\int_0^t \| f(s) \|_{L^2(\Omega)} \ E(s)^{\frac{1}{2}} ds. \tag{4.35}$$

Pour conclure nous utilisons le lemme de Gronwall 3.13, avec :

$$\varphi(t) = E(t), \quad m(t) = \sqrt{2} \| f(t) \|_{L^2(\Omega)}, \quad C = E_0, \quad \gamma = \frac{1}{2}.$$

Il vient :

$$E(t) \leq \{E_0^{\frac{1}{2}} + \frac{1}{\sqrt{2}}\int_0^t \| f(s) \|_{L^2(\Omega)} \ ds\}^2. \tag{4.36}$$

Pour obtenir les deux premières estimations de (4.34) il suffit alors de remarquer que

$$\| \frac{du}{dt}(t) \|_{L^2(\Omega)} \ \text{ et } \ \| \nabla u(t) \|_{L^2(\Omega)^n}$$

sont majorés par $\sqrt{2E(t)}$. Enfin pour obtenir la dernière estimation, il suffit d'écrire que

$$u(t) = u_0 + \int_0^t \frac{du}{dt}(s) ds,$$

et d'utiliser la première estimation. ∎

Remarque 4.15 *Nous avons choisi de passer par la forme intégrée (4.32) de l'identité (4.30) et par le lemme de Gronwall 3.13 en raison du caractère fondamental de ce lemme dans beaucoup d'applications. On peut toutefois utiliser un raccourci en remarquant que (4.30) entraîne :*

$$\frac{d}{dt}E(t) \leq \ \sqrt{2} \ E(t)^{\frac{1}{2}} \ \| f(t) \|_{L^2(\Omega)},$$

ce qui implique aussi, en remarquant que $E = (E^{1/2})^2$,

$$\frac{d}{dt}E(t)^{1/2} \leq \ \frac{\sqrt{2}}{2} \ \| f(t) \|_{L^2(\Omega)},$$

En intégrant cette inéquation différentielle, on obtient (4.36).

Remarque 4.16 *Compte tenu de l'expression de* E_0, *les estimations (4.34) impliquent :*

$$\begin{aligned} &\|\frac{du}{dt}(t)\|_{L^2(\Omega)} + \|c\nabla u(t)\|_{L^2(\Omega)^n} \leq c\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} + \|f\|_{L^1(0,T;L^2(\Omega))}, \\ &\|u(t)\|_{L^2(\Omega)} \leq (1+T)c\|u_0\|_{H^1(\Omega)} + T\|u_1\|_{L^2(\Omega)} + T\|f\|_{L^1(0,T;L^2(\Omega))}. \end{aligned} \tag{4.37}$$

Le résultat de la proposition 4.14 est donc bien un résultat de continuité de la solution par rapport aux données. Si on introduit l'espace fonctionnel

$$W(0,T) = \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega)), \tag{4.38}$$

muni de la norme

$$\| u \|_{W(0,T)} = \sup_{[0,T]} \{ \| u(t) \|_{L^2(\Omega)} + \| \frac{du}{dt}(t) \|_{L^2(\Omega)} + \| \nabla u(t) \|_{L^2(\Omega)^n} \}, \tag{4.39}$$

qui en fait un espace de Banach. Les estimations (4.37) s'expriment de la façon suivante :

$$\|u\|_{W(0,T)} \leq C(T)(\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} + \|f\|_{L^1(0,T;L^2(\Omega))}).$$

Remarque 4.17 *Ce sont les estimations obtenues à partir de l'énergie qui déterminent dans quel espace chercher la solution et qui expliquent la différence d'espaces pour l'équation de la chaleur et pour l'équation des ondes.*
– *Pour l'équation des ondes, l'identité d'énergie permet de contrôler*

$$\|\frac{du}{dt}(t)\|^2_{L^2(\Omega)} + \|\nabla u(t)\|^2_{L^2(\Omega)^n},$$

pour tout t, il est donc naturel de chercher la solution u telle que $\partial_t u$ et les composantes de ∇u appartiennent à $\mathcal{C}^0(0,T;L^2(\Omega))$, soit $u \in W(0,T)$.
– *Pour l'équation de la chaleur, la quantité naturellement contrôlée à partir de l'identité d'énergie est $\|u(t)\|^2_{L^2(\Omega)} + \int_0^T \|\nabla u(s)\|^2_{L^2(\Omega)^n} ds$. L'espace naturel est donc $\mathcal{C}^0(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$.*

Nous avons vu au chapitre 3 que l'équation de la chaleur avait un effet régularisant lié à sa non réversibilité et à la propriété de propagation à vitesse infinie. Pour l'équation des ondes, le comportement des solutions est différent. Il n'y a pas d'effet régularisant (les singularités se propageant le long des courbes caractéristiques), l'équation est réversible en temps et la propagation des ondes s'effectue à vitesse finie. Ces différences de comportement réapparaissent lors de la discrétisation en temps et conduisent à des schémas numériques de nature différente.
La régularité de la solution augmente pour l'équation des ondes avec la régularité des données, à condition que Ω soit très régulier. Plus précisément nous avons les deux résultats suivants que nous admettrons (voir par exemple [34] pour plus de détails).

Proposition 4.18 *On suppose que Ω est un ouvert de classe $\mathcal{C}^\infty$ (non nécessairement borné) et on fait de plus les hypothèses*

$$\begin{cases} (u_0, u_1) \in H^{m+1}(\Omega) \times H^m(\Omega), \\ f \in \mathcal{C}^{m+1}(0,T;L^2(\Omega)) \cap \bigcap_{j=0}^{m-1} \mathcal{C}^j(0,T;H^{m-j}(\Omega)). \end{cases} \tag{4.40}$$

Dans ce cas, la solution u du problème (4.3) possède la régularité

$$u \in \bigcap_{j=0}^{m+1} \mathcal{C}^j(0,T;H^{m+1-j}(\Omega)). \tag{4.41}$$

Proposition 4.19 *On suppose toujours que Ω est un ouvert de classe $\mathcal{C}^\infty$. Si on fait de plus les hypothèses*

$$\begin{cases} (u_0,u_1) \in \mathcal{D}(\Omega)\times\mathcal{D}(\Omega), \\ f \in \mathcal{C}^\infty(0,T;\mathcal{D}(\Omega)), \end{cases} \tag{4.42}$$

la solution u du problème (4.3) vérifie

$$u \in \mathcal{C}^\infty(0,T;\mathcal{D}(\Omega)). \tag{4.43}$$

Autrement dit, lorsque les données du problèmes sont très régulières, la solution est elle-même très régulière.

Remarque 4.20 *Les résultats précédents sont issus d'un résultat abstrait qu'on peut trouver dans [34]. Dans [34], l'application est faite à l'équation des ondes posée dans tout l'espace. Ils sont encore vrais si l'ouvert a la régularité $\mathcal{C}^\infty$. Dans le cas d'un ouvert moins régulier, la régularité du bord intervient et les résultats énoncés aux propositions 4.18 et 4.19 ne sont plus vrais. La régularité de la solution est plus compliquée à exprimer et nous ne l'aborderons pas dans ce cours.*

4.3.2 Propagation à vitesse finie

Nous montrons dans ce paragraphe une des propriétés caractéristiques de l'équation des ondes, et qui diffère de l'équation de la chaleur, la propagation à vitesse finie de la solution. Nous nous plaçons dans le cas où le domaine de propagation est l'espace tout entier, $\Omega = \mathbb{R}^n$.

Théorème 4.21 *Si les données (u_0,u_1,f) ont la régularité (4.16) et satisfont :*

$$supp\ u_0 \cup\ supp\ u_1 \subset K, \quad \forall t \geq 0,\ \ supp\ f(.,t) \subset K,$$

où K est un compact fixe, alors l'unique solution faible du problème (4.15) satisfait :

$$\forall 0 \leq t \leq T, \quad supp\ u(.,t) \subset K + B(0,ct).$$

Démonstration : Nous faisons une démonstration en trois parties : la première repose sur une technique d'estimation d'énergie dans un domaine mobile, les deux dernières sont purement techniques. Nous nous restreignons au cas où les données sont régulières et K est une boule et nous renvoyons à [34] pour les autres cas. Nous supposons que :

$$\begin{cases} (u_0, u_1) \in \mathcal{D}(\mathbb{R}^n) \times \mathcal{D}(\mathbb{R}^n), \\ f \in \mathcal{C}^\infty(0, T; \mathcal{D}(\mathbb{R}^n)). \end{cases}$$

Nous savons alors que la solution est de classe $\mathcal{C}^\infty$ en temps et en espace. Ceci justifie tous les calculs qui vont suivre.

Nous allons de plus supposer que K est une boule, par exemple la boule de centre 0 et de rayon $R > 0$: $K = B(0, R)$. Nous considérons maintenant, ν désignant un vecteur quelconque de la sphère unité de $\mathbb{R}^n$, le demi-espace mobile

$$\Omega_\nu^t = \{x \in \mathbb{R}^n;\ x \cdot \nu > X(t)\},$$

où $X(t)$ est défini par :

$$\begin{cases} \dfrac{dX}{dt}(t) = V > 0 \quad (V \text{ à déterminer}), \\ X(0) = R. \end{cases}$$

Ω_ν^t est donc un demi-espace mobile qui "fuit" dans la direction ν à la vitesse V (voir figure 4.3). Nous désignerons par Γ_ν^t la frontière de Ω_ν^t ($\Gamma_\nu^t = \{x \in \mathbb{R}^n;\ x \cdot \nu = X(t)\}$ est un hyperplan orthogonal à ν) et par $d\sigma$ la mesure surfacique sur Γ_ν^t. On notera que par construction :

$$\begin{cases} \forall\ x \in \Omega_\nu^0, \quad u_0(x) = u_1(x) = 0, \\ \forall\ t > 0,\ \forall\ x \in \Omega_\nu^t, \quad f(x,t) = 0. \end{cases} \tag{4.44}$$

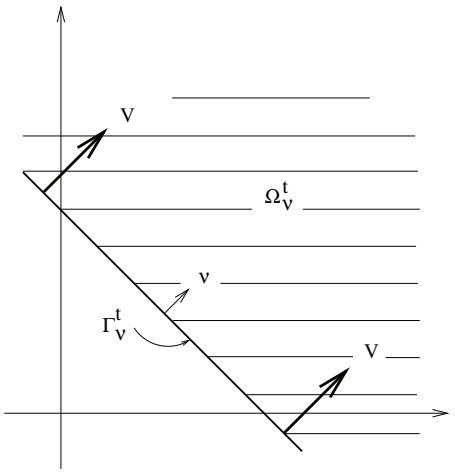


Figure 4.3. Le demi-espace mobile

A l'instant $t = 0$, la solution est nulle dans Ω_ν^t. L'idée de la démonstration est de trouver V (assez grand) pour que la solution u ne "pénètre" jamais Ω_ν^t, autrement dit pour que le demi-espace Ω_ν^t se propage plus vite que la solution dans la direction ν. Ceci passe par une identité d'énergie.

Etape 1 : Obtention d'une identité d'énergie. Nous allons nous intéresser à l'évolution de la fonction du temps :

$$E(\Omega_\nu^t, t) = \int_{\Omega_\nu^t} e(x,t)\ dx, \tag{4.45}$$

où nous avons défini la densité d'énergie :

$$e(x,t) = \frac{1}{2}\left(\left(\frac{\partial u}{\partial t}(x,t)\right)^2 + |c\nabla u(x,t)|^2\right). \tag{4.46}$$

Par définition $E(\Omega_\nu^t, t)$ est l'énergie de la solution contenue à l'instant t dans le domaine Ω_ν^t. Nous allons calculer sa dérivée en temps. Pour cela, nous utilisons le :

Lemme 4.22 *Soit $F \in \mathcal{C}^1(0,T;L^1(\mathbb{R}^n))$, il vient :*

$$\frac{d}{dt}\left(\int_{\Omega_\nu^t} F(x,t)\,dx\right) = \int_{\Omega_\nu^t} \frac{\partial F}{\partial t}(x,t)\,dx - V\int_{\Gamma_\nu^t} F(x,t)\,d\sigma. \tag{4.47}$$

Démonstration : Utilisons un changement de coordonnées par rotation

$$(x_1, x_2, ..., x_n) \mapsto (X_1, X_2, ..., X_n),$$

tel que l'axe $0X_n$ soit orienté par le vecteur ν. Nous avons :

$$(x_1, ...x_n) \in \Omega_\nu^t \Longleftrightarrow X_n > R + Vt.$$

Posons $X = (X', X_n)$ et $F(x,t) = \tilde{F}(X', X_n, t)$. Il vient :

$$\int_{\Omega_\nu^t} F(x,t)\,dx = \int_{\mathbb{R}^{n-1}} \left(\int_{R+Vt}^{+\infty} \tilde{F}(X', X_n, t) dX_n\right) dX',$$

$$\begin{aligned}
\frac{d}{dt}\left(\int_{\Omega_\nu^t} F(x,t)dx\right) &= \int_{\mathbb{R}^{n-1}} \left[\frac{d}{dt}\left(\int_{R+Vt}^{+\infty} \tilde{F}(X', X_n, t)dX_n\right)\right] dX' \\
&= \int_{\mathbb{R}^{n-1}} \int_{R+Vt}^{+\infty} \frac{\partial \tilde{F}}{\partial t}(X', X_n, t)\,dX_n\,dX' - V\int_{\mathbb{R}^{n-1}} \tilde{F}(X', R+Vt, t)\,dX' \\
&= \int_{\Omega_\nu^t} \frac{\partial F}{\partial t}(x,t)\,dx - V\int_{\Gamma_\nu^t} F(x,t)\,d\sigma.
\end{aligned}$$

■

Appliquons le lemme avec $F = e$. Nous obtenons

$$\frac{d}{dt}E(\Omega_\nu^t, t) = \int_{\Omega_\nu^t} \left(\frac{\partial^2 u}{\partial t^2}\frac{\partial u}{\partial t} + c^2\nabla\frac{\partial u}{\partial t}\cdot\nabla u\right) dx - \frac{V}{2}\int_{\Gamma_\nu^t}\left(\left(\frac{\partial u}{\partial t}\right)^2 + |c\nabla u|^2\right) d\sigma.$$

Par la formule de Green, en notant que le vecteur normal à Γ_ν^t extérieur à Ω_ν^t est $-\nu$, on a

$$\int_{\Omega_\nu^t} \nabla\frac{\partial u}{\partial t}\cdot\nabla u = -\int_{\Omega_\nu^t} \Delta u\frac{\partial u}{\partial t}\,dx - \int_{\Gamma_\nu^t} \frac{\partial u}{\partial \nu}\frac{\partial u}{\partial t}\,d\sigma.$$

Par suite, compte tenu de (4.44) et de l'équation satisfaite par u avec $f = 0$ dans Ω_ν^t, il vient

$$\frac{d}{dt}E(\Omega_\nu^t, t) = -\frac{1}{2}\int_{\Gamma_\nu^t}\left\{V\left(\left(\frac{\partial u}{\partial t}\right)^2 + |c\nabla u|^2\right) + 2c^2\frac{\partial u}{\partial \nu}\frac{\partial u}{\partial t}\right\} d\sigma.$$

Décomposons ∇u en la somme de sa partie tangentielle et de sa partie normale à Γ_ν^t :

$$\nabla u = \nabla_\Gamma u + \frac{\partial u}{\partial \nu}\,\nu, \quad \nabla_\Gamma u \perp \nu.$$

Comme $|\nabla u|^2 = |\nabla_\Gamma u|^2 + \left(\frac{\partial u}{\partial \nu}\right)^2$, nous obtenons :

$$\frac{d}{dt}E(\Omega_\nu^t,t) = -\frac{1}{2}\int_{\Gamma_\nu^t} V|c\nabla_\Gamma u|^2\,d\sigma - \frac{1}{2}\int_{\Gamma_\nu^t}\left\{V\left(\left(\frac{\partial u}{\partial t}\right)^2 + \left(c\frac{\partial u}{\partial \nu}\right)^2\right) + 2c^2\frac{\partial u}{\partial \nu}\frac{\partial u}{\partial t}\right\}\,d\sigma.$$

Après avoir remarqué que

$$\begin{aligned} V\left(\left(\frac{\partial u}{\partial t}\right)^2 + \left(c\frac{\partial u}{\partial \nu}\right)^2\right) + 2c^2\frac{\partial u}{\partial \nu}\frac{\partial u}{\partial t} &= V\left(\left(\frac{\partial u}{\partial t}\right)^2 + 2\frac{c^2}{V}\,\frac{\partial u}{\partial \nu}\frac{\partial u}{\partial t} + \left(c\frac{\partial u}{\partial \nu}\right)^2\right), \\ &= V\left(\left(\frac{\partial u}{\partial t} + \frac{c^2}{V}\frac{\partial u}{\partial \nu}\right)^2 + c^2(1-\frac{c^2}{V^2})\left(\frac{\partial u}{\partial \nu}\right)^2\right), \end{aligned}$$

nous aboutissons finalement à l'identité :

$$\begin{aligned} \frac{d}{dt}E(\Omega_\nu^t,t) = -\frac{1}{2}\int_{\Gamma_\nu^t} V|c\nabla_\Gamma u|^2\,d\sigma - \frac{1}{2}\int_{\Gamma_\nu^t} V\left(\frac{\partial u}{\partial t} + \frac{c^2}{V}\frac{\partial u}{\partial \nu}\right)^2\,d\sigma \\ -\frac{1}{2}\int_{\Gamma_\nu^t} c^2V(1-\frac{c^2}{V^2})\,\left(\frac{\partial u}{\partial \nu}\right)^2\,d\sigma. \end{aligned} \tag{4.48}$$

Etape 2 : L'idée est de choisir V assez grand pour que la fonction $t \mapsto E(\Omega_\nu^t,t)$ soit décroissante. D'après (4.48) pour avoir

$$\frac{d}{dt}E(\Omega_\nu^t,t) \leq 0,$$

il suffit que :

$$1 - \frac{c^2}{V^2} \geq 0 \quad \forall x \in \mathbb{R}^n,$$

et il suffit donc de choisir $V = c$.

Etape 3 : Majoration du support de la solution. Avec $V = c$ on a donc en particulier l'inégalité :

$$E(\Omega_\nu^t,t) \leq E(\Omega_\nu^t,0).$$

Mais par construction (si A désigne un sous-ensemble de $\mathbb{R}^n$, A^c désigne son complémentaire dans $\mathbb{R}^n$), on a

$$E(\Omega_\nu^t,0) = 0 \Longrightarrow E(\Omega_\nu^t,t) = 0 \Longrightarrow u(x,t) = 0 \text{ dans } \Omega_\nu^t \Longrightarrow \text{supp } u(.,t) \subset \{\Omega_\nu^t\}^c.$$

Le raisonnement étant valable pour tout $\nu \in S^{n-1}$, sphère unité de $\mathbb{R}^n$, on a :

$$\text{supp } u(.,t) \subset \bigcap_{\nu\in S^{n-1}} \{\Omega_\nu^t\}^c = B(0,R+ct) = B(0,R) + B(ct).$$

Le résultat est donc démontré lorsque K est une boule. ∎

4.3.3 Fonction de Green de l'équation des ondes

Dans ce paragraphe, nous introduisons très brièvement la notion de fonction de Green (ou solution élémentaire). Nous renvoyons pour plus de détails à [34].

La fonction de Green (ou noyau de Green) est une solution particulière de l'équation des ondes posée dans tout l'espace en milieu homogène, grâce à laquelle on peut obtenir une expression explicite de n'importe quelle autre solution (toujours du problème posé dans tout l'espace en milieu homogène). En effet, considérons le système posé dans tout l'espace homogène :

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,t) - c^2 \Delta u(x,t) = f(x,t), & x \in \mathbb{R}^n, t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}^n, \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x), & x \in \mathbb{R}^n. \end{cases} \tag{4.49}$$

Désignons par $\hat{u}(\kappa,t)$ la transformée de Fourier en espace de $u(x,t)$ et appliquons la transformation de Fourier à (4.49) : nous obtenons pour tout $\kappa \in \mathbb{R}^n$ l'équation différentielle ordinaire

$$\begin{cases} \dfrac{d^2 \hat{u}}{dt^2}(\kappa,t) + c^2|\kappa|^2 \hat{u}(\kappa,t) = \hat{f}(\kappa,t), \\ \hat{u}(\kappa,0) = \hat{u}_0(\kappa), \\ \dfrac{d\hat{u}}{dt}(\kappa,0) = \hat{u}_1(\kappa), \end{cases} \tag{4.50}$$

qui se résout aisément :

$$\begin{aligned} \hat{u}(\kappa,t) = \hat{u}_0(\kappa)\cos(c|\kappa|t) + \hat{u}_1(\kappa)\left[\frac{\sin(c|\kappa|t)}{c|\kappa|}\right] \\ + \int_0^t \frac{\sin(c|\kappa|(t-s))}{c|\kappa|}\ \hat{f}(\kappa,s)\ ds. \end{aligned} \tag{4.51}$$

Introduisons alors la fonction définie sur $\mathbb{R}^n \times \mathbb{R}^+$ par

$$\hat{G}(\kappa,t) = \frac{\sin(c|\kappa|t)}{c|\kappa|}, \quad \hat{G} \in \mathcal{C}^\infty(\mathbb{R}^n \times \mathbb{R}^+). \tag{4.52}$$

Il est facile de voir que la formule (4.51) se réécrit

$$\begin{aligned} \hat{u}(\kappa,t) = \frac{\partial}{\partial t}\left(\hat{G}(\kappa,t)\hat{u}_0(\kappa)\right) + \hat{G}(\kappa,t)\hat{u}_1(\kappa) \\ + \int_0^t \hat{G}(\kappa,t-s)\ \hat{f}(\kappa,s)\ ds. \end{aligned} \tag{4.53}$$

On peut alors expliciter la transformée de Fourier inverse de cette expression et obtenir

$$\begin{aligned} u(.,t) = \frac{\partial}{\partial t}\left(G(.,t) * u_0\right) + G(.,t) * u_1 \\ + \int_0^t G(.,t-s) * f(.,s) ds, \end{aligned} \tag{4.54}$$

où $*$ désigne le produit de convolution en espace. Autrement dit, on a

$$u(x,t) = \frac{\partial}{\partial t}\Big(\int G(x-y,t)u_0(y)\,dy\Big) + \int G(x-y,t)u_1(y)\,dy + \int_0^t \int G(x-y,t-s)f(y,s)\,dy\,ds.$$

On retrouve ici la structure de la formule de D'Alembert trouvée en dimension 1(cf. 4.1.1). Ceci montre que, pour $n=1$

$$G(x,t) = \frac{1}{2c}\chi_{[-ct,ct]}(x).$$

En dimensions supérieures, nous renvoyons à [34] pour le calcul de la fonction de Green et donnons ici le résultat :

– en dimension 3, la solution élémentaire est indifféremment définie par l'une de ces deux formules :

$$G(x,t) = \frac{1}{4\pi c^2 t}\,\delta(|x|-ct) = \frac{1}{4\pi c|x|}\,\delta(|x|-ct).$$

– La solution fondamentale de l'equation des ondes en dimension 2 est définie par :

$$G(x,t) = \begin{cases} \dfrac{1}{2\pi c}\dfrac{1}{\sqrt{c^2t^2-|x|^2}}, & \text{si } |x| \le ct, \\ 0, & \text{si } |x| > ct. \end{cases}$$

Nous pouvons remarquer que la singularité de $x \mapsto G(x,t)$ en tant que distribution sur $\mathbb{R}^n$ augmente avec la dimension d'espace n. Ainsi, on a

pour $n=1$: $G(.,t) \in L^2(\mathbb{R})$,
pour $n=2$: $G(.,t) \in L^1(\mathbb{R}^2)$ mais $G(.,t) \notin L^2(\mathbb{R}^2)$,
pour $n=3$: $G(.,t)$ est une mesure sur $\mathbb{R}^3$ mais $G(.,t) \notin L^1(\mathbb{R}^3)$.

On retrouve la propriété de propagation à vitesse finie, le support de la solution fondamentale étant toujours inclus dans $|x| \le ct$.

4.4 Semi-discrétisation en espace de l'équation des ondes

L'opérateur spatial étant le même pour l'équation des ondes que pour l'équation de la chaleur (le laplacien), il est naturel de suivre la même démarche que celle présentée à la section 3.3.1. Nous approchons donc la formulation variationnelle (4.15) en utilisant la méthode des éléments finis et nous renvoyons à la section 3.3.1 pour les définitions et notations. Rappelons que V_h désigne un sous-espace de $H_0^1(\Omega)$ de dimension finie, qui approche $H_0^1(\Omega)$ au sens suivant (cf. 1.54)

$$\lim_{h\to 0} \inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)} = 0, \quad \forall v \in H_0^1(\Omega). \tag{4.55}$$

4.4.1 Problème variationnel approché

L'approximation dans l'espace V_h du problème (4.3) conduit à la formulation variationnelle semi-discrète suivante

$$\begin{cases} \text{trouver } u_h \in \mathcal{C}^1(0,T;V_h) \text{ tel que :} \\ \dfrac{d^2}{dt^2}\displaystyle\int_\Omega u_h(t)\, v_h \, d\Omega + \int_\Omega c^2 \nabla u_h(t) \cdot \nabla v_h \, d\Omega = \int_\Omega f(t)\, v_h \, d\Omega, \qquad \begin{array}{c} \forall v_h \in V_h, \\ \text{p.p. } t \in]0,T[, \end{array} \\ u_h(0) = u_{h,0}, \dfrac{du_h}{dt}(0) = u_{h,1}, \end{cases} \tag{4.56}$$

où $u_{h,0} \in V_h$ approche u_0 dans $H_0^1(\Omega)$ et $u_{h,1} \in V_h$ approche u_1 dans $L^2(\Omega)$, c'est-à-dire :

$$\lim_{h\to 0} \|u_0 - u_{h,0}\|_{H^1(\Omega)} = 0, \quad \lim_{h\to 0} \|u_1 - u_{h,1}\|_{L^2(\Omega)} = 0. \tag{4.57}$$

Remarque 4.23 *On peut par exemple choisir comme conditions initiales approchées :*

$u_{h,0}$ *= la projection orthogonale de* u_0 *sur* V_h *pour le produit scalaire* $H_0^1(\Omega)$ *,*
$u_{h,1}$ *= la projection orthogonale de* u_1 *sur* V_h *pour le produit scalaire* $L^2(\Omega)$*,*

et montrer que ce choix vérifie bien (4.57) en utilisant le résultat suivant :

si V *et* H *sont 2 espaces de Hilbert, tels que* $V \subset H$*,* V *dense dans* H *et avec injection continue alors l'hypothèse (4.55) implique*

$$\lim_{h\to 0} \inf_{v_h \in V_h} |v - v_h|_H = 0, \quad \forall v \in H.$$

Remarque 4.24 (notation) *On utilisera parfois la notation* $a(u,v)$ *pour désigner le produit scalaire* $(\nabla u, \nabla v)_{L^2(\Omega)^n}$ *(notation introduite en (3.53)).*

Interprétation matricielle

Le problème discrétisé (4.56) est un système différentiel d'ordre N. En effet, notons $\overrightarrow{U}(t)$ le vecteur de composantes $U_I(t) = u_h(M_I,t)$, $I = 1,N$, $t \in [0,T[$. Par construction, on a :

$$u_h(t) = \sum_{I=1,N} u_h(M_I,t) w_I = \sum_{I=1,N} U_I(t)\, w_I. \tag{4.58}$$

En substituant cette expression dans (4.56) et en prenant $v_h = w_J$ on obtient :

$$\frac{d^2}{dt^2} \sum_{I=1,N} \left(\int_\Omega w_I \, w_J \, d\Omega \right) U_I(t) + \sum_{I=1,N} \left(\int_\Omega c^2 \nabla w_I \cdot \nabla w_J \, d\Omega \right) U_I(t) = \int_\Omega f(t) w_J \, d\Omega,$$

soit en notant :

– $\mathbb{K}$ la matrice symétrique de $\mathbb{R}^{N\times N}$ définie par $\mathbb{K}_{IJ} = \displaystyle\int_\Omega \nabla w_I \cdot \nabla w_J \, d\Omega$,

– $\mathbb{M}$ la matrice symétrique de $\mathbb{R}^{N\times N}$ définie par $\mathbb{M}_{IJ} = \displaystyle\int_\Omega w_I \, w_J \, d\Omega$,

– $\overrightarrow{F}(t)$ le vecteur de $\mathbb{R}^N$ défini par $F_I(t) = \displaystyle\int_\Omega f(t)\, w_I \, d\Omega$,

le système différentiel équivalent

$$\left\{\begin{array}{l} \text{trouver } \overrightarrow{U} \in \mathcal{C}^1(0,T;\mathbb{R}^N) \text{ tel que :} \\ \\ \dfrac{d^2}{dt^2}\mathbb{M}\overrightarrow{U}(t) + c^2\mathbb{K}\overrightarrow{U}(t) = \overrightarrow{F}(t) \qquad \text{p.p. } t \in]0,T[, \\ \\ \overrightarrow{U}(0) = \overrightarrow{U}_0, \dfrac{d\overrightarrow{U}}{dt}(0) = \overrightarrow{U}_1, \end{array}\right. \tag{4.59}$$

où $\overrightarrow{U}_\alpha$ est le vecteur de composantes $u_{h,\alpha}(M_I),\ I = 1, N$ pour $\alpha = 0, 1$. Rappelons que nous avons les correspondances suivantes

$$(\mathbb{M}\overrightarrow{U}|\overrightarrow{U}) = \int_\Omega u_h^2 \ d\Omega \quad ; \quad (\mathbb{K}\overrightarrow{U}|\overrightarrow{U}) = \int_\Omega |\nabla u_h|^2 \ d\Omega = a(u_h, u_h). \tag{4.60}$$

Existence d'une solution au problème semi-discrétisé

De la même façon que pour l'équation de la chaleur (voir section 3.3.1), on peut décomposer $\overrightarrow{U}(t)$ sur la base de vecteurs propres du problème (3.28) : on écrit

$$\overrightarrow{U}(t) = \sum_{m=1,N} \alpha_{m,h}(t)\overrightarrow{V}_m, \text{ avec } \alpha_{m,h}(t) = (\mathbb{M}\overrightarrow{U}(t)|\overrightarrow{V}_m). \tag{4.61}$$

En introduisant cette expression dans (4.59) et en faisant le produit scalaire usuel avec $\overrightarrow{V}_\ell$, on obtient, compte-tenu de (3.28) et des propriétés d'orthogonalité de la base de vecteurs propres les équations différentielles, $\forall m = 1, N$:

$$\left\{\begin{array}{l} \dfrac{d^2}{dt^2}\alpha_{m,h}(t) + c^2\lambda_{m,h}\alpha_{m,h}(t) = (\overrightarrow{F}(t)|\overrightarrow{V}_m), \\ \\ \alpha_{m,h}(0) = (\mathbb{M}\overrightarrow{U}_0|\overrightarrow{V}_m), \ \dfrac{d}{dt}\alpha_{m,h}(0) = (\mathbb{M}\overrightarrow{U}_1|\overrightarrow{V}_m). \end{array}\right. \tag{4.62}$$

Ces équations admettent pour solution

$$\begin{aligned} \alpha_{m,h}(t) = \cos\left(c\sqrt{\lambda_{m,h}}t\right) \ (\mathbb{M}\overrightarrow{U}_0|\overrightarrow{V}_m) + \frac{\sin\left(c\sqrt{\lambda_{m,h}}t\right)}{c\sqrt{\lambda_{m,h}}} \ (\mathbb{M}\overrightarrow{U}_1|\overrightarrow{V}_m) \\ + \int_0^t (\overrightarrow{F}(s)|\overrightarrow{V}_m)\frac{\sin\left(c\sqrt{\lambda_{m,h}}(t-s)\right)}{c\sqrt{\lambda_{m,h}}}ds, \quad m = 1, N. \end{aligned} \tag{4.63}$$

On en conclut finalement :

Proposition 4.25 *Le problème semi-discrétisé (4.56) admet une unique solution* $u_h \in \mathcal{C}^1(0,T;V_h)$ *donnée par :*

$$u_h(t) = \sum_{m=1,N} \alpha_{m,h}(t) v_{m,h}, \tag{4.64}$$

où $\alpha_{m,h}(t)$ *est donnée par (4.63) et* $v_{m,h}$ *par (3.29).*

Démonstration : elle est identique à celle de la proposition 3.21. ∎

Discrétisation par différences finies

En reprenant la démarche et les notations de la section 3.3.1, il est facile de voir qu'une approximation par différences finies de l'opérateur Δ conduit au système différentiel suivant :

$$\begin{cases} \dfrac{d^2}{dt^2}\underline{\overrightarrow{U}}(t) + c^2 \mathbb{D}\underline{\overrightarrow{U}}(t) = \underline{\overrightarrow{F}}(t) & \forall t \in]0,T[, \\ \underline{\overrightarrow{U}}(0) = \underline{\overrightarrow{U}}_0, \\ \dfrac{d}{dt}\underline{\overrightarrow{U}}(0) = \underline{\overrightarrow{U}}_1. \end{cases} \tag{4.65}$$

Le système différentiel (4.65) a une structure différente du système différentiel (4.59), issu de la discrétisation par éléments finis : on a le terme $\dfrac{d^2}{dt^2}\underline{\overrightarrow{U}}(t)$ au lieu du terme $\dfrac{d^2}{dt^2}\mathbb{M}\overrightarrow{U}(t)$. Toutefois, si on utilise la méthode de condensation de masse ou lumping (voir section 4.5.1) la matrice $\mathbb{M}$ devient diagonale et le système (4.59) retrouve une structure équivalente au système (4.65).

4.4.2 Estimation d'énergie semi-discrète et convergence du schéma

En suivant la même démarche que dans le cas continu (voir section 4.3.1), il est immédiat d'obtenir l'identité d'énergie semi-discrète :

Lemme 4.26 (identité d'énergie semi-discrète) *On a l'identité*

$$\frac{dE_h(t)}{dt} = (f(t), \frac{\partial u_h}{\partial t}(t))_{L^2(\Omega)}, \tag{4.66}$$

où $E_h(t)$ *est l'énergie à l'instant* t *définie par*

$$E_h(t) = \frac{1}{2}\|\frac{\partial u_h}{\partial t}(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|c\nabla u_h(t)\|^2_{L^2(\Omega)^n}. \tag{4.67}$$

Nous présentons ici une démonstration de convergence différente de celle présentée pour l'équation de la chaleur au paragraphe 3.5.1 et qui s'appuyait sur une représentation spectrale discrète de la solution. La démonstration présentée ici utilise comme ingrédients une estimation d'énergie et la projection elliptique définie par (3.52).
Soient u solution de (4.15) et u_h solution de (4.56). L'étude de la convergence suppose la plupart du temps plus de régularité. Nous faisons ici les hypothèses :

$$u \in \mathcal{C}^2(0,T;H_0^1(\Omega)), \quad u_h \in \mathcal{C}^2(0,T;V_h), \tag{4.68}$$

ce qui suppose en particulier

$$u_0 \in H_0^1(\Omega) \text{ et } u_1 \in H_0^1(\Omega). \tag{4.69}$$

Pour $u_{h,0}$ et $u_{h,1}$, nous faisons les hypothèses d'approximation usuelles (4.57). Puisque $\partial_{tt}(u-u_h) \in C^0(0,T;L^2(\Omega))$, l'erreur $u-u_h$ vérifie le problème suivant :

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial^2}{\partial t^2}(u-u_h)(t)v_h d\Omega + \int_\Omega c^2 \nabla(u-u_h)(t)\cdot\nabla v_h d\Omega = 0, \, \forall v_h \in V_h, \\ (u-u_h)(0) = u_0 - u_{h,0} \ \ , \dfrac{d(u-u_h)}{dt}(0) = u_1 - u_{h,1}. \end{cases} \tag{4.70}$$

La première équation peut encore s'écrire en introduisant $w_h \in V_h$ quelconque

$$\begin{aligned} &\int_\Omega \partial_{tt}^2(w_h-u_h)(t)v_h \, d\Omega + \int_\Omega c^2\nabla(w_h-u_h)(t)\cdot\nabla v_h \, d\Omega \\ &\quad = \int_\Omega \partial_{tt}^2(w_h-u)(t)v_h \, d\Omega + \int_\Omega \nabla(w_h-u)\cdot\nabla v_h \, d\Omega. \end{aligned} \tag{4.71}$$

En choisissant $v_h = \partial_t(w_h-u_h)$, cette identité se réécrit :

$$\begin{aligned} \frac{d}{dt}\mathcal{E}_h(t) &= \int_\Omega \partial_{tt}^2(w_h-u)(t)\partial_t(w_h-u_h)(t) \, d\Omega \\ &\quad + \int_\Omega c^2\nabla(w_h-u)(t)\cdot\nabla(\partial_t(w_h-u_h)(t)) \, d\Omega, \end{aligned} \tag{4.72}$$

où on a introduit l'énergie

$$\mathcal{E}_h(t) = \frac{1}{2}\left\|\partial_t(w_h-u_h)\right\|_{L^2(\Omega)}^2 + \frac{1}{2}\left\|c\nabla(w_h-u_h)\right\|_{L^2(\Omega)^n}^2. \tag{4.73}$$

L'idée est maintenant d'obtenir une estimation de cette énergie, ce qui va être possible en choisissant comme w_h particulier la projection elliptique de u. On montre ainsi le :

Lemme 4.27 *On a l'estimation d'énergie suivante :*

$$\mathcal{E}_h^{1/2}(t) \leq \mathcal{E}_h^{1/2}(0) + \frac{\sqrt{2}}{2} \int_0^t \left\| (I - P_h)\partial_{tt}^2 u(s) \right\|_{L^2(\Omega)} ds, \tag{4.74}$$

où $\mathcal{E}_h$ est défini par (4.73) avec $w_h = P_h u$ défini par (3.52).

Démonstration : en choisissant $w_h = P_h u$ dans (4.72), on a par définition de la projection elliptique :

$$\frac{d}{dt}\mathcal{E}_h(t) = (\partial_{tt}^2(w_h - u), \partial_t(w_h - u_h))_{L^2(\Omega)} \leq \left\| \partial_{tt}^2(w_h - u) \right\|_{L^2(\Omega)} \left\| \partial_t(w_h - u_h) \right\|_{L^2(\Omega)}.$$

En notant que $\frac{d}{dt}\mathcal{E}_h(t) = 2\mathcal{E}_h^{1/2}\frac{d}{dt}\mathcal{E}_h^{1/2}$ et $\|\partial_t(w_h - u_h)\|_{L^2(\Omega)} \leq (2\mathcal{E}_h)^{1/2}$ on obtient

$$\frac{d}{dt}\mathcal{E}_h^{1/2} \leq \frac{\sqrt{2}}{2} \left\| \partial_{tt}(w_h - u) \right\|_{L^2(\Omega)}.$$

On en déduit (4.74) en intégrant en temps. ∎

Théorème 4.28 *Si u est solution de (4.15) et u_h solution de (4.56), avec $u \in \mathcal{C}^2(0,T;H_0^1(\Omega))$ et $u_h \in \mathcal{C}^2(0,T;V_h)$, alors on a les estimations d'erreur suivantes :*

$$\begin{aligned}
&\text{(i)} \quad \|\partial_t(u - u_h)(t)\|_{L^2(\Omega)} \leq \sqrt{2}\mathcal{E}_h^{1/2}(0) + \|(I - P_h)\partial_t u(t)\|_{L^2(\Omega)} \\
&\qquad\qquad + \int_0^t \left\| (I - P_h)\partial_{tt}^2 u(s) \right\|_{L^2(\Omega)} ds, \\
&\text{(ii)} \quad \|(u - u_h)(t)\|_{H^1(\Omega)} \leq \sqrt{2}\mathcal{E}_h^{1/2}(0) + \|(I - P_h)u(t)\|_{H^1(\Omega)} \\
&\qquad\qquad + \int_0^t \left\| (I - P_h)\partial_{tt}^2 u(s) \right\|_{L^2(\Omega)} ds, \\
&\qquad \textit{pour tout } t \in]0,T[.
\end{aligned} \tag{4.75}$$

Démonstration : Par l'inégalité triangulaire, on a

$$\|\partial_t u - \partial_t u_h\|_{L^2(\Omega)} \leq \|\partial_t u - \partial_t P_h u\|_{L^2(\Omega)} + \underbrace{\|\partial_t(P_h u - u_h)\|_{L^2(\Omega)}}_{\leq (2\mathcal{E}_h)^{1/2}},$$

et on obtient (4.75)-(i) en utilisant (4.74). De même, on a

$$\|u - u_h\|_{H^1(\Omega)} \leq \|u - P_h u\|_{H^1(\Omega)} + \underbrace{\|P_h u - u_h\|_{H^1(\Omega)}}_{\leq (2\mathcal{E}_h)^{1/2}},$$

qui permet d'obtenir (4.75)- (ii). ∎

Les estimations d'erreur (4.75) montrent donc que la convergence de u_h vers u est liée, d'une part, à la bonne approximation des données initiales (qui vont jouer sur la convergence vers 0 de $\mathcal{E}_h(0)$) et d'autre part, à l'erreur de projection elliptique de u ou de ses dérivées en temps, en norme L^2 et en norme H^1. En ce qui concerne les données initiales, nous avons le :

Lemme 4.29
– Si on choisit comme conditions initiales approchées

$$u_{h,0} = P_h u_0, \quad et \quad u_{h,1} = P_h u_1, \tag{4.76}$$

alors $\mathcal{E}_h(0) = 0$.
– Pour d'autres choix de conditions initiales approchées on a

$$\mathcal{E}_h^{1/2}(0) \leq C \left(\|(I - P_h)u_1\|_{L^2(\Omega)} + \|u_1 - u_{h,1}\|_{L^2(\Omega)} + \|u_0 - u_{h,0}\|_{H^1(\Omega)} \right). \tag{4.77}$$

Démonstration : l'énergie discrète initiale s'écrit :

$$\mathcal{E}_h(0) = \frac{1}{2} \|P_h u_1 - u_{h,1}\|^2_{L^2(\Omega)} + \frac{1}{2} a(P_h u_0 - u_{h,0}, P_h u_0 - u_{h,0}),$$

on obtient donc immédiatement le premier point. Sinon, on utilise l'inégalité triangulaire pour majorer :

$$\|P_h u_1 - u_{h,1}\|_{L^2(\Omega)} \leq \|(I - P_h)u_1\|_{L^2(\Omega)} + \|u_1 - u_{h,1}\|_{L^2(\Omega)}.$$

D'autre part, on a vu que (cf. (3.54)

$$\|P_h u_0 - u_{h,0}\|_{H^1(\Omega)} \leq \frac{C_a}{\alpha} \|v_h - u_0\|_{H^1(\Omega)}, \quad \forall v_h \in V_h,$$

donc en particulier pour $v_h = u_{h,0}$:

$$\|P_h u_0 - u_{h,0}\|_{H^1(\Omega)} \leq \frac{C_a}{\alpha} \|u_{h,0} - u_0\|_{H^1(\Omega)}.$$

■

En ce qui concerne l'erreur de projection elliptique, nous savons d'après (3.55) que pour tout $v \in H_0^1(\Omega)$, $P_h v$ approche v dans $H_0^1(\Omega)$. Il est en fait possible d'obtenir une estimation en norme $L^2(\Omega)$, en faisant une hypothèse supplémentaire, en introduisant le problème adjoint (voir [15], définition 2.5). Rappelons que le problème adjoint est défini, pour tout $g \in L^2(\Omega)$, par :

$$\begin{cases} \text{trouver} \quad \varphi_g \in H_0^1(\Omega) \text{ tel que} \\ a(v, \varphi_g) = (g, v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega), \end{cases} \tag{4.78}$$

et qu'il est dit *régulier* si :

Définition 4.30 *Le problème adjoint (4.78) est dit régulier s'il existe une constante* $C > 0$ *telle que pour tout* $g \in L^2(\Omega)$*, la solution* φ_g *appartient à* $H_0^1(\Omega) \cap H^2(\Omega)$ *avec*

$$\|\varphi_g\|_{H^2(\Omega)} \leq C \|g\|_{L^2(\Omega)}. \tag{4.79}$$

Pour estimer $P_h v - v$ en norme L^2, nous admettons le lemme suivant (voir [6]).

Lemme 4.31 *Si le problème adjoint est régulier, on a l'estimation suivante :*

$$\|P_h v - v\|_{L^2(\Omega)} \leq Ch \|P_h v - v\|_{H^1(\Omega)}, \quad \forall v \in H_0^1(\Omega). \tag{4.80}$$

Nous pouvons maintenant regrouper ces résultats :

Théorème 4.32 (convergence en espace) *Selon la régularité de la solution du problème (4.3), on a les résultats suivants :*
(i) *Sous les hypothèses (4.68) et (4.69) on a*

$$\begin{aligned} &\lim_{h\to 0} \sup_{t\in]0,T[} \|\partial_t(u_h(t)-u(t))\|_{L^2(\Omega)} = 0, \\ &\lim_{h\to 0} \sup_{t\in]0,T[} \|u_h(t)-u(t)\|_{H^1(\Omega)} = 0, \end{aligned} \tag{4.81}$$

c'est-à-dire la convergence dans $\mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$.
(ii) *Si on suppose de plus que le problème adjoint est régulier et qu'il existe une constante* C *indépendante de* h *telle que*

$$\|u_{h,0}-u_0\|_{H^1(\Omega)} \le Ch, \quad \|u_{h,1}-u_1\|_{L^2(\Omega)} \le Ch, \tag{4.82}$$

alors il existe une constante C *indépendante de* h *telle que*

$$\forall t\in[0,T], \quad \|\partial_t(u_h(t)-u(t))\|_{L^2(\Omega)} \quad \le Ch. \tag{4.83}$$

(iii) *Si on suppose que* $u \in \mathcal{C}^2(0,T;H_0^1(\Omega)\cap H^2(\Omega))$ *(ce qui suppose que* $u_\alpha \in H_0^1(\Omega)\cap H^2(\Omega)$ *pour* $\alpha = 0,1$*) et qu'il existe une constante* C *indépendante de* h *telle que*

$$\|u_{h,0}-u_0\|_{H^1(\Omega)} \le Ch^2, \quad \|u_{h,1}-u_1\|_{L^2(\Omega)} \le Ch^2, \tag{4.84}$$

alors il existe une constante C *indépendante de* h *telle que*

$$\begin{aligned} \forall t\in[0,T], \quad &\|\partial_t(u_h(t)-u(t))\|_{L^2(\Omega)} \quad \le Ch^2, \\ &\|u_h(t)-u(t)\|_{H^1(\Omega)} \quad \le Ch^2. \end{aligned} \tag{4.85}$$

Démonstration : (i) Si on suppose seulement que $u \in \mathcal{C}^2(0,T;H_0^1(\Omega))$: d'après (3.55) on sait que pour tout $v \in H_0^1(\Omega)$ on a $\|v-P_hv\|_{H^1(\Omega)} \underset{h\to 0}{\longrightarrow} 0$ et par conséquent $\|v-P_hv\|_{L^2(\Omega)} \underset{h\to 0}{\longrightarrow} 0$. En appliquant ce résultat à $v = u_\alpha$, pour $\alpha = 0,1$ et $v = \partial_t^\beta$ pour $\beta = 0,1,2$, et en utilisant les estimations (4.75) et (4.77), on déduit aisément (4.81) à l'aide du théorème d'Ascoli (cf. note de bas de page 175-4).
(ii) Si on suppose de plus que le problème adjoint est régulier, l'estimation (4.80) du lemme 4.31 montre qu'il existe une constante C indépendante de h telle que $\|v-P_hv\|_{L^2(\Omega)} \le Ch$ pour tout $v \in H_0^1(\Omega)$. On voit donc que si les conditions initiales sont approchées à l'ordre 1 en h, i.e. si on a (4.82), l'estimation (4.75)- (i) donne aussi de l'ordre 1 en h, on obtient ainsi (4.83).
(iii) Si on suppose maintenant que $u \in \mathcal{C}^2(0,T;H_0^1(\Omega)\cap H^2(\Omega))$ et que les conditions initiales sont approchées à l'ordre 2 en h, i.e. si on a (4.84), alors en utilisant (3.56) et (3.57), on obtient finalement (4.85). ∎

4.5 Discrétisation totale

Nous adoptons ici les notations introduites à la section 3.3.2. Ainsi, on note $\overrightarrow{U}^k \in \mathbb{R}^N$ le vecteur approché à l'instant t_k du vecteur $\overrightarrow{U}(t_k)$. En d'autres termes, la $I^{ème}$ composante U_I^k du vecteur $\overrightarrow{U}^k \in \mathbb{R}^N$ représente une approximation de la solution u du problème (4.3) à l'instant t_k et au nœud M_I :

$$U_I^k \simeq U_I(t_k) \simeq u(M_I, t_k). \tag{4.86}$$

Et nous notons u_h^k la fonction de V_h correspondante, c'est-à-dire définie par

$$u_h^k = \sum_{I=1}^{N} U_I^k w_I, \tag{4.87}$$

soit encore telle que $u_h^k(M_I) = U_I^k$. Les schémas les plus utilisés en pratique pour approcher le système différentiel (4.59) sont les schémas explicites et dans toute la suite, nous nous focalisons sur le schéma *saute-mouton*, schéma explicite centré d'ordre deux :

$$\begin{cases} \mathbb{M}\dfrac{(\overrightarrow{U}^{k+1} - 2\overrightarrow{U}^k + \overrightarrow{U}^{k-1})}{\Delta t^2} + c^2\mathbb{K}\overrightarrow{U}^k = \overrightarrow{F}^k, \quad k = 1, \ldots, K-1 \\ \overrightarrow{U}^0, \overrightarrow{U}^1 \ \text{ donnés,} \end{cases} \tag{4.88}$$

qui s'écrit encore variationnellement : trouver $(u_h^k)_{k=0,K} \in V_h^{K+1}$ tel que

$$\begin{cases} (\dfrac{u_h^{k+1} - 2u_h^k + u_h^{k-1}}{\Delta t^2}, v_h)_{L^2(\Omega)} + a(u_h^k, v_h) = (f(t_k), v_h), \\ \qquad\qquad \forall v_h \in V_h, \ k = 1, \ldots, K-1 \\ u_h^0, \ \text{et } u_h^1 \text{ donnés dans } V_h. \end{cases} \tag{4.89}$$

Remarque 4.33 *Le schéma saute-mouton entre dans une classe de schémas plus générale, les schémas de Newmark à deux paramètres θ et δ, qui consistent à approcher le système différentiel (4.59) par :*

$$\begin{cases} \mathbb{M}\dfrac{(\overrightarrow{U}^{k+1} - 2\overrightarrow{U}^k + \overrightarrow{U}^{k-1})}{\Delta t^2} + c^2\mathbb{K}\left[\theta\overrightarrow{U}^{k+1} + \left(\frac{1}{2} + \delta - 2\theta\right)\overrightarrow{U}^k + \left(\frac{1}{2} + \theta - \delta\right)\overrightarrow{U}^{k-1}\right] \\ \qquad = \theta\overrightarrow{F}^{k+1} + \left(\dfrac{1}{2} + \delta - 2\theta\right)\overrightarrow{F}^k + \left(\dfrac{1}{2} + \theta - \delta\right)\overrightarrow{F}^{k-1}, \\ \qquad\qquad k = 1, \ldots, K-1 \\ \overrightarrow{U}^0, \overrightarrow{U}^1 \ \text{ donnés.} \end{cases}$$

Par des techniques similaires à celles utilisées au §3.3.2, on montre (voir par exemple [45, 23]) que si :

– $\delta \geq \frac{1}{2}$ *et si* $2\theta \geq \delta$ *le schéma est inconditionnellement stable ;*
– $\delta \geq \frac{1}{2}$ *et si* $2\theta < \delta$ *le schéma est stable sous la condition*

$$c^2 \left(\max_{m=1,N} \lambda_{m,h} \right) \Delta t^2 \leq 4/(1-4\theta);$$

– $\delta < \frac{1}{2}$ *le schéma est toujours instable ;*
– $\delta \neq \frac{1}{2}$ *le schéma est d'ordre 1 ;*
– $\delta = \frac{1}{2}$ *le schéma est d'ordre 2 ;*
– $\delta = \frac{1}{2}$ *et si* $\theta = \frac{1}{2}$ *le schéma est d'ordre 4.*

Evidemment le schéma obtenu est explicite *si, et seulement si,* $\theta = 0$*, et le schéma explicite centré d'ordre 2 (saute-mouton) correspond à* $\theta = 0$ *et* $\delta = 1/2$*.*

Initialisation du schéma

Pour le problème continu, comme pour le problème semi-discrétisé, les données initiales sont définies par les valeurs de la solution et de sa dérivée en temps à $t = 0$, c'est-à-dire $\partial_t^\alpha u(0) = u_\alpha,\ \alpha = 0, 1$. Pour le problème totalement discrétisé, $u_h^k = (u_J^k)$ est une approximation de $u(M_J, t_k)$ et on a besoin de u_h^0 et de u_h^1 pour pouvoir démarrer le schéma. On doit donc approcher $u(0) = u_0$ et $u(\Delta t)$. Pour le premier instant il est naturel de choisir

$$u_h^0 = u_{h,0}. \tag{4.90}$$

Pour approcher $u(\Delta t)$, on peut utiliser la formule de Taylor :

$$u(\Delta t) = u(0) + \Delta t \frac{\partial u}{\partial t}(0) + O(\Delta t^2),$$

ce qui conduit à l'approximation :

$$u_h^1 = u_{h,0} + \Delta t u_{h,1}.$$

Ce choix correspond alors à une approximation d'ordre 1 en temps de $\partial u/\partial t$. Pour avoir une approximation d'ordre 2 en temps, cohérente avec le schéma saute-mouton, on s'appuie sur le développement à l'ordre 3 :

$$u(\Delta t) = u_0 + \Delta t u_1 + \frac{\Delta t^2}{2} \frac{\partial^2 u}{\partial t^2}(0) + O(\Delta t^3),$$

et on utilise l'équation : $\dfrac{\partial^2 u}{\partial t^2} = c^2 \Delta u + f$, d'où

$$u(\Delta t) = u_0 + \Delta t u_1 + \frac{c^2 \Delta t^2}{2} \Delta u(0) + \frac{\Delta t^2}{2} f(0) + O(\Delta t^3),$$

qui s'écrit variationnellement :

$$(u(\Delta t), v)_{L^2(\Omega)} = (u_0, v)_{L^2(\Omega)} + \Delta t(u_1, v)_{L^2(\Omega)} - \frac{\Delta t^2}{2} a(u(0), v) + \frac{\Delta t^2}{2}(f(0), v)_{L^2(\Omega)} + O(\Delta t^3), \quad \forall v \in H_0^1(\Omega).$$

Ce choix conduit cette fois à l'approximation :

$$\begin{aligned}(u_h^1, v_h)_{L^2(\Omega)} = (u_{h,0}, v_h)_{L^2(\Omega)} - \frac{\Delta t^2}{2} a(u_{h,0}, v_h) + \Delta t(u_{h,1}, v_h)_{L^2(\Omega)} \\ + \frac{\Delta t^2}{2}(f(0), v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h.\end{aligned} \tag{4.91}$$

C'est ce dernier choix que nous adoptons dans la suite du cours. Il s'écrit matriciellement

$$\begin{cases} \overrightarrow{U}^0 = \overrightarrow{U}_0, \\ \mathbb{M}\overrightarrow{U}^1 = \mathbb{M}\overrightarrow{U}_0 - \dfrac{\Delta t^2}{2} c^2 \mathbb{K}\overrightarrow{U}_0 + \Delta t \mathbb{M}\overrightarrow{U}_1 + \dfrac{\Delta t^2}{2}\overrightarrow{F}^0, \end{cases} \tag{4.92}$$

le schéma est donc défini par (4.88) et (4.92) (ou encore par (4.89) et (4.90)-(4.91)).

4.5.1 Condensation de masse

Le schéma (4.88) n'est pas explicite : pour calculer $\overrightarrow{U}^{k+1}$, on doit inverser la matrice de masse $\mathbb{M}$:

$$\overrightarrow{U}^{k+1} = \mathbb{M}^{-1}(\Delta t^2 \overrightarrow{F}^k - c^2 \Delta t^2 \mathbb{K}\overrightarrow{U}^k + 2\overrightarrow{U}^k - \overrightarrow{U}^{k-1}).$$

La condensation de masse (ou lumping) consiste à approcher $\mathbb{M}$ par une matrice diagonale, en utilisant une formule de quadrature. On peut montrer que si on utilise une formule de quadrature assez précise, l'ordre du schéma n'est pas détérioré. On ne perd donc pas en précision, et on gagne en coût puisque le schéma devient alors explicite. Une analyse par ondes planes, dite *analyse de dispersion* (voir section 4.6) montre même que dans certains cas le schéma devient plus précis ! (mais du même ordre bien sûr)...

Expliquons maintenant le principe, lorsqu'on utilise des éléments finis de Lagrange. Si on dispose d'une formule de quadrature dont les nœuds de quadrature coïncident avec les positions des degrés de liberté M_I, c'est-à-dire d'une formule d'intégration numérique qui peut s'écrire formellement pour une fonction g régulière

$$\int g(M)dM \approx \sum_q \omega_q g(M_q),$$

alors, en appliquant cette formule pour approcher la matrice de masse, on obtient

$$\mathbb{M}_{IJ} = \int w_I w_J dx \approx \sum_q \omega_q w_I(M_q) w_J(M_q).$$

Or les fonctions de base vérifient $w_I(M_q) = \delta_{qI}$. On voit donc que les termes non diagonaux deviennent nuls et $\mathbb{M}$ est approchée par une matrice diagonale.
Dans ce qui suit nous donnons quelques exemples de schémas avec condensation, en dimension 1 puis en dimension 2.

a - Schéma P^1 avec et sans condensation en 1D

Schéma sans condensation. Nous présentons les schémas obtenus sur un maillage uniforme sur $\mathbb{R}$. Il est alors facile de calculer les matrices de masse et de rigidité, qui sont toutes les deux symétriques positives et tridiagonales :

$$\mathbb{M}_{I,I} = \frac{2}{3}h \ ; \quad \mathbb{M}_{I,I+1} = \frac{1}{6}h \ ; \ \mathbb{K}_{I,I} = \frac{2}{h} \ ; \ \mathbb{K}_{I,I+1} = -\frac{1}{h}. \tag{4.93}$$

Le schéma (4.88) se réinterprète comme un schéma aux différences finies

$$\begin{aligned} &\frac{2}{3}\frac{U_I^{k+1} - 2U_I^k + U_I^{k-1}}{\Delta t^2} + \frac{1}{6}\frac{U_{I+1}^{k+1} - 2U_{I+1}^k + U_{I+1}^{k-1}}{\Delta t^2} \\ &+\frac{1}{6}\frac{U_{I-1}^{k+1} - 2U_{I-1}^k + U_{I-1}^{k-1}}{\Delta t^2} - c^2\frac{U_{I+1}^k - 2U_I^k + U_{I-1}^k}{h^2} = \frac{F_I^k}{h}. \end{aligned} \tag{4.94}$$

Schéma avec condensation. Pour condenser la matrice de masse, nous utilisons la formule de quadrature (formule des trapèzes) :

$$\int_a^b g(x)dx \approx (b-a)\frac{g(a)+g(b)}{2}, \tag{4.95}$$

et nous obtenons alors la matrice de masse condensée $\mathbb{M}_{cond} = h\,\mathbb{I}$, ce qui conduit au schéma suivant :

$$\frac{U_I^{k+1} - 2U_I^k + U_I^{k-1}}{\Delta t^2} - c^2\frac{U_{I+1}^k - 2U_I^k + U_{I-1}^k}{h^2} = \frac{F_I^k}{h}. \tag{4.96}$$

Le schéma P^1 avec condensation de masse coïncide avec le schéma centré d'ordre 2 aux différences finies. Ce résultat est encore vrai en dimensions supérieures : sur une *grille régulière* et avec les éléments finis P^1, on peut réinterpréter le schéma variationnel comme un schéma aux différences finies. Si de plus on utilise une formule de quadrature pour condenser la masse, on retrouve également le schéma aux différences finies usuel.

Remarque 4.34 *On peut remarquer que la matrice de masse condensée s'obtient à partir de la matrice de masse sans condensation en sommant les éléments de chaque ligne et en les affectant à la diagonale. Cependant ce procédé n'est valide qu'avec des éléments finis P^1. Si on utilise des éléments finis d'ordre supérieur, l'utilisation d'une formule de quadrature adéquate n'est plus équivalente.*

b - Schéma P^1 avec condensation en 2D (ou schéma à 5 points)

Nous ne détaillons pas ici les calculs des matrices de masse et de rigidité. Il est cependant facile de vérifier que, si on applique la formule de quadrature suivante :

$$\int_T g(x)dx \approx \frac{|T|}{3}\sum_{i=1}^{3} g(S_i), \tag{4.97}$$

où T est un triangle de la triangulation, S_i ses sommets, $|T|$ sa mesure, alors la matrice de masse condensée s'exprime simplement comme $\mathbb{M} = h^2\mathbb{I}$ lorsqu'on la calcule sur un maillage uniforme de $\mathbb{R}^2$. Comme mentionné en dimension 1, le schéma obtenu est alors le schéma aux différences finies usuel, qui s'écrit en notant (i, j) les coordonnées du point I :

$$\frac{U_{i,j}^{k+1} - 2U_{i,j}^k + U_{i,j}^{k-1}}{\Delta t^2} - \frac{c^2}{h^2}(U_{i+1,j}^k - 2U_{i,j}^k + U_{i-1,j}^k + U_{i,j+1}^k - 2U_{i,j}^k + U_{i,j-1}^k) = 0. \tag{4.98}$$

c - Schéma Q^1 avec condensation en 2D (ou schéma à 9 points)

La formule de quadrature utilisée dans ce cas est la suivante : :

$$\int_K g(x)dx = \frac{|K|}{4}\sum_{i=1}^{4} g(S_i), \tag{4.99}$$

où S_i sont les sommets de l'élément K, quadrilatère de mesure $|K|$. La matrice de masse condensée s'exprime de nouveau comme $\mathbb{M} = h^2\mathbb{I}$ lorsqu'on la calcule sur un maillage uniforme de $\mathbb{R}^2$ et le schéma aux différences finies correspondant s'écrit alors

$$\begin{aligned}\frac{U_{i,j}^{k+1} - 2U_{i,j}^k + U_{i,j}^{k-1}}{\Delta t^2} - \frac{c^2}{3h^2}(&U_{i+1,j+1}^k + U_{i+1,j}^k + U_{i+1,j-1}^k \\ &+U_{i-1,j-1}^k + U_{i-1,j}^k + U_{i-1,j+1}^k + U_{i,j+1}^k + U_{i,j-1}^k - 8U_{i,j}^k) = 0.\end{aligned} \tag{4.100}$$

4.5.2 Stabilité par techniques d'énergie

Nous présentons ici une technique pour analyser la stabilité du schéma saute-mouton (4.88) très différente de celle suivie pour l'équation de la chaleur et basée sur des techniques d'énergie. Pour simplifier, nous nous plaçons dans le cas où la source est nulle ($f = 0$), mais les résultats de stabilité s'étendent aisément au cas d'une source non nulle. Nous définissons une *énergie discrète* du schéma, qui représente une approximation de l'énergie continue (4.31) définie à la section 4.3.1 :

$$\mathcal{E}^{k+1/2} = \frac{1}{2}\left|\frac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^{k}}{\Delta t}\right|_{\mathbb{M}}^{2} + \frac{1}{2}(\mathbb{K}\overrightarrow{U}^{k}|\overrightarrow{U}^{k+1}), \tag{4.101}$$

où $|.|_{\mathbb{M}}^2$ est la norme associée à la matrice $\mathbb{M}$, c'est-à-dire $\left|\overrightarrow{U}\right|_{\mathbb{M}}^2 = (\mathbb{M}\overrightarrow{U}, \overrightarrow{U})$. Notons que cette quantité peut également s'écrire

$$\mathcal{E}^{k+1/2} = \frac{1}{2}\left\|\frac{u_h^{k+1} - u_h^k}{\Delta t}\right\|_{L^2(\Omega)}^{2} + \frac{1}{2}a(u_h^k, u_h^{k+1}).$$

Nous montrons que cette quantité se conserve en l'absence de source :

Lemme 4.35 (conservation de l'énergie discrète) *En l'absence de source ($f = 0$), la quantité définie par (4.101) se conserve :*

$$\mathcal{E}^{k+1/2} = \mathcal{E}^{k-1/2}, \quad \forall k \geq 1. \tag{4.102}$$

Démonstration : le résultat s'obtient très facilement en multipliant scalairement l'équation (4.88) par $\dfrac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^{k-1}}{2\Delta t}$. ∎

Pour le problème continu, la conservation d'énergie suffit à déduire des résultats de stabilité, car l'énergie continue est une forme quadratique définie positive, donc contrôler l'énergie revient à contrôler certaines normes de la solution. En ce qui concerne le schéma totalement discrétisé, il en va autrement, du fait du terme $(\mathbb{K}\overrightarrow{U}^{k}|\overrightarrow{U}^{k+1})$ qui fait intervenir la solution à deux instants différents et qui n'a donc a priori pas de signe. On va chercher sous quelle condition l'énergie discrète définit encore une forme quadratique positive, auquel cas nous pourrons conclure. Pour ce faire, nous utilisons l'identité du parallélogramme

$$(\mathbb{K}\overrightarrow{U}|\overrightarrow{V}) = \frac{1}{4}(\mathbb{K}(\overrightarrow{U} + \overrightarrow{V})|\overrightarrow{U} + \overrightarrow{V}) - \frac{1}{4}(\mathbb{K}(\overrightarrow{U} - \overrightarrow{V})|\overrightarrow{U} - \overrightarrow{V}),$$

qui permet de réécrire l'énergie discrète sous la forme :

$$\begin{aligned}\mathcal{E}^{k+1/2} = &\frac{1}{2}\left(\mathcal{M}\frac{\overrightarrow{U}^{k+1}-\overrightarrow{U}^{k}}{\Delta t}|\frac{\overrightarrow{U}^{k+1}-\overrightarrow{U}^{k}}{\Delta t}\right)\\ &+\frac{1}{2}\left(c^2\mathbb{K}\frac{\overrightarrow{U}^{k}+\overrightarrow{U}^{k+1}}{2}|\frac{\overrightarrow{U}^{k}+\overrightarrow{U}^{k+1}}{2}\right),\end{aligned} \tag{4.103}$$

avec

$$\mathcal{M} = \mathbb{M} - \frac{c^2\Delta t^2}{4}\mathbb{K}.$$

Si la matrice $\mathcal{M}$ est définie positive, cette quantité est positive et on a la propriété de stabilité. Cette condition est une condition de stabilité de type CFL (Courant-Friedrichs-Levy) et s'écrit sous la forme :

$$\gamma_{cfl} \equiv \frac{c^2\Delta t^2}{4}\sup_{V\neq 0}\frac{(\mathbb{K}\overrightarrow{V}|\overrightarrow{V})}{(\mathbb{M}\overrightarrow{V}|\overrightarrow{V})} \leq 1, \tag{4.104}$$

ou encore

$$\gamma_{cfl} \equiv \frac{\Delta t^2}{4}\sup_{v_h\neq 0}\frac{a(v_h,v_h)}{\|v_h\|^2_{L^2(\Omega)}} \leq 1. \tag{4.105}$$

On reconnaît le quotient de Rayleigh (voir (1.30))

$$\mathcal{R}(v_h) = \frac{a(v_h,v_h)}{\|v_h\|^2_{L^2(\Omega)}}.$$

Chercher la borne supérieure de ce quotient revient à trouver le maximum des valeurs propres du problème (voir chapitre 1) :

$$\text{trouver } u_h \in V_h \text{ et } \lambda_h \text{ tels que } a(u_h,v_h) = \lambda_h(u_h,v_h)_{L^2(\Omega)},$$

ou encore sous forme matricielle (voir aussi §3.3.1)

$$c^2\mathbb{K}\overrightarrow{U} = \lambda_h\mathbb{M}\overrightarrow{U}.$$

Notation. *Il est courant d'exprimer la condition CFL sous la forme*

$$\alpha_{cfl} \leq \alpha_{max} \tag{4.106}$$

où α_{cfl} *désigne*

$$\alpha_{cfl} = \frac{c\Delta t}{h}, \tag{4.107}$$

et α_{max} *est la quantité bornée indépendemment de* h *et* Δt *définie par*

$$\frac{1}{\alpha^2_{max}} = \frac{h^2}{4}\sup_{V\neq 0}\frac{(\mathbb{K}\overrightarrow{V}|\overrightarrow{V})}{(\mathbb{M}\overrightarrow{V}|\overrightarrow{V})}$$

Nous sommes maintenant en mesure d'énoncer le

Théorème 4.36 *Si les paramètres de discrétisation vérifient la condition CFL (4.104), le schéma (4.88), (4.92) est stable. Si on suppose de plus que $\gamma_{cfl} < 1$, la solution discrète vérifie les estimations suivantes :*

$$\begin{aligned}
&\text{(i)} \quad \left\| \frac{u_h^{k+1} - u_h^k}{\Delta t} \right\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{1-\gamma_{cfl}}} \left(\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} \right), \\
&\text{(ii)} \quad \left\| \nabla \left(\frac{u_h^{k+1} + u_h^k}{\Delta t} \right) \right\|_{L^2(\Omega)} \leq C \left(\|u_0\|_{H^1(\Omega)} + \|u_1\|_{L^2(\Omega)} \right), \\
&\text{(iii)} \quad \left\| u_h^k \right\|_{L^2(\Omega)} \leq \frac{C}{\sqrt{1-\gamma_{cfl}}} \left((1+T) \|u_0\|_{H^1(\Omega)} + T \|u_1\|_{L^2(\Omega)} \right).
\end{aligned} \tag{4.108}$$

Remarque 4.37 *Ces estimations de continuité de la solution discrète par rapport aux données initiales sont à rapprocher des estimations de la solution continue (4.34). Notons qu'elles ne permettent pas de contrôler la solution discrète dans le cas limite $\gamma_{cfl} = 1$. Il est néanmoins encore possible dans ce cas d'obtenir des estimations de stabilité par d'autres techniques (voir [6] pour plus de détails).*

Démonstration : Remarquons tout d'abord que la condition CFL (4.104) implique que

$$(\mathcal{M}\overrightarrow{V}|\overrightarrow{V}) = (\mathbb{M}\overrightarrow{V}|\overrightarrow{V}) - \frac{\Delta t^2}{4}(c^2\mathbb{K}\overrightarrow{V}|\overrightarrow{V}) \geq (1-\gamma_{cfl})\ (\mathbb{M}\overrightarrow{V}|\overrightarrow{V}).$$

Compte tenu de l'expression de l'énergie discrète (4.103), on a donc

$$2\mathcal{E}^{k+1/2} \geq (1-\gamma_{cfl}) \left| \frac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^k}{\Delta t} \right|_{\mathbb{M}}^2 + c^2 \left| \frac{\overrightarrow{U}^{k+1} + \overrightarrow{U}^k}{2} \right|_{\mathbb{K}}^2.$$

La conservation de l'énergie discrète (4.102) montre alors que

$$(1-\gamma_{cfl}) \left| \frac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^k}{\Delta t} \right|_{\mathbb{M}}^2 + c^2 \left| \frac{\overrightarrow{U}^{k+1} + \overrightarrow{U}^k}{2} \right|_{\mathbb{K}}^2 \leq 2\mathcal{E}^{1/2}. \tag{4.109}$$

Cherchons à estimer $\mathcal{E}^{1/2}$ en fonction des données initiales exactes. Rappelons que

$$\mathcal{E}^{1/2} = \frac{1}{2} \left| \frac{\overrightarrow{U}^1 - \overrightarrow{U}^0}{\Delta t} \right|_{\mathbb{M}}^2 + \frac{1}{2}(c^2\mathbb{K}\overrightarrow{U}^0|\overrightarrow{U}^1),$$

qui, compte tenu de l'initialisation du schéma (4.92), et en posant $\mathbb{K}^c = c^2\mathbb{K}$, s'écrit encore :

$$\begin{aligned}
\mathcal{E}^{1/2} &= \frac{1}{2} \left| -\frac{\Delta t}{2}\mathbb{M}^{-1}\mathbb{K}^c\overrightarrow{U}_0 + \overrightarrow{U}_1 \right|_{\mathbb{M}}^2 + \frac{1}{2}(\mathbb{K}^c\overrightarrow{U}_0|\mathbb{M}^{-1}(\mathbb{M}\overrightarrow{U}_0 - \frac{\Delta t^2}{2}\mathbb{K}^c\overrightarrow{U}_0 + \Delta t\mathbb{M}\overrightarrow{U}_1)) \\
&= \frac{1}{2}\left(\left|\overrightarrow{U}_1\right|_{\mathbb{M}}^2 + \frac{\Delta t^2}{4} \left|\mathbb{M}^{-1}\mathbb{K}^c\overrightarrow{U}_0\right|_{\mathbb{M}}^2 - \Delta t(\overrightarrow{U}_1|\mathbb{M}^{-1}\mathbb{K}^c\overrightarrow{U}_0)_{\mathbb{M}} \right) \\
&\quad + \frac{1}{2}(\mathbb{K}^c\overrightarrow{U}_0|\overrightarrow{U}_0) - \frac{\Delta t^2}{4}(\mathbb{K}^c\overrightarrow{U}_0|\mathbb{M}^{-1}\mathbb{K}^c\overrightarrow{U}_0) + \frac{\Delta t}{2}(\mathbb{K}^c\overrightarrow{U}_0|\overrightarrow{U}_1) \\
&= \frac{1}{2}\left|\overrightarrow{U}_1\right|_{\mathbb{M}}^2 - \frac{\Delta t^2}{8}\left|\mathbb{M}^{-1}\mathbb{K}^c\overrightarrow{U}_0\right|_{\mathbb{M}}^2 + \frac{1}{2}(\mathbb{K}^c\overrightarrow{U}_0|\overrightarrow{U}_0) \leq \frac{1}{2}\left|\overrightarrow{U}_1\right|_{\mathbb{M}}^2 + \frac{1}{2}(\mathbb{K}^c\overrightarrow{U}_0|\overrightarrow{U}_0),
\end{aligned}$$

ou de manière équivalente :

$$\mathcal{E}^{1/2} \leq \frac{1}{2} \|u_{h,1}\|^2_{L^2(\Omega)} + \frac{c^2}{2} \|u_{h,0}\|^2_{H^1(\Omega)} .$$

Les hypothèses de convergence (4.57) entraînent en particulier :

$$\|u_{h,1}\|_{L^2(\Omega)} \leq C \|u_1\|_{L^2(\Omega)} , \quad \|u_{h,0}\|_{H^1(\Omega)} \leq C \|u_0\|_{H^1(\Omega)} ,$$

et par conséquent

$$\mathcal{E}^{1/2} \leq CE(0) \leq C(\|u_1\|^2_{L^2(\Omega)} + \|u_0\|^2_{H^1(\Omega)}).$$

On déduit donc de l'estimation (4.109) que

$$(1 - \gamma_{cfl}) \left| \frac{\overrightarrow{U}^{k+1} - \overrightarrow{U}^k}{\Delta t} \right|^2_{\mathbb{M}} + c^2 \left| \frac{\overrightarrow{U}^{k+1} + \overrightarrow{U}^k}{2} \right|^2_{\mathbb{K}} \leq C((\|u_1\|^2_{L^2(\Omega)} + \|u_0\|^2_{H^1(\Omega)}),$$

dont découlent immédiatement les estimations (4.108)- (i) et (ii).

L'estimation (4.108)- (iii) se déduit de (4.108)- (i) en utilisant :

$$\left\|u_h^k\right\|_{L^2(\Omega)} \leq \left\|u_h^k - u_h^{k-1}\right\|_{L^2(\Omega)} + \left\|u_h^{k-1}\right\|_{L^2(\Omega)} \leq \frac{C\Delta t}{\sqrt{1-\gamma_{cfl}}} \sqrt{E(0)} + \left\|u_h^{k-1}\right\|_{L^2(\Omega)} .$$

Par récurrence, on obtient donc

$$\left\|u_h^k\right\|_{L^2(\Omega)} \leq \frac{CT}{\sqrt{1-\gamma_{cfl}}} \sqrt{E(0)} + \left\|u_h^0\right\|_{L^2(\Omega)} .$$

■

Illustration dans le cas 1D, avec $\Omega =]0, L[$

Nous pouvons déduire des résultats précédents une condition suffisante de stabilité pour le schéma P^1, avec ou sans condensation de masse, lorsqu'on se place sur un maillage uniforme en dimension 1, de pas $h = L/(N+1)$, avec $U_0 = U_{N+1} = 0$.

Lemme 4.38

(i) *Le schéma P^1 avec condensation (4.96) est stable sous la condition*

$$\frac{c\Delta t}{h} \leq 1. \tag{4.110}$$

(ii) *Le schéma P^1 sans condensation (4.94) est stable sous la condition*

$$\frac{c\Delta t}{h} \leq \frac{\sqrt{3}}{3}. \tag{4.111}$$

Démonstration : (i) Compte tenu de l'expression de $\mathbb{K}$ (voir (4.93)), il est aisé de montrer que

$$(\mathbb{K}\overrightarrow{U}\,|\,\overrightarrow{U}) = \frac{1}{h} \sum_{j=1,N} |U_{j+1} - U_j|^2 . \tag{4.112}$$

La matrice de masse condensée étant égale à $\mathbb{M} = h\mathbb{I}$, on a

$$\frac{(\mathbb{K}\overrightarrow{U}|\overrightarrow{U})}{(\mathbb{M}\overrightarrow{U}|\overrightarrow{U})} = \frac{1}{h^2}\frac{\sum_{j=1,N}|U_{j+1}-U_j|^2}{\sum_{j=1,N}|U_j|^2}.$$

En utilisant l'inégalité $(a-b)^2 \leq 2(a^2+b^2)$, on majore cette quantité par

$$\frac{1}{h^2}\frac{2\sum_{j=1,N}(|U_{j+1}|^2+|U_j|^2)}{\sum_{j=1,N}|U_j|^2} \leq \frac{4}{h^2}\frac{\sum_{j=1,N}|U_j|^2}{\sum_{j=1,N}|U_j|^2} = \frac{4}{h^2}.$$

On en déduit que sous la condition (4.110), la condition (4.104) est bien satisfaite.
(ii) Dans le cas du schéma sans condensation, la matrice de masse est donnée par (4.93) et le problème aux valeurs propres $c^2\mathbb{K}\overrightarrow{U} = \lambda\mathbb{M}\overrightarrow{U}$ revient à trouver λ et $\overrightarrow{U} \neq 0$ tels que

$$\begin{aligned} &-c^2\frac{U_{i+1}-2U_i+U_{i-1}}{h^2} = \lambda(\frac{2}{3}U_i + \frac{1}{6}U_{i-1} + \frac{1}{6}U_{i+1}),\ 1 \leq i \leq N \\ &U_0 = U_{N+1} = 0. \end{aligned}$$

Après quelques calculs (voir par exemple [6]), on trouve

$$\lambda_{m,h} = \frac{6c^2}{h^2}\frac{1-\cos\frac{m\pi}{N+1}}{2+\cos\frac{m\pi}{N+1}}, \quad m = 1, N.$$

La fonction $x \in [-1,1] \mapsto (1-x)/(2+x)$ est strictement décroissante ; le max est donc obtenu pour la plus petite valeur du cos c'est-à-dire pour $m = N$ et cette valeur est inférieure à celle obtenue pour $x = -1$ c'est-à-dire 2 :

$$\max_m \lambda_{m,h} = \lambda_{N,h} = \frac{6c^2}{h^2}\frac{1-\cos\frac{N\pi}{N+1}}{2+\cos\frac{N\pi}{N+1}} \leq \frac{12c^2}{h^2}.$$

Pour vérifier la condition CFL :

$$\frac{\Delta t^2}{4}\frac{6c^2}{h^2}\frac{1-\cos\frac{N\pi}{N+1}}{2+\cos\frac{N\pi}{N+1}} \leq 1,$$

il suffit que

$$\frac{\Delta t^2}{4}\frac{12c^2}{h^2} \leq 1 \Longleftrightarrow \frac{\sqrt{3}c\Delta t}{h} \leq 1.$$

La condition donnée en (4.111) est une approximation à l'ordre 2 en h de la CFL exacte et correspond à la condition CFL du problème posé sur un domaine infini ou semi-infini (faire $L \to +\infty$). ∎

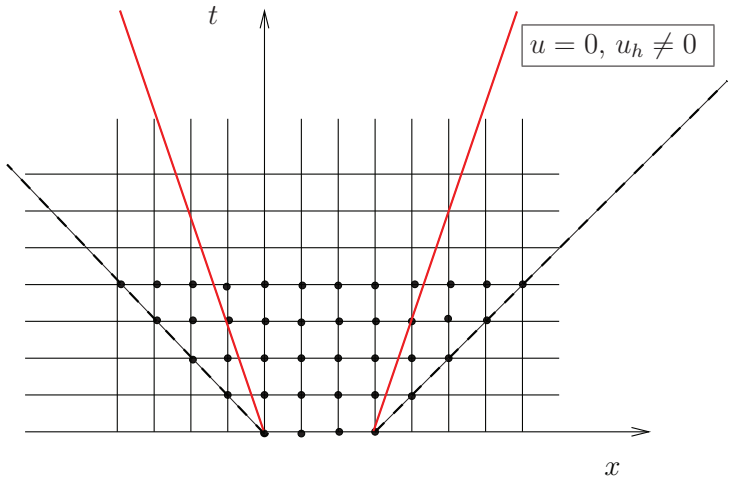
Remarque 4.39 *Le schéma P^1 avec condensation de masse, aussi précis que le schéma sans condensation (voir section 4.5.1), est en outre moins coûteux non seulement par son caractère explicite, mais également grâce à sa condition CFL moins contraignante.*

Interprétation géométrique de la condition de stabilité en 1D sur $\Omega = \mathbb{R}$ du schéma P^1 avec condensation

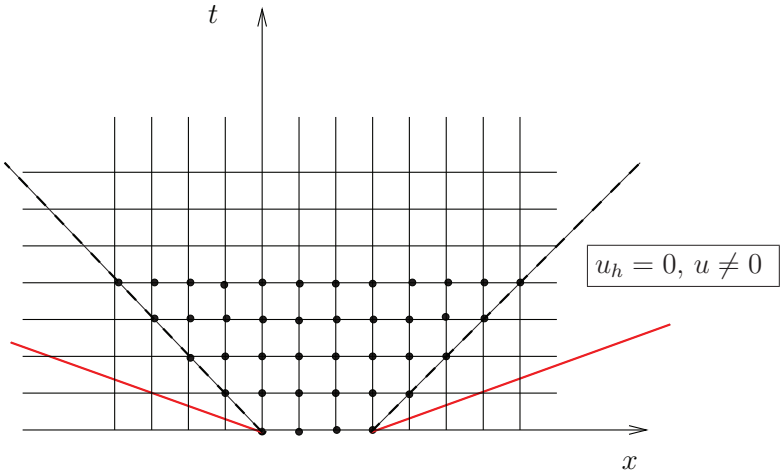
On a vu au §4.1.2 que, pour des données initiales à support dans un intervalle $[a, b]$, la solution exacte est à support dans le cône de dépendance $\cup_t K_t \times \{t\}$ où $K_t = [a - ct, b + ct]$. À chaque itération en temps du schéma (4.96), la solution discrète se propage sur un nœud du maillage supplémentaire, ce qui définit la vitesse numérique :

$$V_{num} = \frac{h}{\Delta t}.$$

La solution discrète a ainsi pour support le cône de dépendance numérique situé entre les deux droites limites de pentes $\pm 1/V_{num}$ (voir figure 4.4). Si la pente



Cas 1 : $1/c > 1/V_{num}$, le schema peut converger



Cas 2 : $1/c < 1/V_{num}$, le schema ne converge pas

Figure 4.4. Condition nécessaire de convergence : $1/V_{num} \leq 1/c \Longleftrightarrow \alpha_{cfl} = \dfrac{c\Delta t}{h} \leq 1$

$1/V_{num} > 1/c$, alors u_h est nulle dans la zone comprise entre les deux cônes, alors que u ne l'est pas. Il ne peut pas y avoir convergence. Une *condition nécessaire de convergence* est donc :

$$\frac{c\Delta t}{h} \leq 1. \tag{4.113}$$

Nous retrouvons la condition de stabilité du schéma. Cette condition est en fait une condition nécessaire et suffisante de stabilité. En effet, le schéma étant consistant, la convergence dépend seulement de sa stabilité, d'après le résultat classique qui dit que la stabilité et la consistance entraînent la convergence (voir théorème 3.28).

4.5.3 Convergence du schéma totalement discrétisé

Nous considérons la solution $u \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H_0^1(\Omega))$ du problème continu (4.15) et $u_h^k \in V_h$ la solution approchée solution de (4.89)-(4.90)-(4.91). Pour établir un résultat de convergence, il est nécessaire de supposer plus de régularité pour u, comme nous le verrons par la suite. Afin d'alléger les notations, nous notons dans cette section $(\cdot,\cdot)$ le produit scalaire dans $L^2(\Omega)$ et $\|\cdot\|$ la norme de $L^2(\Omega)$. Nous posons $\overline{u}^k = u(t_k)$ et introduisons l'erreur de convergence

$$e_h^k = u_h^k - \overline{u}^k.$$

Cette erreur vérifie le problème suivant :

$$\begin{cases} (\dfrac{e_h^{k+1} - 2e_h^k + e_h^{k-1}}{\Delta t^2}, v_h) + a(e_h^k, v_h) = (\varepsilon_h^k, v_h),\ \forall v_h \in V_h, \\ e_h^0 = u_h^0 - \overline{u}_h^0, \quad e_h^1 = u_h^1 - \overline{u}_h^1, \end{cases} \tag{4.114}$$

où

$$\varepsilon_h^k = \frac{\partial^2 u}{\partial t^2}(t_k) - \frac{\overline{u}_h^{k+1} - 2\overline{u}_h^k + \overline{u}_h^{k-1}}{\Delta t^2}.$$

L'idée est d'obtenir une estimation de l'erreur en fonction de ε_h^k, qui est "petit" (de l'ordre de Δt^2), ce qui permettra de montrer la convergence (et d'obtenir l'ordre d'approximation). L'erreur ε_h^k représente une erreur de troncature (ou erreur de consistance). Si on pouvait choisir $v_h = (e_h^{k+1} - e_h^{k-1})/(2\Delta t)$ dans (4.114), on obtiendrait une estimation d'énergie sur l'erreur et ce serait fini... Mais ce n'est pas possible car $e_h^k \notin V_h$. Comme pour le problème semi-discrétisé, on introduit la projection elliptique $P_h\overline{u}^k \in V_h$ définie par (3.52), et on décompose l'erreur sous la forme

$$e_h^k = \underbrace{u_h^k - P_h\overline{u}^k}_{\delta_h^k \in V_h} + \underbrace{P_h\overline{u}^k - \overline{u}^k}_{-r_h^k}.$$

On remarque maintenant que $\delta_h^k = u_h^k - P_h\overline{u}^k$ appartient à V_h et que $r_h^k = \overline{u}^k - P_h\overline{u}^k$ représente l'erreur de projection elliptique qui est "petite". On réécrit le problème (4.114) sous la forme

$$\begin{cases} (\dfrac{\delta_h^{k+1} - 2\delta_h^k + \delta_h^{k-1}}{\Delta t^2}, v_h) + a(\delta_h^k, v_h) \\ \qquad = (\varepsilon_h^k, v_h) + (\dfrac{r_h^{k+1} - 2r_h^k + r_h^{k-1}}{\Delta t^2}, v_h) \quad + a(r_h^k, v_h), \quad \forall v_h \in V_h, \\ e_h^0 = u_h^0 - \overline{u}_h^0, \quad e_h^1 = u_h^1 - \overline{u}_h^1. \end{cases}$$

Par définition de la projection elliptique, on a

$$a(r_h^k, v_h) = 0, \qquad \forall v_h \in V_h,$$

ce qui permet finalement de réécrire le problème sous la forme

$$\begin{cases} (\dfrac{\delta_h^{k+1} - 2\delta_h^k + \delta_h^{k-1}}{\Delta t^2}, v_h) + a(\delta_h^k, v_h) \\ \qquad = (\varepsilon_h^k, v_h) + (\dfrac{r_h^{k+1} - 2r_h^k + r_h^{k-1}}{\Delta t^2}, v_h) \quad \forall v_h \in V_h, \\ e_h^0 = u_h^0 - \overline{u}_h^0, \quad e_h^1 = u_h^1 - \overline{u}_h^1. \end{cases} \tag{4.115}$$

En choisissant $v_h = \dfrac{\delta_h^{k+1} - \delta_h^{k-1}}{2\Delta t}$, on obtient

$$\frac{\mathcal{E}_h^{k+1/2} - \mathcal{E}_h^{k-1/2}}{\Delta t} = (\varepsilon_h^k, \frac{\delta_h^{k+1} - \delta_h^{k-1}}{2\Delta t}) + (\frac{r_h^{k+1} - 2r_h^k + r_h^{k-1}}{\Delta t^2}, \frac{\delta_h^{k+1} - \delta_h^{k-1}}{2\Delta t}),$$

avec

$$\mathcal{E}_h^{k+1/2} = \frac{1}{2}\left\|\frac{\delta_h^{k+1} - \delta_h^k}{\Delta t}\right\|^2 + \frac{1}{2}a(\delta_h^k, \delta_h^{k+1}). \tag{4.116}$$

Le lemme suivant fournit une l'estimation de l'erreur de projection elliptique.

Lemme 4.40 *Si la solution la solution du problème continu (4.15) a la régularité $u \in \mathcal{C}^4(0,T;V)$ (ce qui suppose en particulier que u_0 et u_1 appartiennent à V) et si on suppose la condition CFL stricte satisfaite, i.e. $|\gamma_{cfl}| < 1$, alors l'énergie définie par (4.116) vérifie l'estimation suivante :*

$$\begin{aligned} \sqrt{\mathcal{E}_h^{k+1/2}} \leq \sqrt{\mathcal{E}_h^{1/2}} + \frac{Ct_k}{\sqrt{1-\gamma_{cfl}}} \sup_{s\in[0,T]} \Bigg(& \Delta t^2 \left\|\frac{\partial^4 u}{\partial t^4}(s)\right\|_{L^2(\Omega)} \\ & + \left\|(I-P_h)\frac{\partial^2 u}{\partial t^2}(s)\right\|_{L^2(\Omega)} + \Delta t^2 \left\|(I-P_h)\frac{\partial^4 u}{\partial t^4}(s)\right\|_{L^2(\Omega)} \Bigg). \end{aligned} \tag{4.117}$$

Si de plus on a $u \in \mathcal{C}^4(0,T;V) \cap \mathcal{C}^2(0,T;H^2(\Omega))$ *(ce qui suppose en particulier que* $(u_0,u_1) \in (V \cap H^2(\Omega))^2$*), alors l'énergie vérifie*

$$\sqrt{\mathcal{E}_h^{k+1/2}} \leq \sqrt{\mathcal{E}_h^{1/2}} + \frac{CT}{\sqrt{1-\gamma_{cfl}}}\left(\Delta t^2 + h^2\right). \tag{4.118}$$

Démonstration : Nous ne rentrons pas dans le détail de la démonstration. Nous indiquons les grandes étapes de la preuve de l'estimation (4.117). Quant à l'estimation (4.118), elle découle des propriétés sur la projection elliptique présentées à la section 3.5.1.

1) On suppose que la condition de stabilité (4.105) est satisfaite de façon stricte (i.e. $\gamma_{cfl} < 1$). Sous cette condition on montre que

$$\mathcal{E}_h^{k+1/2} \geq \frac{1}{2}(1-\gamma_{cfl}) \left\| \frac{\delta_h^{k+1} - \delta_h^k}{\Delta t} \right\|^2 > 0. \tag{4.119}$$

2) A partir de l'identité d'énergie :

$$\frac{\mathcal{E}_h^{k+1/2} - \mathcal{E}_h^{k-1/2}}{\Delta t} \leq \underbrace{\left(\left\| \varepsilon_h^k \right\| + \left\| \frac{r_h^{k+1} - 2r_h^k + r_h^{k-1}}{\Delta t^2} \right\| \right)}_{=\mu_k} \left\| \frac{\delta_h^{k+1} - \delta_h^{k-1}}{2\Delta t} \right\|,$$

en utilisant l'inégalité triangulaire et (4.119) on a

$$\begin{aligned} \left\| \frac{\delta_h^{k+1} - \delta_h^{k-1}}{2\Delta t} \right\| &\leq \frac{1}{2} \left\| \frac{\delta_h^{k+1} - \delta_h^k}{\Delta t} \right\| + \frac{1}{2} \left\| \frac{\delta_h^k - \delta_h^{k-1}}{\Delta t} \right\| \\ &\leq \frac{1}{2} \frac{\sqrt{2}}{\sqrt{1-\gamma_{cfl}}} \sqrt{\mathcal{E}_h^{k+1/2}} + \frac{1}{2} \frac{\sqrt{2}}{\sqrt{1-\gamma_{cfl}}} \sqrt{\mathcal{E}_h^{k-1/2}}, \end{aligned}$$

qui implique que

$$\begin{aligned} &\mathcal{E}_h^{k+1/2} - \mathcal{E}_h^{k-1/2} \leq \Delta t \mu_k \frac{\sqrt{2}}{2\sqrt{1-\gamma_{cfl}}} (\sqrt{\mathcal{E}_h^{k+1/2}} + \sqrt{\mathcal{E}_h^{k-1/2}}) \\ &\Longrightarrow (\sqrt{\mathcal{E}_h^{k+1/2}} + \sqrt{\mathcal{E}_h^{k-1/2}})(\sqrt{\mathcal{E}_h^{k+1/2}} - \sqrt{\mathcal{E}_h^{k-1/2}}) \\ &\qquad\qquad \leq \Delta t \mu_k \frac{\sqrt{2}}{2\sqrt{1-\gamma_{cfl}}} (\sqrt{\mathcal{E}_h^{k+1/2}} + \sqrt{\mathcal{E}_h^{k-1/2}}) \\ &\Longrightarrow \sqrt{\mathcal{E}_h^{k+1/2}} \leq \sqrt{\mathcal{E}_h^{k-1/2}} + \Delta t \mu_k \frac{\sqrt{2}}{2\sqrt{1-\gamma_{cfl}}}, \end{aligned}$$

d'où finalement :

$$\sqrt{\mathcal{E}_h^{k+1/2}} \leq \sqrt{\mathcal{E}_h^{1/2}} + \Delta t \frac{\sqrt{2}}{2\sqrt{1-\gamma_{cfl}}} \sum_{m=1}^{k} \mu_m. \tag{4.120}$$

D'après la définition de μ_m, le terme de droite contient deux erreurs : l'erreur de projection elliptique $\Delta t \sum_{m=1}^k \left\| \frac{r_h^{m+1} - 2r_h^m + r_h^{m-1}}{\Delta t^2} \right\|$ et l'erreur de troncature $\Delta t \sum_{m=1}^k \|\varepsilon_h^m\|$ que nous estimons dans les deux prochains points.

3) (Erreur de troncature) Sous l'hypothèse que $u \in \mathcal{C}^4(0,T;L^2)$ on montre que

$$\Delta t \sum_{m=1}^{k} \|\varepsilon_h^m\| \leq C t_k \Delta t^2 \sup_{s\in[0,T]} \left\| \frac{\partial^4 u}{\partial t^4}(s) \right\|. \tag{4.121}$$

4) (Erreur de projection elliptique) On rappelle que $r_h^k = \overline{u}^k - P_h \overline{u}^k$. On montre que si $u \in \mathcal{C}^4(0,T;V)$, on a

$$\begin{aligned} \Delta t \sum_{m=1}^{k} \left\| \frac{r_h^{m+1} - 2r_h^m + r_h^{m-1}}{\Delta t^2} \right\| \leq C t_k \sup_{s\in[0,T]} \Bigg(& \left\| (I-P_h) \frac{\partial^2 u}{\partial t^2}(s) \right\| \\ & + \Delta t^2 \left\| (I-P_h) \frac{\partial^4 u}{\partial t^4}(s) \right\| \Bigg). \end{aligned} \tag{4.122}$$

En regroupant ces résultats, on obtient l'estimation annoncée. ∎

Par récurrence, nous pouvons obtenir une estimation sur δ_h^k, que nous énonçons pour plus de simplicité dans un cas où la solution est très régulière :

Lemme 4.41 *Si $u \in \mathcal{C}^4(0,T;V) \cap \mathcal{C}^2(0,T;H^2(\Omega))$ (ce qui suppose en particulier que $(u_0,u_1) \in (V \cap H^2(\Omega))^2$), et si on suppose la condition CFL stricte satisfaite ($\gamma_{cfl} < 1$), alors il existe une constante $C > 0$ indépendante de h et de Δt telle que*

$$\left\| \delta_h^k \right\|_{L^2(\Omega)} \leq \frac{CT}{\sqrt{1-\gamma_{cfl}}} \left(\sqrt{\mathcal{E}_h^{1/2}} + \frac{T}{\sqrt{1-\gamma_{cfl}}} (h^2 + \Delta t^2) \right) + \left\| \delta_h^0 \right\|_{L^2(\Omega)}. \tag{4.123}$$

Démonstration : En utilisant l'inégalité triangulaire, puis une récurrence, on a

$$\begin{aligned} \|\delta_h^{k+1}\| &\leq \|\delta_h^{k+1} - \delta_h^k\| + \|\delta_h^k\| \\ &\leq \sum_{m=0,k} \|\delta_h^{m+1} - \delta_h^m\| + \|\delta_h^0\|, \end{aligned}$$

et en utilisant l'estimation (4.119)

$$\|\delta_h^{k+1}\| \leq \sum_{m=0,k} \frac{2\Delta t}{\sqrt{1-\gamma_{cfl}}} \sqrt{\mathcal{E}_h^{m+1/2}} + \|\delta_h^0\|.$$

Nous utilisons maintenant l'estimation sur l'énergie (4.118) obtenue au lemme précédent

$$\begin{aligned} \|\delta_h^{k+1}\| &\leq \sum_{m=0,k} \frac{2\Delta t}{\sqrt{1-\gamma_{cfl}}} \left(\sqrt{\mathcal{E}_h^{1/2}} + \frac{CT}{\sqrt{1-\gamma_{cfl}}} \left(\Delta t^2 + h^2\right) \right) + \|\delta_h^0\| \\ &\leq \frac{2(k+1)\Delta t}{\sqrt{1-\gamma_{cfl}}} \left(\sqrt{\mathcal{E}_h^{1/2}} + \frac{CT}{\sqrt{1-\gamma_{cfl}}} \left(\Delta t^2 + h^2\right) \right) + \|\delta_h^0\|, \end{aligned}$$

d'où le résultat. ∎

L'erreur de convergence peut alors être estimée, encore une fois grâce à l'inégalité triangulaire :

$$\|e_h^k\| \leq \|\delta_h^k\| + \|r_h^k\|.$$

Le deuxième terme représente l'erreur de projection elliptique que nous avons déjà estimée. Finalement nous énonçons le résultat de convergence :

Proposition 4.42 *Si $u \in \mathcal{C}^4(0,T;V) \cap \mathcal{C}^2(0,T;H^2(\Omega))$ (ce qui suppose en particulier que $(u_0,u_1) \in (V \cap H^2(\Omega))^2$), et si on suppose la condition CFL stricte satisfaite ($\gamma_{cfl} < 1$), alors il existe une constante $C > 0$ indépendante de h et de Δt (mais qui dépend de $T/\sqrt{1-\gamma_{cfl}}$) telle que*

$$\left\| u_h^k - \overline{u}^k \right\|_{L^2(\Omega)} \leq C \left(h^2 + \Delta t^2 + \sqrt{\mathcal{E}_h^{1/2}} + \left\| \delta_h^0 \right\|_{L^2(\Omega)} \right). \tag{4.124}$$

Nous voyons l'influence du choix d'une approximation consistante des conditions initiales sur les deux derniers termes de l'estimation de l'erreur. Si nous ne faisons pas le bon choix (4.76) au démarrage, ces termes ne donnent qu'une approximation à l'ordre 1 et par conséquent détériorent l'ordre d'erreur pour tous les instants.

Lemme 4.43 *Nous nous plaçons sous les hypothèses de la proposition 4.42 et nous faisons le choix (4.76) et (4.90)-(4.91) pour le démarrage du schéma. Alors il existe une constante C indépendante de h et de Δt telle que*

$$\sqrt{\mathcal{E}_h^{1/2}} + \left\| \delta_h^0 \right\|_{L^2(\Omega)} \leq C(h^2 + \Delta t^2),$$

et par conséquent l'erreur de convergence est d'ordre 2 :

$$\left\| u_h^k - \overline{u}^k \right\|_{L^2(\Omega)} \leq C \left(h^2 + \Delta t^2 \right). \tag{4.125}$$

Démonstration : par hypothèse on a $\delta_h^0 = u_h^0 - P_h \overline{u}^0 = 0$ par conséquent, le deuxième terme est nul et il nous reste seulement à estimer

$$\sqrt{\mathcal{E}_h^{1/2}} = \frac{1}{\sqrt{2}} \left\| \frac{\delta_h^1}{\Delta t} \right\| = \frac{1}{\sqrt{2}} \left\| \frac{u_h^1 - P_h \overline{u}^1}{\Delta t} \right\|,$$

avec $\delta_h^1 = u_h^1 - P_h \overline{u}^1 = u_h^1 - P_h u(\Delta t)$. La solution étant très régulière, on peut écrire le développement de Taylor (dans lequel on a utilisé l'équation dont u est solution)

$$\overline{u}^1 = u(\Delta t) = u_0 + \Delta t u_1 + \frac{\Delta t^2 c^2}{2} \Delta u_0 + \frac{\Delta t^2}{2} f(0) + \frac{\Delta t^3}{6} \frac{\partial^3 u}{\partial t^3}(s), \quad \text{pour un } s \in [0, \Delta t],$$

ce qui implique en particulier que pour tout $v_h \in V_h$:

$$(\overline{u}^1, v_h) = (u_0, v_h) + \Delta t (u_1, v_h) - \frac{\Delta t^2}{2} a(u_0, v_h) + \frac{\Delta t^2}{2} (f(0), v_h) + \frac{\Delta t^3}{6} (\frac{\partial^3 u}{\partial t^3}(s), v_h).$$

En faisant la différence avec (4.91), on obtient donc

$$\begin{aligned}(u_h^1 - \overline{u}^1, v_h) = {} & (u_{h,0} - u_0, v_h) + \Delta t (u_{h,1} - u_1, v_h) \\ & - \frac{\Delta t^2}{2} a(u_{h,0} - u_0, v_h) - \frac{\Delta t^3}{6} (\frac{\partial^3 u}{\partial t^3}(s), v_h),\end{aligned}$$

En intercalant la projection elliptique, on obtient, puisque $u_{h,0} = P_h u_0$

$$
\begin{aligned}
(\frac{u_h^1 - P_h\overline{u}^1}{\Delta t}, v_h) = (\frac{\overline{u}^1 - u_0}{\Delta t} - P_h(\frac{\overline{u}^1 - u_0}{\Delta t}), v_h) \\
+(u_{h,1} - u_1, v_h) - \frac{\Delta t^2}{6}(\frac{\partial^3 u}{\partial t^3}(s), v_h).
\end{aligned}
\tag{4.126}
$$

Le choix (4.76) et l'hypothèse $u_1 \in H^2(\Omega)$ nous indiquent que

$$\|u_{h,1} - u_1\| = \|P_h u_1 - u_1\| \leq Ch^2 |u_1|_{H^2(\Omega)}.$$

Par ailleurs, on remarque que d'après les hypothèses de régularité :

$$\frac{\overline{u}^1 - u_0}{\Delta t} = u_1 + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(s') \in H^2(\Omega), \quad \text{pour un } s' \in [0, \Delta t],$$

d'où

$$\left\| \frac{\overline{u}^1 - u_0}{\Delta t} - P_h(\frac{\overline{u}^1 - u_0}{\Delta t}) \right\| \leq Ch^2 \left(|u_1|_{H^2(\Omega)} + \Delta t \sup_{s\in[0,T]} \left| \frac{\partial^2 u}{\partial t^2}(s) \right|_{H^2(\Omega)} \right).$$

En choisissant $v_h = \dfrac{u_h^1 - P_h\overline{u}^1}{\Delta t}$ dans (4.126), on obtient finalement

$$
\begin{aligned}
\left\| \frac{u_h^1 - P_h\overline{u}^1}{\Delta t} \right\| &\leq \left\| \frac{\overline{u}^1 - u_0}{\Delta t} - P_h(\frac{\overline{u}^1 - u_0}{\Delta t}) \right\| + \|u_{h,1} - u_1\| + \frac{\Delta t^2}{6} \sup_{s\in[0,T]} \left\| \frac{\partial^3 u}{\partial t^3}(s) \right\| \\
&\leq C(h^2 + \Delta t^2).
\end{aligned}
$$

■

4.6 Analyse de dispersion

4.6.1 Introduction

Un des outils importants de l'analyse des schémas d'approximation des équations d'ondes est l'analyse de dispersion. Cette analyse consiste à étudier et comparer les ondes planes qui se propagent au niveau discret et continu. Cette analyse revient en fait à faire une transformée de Fourier (ou Fourier discrète) des équations (ou schémas), et suppose donc qu'on se place dans un milieu homogène et qu'on utilise un maillage uniforme. La notion de *relation de dispersion* a été abordée sommairement dans le cas momodimensionnel continu à la section 4.1.2.
Dans le cas bidimensionnel, les *ondes planes du problème continu* sont des solutions de la forme :

$$u(x,t) = e^{i(\kappa\cdot x - \omega t)}, \tag{4.127}$$

où $\kappa = (\kappa_1, \kappa_2)^t \in \mathbb{R}^2$ est le vecteur d'onde et ω, la pulsation. La fonction u est une solution de l'équation des ondes (4.3) si la *relation de dispersion* est vérifiée :

$$\omega^2 = c^2|\kappa|^2. \tag{4.128}$$

La quantité $\omega/|\kappa|$ est la vitesse de phase de l'onde et peut prendre deux valeurs $\pm c$.

Les *ondes planes du schéma semi-discrétisé* : ce sont cette fois-ci les solutions $u_h(t)$ de composantes $U_{jm}(t) = e^{i(jh\kappa_1+mh\kappa_2-\omega_h t)}$ ($(j,m) \in \mathbb{Z}^2$ désigne les indices d'un point du maillage), qui vérifient le schéma semi-discrétisé, réinterprété comme un schéma aux différences finies. La *relation de dispersion du schéma semi-discrétisé* est alors de la forme

$$\omega_h^2 = D_h(\kappa),$$

le terme $D_h(\kappa)$ provenant de la discrétisation en espace (correspondant en fait à l'opérateur $\mathbb{M}^{-1}\mathbb{K}$) et le ω_h^2 venant de l'opérateur $\partial^2/\partial t^2$. Nous ne détaillons pas ici le cas semi-discrétisé (pour plus de détails, voir [6]).

Enfin, les ondes planes du schéma totalement discrétisé sont les solutions u_h^k du schéma (toujours réinterprété comme un schéma aux différences finies) de la forme :

$$U_{jm}^k = e^{i(jh\kappa_1+mh\kappa_2-\omega_{h,\Delta t}k\Delta t)}.$$

La *relation de dispersion du schéma totalement discrétisé* peut alors s'écrire sous la forme

$$D_{\Delta t}(\omega_{h,\Delta t}) = D_h(\kappa),$$

où cette fois-ci $D_{\Delta t}(\omega_{h,\Delta t})$ provient de la discrétisation en temps. Nous donnons quelques exemples dans ce qui suit. Notons toutefois que cette relation permet de déterminer deux valeurs de $\omega_{h,\Delta t}$ et dans les exemples qui suivent, on peut toujours ramener cette relation sous la forme du rapport entre la vitesse de phase numérique que nous notons $V = \omega_{h,\Delta t}/|\kappa|$ et la vitesse de phase continue c :

$$\frac{V}{c} = \frac{1}{c}\frac{\omega_{h,\Delta t}}{|\kappa|} = \pm q(\alpha_{cfl}, G, \theta), \tag{4.129}$$

où on définit :
– le nombre CFL $\alpha_{cfl} = c\Delta t/h$,
– la longueur d'onde $\lambda = 2\pi/|\kappa|$,
– l'inverse du nombre de points par longueur d'onde $G = K/(2\pi) = h/\lambda$ où $K = |\kappa|h$,
– l'angle d'incidence $\theta = \arctan(\kappa_2/\kappa_1)$ (i.e., $\kappa = |\kappa|(\cos\theta, \sin\theta)^t$).

Pour un choix de paramètres de discrétisation donné (donc α_{cfl} donné) et un angle d'incidence donné, un schéma produit une erreur entre la vitesse de phase numérique et la vitesse de phace continue c, ce qui se traduit par l'écart que fait $q(\alpha_{cfl}, G, \theta)$ avec 1. C'est ce qu'on appelle *l'erreur de dispersion* qui s'observe sur les *courbes de dispersion numérique* obtenues en traçant la fonction

$$G \mapsto q(\alpha_{cfl}, G, \theta).$$

Nous observons également que la vitesse de phase numérique V dépend de l'angle θ contrairement au cas continu, ce qui provient de l'anisotropie introduite par le maillage (deux directions de maillage). C'est ce qu'on appelle *l'anistropie numérique*. On fixe cette fois-ci α_{cfl} ainsi que $N = 1/G$ le nombre de points par longueur d'ondes (supposé supérieur ou égal à 2) et on représente les *courbes d'anisotropie numérique*

$$\theta \mapsto q(\alpha_{cfl}, G, \theta),$$

ce qui montre dans quelles directions l'erreur est la moins ou la plus importante.

Condition de stabilité à partir de la relation de dispersion

Nous indiquons ici une technique très pratique pour déterminer la condition de stabilité d'un schéma. Dans les exemples que nous allons voir, la relation de dispersion admet deux solutions $\omega^{\pm}$ opposées l'une de l'autre. Une condition nécessaire de stabilité est donc que ces solutions soient réelles :

$$\omega^{\pm}_{h,\Delta t} \in \mathbb{R}. \tag{4.130}$$

En effet, si $\omega^{\pm} = \omega^{\pm}_R + i\omega^{\pm}_I$ était complexe, cela signifierait qu'il existe des solutions de la forme

$$U^k_{jm} = e^{i(jh\kappa_1 + mh\kappa_2 - \omega^{\pm} k\Delta t)} = e^{i(jh\kappa_1 + mh\kappa_2 - \omega^{\pm}_R k\Delta t)} e^{\omega^{\pm}_I k\Delta t}.$$

Or les deux solutions $\omega^{\pm}$ étant opposées l'une de l'autre, l'une des deux est telle que $\omega_I > 0$, ce qui correspond à une solution exponentiellement croissante en temps.

Remarque 4.44 *Cette technique pour retrouver la condition de stabilité ne s'applique plus telle quelle pour des problèmes qui admettent de la dissipation, car dans ce cas les solutions de la relation de dispersion deviennent complexes.*

4.6.2 Analyse de dispersion des schémas en dimension 1

Nous menons ici l'analyse de dispersion des schémas P^1 avec et sans condensation de masse (resp. (4.96) et (4.94)) qui ont été présentés en section 4.5.1. En dimension 1, les ondes planes numériques sont les solutions particulières de la forme

$$U^k_j = e^{i(jh\kappa - \omega_{h,\Delta t} k\Delta t)}$$

où $\kappa \in \mathbb{R}$ est le nombre d'onde. En injectant cette expression dans chacun des schémas, nous obtenons leur relation de dispersion et nous pouvons en faire l'analyse. Notons qu'en dimension 1, il n'y a pas d'angle θ, donc le rapport entre la vitesse de phase numérique et la vitesse de phase continue ne dépend que des deux paramètres α_{cfl} et G ; par conséquent, nous ne présentons que les courbes de dispersion numérique.

Relation de dispersion du schéma P^1 avec condensation de masse (4.96)

On a

$$\frac{4}{\Delta t^2}\sin^2\frac{\omega_{h,\Delta t}\Delta t}{2} = \frac{4c^2}{h^2}\sin^2\frac{\kappa h}{2}, \tag{4.131}$$

ce qui donne

$$\omega_{h,\Delta t} = \pm\frac{2}{\Delta t}\arcsin\left(\frac{c\Delta t}{h}\sin\frac{\kappa h}{2}\right) \equiv \pm\frac{2}{\Delta t}\arcsin\left(\alpha_{cfl}\sin\frac{\kappa h}{2}\right), \tag{4.132}$$

d'où

$$q(\alpha_{cfl}, G) = \frac{1}{c}\frac{\omega^+_{h,\Delta t}}{|\kappa|} = \frac{1}{\alpha_{cfl}\pi G}\arcsin(\alpha_{cfl}\sin\pi G).$$

On remarque que :

- $q(\alpha_{cfl}, G) \leq 1$ pour tout α_{cfl} et G, ce qui veut dire que les ondes numériques se propagent plus lentement que les ondes du modèle continu. Nous représentons sur la figure 4.5 les courbes de dispersion. Celles du schéma (4.96) se situent donc en dessous de (ou sur) la droite $q = 1$.
- Si α_{cfl} est fixé et $h \to 0$, on a

  $$q(\alpha_{cfl}, G) = 1 - 2\pi^2(1 - \alpha^2_{cfl})G^2 + ...,$$

  ce qui montre que le schéma est d'ordre 2 en G. On peut montrer qu'il est d'ordre infini si $\alpha_{cfl} = \alpha_{max} = 1$, ceci est particulier au cas monodimensionnel.
- Pour α_{cfl} fixé, la fonction $G \mapsto q(\alpha_{cfl}, G)$ est décroissante (plus on a de points par longueur d'onde, plus q est proche de 1).
- Pour G fixé, la fonction $\alpha_{cfl} \mapsto q(\alpha_{cfl}, G)$ est une fonction croissante. Le meilleur schéma est donc obtenu pour le plus grand α_{cfl} possible (limité par la CFL) c'est-à-dire pour $\alpha_{cfl} = \alpha_{max} = 1$.
- En appliquant le critère de stabilité (4.130), nous retrouvons que la CFL du schéma (4.96) s'écrit $\alpha_{cfl} \leq 1$.
- La limite quand $\alpha_{cfl} \to 0$ correspond à

  $$q(\alpha_{cfl} \to 0, G) \sim \frac{1}{\pi G},$$

  qui correspond à l'erreur de dispersion du schéma semi-discrétisé. On voit donc que sa dispersion est plus grande que celle du schéma totalement discrétisé.

Relation de dispersion du schéma P^1 sans condensation de masse (4.94)

On trouve dans ce cas

$$\frac{4}{\Delta t^2}\sin^2\frac{\omega_{h,\Delta t}\Delta t}{2} = \frac{4c^2}{h^2}\sin^2\frac{\kappa h}{2}\frac{1}{1 - \frac{2}{3}\sin^2\frac{\kappa h}{2}}, \tag{4.133}$$

ce qui donne cette fois-ci

$$q(\alpha_{cfl}, G) = \frac{1}{\alpha_{cfl}\pi G} \arcsin\left(\alpha_{cfl} \frac{\sin \pi G}{\sqrt{1 - \frac{2}{3}\sin^2 \pi G}}\right)$$
$$= 1 + \frac{\pi^2}{6}(1 + \alpha_{cfl}^2)G^2 + O(G^4).$$

On remarque que la vitesse est approchée ici par le haut alors qu'elle l'est par le bas pour le schéma avec condensation (voir figure 4.5) : les ondes numériques se propagent plus rapidement que celles du modèle continu. D'autre part, le développement de Taylor indique que plus α_{cfl} est grand, plus la vitesse numérique s'éloigne de la vitesse continue. La meilleure valeur est donc obtenue cette fois-ci pour $\alpha_{cfl} = 0$, pour laquelle on a :

$$q(0, G) = \frac{1}{\pi G} \frac{\sin \pi G}{\sqrt{1 - \frac{2}{3}\sin^2 \pi G}},$$

qui est différent de 1. Le schéma est donc plus dispersif que le schéma avec condensation, comme on peut le voir sur la figure 4.5 en comparant les courbes optimales pour les deux schémas (i.e. $\alpha_{cfl} = \alpha_{max} = 1$ pour le schéma avec condensation et $\alpha_{cfl} = 0$ pour le schéma sans condensation). En pratique, pour des raisons de coût, on préfère choisir le plus grand Δt autorisé par la condition de stabilité, ce qui revient à choisir $\alpha_{cfl} = \alpha_{max}$. Rappelons que cette condition de stabilité est plus contraignante pour le schéma sans condensation ($\alpha_{max} = \sqrt{3}/3$) que pour le schéma avec condensation ($\alpha_{max} = 1 > \sqrt{3}/3$). La courbe, représentée sur la figure 4.5, correspondant à $\alpha_{max} = \sqrt{3}/3$ pour le schéma sans condensation, montre en outre que l'erreur de dispersion est toujours plus forte que celle du schéma avec condensation. En conclusion, le schéma sans condensation est beaucoup plus coûteux que le schéma avec condensation (système linéaire à résoudre à chaque itération en temps, et pas de temps plus petit) et il est moins précis.

4.6.3 Analyse des schémas en dimension 2

Nous ne considérons ici que les schémas P^1 (4.98) et Q^1 (4.100), obtenus après condensation de masse (l'étude 1D nous a montré que le schéma sans condensation était moins efficace sur tous les plans).

Le schéma à 5 points, P^1 (4.98)

Rappelons que ce schéma, lorsqu'on l'écrit sur un maillage uniforme, coïncide avec le schéma aux différences finies usuel. Sa relation de dispersion s'écrit :

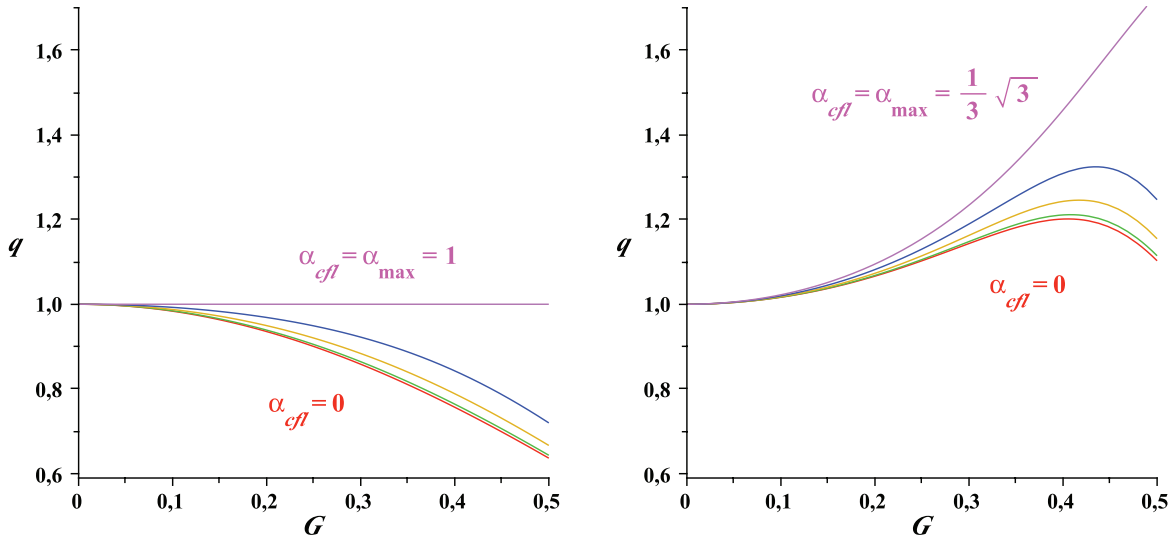


Figure 4.5. Courbes de dispersion du schéma P^1 avec condensation (à gauche) et sans condensation (à droite) en dimension 1, pour différentes valeurs de $\alpha_{cfl} = \gamma\alpha_{max}$ avec $\gamma = 0,\ 0.25,\ 0.5,\ 0.75,\ 1$.

$$\frac{4}{\Delta t^2}\sin^2\frac{\omega\Delta t}{2} = \frac{4c^2}{h^2}(\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2}).$$

Stabilité : on applique le critère (4.130) : le schéma est stable si $\forall(\kappa_1,\kappa_2)$, la solution ω est réelle i.e., si $\forall(\kappa_1,\kappa_2)$

$$0 \le \alpha_{cfl}^2(\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2}) \le 1.$$

Or

$$\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2} \le 2, \quad \forall(\kappa_1,\kappa_2),$$

et pour $\kappa_1 = \kappa_2 = \pi/h$, $\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2} = 2$, par conséquent on a

$$\min(\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2}) = 0 \quad \text{et} \quad \max(\sin^2\frac{\kappa_1 h}{2} + \sin^2\frac{\kappa_2 h}{2}) = 2.$$

La condition de stabilité s'écrit finalement $2\alpha_{cfl}^2 \le 1$ soit

$$\alpha_{cfl} \le \alpha_{max} = \frac{\sqrt{2}}{2}. \tag{4.134}$$

Interprétation géométrique de la condition de stabilité du schéma P^1

Comme en dimension 1, nous pouvons interpréter cette condition de la façon suivante : en dimension 2, la solution élémentaire (fonction de Green) est :

$$G(x,t) = \begin{cases} \dfrac{1}{2\pi\sqrt{t^2 - |x|^2/c^2}} & \text{si } |x| < ct, \\ 0 & \text{sinon.} \end{cases}$$

A l'instant $t_k = k\Delta t$, elle a atteint tous les points $|x| \leq ck\Delta t$. La solution numérique, quant à elle, se propage sur un losange de demi-diagonale kh et donc de côté $l = kh/\sqrt{2}$ (voir figure 4.6). La condition CFL exprime le fait que le losange

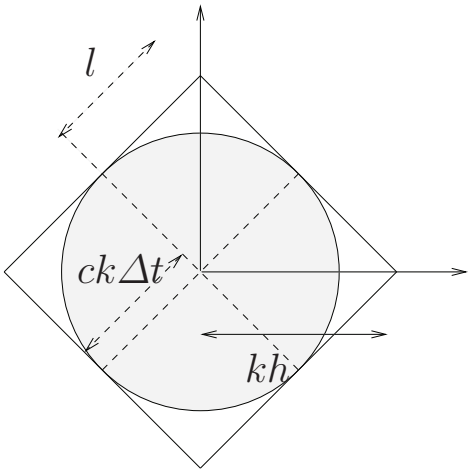


Figure 4.6. Support de la solution élémentaire exacte (disque) et approchée par le schéma à 5 points (losange).

doit strictement contenir le disque $\{|x| \leq ct\}$. S'il existait des points du disque à l'extérieur du losange, cela signifierait qu'en ces points la solution exacte est non nulle alors que la solution approchée l'est. Cette condition apparaît de nouveau comme une condition nécessaire de convergence du schéma. Le losange contient strictement le disque si $l \geq ck\Delta t$ et donc si

$$kh/\sqrt{2} \geq ck\Delta t \Longleftrightarrow c\Delta t/h \leq \sqrt{2}/2.$$

Courbes de dispersion et d'anisotropie

La relation de dispersion fournit $\omega_{h,\Delta t}(\kappa_1, \kappa_2, \alpha_{cfl})$. En utilisant les notations introduites au paragraphe 4.6.1, nous obtenons le rapport entre vitesse de phase numérique et continue :

$$\begin{aligned} q(\alpha_{cfl}, G, \theta) &= V(\alpha_{cfl}, G, \theta)/c \\ &= \frac{1}{\alpha_{cfl}\pi G} \arcsin\left(\alpha_{cfl}\left(\sin^2(\pi G\cos\theta) + \sin^2(\pi G \sin\theta)\right)^{1/2}\right). \end{aligned}$$

On peut faire un développement de Taylor pour G petit :

$$q(\alpha_{cfl}, G, \theta) = 1 - \frac{\pi^2 G^2}{6}(1 - \alpha_{cfl}^2 - \frac{1}{2}\sin^2 2\theta) + O(G^4),$$

ce qui montre que la vitesse est approchée à l'ordre 2. Contrairement au cas monodimensionnel, le schéma n'est plus exact pour $\alpha_{cfl} = \alpha_{max}$ ou pour toute autre valeur de α_{cfl}. On voit que la vitesse est approchée par le bas ($V(\alpha_{cfl}, G, \theta) \leq c$). Pour $\alpha_{cfl} = \alpha_{max} = \sqrt{2}/2$ et $\theta = \pi/4$ (modulo $\pi/2$), on obtient la vitesse exacte

$$V(\alpha_{cfl} = \frac{\sqrt{2}}{2}, G, \theta = \pi/4) = c,$$

ce qui montre que, pour cette valeur de α_{cfl} le schéma est exact seulement dans les directions diagonales. Pour α_{cfl} fixé, la dispersion décroît pour θ variant de 0 à $\pi/4$, ce qui confirme que l'erreur est plus mauvaise dans les directions du maillage et est meilleure dans les directions diagonales. Sur les figures qui suivent, on représente :

• *les courbes de dispersion* (figures 4.7) : sur chaque figure, on fixe l'angle θ (de 0 à $\pi/4$), et on représente les courbes de dispersion obtenues pour différentes valeurs du paramètre $\gamma = \alpha_{cfl}/\alpha_{max}$ (avec ici $\alpha_{max} = \sqrt{2}/2$) comprises entre 0 et 1 (plus précisément $\gamma = 0,\ 0.25,\ 0.5,\ 0.75,\ 1$). Les courbes sont tracées en fonction de G.

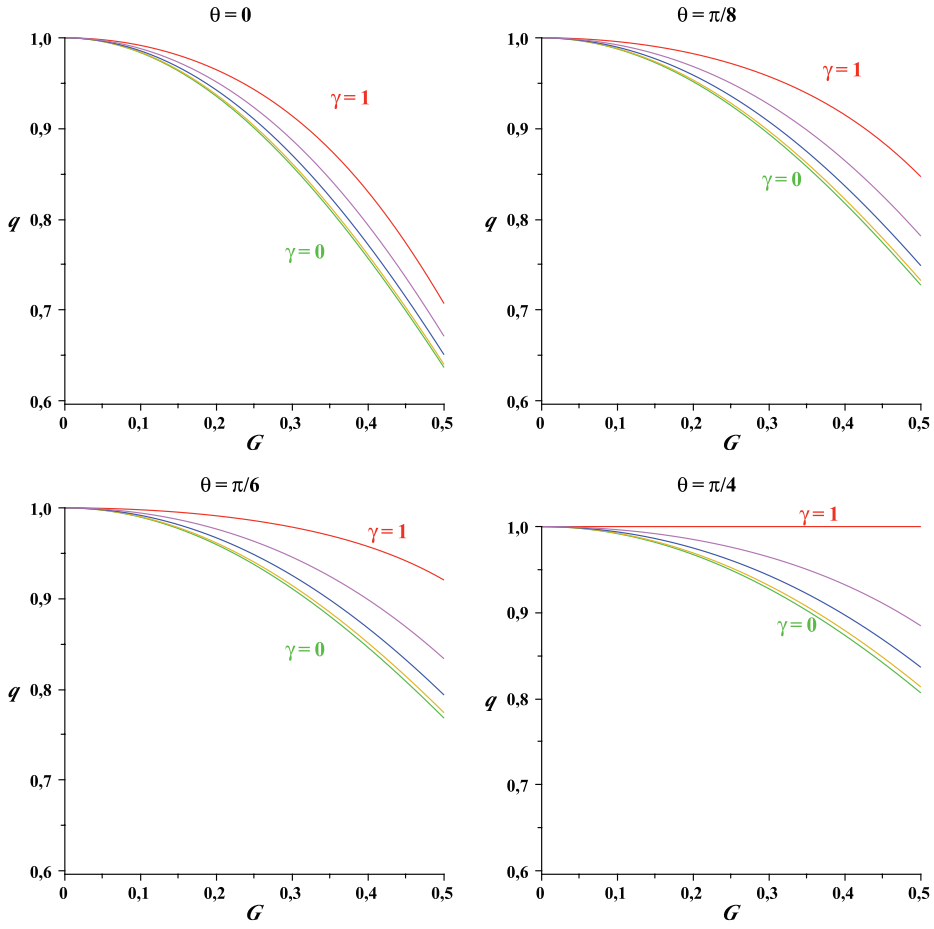


Figure 4.7. Courbes de dispersion du schéma P^1 à 5 points pour $\theta = 0,\ \pi/8,\ \pi/6,\ \pi/4$. Sur chaque figure, les courbes correspondent aux différentes valeurs de $\gamma = 0,\ 0.25,\ 0.5,\ 0.75,\ 1$.

• *les courbes d'anisotropie* (figure 4.8) : on fixe $\gamma = \alpha_{cfl}/\alpha_{max}$ ($\gamma = 0, 0.5, 1$), et on représente la vitesse de phase numérique en fonction de θ, pour différentes valeurs du nombre de points par longueur d'ondes $N = 1/G$ ($N = 2, 3, 4, 5, 10, 20, 100$).

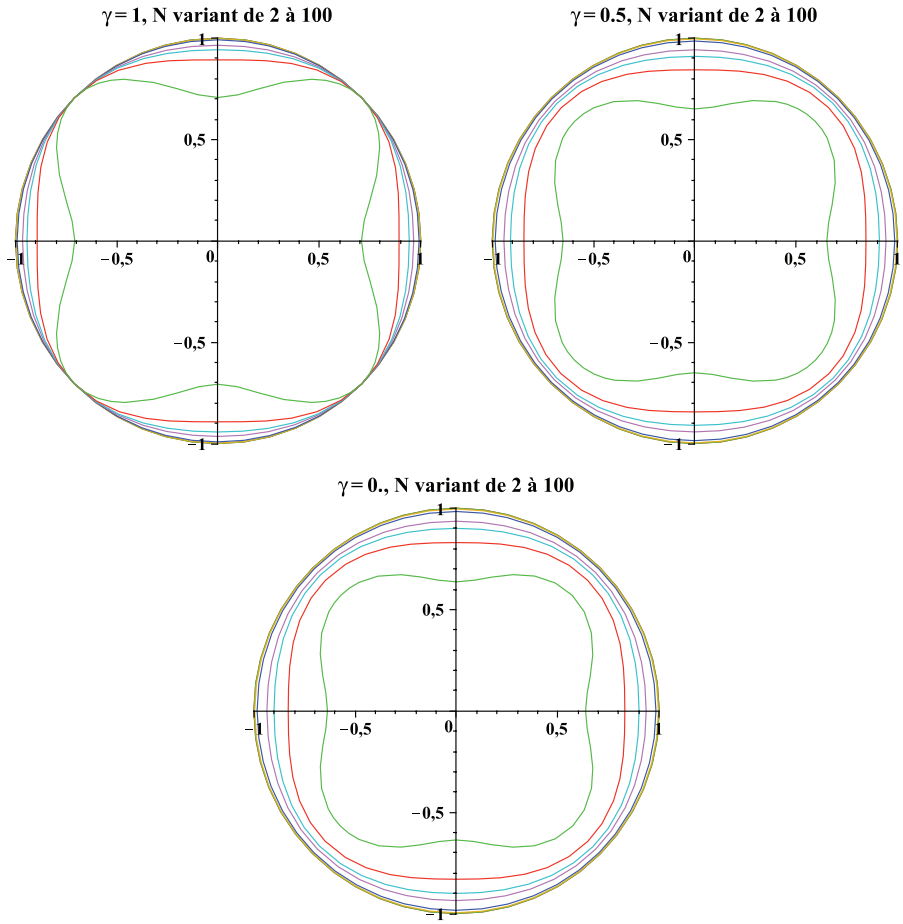


Figure 4.8. Courbes d'anisotropie du schéma P^1 à 5 points pour différentes valeurs du nombre de points par longueur d'ondes $N = 1/G$ et pour $\gamma = \alpha_{cfl}/\alpha_{max} = 1.$, 0.5 et 0 (ici $\alpha_{max} = \sqrt{2}/2$).

Le schéma Q^1 à 9 points (4.100)

Nous menons la même étude que précédemment. La relation de dispersion s'écrit

$$\frac{4}{\Delta t^2}\sin^2\frac{\omega\Delta t}{2} = \frac{2c^2}{3h^2}(4 - \cos\kappa_1 h - \cos\kappa_2 h - \cos(\kappa_1+\kappa_2)h - \cos(\kappa_1-\kappa_2)h).$$

Le schéma est stable si $\forall(\kappa_1, \kappa_2)$, la solution ω est réelle i.e., si

$$0 \le F(\kappa_1, \kappa_2) \le 1 \quad \forall(\kappa_1, \kappa_2),$$

$$\text{où } F(\kappa_1, \kappa_2) = \frac{\alpha_{cfl}^2}{6}(4 - \cos\kappa_1 h - \cos\kappa_2 h - \cos(\kappa_1+\kappa_2)h - \cos(\kappa_1-\kappa_2)h).$$

Lemme 4.45 *On a*

$$\min_{(\kappa_1,\kappa_2)\in\mathbb{R}^2} F(\kappa_1,\kappa_2) = 0 \quad et \quad \max_{(\kappa_1,\kappa_2)\in\mathbb{R}^2} F(\kappa_1,\kappa_2) = \alpha_{cfl}^2.$$

Démonstration du lemme : cela découle de l'identité

$$\cos(\kappa_1+\kappa_2)h + \cos(\kappa_1-\kappa_2)h = 2\cos\kappa_1 h\cos\kappa_2 h,$$

qui donne

$$F(\kappa_1,\kappa_2) = \frac{\alpha_{cfl}^2}{12}(9-(1+2\cos\kappa_1 h)(1+2\cos\kappa_2 h)).$$

Or on a

$$-1 \le 1+2\cos\eta \le 3, \quad \forall\eta\in\mathbb{R},$$

d'où

$$-3 \le (1+2\cos\kappa_1 h)(1+2\cos\kappa_2 h) \le 9,$$

ce qui montre que

$$0 \le F(\kappa_1,\kappa_2) \le \frac{\alpha_{cfl}^2}{12}(9+3) = \alpha_{cfl}^2.$$

Comme $F(0,0)=0$ et $F(0,\pi/h)=\alpha_{cfl}^2$, le minimum et le maximum sont atteints. ∎

On en déduit aussitôt le

Théorème 4.46 *Le schéma Q^1 est stable sous la condition CFL*

$$\alpha_{cfl} \le 1. \tag{4.135}$$

Interprétation géométrique de la condition de stabilité du schéma Q^1

L'interprétation est la même que pour le schéma à 5 points. La différence ici provient du support de la solution élémentaire approchée qui est cette fois-ci un carré de côté $l = kh$. La condition CFL exprime le fait que le carré doit strictement contenir le disque, support de la solution élémentaire exacte (voir figure 4.9), c'est-à-dire $kh \ge ck\Delta t$ d'où la condition CFL : $c\Delta t/h \le 1$.

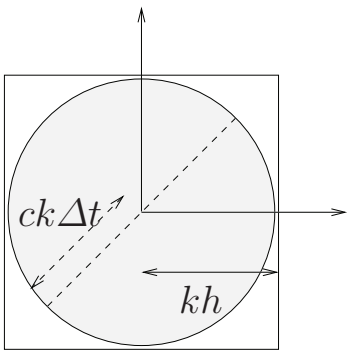


Figure 4.9. Support de la solution élémentaire exacte (disque) et approchée par le schéma à 9 points (carré).

Courbes de dispersion et d'anisotropie

La relation de dispersion conduit à

$$q(\alpha_{cfl}, G, \theta) = \frac{1}{\alpha_{cfl}\pi G} \arcsin \sqrt{F(\frac{2\pi G}{h}\cos\theta, \frac{2\pi G}{h}\sin\theta)}.$$

La vitesse numérique dépend de l'angle θ, ce qui traduit l'anisotropie numérique

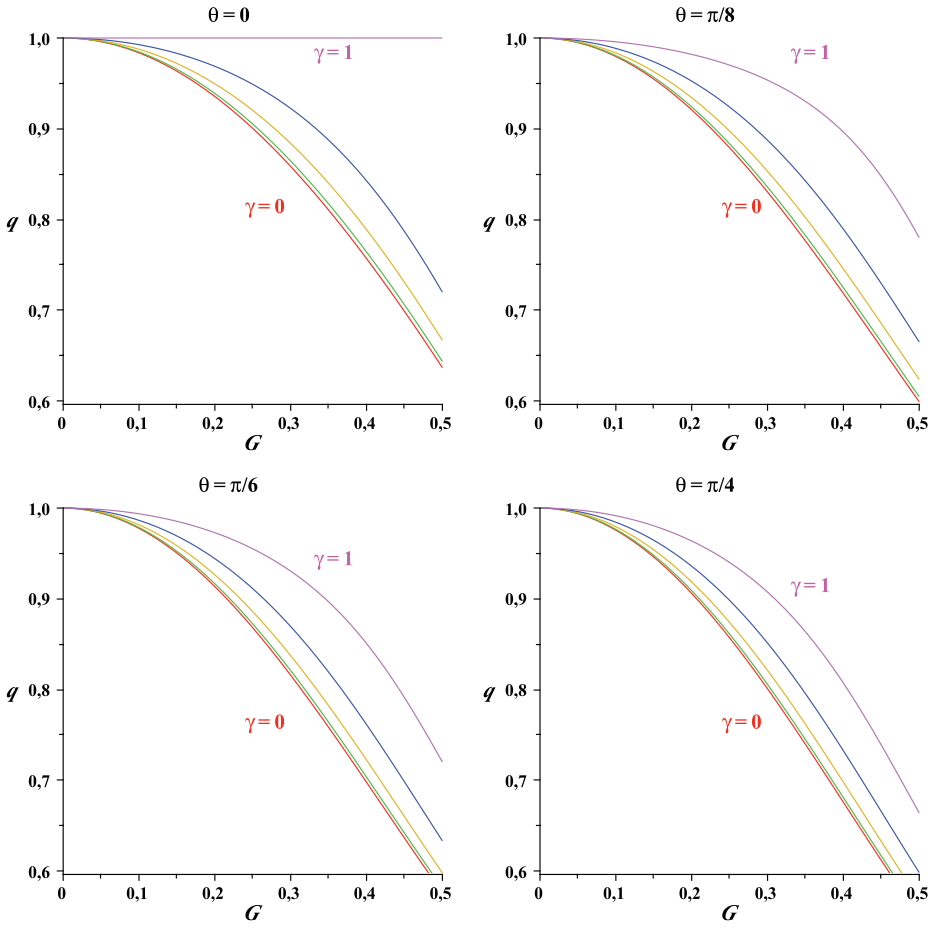


Figure 4.10. Courbes de dispersion du schéma Q^1 à 9 points pour $\theta = 0,\ \pi/8,\ \pi/6,\ \pi/4$. Sur chaque figure, les courbes correspondent aux différentes valeurs de $\gamma = 0,\ 0.25,\ 0.5,\ 0.75,\ 1$.

du schéma. On peut faire un développement de Taylor pour G petit :

$$q(\alpha_{cfl}, G, \theta) = 1 - \frac{\pi^2 G^2}{6}(1 - \alpha_{cfl}^2 + \frac{1}{2}\sin^2 2\theta) + O(G^4),$$

qui montre que l'approximation est d'ordre 2 pour la vitesse. On note que la vitesse est approchée par le bas car $1 - \alpha_{cfl}^2 + \frac{1}{2}\sin^2 2\theta \geq 1 - \alpha_{cfl}^2 \geq 0$ ($V(\alpha_{cfl}, G, \theta) \leq c$). Pour $\alpha_{cfl} = \alpha_{max} = 1$ et $\theta = 0$ (modulo $\pi/2$), on obtient la vitesse exacte

$$V(\alpha_{cfl} = 1, K, \theta = 0) = c,$$

ce qui indique que le schéma est seulement exact dans les directions du maillage. Pour α_{cfl} fixé, la fonction $\theta \in [0, \pi/4] \mapsto 1 - \alpha_{cfl}^2 + \frac{1}{2}\sin^2 2\theta$ est une fonction croissante ce qui montre que la dispersion augmente avec l'angle θ entre la direction de propagation et les directions du maillage. Ceci confirme que l'erreur est plus mauvaise dans les directions diagonales et est meilleure dans les directions du maillage. Sur les figures qui suivent, nous représentons de nouveau les courbes de dispersion (figures 4.10) et d'anisotropie (figures 4.11), le paramètre $\gamma = \alpha_{cfl}/\alpha_{max}$ étant défini cette fois-ci par $\alpha_{max} = 1$.

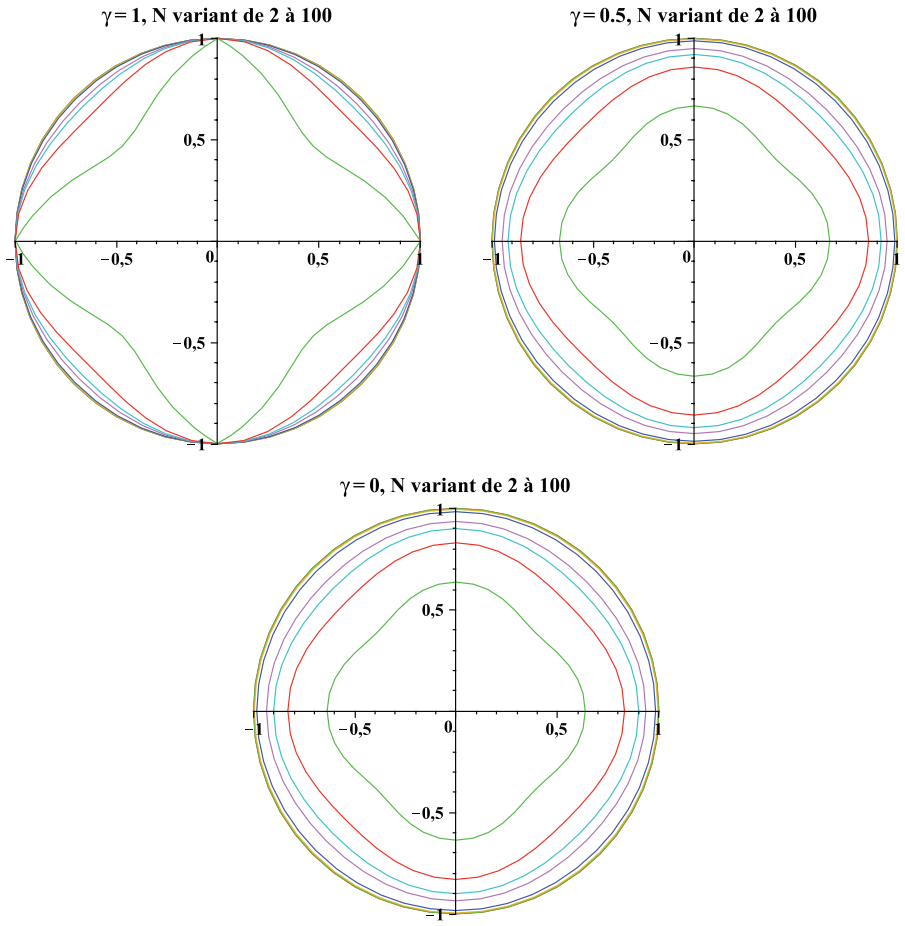


Figure 4.11. Courbes d'anisotropie du schéma Q^1 à 9 points pour différentes valeurs du nombre de points par longueur d'ondes $N = 1/G$ et pour $\gamma = 1$, 0.5 et 0.

4.7 Introduction aux Conditions aux Limites Absorbantes

Les problèmes de propagation d'ondes sont souvent posés dans des domaines non bornés. En prospection sismique le domaine d'étude est le sous-sol c'est-à-dire un demi-espace. En "scattering", on s'intéresse à la diffraction par un obstacle, le domaine d'étude est dans ce cas l'extérieur de l'obstacle. Numériquement on doit

résoudre ce problème dans un domaine borné. On doit donc borner le domaine de calcul de manière artificielle. Il existe essentiellement deux techniques pour le faire. La première, que nous n'étudierons pas ici, consiste à introduire des *couches absorbantes* qui entourent le domaine de calcul et dans lesquelles les ondes sont absorbées. La deuxième consiste à introduire des *frontières artificielles* sur lesquelles sont imposées des conditions aux limites qui *laissent passer les ondes*.
La première question est de savoir comment déterminer ces conditions aux limites artificielles. Considérons le problème de propagation d'ondes dans un ouvert $\Omega \subset \mathbb{R}^n$

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,t) - c^2 \Delta u(x,t) = f(x,t), \quad (x,t) \in \Omega \times \mathbb{R}^+, \\ u(x,0) = u_0(x), \quad x \in \Omega, \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x), \quad x \in \Omega. \end{cases}$$

Il est nécessaire de lui adjoindre des conditions aux limites sur $\partial\Omega$ pour que ce problème soit bien posé. On peut rajouter par exemple des conditions de Dirichlet ou de Neumann homogènes sur la frontière artificielle, mais ces conditions conduisent à une réflexion totale d'une onde incidente (voir l'analyse énergétique au §4.7.3). On doit construire des conditions qui *absorbent* les ondes, conditions qui sont appelées des Conditions aux Limites Absorbantes (C.L.A.). Elles devront assurer que la solution du problème en domaine borné est *proche* de la restriction à ce domaine de la solution du problème initial, ce qui suppose en particulier que la frontière ne *génère pas d'énergie*.
La première analyse du caractère bien posé de plusieurs familles de C.L.A. est due à B. Engquist et A. Majda [24] dans le contexte des ondes acoustiques en transitoire. Depuis, des travaux ont été réalisés dans plusieurs directions : développement de familles de conditions d'ordre arbitrairement élevé et stables, question du traitement des coins, C.L.A. pour des modèles plus compliqués (e.g. [49, 28, 18, 30]...).

4.7.1 Construction de la condition à la limite transparente en 2D

On s'interesse maintenant à l'équation des ondes posée dans $\Omega = \mathbb{R}^2$:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(x,y,t) - c^2 \left(\dfrac{\partial^2 u}{\partial x^2}(x,y,t) + \dfrac{\partial^2 u}{\partial y^2}(x,y,t) \right) = f(x,y,t), \ (x,y) \in \mathbb{R}^2, \ t > 0, \\ u(x,y,0) = u_0(x,y), \ (x,y) \in \mathbb{R}^2, \\ \dfrac{\partial u}{\partial t}(x,y,0) = u_1(x,y), \ (x,y) \in \mathbb{R}^2. \end{cases}$$

On suppose que les conditions initiales et le second membre de l'équation des ondes sont à support compact dans le demi-espace des x négatifs, $\Omega^- = \mathbb{R}^- \times \mathbb{R}$. Notre objectif est de ramener ce problème à un problème posé dans le demi-espace Ω^-, du type :

$$\begin{cases} \dfrac{\partial^2 v}{\partial t^2}(x,y,t) - c^2 \left(\dfrac{\partial^2 v}{\partial x^2}(x,y,t) + \dfrac{\partial^2 v}{\partial y^2}(x,y,t) \right) = f(x,y,t),\ (x,y) \in \Omega^-,\ t>0, \\ v(x,y,0) = u_0(x,y),\ (x,y) \in \Omega^-, \\ \dfrac{\partial v}{\partial t}(x,y,0) = u_1(x,y),\ (x,y) \in \Omega^-, \\ Bv(0,y,t) = 0,\ (0,y) \in \Gamma, \end{cases} \tag{4.136}$$

où Γ est la frontière artificielle $x = 0$ et B est l'opérateur de bord décrivant la C.L.A. à déterminer.
On dit que la condition à la limite artificielle est *exacte* (C.L.A.E.) ou encore qu'il s'agit d'une *condition à la limite transparente*, si la solution calculée dans le domaine restreint coïncide avec la restriction de la solution du problème dans le domaine initial, c'est-à-dire si

$$v = u_{/\Omega^-}.$$

L'idée pour la construction de la C.L.A.E. est de trouver une relation vérifiée par la solution sur Γ indépendamment des valeurs des seconds membres. Cette relation satisfaite de manière exacte, et définissant donc la condition à la limite transparente, est en général (à part en dimension 1) une relation intégro-différentielle difficile à utiliser numériquement. Elle est donc approchée par une équation aux dérivées partielles.
On cherche une relation satisfaite par la solution qui relie les traces de la solution à ses traces normales, c'est-à-dire ici entre $\partial u/\partial t$ et $\partial u/\partial n$, n étant la normale extérieure au domaine Ω^-. L'idée est d'écrire le problème satisfait par u dans le demi-espace $\Omega^+ = \mathbb{R}^+ \times \mathbb{R}$ où toutes les sources sont nulles, d'en déduire (en utilisant par exemple une transformée de Fourier en t et en y) d'abord une expression exacte de la solution dans $x > 0$ en fonction de sa trace en $x = 0^+$, puis une relation entre ses traces en $x = 0^+$, et enfin d'utiliser les relations de continuité de la solution en $x = 0$ pour déterminer la relation analogue entre ses traces en $x = 0^-$.
On se place dans le domaine Ω^+ où la solution u vérifie l'équation des ondes homogène. Pour pouvoir utiliser la transformée de Fourier en temps, on prolonge u par zéro pour les temps négatifs. La nullité des conditions initiales dans Ω^+ fait que la fonction prolongée $\widetilde{u}$ vérifie encore l'équation des ondes homogène :

$$\frac{\partial^2 \widetilde{u}}{\partial t^2}(x,y,t) - c^2 \left(\frac{\partial^2 \widetilde{u}}{\partial x^2}(x,y,t) + \frac{\partial^2 \widetilde{u}}{\partial y^2}(x,y,t) \right) = 0, \quad (x,y) \in \Omega^+,\ t \in \mathbb{R}. \tag{4.137}$$

Nous définissons la transformée de Fourier en t et en y (variables duales ω et κ_y respectivement) par :

$$\widehat{\widetilde{u}}(x,\kappa_y,\omega) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}} \int_{\mathbb{R}} u(x,y,t) e^{-i(\omega t + \kappa_y y)} dy\, dt.$$

On rappelle qu'on a la formule d'inversion :

$$\widetilde{u}(x,y,t) = \int_{\mathbb{R}}\int_{\mathbb{R}} \widehat{\widetilde{u}}(x,\kappa_y,\omega) e^{i(\omega t+\kappa_y y)} d\kappa_y\, d\omega.$$

La transformée de Fourier en t et en y de l'équation (4.137) conduit à l'équation différentielle ordinaire suivante :

$$\frac{d^2\widehat{\widetilde{u}}}{dx^2} + (\frac{\omega^2}{c^2} - \kappa_y^2)\widehat{\widetilde{u}} = 0, \text{ sur } \mathbb{R}^+,$$

dont les solutions sont de la forme

$$\widehat{\widetilde{u}}(x,\kappa_y,\omega) = \alpha(\kappa_y,\omega) e^{z(\kappa_y,\omega)x},$$

où

$$z(\kappa_y,\omega)^2 = \kappa_y^2 - \frac{\omega^2}{c^2}.$$

On ne retient que les solutions *admissibles*, qui correspondent aux ondes sortantes du domaine Ω^-, c'est-à-dire soit des ondes planes harmoniques se propageant dans la direction des $x > 0$, soit des ondes évanescentes, exponentiellement décroissantes dans la direction des $x > 0$:

$$z(\kappa_y,\omega) = \begin{cases} -\sqrt{\kappa_y^2 - \dfrac{\omega^2}{c^2}}, & \text{si } \kappa_y^2 > \dfrac{\omega^2}{c^2} \quad \text{(ondes évanescentes)}, \\ -i\dfrac{\omega}{|\omega|}\sqrt{\dfrac{\omega^2}{c^2} - \kappa_y^2}, & \text{si } \kappa_y^2 < \dfrac{\omega^2}{c^2} \quad \text{(ondes propagatives)}. \end{cases} \tag{4.138}$$

Il est alors facile d'obtenir une relation entre la trace de u et sa trace normale dans l'espace de Fourier :

$$\begin{aligned} &\widehat{\widetilde{u}}(x,\kappa_y,\omega) = \widehat{\widetilde{u}}(0^+,\kappa_y,\omega) e^{z(\kappa_y,\omega)x}, \\ &\frac{\partial\widehat{\widetilde{u}}}{\partial x}(x,\kappa_y,\omega) = z(\kappa_y,\omega)\widehat{\widetilde{u}}(0^+,\kappa_y,\omega) e^{z(\kappa_y,\omega)x}, \end{aligned}$$

d'où la relation

$$\frac{\partial\widehat{\widetilde{u}}}{\partial x}(0^+,\kappa_y,\omega) = z(\kappa_y,\omega)\widehat{\widetilde{u}}(0^+,\kappa_y,\omega).$$

Le terme $z(\kappa_y,\omega)$ est appelé le symbole de l'opérateur Dirichlet to Neumann. En utilisant maintenant les conditions de raccord de u en $x = 0$:

$$\begin{aligned} &\frac{\partial\widehat{\widetilde{u}}}{\partial x}(0^+,\kappa_y,\omega) = \frac{\partial\widehat{\widetilde{u}}}{\partial x}(0^-,\kappa_y,\omega), \\ &\widehat{\widetilde{u}}(0^+,\kappa_y,\omega) = \widehat{\widetilde{u}}(0^-,\kappa_y,\omega), \end{aligned}$$

on obtient la condition transparente exprimée dans l'espace de Fourier :

$$\frac{\partial \widehat{\widetilde{u}}}{\partial x}(0^-, \kappa_y, \omega) = z(\kappa_y, \omega)\widehat{\widetilde{u}}(0^-, \kappa_y, \omega). \tag{4.139}$$

Cette relation peut s'écrire dans le domaine espace-temps à l'aide de la transformation de Fourier inverse :

$$\frac{\partial u}{\partial x}(0^-, \kappa_y, \omega) = \int_{\mathbb{R}}\int_{\mathbb{R}} z(\kappa_y, \omega)\widehat{\widetilde{u}}(0^-, \kappa_y, \omega)e^{i(\omega t + \kappa_y y)}\, d\omega\, d\kappa_y, \tag{4.140}$$

soit encore :

$$\frac{\partial u}{\partial x}(0^-, \kappa_y, \omega) = \mathcal{F}^{-1}_{y,t}(z(\kappa_y, \omega)) \overset{(y,t)}{*} u(0^-, y, t). \tag{4.141}$$

En utilisant l'expression du symbole $z(\kappa_y, \omega)$, nous pouvons l'exprimer sous la forme

$$\begin{aligned}\frac{\partial u}{\partial x}(0^-, \kappa_y, \omega) = &\int_{\mathbb{R}}\int_{|\kappa_y|<|\omega/c|} -i\frac{\omega}{|\omega|}\sqrt{\frac{\omega^2}{c^2} - \kappa_y^2}\ \widehat{\widetilde{u}}(0^-, \kappa_y, \omega)e^{i(\omega t + \kappa_y y)}\, d\omega\, d\kappa_y, \\ &+ \int_{\mathbb{R}}\int_{|\kappa_y|>|\omega/c|} -\sqrt{\kappa_y^2 - \frac{\omega^2}{c^2}}\ \widehat{\widetilde{u}}(0^-, \kappa_y, \omega)e^{i(\omega t + \kappa_y y)}\, d\omega\, d\kappa_y.\end{aligned} \tag{4.142}$$

La présence de la racine carrée dans la définition de $z(\kappa_y, \omega)$ conduit donc à une condition transparente (exacte) de type intégro-différentielle (non locale). En pratique, cette condition est difficile à discrétiser, c'est pourquoi on va l'approcher par des conditions locales.

4.7.2 Approximations de la condition transparente

La condition transparente est non locale du fait de la racine carrée présente dans la définition du symbole $z(\kappa_y, \omega)$. L'idée de base pour la construction de conditions approchées locales, est de remarquer que si $z(\kappa_y, \omega)$ était une fraction rationnelle en κ_y et ω, on obtiendrait alors une condition locale, exprimée simplement à l'aide d'opérateurs aux dérivées partielles.
Les conditions approchées sont alors construites de façon à approcher la condition transparente pour des ondes propagatives, c'est-à-dire pour

$$\varepsilon := \frac{c^2\kappa_y^2}{\omega^2} < 1.$$

Le paramètre ε est lié à l'incidence de l'onde sur la frontière artificielle. Ainsi, $\varepsilon = 0$ correspond à des ondes propagatives d'incidence normale et la limite ε proche de 1 correspond à des ondes rasantes. Ces conditions approchées seront

appelées *Conditions aux Limites Absorbantes* (C.L.A.). Rappelons que pour $\varepsilon < 1$, le symbole exprimé en (4.138) a pour expression

$$z(\kappa_y, \omega) = -i\frac{\omega}{c}\sqrt{1 - \frac{c^2\kappa_y^2}{\omega^2}} = -i\frac{\omega}{c}\sqrt{1-\varepsilon}.$$

Condition à la limite absorbante d'ordre 1

La première approximation de la racine, à l'ordre 1

$$\sqrt{1-\varepsilon} = 1 + O(\varepsilon),$$

conduit à la condition à la limite écrite dans l'espace de Fourier :

$$\frac{\partial \widehat{\widetilde{u}}}{\partial x}(0^-, \kappa_y, \omega) = -i\frac{\omega}{c}\widehat{\widetilde{u}}(0^-, \kappa_y, \omega).$$

En prenant la transformée de Fourier inverse ($i\omega \longleftrightarrow \partial_t$) on obtient la C.L.A. d'ordre 1 :

$$\frac{\partial u}{\partial n}(0, y, t) + \frac{1}{c}\frac{\partial u}{\partial t}(0, y, t) = 0, \quad (0, y) \in \Gamma,\ t \in \mathbb{R}, \tag{4.143}$$

où n est la normale extérieure à Ω^-.

Condition à la limite absorbante d'ordre 2

En utilisant le développement de Taylor à l'ordre 2 de la racine, nous obtenons

$$\sqrt{1-\varepsilon} = 1 - \frac{1}{2}\varepsilon + O(\varepsilon^2),$$

qui conduit à

$$\frac{\partial \widehat{\widetilde{u}}}{\partial x}(0^-, \kappa_y, \omega) = -i\frac{\omega}{c}\ \widehat{\widetilde{u}}(0^-, \kappa_y, \omega) - \frac{c}{2}\frac{\kappa_y^2}{i\omega}\ \widehat{\widetilde{u}}(0^-, \kappa_y, \omega).$$

Le terme $c\kappa_y^2/(2i\omega)$ ne correspond pas à un opérateur aux dérivées partielles, car il n'est pas polynomial en (κ_y, ω). À ce stade, on a deux possibilités :

– On multiplie tout par $i\omega$, ce qui nous donne de part et d'autre des polynômes. En prenant la transformée de Fourier inverse, on obtient la C. L. A. d'ordre 2 :

$$\frac{1}{c}\frac{\partial^2 u}{\partial t \partial n}(0, y, t) + \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(0, y, t) - \frac{1}{2}\frac{\partial^2 u}{\partial \tau^2}(0, y, t) = 0, \quad (0, y) \in \Gamma,\ t \in \mathbb{R}, \tag{4.144}$$

où $\partial/\partial\tau$ désigne la dérivée tangentielle sur Γ (ici $\partial/\partial y$).

– On introduit la fonction auxiliaire

$$\widehat{\phi}(\kappa_y,\omega)=\frac{c}{2}\frac{\kappa_y^2}{i\omega}\,\widehat{\widetilde{u}}(0^-,\kappa_y,\omega),$$

ce qui permet de réécrire la C. L. A. d'ordre 2 sous la forme d'un système du premier ordre :

$$\begin{cases}\dfrac{\partial u}{\partial n}(0,y,t)+\dfrac{1}{c}\dfrac{\partial u}{\partial t}(0,y,t)+\phi(y,t)=0,\\[2ex] \dfrac{1}{c}\dfrac{\partial \phi}{\partial t}(y,t)=-\dfrac{1}{2}\dfrac{\partial^2 u}{\partial \tau^2}(0,y,t),\end{cases}\qquad (0,y)\in\Gamma,\ t\in\mathbb{R}. \tag{4.145}$$

On peut vérifier que les deux formulations (4.144) et (4.145) sont équivalentes. L'avantage de cette dernière formulation est qu'elle s'adapte plus facilement à une mise en œuvre numérique et est aisément généralisable aux ordres plus élevés.

Condition à la limite absorbante d'ordre plus élevé

L'idée naturelle pour obtenir des approximations d'ordre plus élevé de la condition transparente est d'utiliser un développement de Taylor d'ordre plus élevé de la racine. Mais à partir de l'ordre 3, il a été montré (voir [24]) que les développements de Taylor conduisent à des conditions aux limites instables, c'est-à-dire que le problème aux limites associé devient mal posé (voir section suivante). Une alternative est alors d'approcher la racine carrée par des fractions rationnelles, par exemple grâce à des développements de Padé, qui cette fois-ci conduisent à des problèmes bien posés. Nous ne développons pas ici la construction et l'analyse de ces conditions et renvoyons le lecteur intéressé aux travaux de Engquist et Majda [24] et de Collino [18].

4.7.3 Questions de stabilité. Critère de Kreiss

La question du caractère bien posé du problème aux limites (4.136) est délicate et nous nous contentons de l'aborder très brièvement ici. Il existe essentiellement deux méthodes pour analyser le caractère bien posé d'un problème aux limites : des techniques d'énergie et la théorie de Kreiss.

Techniques d'énergie

Dans certains cas, il est possible d'exhiber une estimation d'énergie pour le problème aux limites. Si cette estimation montre que l'énergie est soit conservée, soit décroissante, alors le caractère bien posé du problème en découle de façon immédiate (voir théorème 4.8). L'avantage de cette approche est que, la plupart du temps, elle ne se limite pas aux milieux homogènes. Mais à part pour la condition

aux limites absorbante d'ordre 1 (voir exemple ci-après), il n'est souvent pas facile de trouver des estimations d'énergie (voir [30]).

Théorie de Kreiss

Nous n'entrerons pas dans les détails de cette théorie mais en donnons un bref aperçu (voir [36]). Elle repose sur l'analyse des *ondes planes entrantes* dans le domaine Ω^-, c'est-à-dire sur les solutions du problème posé dans Ω^- :

$$\begin{cases} \dfrac{\partial^2 v}{\partial t^2}(x,y,t) - c^2 \left(\dfrac{\partial^2 v}{\partial x^2}(x,y,t) + \dfrac{\partial^2 v}{\partial y^2}(x,y,t) \right) = 0, \ (x,y) \in \Omega^-, \ t > 0, \\ Bv(0,y,t) = 0, \ (0,y) \in \Gamma, \ t > 0, \end{cases}$$

de la forme

$$v(x,y,t) = e^{i(\omega t + \kappa_x x + \kappa_y y)}, \tag{4.146}$$

et qui sont admissibles, c'est-à-dire, soit des ondes planes évanescentes dans la direction des $x < 0$, soit des ondes planes harmoniques se propageant dans la direction des $x < 0$.
De même que pour le problème posé dans tout l'espace, la recherche de ces solutions particulières conduit à une *relation de dispersion* constituée ici de deux équations scalaires, la première équation provenant de l'équation intérieure et la seconde, de la condition à la limite, écrite sous la forme vectorielle

$$\boldsymbol{F}(\kappa_x, \kappa_y, \omega) = 0. \tag{4.147}$$

Énonçons maintenant une condition nécéssaire du caractère bien posé.

Définition 4.47 *Si il existe des solutions $(\kappa_x, \kappa_y, \omega)$ de (4.147) qui vérifient*

$$\kappa_y \in \mathbb{R}, \quad \mathbf{Im}\ \kappa_x < 0, \quad \mathbf{Im}\ \omega < 0, \tag{4.148}$$

alors le problème (4.136) est mal posé. Dans le cas contraire, on dit qu'il est bien posé au sens de Kreiss.

Remarque 4.48 *Une interprétation de cette condition est que le problème est bien posé s'il n'existe pas de solution entrante (c'est-à-dire de solutions admissibles se propageant dans le demi-espace $x < 0$) qui soit exponentiellement croissante en temps. La théorie de Kreiss distingue plusieurs notions de problème bien posé : faiblement ou fortement bien posé. La définition 4.47 caractérise un problème faiblement bien posé. Nous n'aborderons pas ici la notion de fortement bien posé qui est plus délicate et plus restrictive.*

Illustrons cette question de la stabilité au travers de deux exemples.

Condition à la limite absorbante d'ordre 1 : technique d'énergie

Pour cette condition, il est facile de montrer la décroissance de l'énergie $E(t)$ définie par

$$E(t) = \frac{1}{2}\int_{\Omega^-} \left(\left(\frac{\partial v}{\partial t} \right)^2 + |c\nabla v|^2 \right) d\Omega.$$

En effet, si v est solution de (4.136) où on suppose $f = 0$, avec la condition à la limite (4.143), on montre que l'énergie vérifie

$$\frac{dE}{dt}(t) = \int_{\Gamma} \frac{\partial v}{\partial n}\frac{\partial v}{\partial t} d\gamma,$$

soit en utilisant la condition à la limite :

$$\frac{dE}{dt}(t) = -\frac{1}{c}\int_{\Gamma} \left(\frac{\partial v}{\partial t} \right)^2 d\gamma \leq 0.$$

L'énergie étant décroissante, le problème est bien posé.

Condition à la limite absorbante d'ordre 2 : théorie de Kreiss

On s'intéresse ici au problème (4.136), avec la condition à la limite (4.144) et on va appliquer le critère de Kreiss. On cherche donc la relation de dispersion du système, obtenue en cherchant les ondes planes de la forme (4.146) :

$$\begin{aligned} &\text{(i)} \;\; \omega^2 = c^2(\kappa_x^2 + \kappa_y^2), \\ &\text{(ii)} \; \omega c\kappa_x + \omega^2 - \frac{1}{2}c^2\kappa_y^2 = 0. \end{aligned} \tag{4.149}$$

D'après (i) on a $c^2\kappa_y^2 = \omega^2 - c^2\kappa_x^2$, qui permet d'éliminer κ_y dans (ii), ce qui montre finalement que les solutions vérifient

$$(\omega + c\kappa_x)^2 = 0 \Longrightarrow \omega = -c\kappa_x.$$

En particulier, on en déduit que $\mathbf{Im}\ \omega = -c\mathbf{Im}\ \kappa_x$. Les deux parties imaginaires ne peuvent donc pas être strictement négatives en même temps, ce qui montre que le problème est bien posé au sens de Kreiss.

4.7.4 Analyse de la précision des C. L. A.

Une façon de quantifier la précision d'une condition à la limite absorbante est d'étudier la réflexion d'une onde plane harmonique arrivant sur la frontière Γ. On considère une onde incidente

$$u^I(x, y, t) = e^{i(\omega t - \kappa_x x + \kappa_y y)},$$

avec $\omega/\kappa_x > 0$ (onde se propageant vers la droite), et $\omega^2 = c^2|\kappa|^2$ (solution de l'équation des ondes). On suppose ici que $\omega \in \mathbb{R}$, et $(\kappa_x, \kappa_y) \in \mathbb{R}^2$. Dans ce qui suit, nous allons supposer que $\omega > 0$, et donc que $\kappa_x > 0$. L'angle d'incidence sur la frontière Γ est défini par

$$\kappa = |\kappa|(\cos\theta, \sin\theta) = \frac{\omega}{c}(\cos\theta, \sin\theta).$$

L'onde totale dans Ω^- se décompose alors en :

$$u(x, y, t) = u^I + u^R,$$

où u^R est l'onde réfléchie par Γ, de la forme

$$u^R = R(\theta)e^{i(\omega t + \kappa_x x + \kappa_y y)},$$

soit encore :

$$u(x, y, t) = e^{i\omega(t+\sin\theta/c)}(e^{i\omega\cos\theta\ x/c} + R(\theta)e^{-i\omega\cos\theta\ x/c}).$$

Le coefficient de réflexion $R(\theta)$ est déterminé par la C.L.A..

Condition à la limite absorbante d'ordre 1

Pour la C.L.A. d'ordre 1 (4.143), on obtient :

$$R(\theta) = \frac{1 - \cos\theta}{1 + \cos\theta}. \tag{4.150}$$

Pour $\theta = 0$, on a $R(\theta) = 0$, ce qui signifie que l'onde incidente à incidence normale est totalement absorbée. Pour θ petit, le coefficient de réflexion devient

$$R(\theta) \sim \frac{1}{2}\theta^2 \text{ et } |R(\theta)| < 10^{-2} \text{ pour } |\theta| \leq 11°,$$

donc $R(\theta)$ est d'ordre 2 en θ.

Condition à la limite absorbante d'ordre 2

Pour la C.L.A. d'ordre 2 (4.144), on obtient :

$$R(\theta) = \left(\frac{1 - \cos\theta}{1 + \cos\theta}\right)^2. \tag{4.151}$$

On a toujours absence de réflexion pour $\theta = 0$, on a $R(\theta) = 0$, et pour θ petit, le coefficient de réflexion devient d'ordre 4 :

$$R(\theta) \sim \frac{1}{4}\theta^4 \text{ et } |R(\theta)| < 10^{-2} \text{ pour } |\theta| \leq 35°,$$

ce qui montre que l'approximation est bien meilleure que la précédente pour de plus grands angles d'incidence.

4.8 Illustrations numériques

Nous présentons dans cette section quelques exemples de la résolution de l'équation des ondes. Nous nous sommes intéressés, d'une part, au cas de la propagation des ondes acoustiques dans un carré constitué d'un milieu homogène et, d'autre part, au problème de la diffraction des ondes dans un milieu stratifié qui correspond par exemple à la propagation acoustique sous-marine en présence d'une interface air-mer. Dans ce dernier cas, nous utilisons des conditions absorbantes du premier ordre pour simuler la propagation dans un milieu non borné.

4.8.1 Résolution de l'équation des ondes

On considère l'équation des ondes acoustiques sur le carré unité $\Omega =]0,1[\times]0,1[$ où $p(x,y,t)$ désigne la pression :

$$\begin{cases} \dfrac{\partial^2 p}{\partial t^2}(x,t) - c^2 \Delta p(x,t) = f(x,t), & (x,t) \in \Omega,\ t > 0, \\ \dfrac{\partial p}{\partial n}(x,t) = 0, & (x,t) \in \partial\Omega,\ t > 0, \\ p(x,0) = p_0(x),\ \dfrac{\partial p}{\partial t}(x,t) = p_1(x), & x \in \Omega. \end{cases}$$

On vérifie facilement que la formulation variationnelle de ce problème est ($f \in L^1(0,T;L^2(\Omega))$, $p_0 \in H^1(\Omega)$ et $p_1 \in L^2(\Omega)$) :

$$\begin{cases} \text{trouver}\ \ p \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H^1(\Omega)) \text{ tel que} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega p(x,t)\, v(x)\, dx\right) + c^2 \int_\Omega \nabla p(x,t) \cdot \nabla v(x)\, dx \\ \qquad\qquad = \displaystyle\int_\Omega f(x,t)\, v(x)\, dx, \quad \forall v \in H^1(\Omega),\ t \in]0,T[, \\ p(x,0) = p_0(x), \dfrac{dp}{dt}(x,0) = p_1(x),\ x \in \Omega. \end{cases} \tag{4.152}$$

La théorie de l'existence et de l'unicité d'une solution faible de l'équation des ondes a été développée précédemment dans le cas où on impose une condition de Dirichlet homogène. Ici on considère le cas où une condition de Neumann homogène est imposée. La théorie s'étend sans difficulté à cette situation.
On considère dans la suite une approximation par éléments finis de Lagrange (en pratique d'ordre 1 ou 2). On note $\mathcal{T} = \bigcup_{\ell=1,L} T_\ell$ une triangulation conforme du domaine Ω, $(w_I)_{I=1,N}$ les fonctions de base globales de Lagrange (P^1 ou P^2) associées aux nœuds $(S_I)_{I=1,N}$ du maillage et on définit l'espace d'approximation de $H^1(\Omega)$:

$$V_h = \text{vect}(w_I)_{I=1,N}.$$

La discrétisation de la formulation variationnelle (4.152) dans l'espace V_h conduit à la formulation semi-discrétisée :

$$\left\{\begin{array}{l} \text{trouver } \; p_h \in \mathcal{C}^1(0,T;V_h) \text{ tel que} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega p_h(x,t)\, v_h(x)\, dx\right) + c^2 \displaystyle\int_\Omega \nabla p_h(x,t)\cdot\nabla v_h(x)\, dx \\ \qquad\qquad = \displaystyle\int_\Omega f(x,t)\, v_h(x)\, dx, \;\; \forall v_h \in V_h, \;\; t \in]0,T[, \\ p_h(x,0) = \Pi_h p_0(x) = p_{h0}(x), \dfrac{dp}{dt}(x,0) = \Pi_h p_1(x) = p_{h1}(x), \quad x \in \Omega \end{array}\right. \tag{4.153}$$

où Π_h désigne une projection sur V_h (par exemple l'opérateur d'interpolation si les fonctions p_0 et p_1 sont continues). Cette formulation s'écrit sous forme matricielle :

$$\left\{\begin{array}{l} \text{trouver } \; \overrightarrow{P} \in \mathcal{C}^1(0,T;\mathbb{R}^N) \text{ tel que} \\ \mathbb{M}\dfrac{d^2}{dt^2}\overrightarrow{P}(t) + c^2\mathbb{K}\overrightarrow{P}(t) = \overrightarrow{F}(t), \;\; t \in]0,T[\\ \overrightarrow{P}(0) = \overrightarrow{P}_0, \dfrac{d\overrightarrow{P}}{dt}(0) = \overrightarrow{P}_1, \end{array}\right. \tag{4.154}$$

où $\overrightarrow{P}(t)$ est le vecteur de $\mathbb{R}^N$ de composantes $p_h(S_I,t)$, $\overrightarrow{P}_0$ et $\overrightarrow{P}_1$ les vecteurs de $\mathbb{R}^N$ de composantes $p_{h0}(S_I)$ et $p_{h1}(S_I)$, $\mathbb{M}$, $\mathbb{K}$ les matrices d'ordre N

$$\mathbb{M}_{IJ} = \int_\Omega w_I \; w_J dx, \quad \mathbb{K}_{IJ} = \int_\Omega \nabla w_I \cdot \nabla w_J dx, \quad I,J = 1,N$$

et $\overrightarrow{F}(t)$ le vecteur de $\mathbb{R}^N$ de composantes

$$F_I(t) = \int_\Omega f(x,t)\, w_I(x)\, dx, \quad I = 1,N.$$

Par la suite, on suppose que la fonction f est suffisamment régulière de telle sorte que l'on puisse l'approcher par son interpolée $\pi_h f$. On remplace alors $\overrightarrow{F}(t)$ par :

$$\mathbb{M}\,\overrightarrow{f}(t) \text{ avec } f_I(t) = f(S_I,t), \;\; I = 1,N.$$

En pratique, on utilise à la place de la matrice de masse la matrice de masse condensée $\mathbb{D}$ (cf. §4.5.1) qui, dans le cas P^1, est la matrice diagonale définie par :

$$\mathbb{D}_{II} = \sum_{J=1,N} \mathbb{M}_{IJ}.$$

Le nouveau système différentiel s'écrit alors

$$\begin{cases} \text{trouver } \overrightarrow{P} \in \mathcal{C}^1(0,T;\mathbb{R}^N) \text{ tel que} \\ \dfrac{d^2}{dt^2}\overrightarrow{P}(t) + c^2\mathbb{D}^{-1}\mathbb{K}\overrightarrow{P}(t) = \overrightarrow{f}(t), \quad t \in]0,T[\\ \overrightarrow{P}(0) = \overrightarrow{P}_0, \dfrac{d\overrightarrow{P}}{dt}(0) = \overrightarrow{P}_1. \end{cases} \tag{4.155}$$

Sur une discrétisation $(t_k)_{k=0,K}$ de l'intervalle de temps $[0,T]$, que l'on suppose uniforme pour simplifier ($t_k = k\Delta t$ et $\Delta t = T/K$), on considère le schéma saute-mouton associé au système différentiel (4.155), $\overrightarrow{P}^k$ désignant une approximation au temps t_k de $\overrightarrow{P}(t_k)$:

$$\begin{cases} \overrightarrow{P}^0 = \overrightarrow{P}_0,\ \overrightarrow{P}^1 = \overrightarrow{P}_0 + \Delta t\overrightarrow{P}_1 + \dfrac{\Delta t^2}{2}(\overrightarrow{f}(0) - c^2\mathbb{D}^{-1}\mathbb{K}\overrightarrow{P}_0), \\ \overrightarrow{P}^{k+1} = 2\overrightarrow{P}^k - \overrightarrow{P}^{k-1} - c^2\Delta t^2\mathbb{D}^{-1}\mathbb{K}\overrightarrow{P}^k + \Delta t^2\overrightarrow{f}(t_k)\ \forall k = 1, K-1. \end{cases} \tag{4.156}$$

Ici on utilise une approximation consistante à l'ordre 2 de la condition initiale sur la vitesse (cf. 4.92).

Mise en œuvre

La mise en œuvre ne soulève aucune difficulté. On réutilise les fonctions Matlab permettant de calculer les matrices éléments finis P^1 $\mathbb{M}$ et $\mathbb{K}$ (calcul_EF_2D.m). Le script Matlab suivant implémente le schéma (4.156) :

```
p=100;c=1;
[S,T,BR,RT]=triangule_rectangle([0 1 0 1],p,p,1);   %maillage P1 du carré
[K,M]=calcul_EF_2D(S,T,RT);                          %matrices EF
ns=size(M,1);
Q = spdiags(1./sum(M)', 0, ns, ns)*K; clear K;       %condensation masse
F=f(S);                                              %données des fonctions
Pk=zeros(ns,1);Pkm=zeros(ns,1);Pkp=zeros(ns,1);
A=100000;gamma=-2000;tg=0.05;                        %données de la source
t=0;tf=1; dt=0.005;
while(t<tf)
    gt=A*(t<2*tg)*exp(-gamma*(t-tg)*(t-tg));         %gaussienne en temps
    Z=gt*F-Q*Pk;
    Pkp=2*Pk-Pkm+dt*dt*Z;
    t=t+dt;Pkm=Pk;Pk=Pkp;
end,
trisurf(T,S(:,1),S(:,2),Pkp);                        %représentation graphique
axis image;colormap bone;shading interp; view(2);

function z=f(X)  %indicatrice d'un disque
a0=0.5;b0=0.5;r0=0.01;
x=X(:,1)-a0; y=X(:,2)-b0;
z=(sqrt(x.*x+y.*y)<=r0);
```

Ce script correspond au cas de la résolution de l'équation des ondes avec une source localisée sur le disque D_0 de centre (a_0, b_0) et de rayon r_0 et suivant une loi gaussienne centrée en t_g :

$$f(x,t) = \begin{cases} Ae^{-\gamma(t-t_g)^2} & \text{si } t < 2t_g \text{ et } x \in D_0 \\ 0 & \text{sinon} \end{cases} . \tag{4.157}$$

On a considéré le cas de conditions initiales nulles ($p_0 = p_1 = 0$).

Essais numériques

Nous avons réalisé des simulations numériques pour divers choix du pas d'espace et du pas de temps avec les paramètres suivants : $a_0 = b_0 = 0.01$, $r_0 = 0.5$, l'amplitude $A = 10^5$, la rampe $\gamma = 2000$, et $t_g = 0.05$. La première série de graphes (fig. 4.12) représente la pression obtenue à l'instant $t = 1$ avec un pas d'espace fixé à $h = 10^{-2}$ (maillage régulier) et pour différentes valeurs du pas de temps $\Delta t = 10^{-3}$, $\Delta t = 0.00666$ et $\Delta t = 0.0067$.
On observe que le schéma est stable pour les deux premières valeurs du pas de temps et qu'il est instable pour la valeur $\Delta t = 0.0067$. Nous avons déterminé cette valeur de façon expérimentale. Elle est à rapprocher de la valeur théorique estimée à partir du schéma aux différences finies (4.134) qui conduit au pas de temps limite $h/\sqrt{2} \approx 0.00707$. Nous n'observons pas tout à fait la même limite de stabilité car le maillage utilisé (bien que régulier) ne coïncide pas exactement avec le schéma à 5 points du laplacien. Afin de retrouver le schéma à 5 points il aurait fallu choisir de découper tous les quadrangles de côtés h du maillage suivant la même direction ! Notons que la solution obtenue pour des valeurs stables du pas de temps n'est pas excellente. En effet, on observe un fort effet de dispersion (risées derrière le front d'onde).
Nous représentons sur la figure 4.13 la pression acoustique obtenue pour un pas

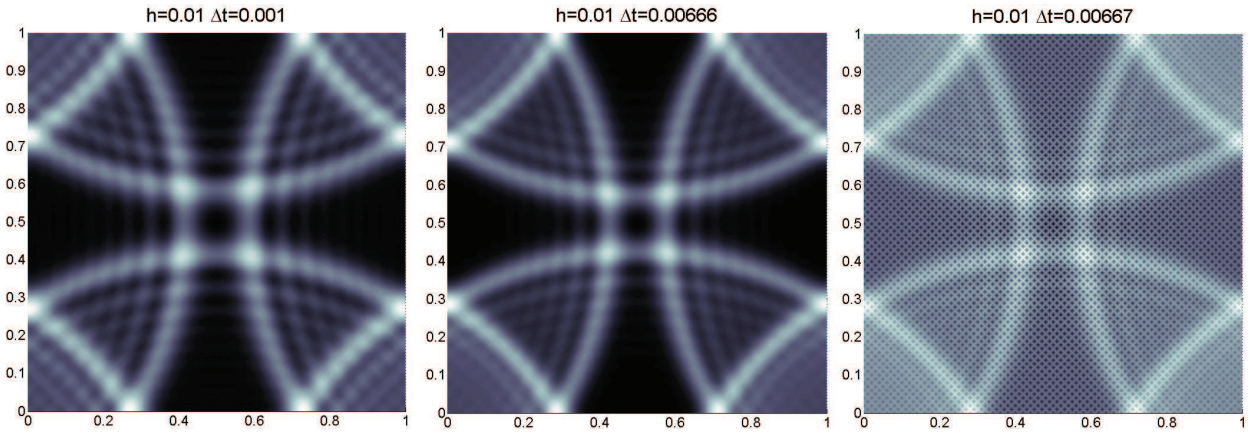


Figure 4.12. Solution à l'instant $t = 1$ pour différentes valeurs du pas de temps Δt.

de temps fixé $\Delta t = 10^{-3}$ et pour les différentes valeurs du pas d'espace $h = 10^{-2}$, $h = 5.\,10^{-3}$ et $h = 3.33\,10^{-3}$. On observe ainsi que l'erreur de dispersion diminue rapidement avec le pas de maillage ; cette erreur n'étant plus discernable pour le maillage le plus fin. Cela rejoint l'analyse de dispersion théorique qui a

été préalablement réalisée. Remarquons qu'il est nécessaire de diminuer de façon

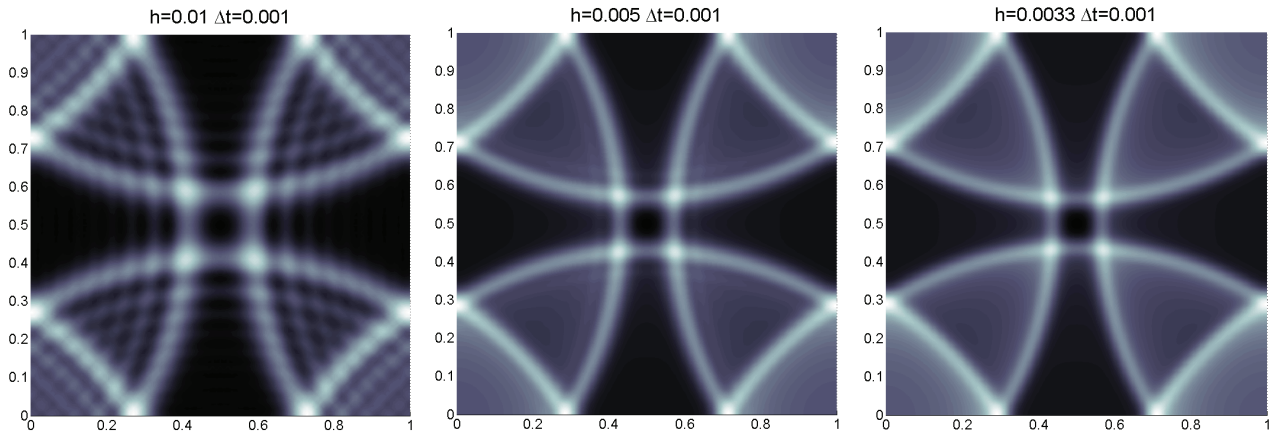


Figure 4.13. Solution à l'instant $t = 1$ pour différentes valeurs du pas d'espace h.

importante le pas d'espace pour obtenir une faible erreur de dispersion. Ici le maillage P^1 le plus fin ($h = 3.33^{-3}$) requiert de l'ordre de 90000 degrés de liberté ! C'est une des raisons qui motive l'utilisation de schémas d'ordre plus élevé en espace et en temps.

4.8.2 Résolution d'un problème de diffraction

On s'intéresse maintenant à un problème de diffraction d'une onde par un obstacle borné D plongé dans un milieu stratifié non borné. Cet exemple va nous permettre d'illustrer la prise en compte de condition absorbante permettant de limiter le domaine de calcul à un domaine borné. On considère un milieu stratifié constitué des domaines non bornés $\Omega_a = \mathbb{R} \times \mathbb{R}_*^+$ et $\Omega_m = \mathbb{R}\times]-h, 0[\backslash\overline{D}$ séparés par l'interface Γ_{am} (voir figure 4.14). Dans la suite, on pose $\Omega = \mathbb{R}\times]-h, +\infty[\backslash\overline{D}$.
La pression acoustique $p(x,t)$ satisfait les équations suivantes :

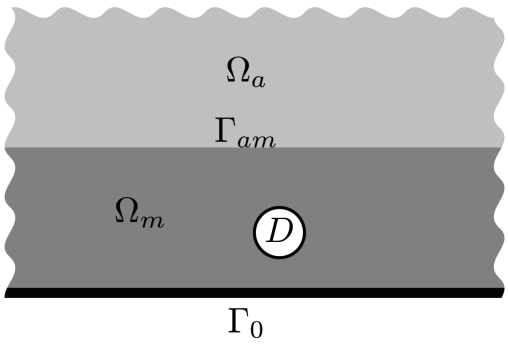


Figure 4.14. Domaine stratifié

$$\begin{cases} \dfrac{\partial^2 p}{\partial t^2}(x,t) - c_a^2 \Delta p(x,t) = f_a(x,t), & (x,t) \in \Omega_a \times]0,T[, \\ \dfrac{\partial^2 p}{\partial t^2}(x,t) - c_m^2 \Delta p(x,t) = f_m(x,t), & (x,t) \in \Omega_m \times]0,T[, \\ [p]\,(x,t) = \left[c^2 \dfrac{\partial p}{\partial n}\right](x,t) = 0, & (x,t) \in \Gamma_{am} \times]0,T[, \\ \dfrac{\partial p}{\partial n}(x,t) = 0, & (x,t) \in (\Gamma_0 \cup \partial D) \times]0,T[, \\ p(x,0) = p_0(x), \dfrac{dp}{dt}(x,0) = p_1(x), & x \in \Omega, \end{cases}$$

où $[q] = (q_{|\Omega_a})_{|\Gamma_{am}} - (q_{|\Omega_m})_{|\Gamma_{am}}$ représente le saut à travers l'interface Γ_{am} et f_a et f_m des sources de support borné, avec :

$$c(x) = \begin{cases} c_a \text{ si } x \in \Omega_a \\ c_m \text{ si } x \in \Omega_m \end{cases} \text{ et } f(x) = \begin{cases} f_a(x) \text{ si } x \in \Omega_a \\ f_m(x) \text{ si } x \in \Omega_m \end{cases}.$$

Les conditions de sauts nuls sur l'interface Γ_{am} sont les conditions usuelles de transmission qui traduisent la continuité de la pression ainsi que la continuité du flux acoustique. Nous imposons des conditions de réflexion parfaite sur le fond Γ_0 et le bord de l'obstacle ∂D.
Afin de mener une simulation numérique, nous allons introduire un problème posé dans un domaine borné obtenu en introduisant des frontières artificielles $\Sigma_a^{\pm}$, $\Sigma_m^{\pm}$ et Γ_a (cf. figure 4.15) sur lesquelles nous imposons des conditions absorbantes du premier ordre (par abus de notation, on note les nouveaux domaines bornés de la même façon qu'auparavant) :

$$\begin{cases} \dfrac{\partial^2 p}{\partial t^2}(x,t) - c_a^2 \Delta p(x,t) = f_a(x,t), & (x,t) \in \Omega_a \times]0,T[, \\ \dfrac{\partial^2 p}{\partial t^2}(x,t) - c_m^2 \Delta p(x,t) = f_m(x,t), & (x,t) \in \Omega_m \times]0,T[, \\ [p]\,(x,t) = \left[c^2 \dfrac{\partial p}{\partial n}\right](x,t) = 0, & (x,t) \in \Gamma_{am} \times]0,T[, \\ \dfrac{\partial p}{\partial n}(x,t) = 0, & x \in \Gamma_0 \cup \partial D,\ t \in]0,T[\\ \dfrac{\partial p}{\partial n}(x,t) + \dfrac{1}{c_a}\dfrac{\partial p}{\partial t}(x,t) = 0, & x \in \Sigma_a^+ \cup \Sigma_a^- \cup \Gamma_a \overset{def}{=} \Sigma_a,\ t \in]0,T[\\ \dfrac{\partial p}{\partial n}(x,t) + \dfrac{1}{c_m}\dfrac{\partial p}{\partial t}(x,t) = 0, & x \in \Sigma_m^+ \cup \Sigma_m^- \overset{def}{=} \Sigma_m,\ t \in]0,T[\\ p(x,0) = p_0(x), \dfrac{dp}{dt}(x,0) = p_1(x), & x \in \Omega. \end{cases}$$

De façon classique, en multipliant par une fonction test $v \in H^1(\Omega)$ (en particulier

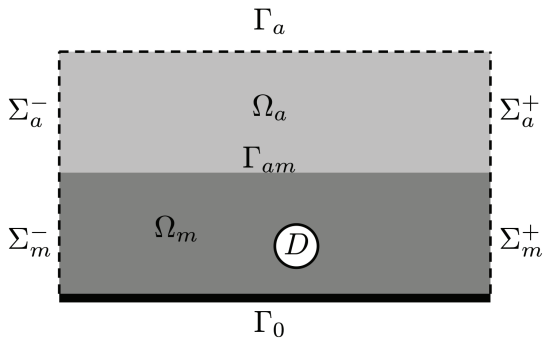


Figure 4.15. Domaine stratifié borné par des frontières artificielles

continue sur l'interface Γ_{am}), on obtient la formulation variationnelle suivante :

$$\begin{cases} \text{trouver } \ p \in \mathcal{C}^1(0,T;L^2(\Omega)) \cap \mathcal{C}^0(0,T;H^1(\Omega)) \text{ tel que} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega p(x,t)\, v(x)\, dx\right) + \displaystyle\int_\Omega c^2(x)\nabla p(x,t)\cdot\nabla v(x)\, dx \\ \qquad\qquad + c_a \dfrac{d}{dt}\displaystyle\int_{\Sigma_a} p(x,t)v(x)\, dx + c_m \dfrac{d}{dt}\displaystyle\int_{\Sigma_m} p(x,t)v(x)\, dx \\ \qquad\qquad = \displaystyle\int_\Omega f(x,t)\, v(x)\, dx, \quad \forall v \in H^1(\Omega),\ t \in]0,T[\\ p(x,0) = p_0(x), \dfrac{dp}{dt}(x,0) = p_1(x), \quad x \in \ \Omega. \end{cases} \tag{4.158}$$

On considère naturellement la formulation approchée par éléments finis conforme dans $H^1(\Omega)$:

$$\begin{cases} \text{trouver } \ p_h \in \mathcal{C}^1(0,T;V_h) \text{ tel que} \\ \dfrac{d^2}{dt^2}\left(\displaystyle\int_\Omega p_h(x,t)\, v_h(x)\, dx\right) + \displaystyle\int_\Omega c^2(x)\nabla p_h(x,t)\cdot\nabla v_h(x)\, dx \\ \qquad\qquad + c_a \dfrac{d}{dt}\displaystyle\int_{\Sigma_a} p_h(x,t)v_h(x)\, dx + c_m \dfrac{d}{dt}\displaystyle\int_{\Sigma_m} p_h(x,t)v_h(x)\, dx \\ \qquad\qquad = \displaystyle\int_\Omega f(x,t)\, v_h(x)\, dx, \quad \forall v_h \in V_h,\ t \in]0,T[\\ p_h(x,0) = \Pi_h p_0(x) = p_{h0}(x), \dfrac{dp}{dt}(x,0) = \Pi_h p(x) = p_{h1}(x), \quad x \in \ \Omega \end{cases}$$

qui s'écrit sous la forme matricielle :

$$\begin{cases} \text{trouver } \overrightarrow{P} \in \mathcal{C}^1(0,T;\mathbb{R}^N) \text{ tel que} \\ \mathbb{M}\dfrac{d^2}{dt^2}\overrightarrow{P}(t) + \mathbb{K}\overrightarrow{P}(t) + (c_a\mathbb{M}_a + c_m\mathbb{M}_m)\dfrac{d}{dt}\overrightarrow{P}(t) = \mathbb{M}\overrightarrow{f}(t),\ t \in]0,T[\\ \overrightarrow{P}(0) = \overrightarrow{P}_0, \dfrac{d\overrightarrow{P}}{dt}(0) = \overrightarrow{P}_1 \end{cases} \tag{4.159}$$

avec, pour $I, J = 1, N$:

$$(\mathbb{M}_a)_{IJ} = \int_{\Sigma_a} w_I w_J \, dx, \; (\mathbb{M}_m)_{IJ} = \int_{\Sigma_m} w_I w_J \, dx \text{ et } \mathbb{K}_{IJ} = \int_{\Omega} c^2 \nabla w_I \cdot \nabla w_J \, dx.$$

On utilise le schéma saute-mouton combiné avec une approximation centrée à l'ordre 2 de la dérivée première :

$$\frac{d}{dt}\overrightarrow{P}(t_k) = \frac{\overrightarrow{P}(t_{k+1}) - \overrightarrow{P}(t_{k-1})}{2\Delta t} + O(\Delta t^2)$$

et on substitue les matrices de masse $\mathbb{M}$, $\mathbb{M}_a$ et $\mathbb{M}_m$ par leur versions condensées $\mathbb{D}$, $\mathbb{D}_a$ et $\mathbb{D}_m$. Ce qui nous conduit finalement au schéma explicite d'ordre 2 suivant :

$$\begin{aligned} &\overrightarrow{P}^0 = \overrightarrow{P}_0, \; \overrightarrow{P}^1 = \overrightarrow{P}_0 + \Delta t \overrightarrow{P}_1 + \tfrac{\Delta t^2}{2}(\overrightarrow{f}(0) - c^2 \mathbb{D}^{-1}\mathbb{K}\overrightarrow{P}_0), \\ &\overrightarrow{P}^{k+1} = (\mathbb{D} + \frac{\Delta t}{2}\mathbb{E})^{-1}[(2\mathbb{D} - \Delta t^2 \mathbb{K})\overrightarrow{P}^k \\ &\qquad -(\mathbb{D} - \frac{\Delta t}{2}\mathbb{E})\overrightarrow{P}^{k-1} + \Delta t^2 \mathbb{D}\overrightarrow{f}(t_k)] \qquad \forall k = 1, K-1, \end{aligned} \tag{4.160}$$

où $\mathbb{E}$ désigne la matrice diagonale associée aux termes de bord :

$$\mathbb{E} = c_a \mathbb{D}_a + c_m \mathbb{D}_m.$$

Mise en œuvre

La mise en œuvre s'appuie principalement sur les scripts de calcul des matrices éléments finis $\mathbb{M}$ et $\mathbb{K}$ (à coefficient variable) (**calcul_EF_2D_var.m**) et des matrices de bord $\mathbb{M}_a$ et $\mathbb{M}_b$ ((**calcul_EF_1D**). Ces scripts donnés ci-après s'inspirent de scripts que nous avons présentés au cours d'illustrations numériques précédentes.

```
function [K,M]=calcul_EF_2D_var(S,T,R,fK)
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;pp2 pp2;pp2 pp4;pp4 pp2];
pp1=(155.-os)/2400.;pp2=(155.+os)/2400.;
pds_quadT=[9./80.;pp1;pp1;pp1;pp2;pp2;pp2];nbq=length(pds_quadT);
nt=size(T,1);ns=size(S,1);q=size(T,2);                % 3 en P1 ou 6 en P2
K=sparse(ns,ns);M=sparse(ns,ns);
for t=1:nt,                              %BOUCLE PRINCIPALE SUR LES TRIANGLES
    St=[S(T(t,1),:);S(T(t,2),:);S(T(t,3),:)];
    S21=St(2,:)-St(1,:);S31=St(3,:)-St(1,:);
    delta=S21(1)*S31(2)-S21(2)*S31(1);
    Jflmt=[S31(2) -S21(2);-S31(1) S21(1)]/delta;          %transfo affine
    Mt=zeros(q,q);Kt=zeros(q,q);
    for k=1:nbq,                                        %boucle quadrature
        x=pts_quadT(k,1);y=pts_quadT(k,2);  %calcul des fonctions de base
        w=[1-x-y x y];gw=[-1 1 0;-1 0 1];               %fonctions de base
        P=St'*[1-x-y; x; y];                        %coordonnées physiques
        pk=pds_quadT(k)*abs(delta);jg=Jflmt*gw;     %calcul des intégrands
        Mt=Mt+pk*w'*w; Kt=Kt+fK(P,R(t))*pk*jg'*jg; %matrices élémentaires
```

```
    end,
    In=T(t,:);
    K(In,In)=K(In,In)+Kt;M(In,In)=M(In,In)+Mt;          %assemblage de K,M
end,

function [Ma,Mm]=calcul_EF_1D(S,T,B,ref_a,ref_m)
s35=sqrt(3./5);pts_quadS=0.5*[1-s35 1 1+s35];
os=1/18;pds_quadS=os*[5 8 5];nbq=length(pds_quadS);
nt=size(T,1);ns=size(S,1);q=size(T,2);
Ma=sparse(ns,ns);Mm=sparse(ns,ns);
for t=1:nt,                              %BOUCLE PRINCIPALE SUR LES TRIANGLES
    for a=1:3,                                          %boucle sur les arêtes
        as=mod(a,3)+1;
        I=T(t,a);J=T(t,as);
        if((B(t,a)==ref_a)||(B(t,a)==ref_m) )   %arête sur le bord ref_am
            L=norm(S(I,:)-S(J,:));                          %longueur arête
            for k=1:nbq,                                %boucle quadrature 1D
                x=pts_quadS(k);c=L*pds_quadS(k);
                w=[1-x x];Mt=c*w'*w; in=[I J];
                if(B(t,a)==ref_a) Ma(in,in)=Ma(in,in)+Mt;      %assemblage
                else              Mm(in,in)=Mm(in,in)+Mt;end, % Ma ou Mm
            end,
        end,
    end,
end,
```

Le script principal suivant implémente le schéma (4.160). Dans un premier temps, on réalise le maillage de la zone stratifiée en fusionnant les maillages P^1 des deux rectangles $]-a,a[\times]0,b[$ et $]-a,a[\times]-b,0[$, maillage duquel on élimine à l'aide de la fonction **def_trou.m** les triangles appartenant à la zone définie par la fonction **obstacle.m**. On calcule ensuite les matrices éléments finis. Après avoir initialisé à la valeur 0 les vecteurs Pk et Pkm, on effectue la boucle en temps qui met à jour le vecteur Pkp suivant la relation itérative (4.160).

```
global ca cm;
p=300;q=p/4;ca=0.7;cm=1.5;a=1;b=0.5
t=0;tf=2; dt=0.001;
[S,T,BR,RT]=triangle_rectangle([-a a 0 b],p,q,1,[0 3 3 3]);      %Omega_a
[S2,T2,BR2,RT2]=triangle_rectangle([-a a -b 0],p,q,2,[0 4 0 4]);%Omega_m
[S,T,BR,RT]=fusion_triangulation(S,T,BR,RT,S2,T2,BR2,RT2);
[S,T,BR,RT]=def_trou(S,T,BR,RT,@obstacle);
[K,M]=calcul_EF_2D_var(S,T,RT,@fK);                          %matrices EF
[Ma,Mm]=calcul_EF_1D(S,T,BR,3,4);                  %matrices EF de bord
E=ca*sum(Ma)+cm*sum(Mm);E=E';D=sum(M)';             %condensation de masse
E1=1./(D+0.5*dt*E);E2=D-0.5*dt*E;
F=f(S); ns=size(S,1);                                %données des fonctions
Pkm=zeros(ns,1);Pkp=zeros(ns,1);                     %p0=p1=0
t=dt;
while(t<tf)
  gt=100000*(t<0.1)*exp(-2000*(t-0.05)*(t-0.05)); %dép. en t de la source
  Pkp=E1.*(2*D.*Pk-dt*dt*K*Pk-E2.*Pkm+dt*dt*gt*D.*F);              %schema
  t=t+dt;Pkm=Pk;Pk=Pkp;
end,

function o=obstacle(M)
a=-0.15;b=0.15;c=-0.4;d=-0.3;e=0.0001;x=M(1);y=M(2);
```

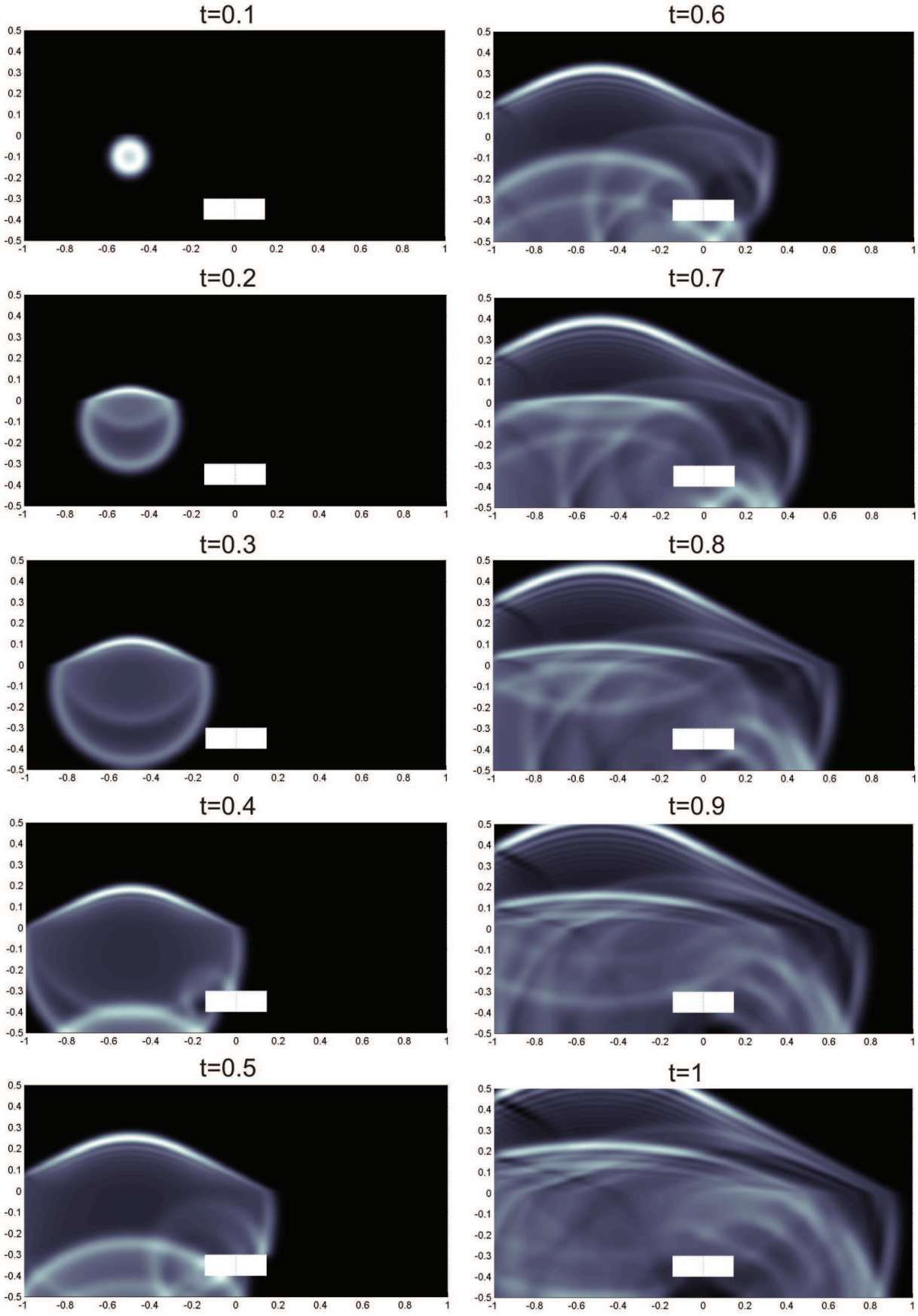


Figure 4.16. Pression acoustique à différents instants.

```
o=(x>a-e) && (x<b+e) && (y>c-e) && (y<d+e);

function z=f(X)
x=X(:,1)+0.5; y=X(:,2)+0.1;
z=(sqrt(x.*x+y.*y)<=0.01);
```

Exemple de simulation

Nous avons considéré la zone de calcul $]-1,1[\times]-0.5,0.5[$ ($a=1,\ b=0.5$) et l'obstacle $]-0.15,0.15[\times]-0.4,-0.3[$ se situant dans le domaine Ω_m. Nous avons choisi $c_a=0.7$ et $c_m=1.5$ de telle sorte que les ondes se propagent environ deux fois plus vite dans le domaine Ω_m que dans le domaine Ω_a. La source f est du type gaussien (4.157) : localisée dans la zone Ω_m au voisinage de l'interface Γ_{am} en prenant D_0 le disque de centre $(-0.5,-0.1)$ et $t_g=0.05$. L'amplitude est $A=100000$ et la rampe est $\gamma=2000$.

Nous donnons sur la figure 4.16, l'allure de la pression acoustique obtenue avec un pas de discrétisation $h=1/150$ et un pas de temps $\Delta t=0.001$ aux instants $t_k=k/10,\ k=1,10$. Ce choix de paramètres de discrétisation satisfait la condition CFL du schéma de saute-mouton.
A l'instant $t=1$, l'onde s'est propagée jusqu'au bord du domaine, hormis le bord de droite $\Sigma_m^+\cup\Sigma_a^+$. Ce qui est compatible ave les vitesses de propagation choisies. On perçoit clairement dans les premiers instants l'effet de l'interface où se produit un changement de vitesse de propagation et une "compression" du front d'onde. A l'instant $t=0.4$ on observe les effets importants de réflexion de l'onde sur le fond et sur l'obstacle. Aux instants suivants lorsque l'onde rencontre les bords artificiels où on a imposé une condition absorbante du premier ordre, on ne note pas de réflexion notable ; preuve de l'efficacité de la condition absorbante, qui rappelons-le n'est pas une condition de transparence parfaite !

Références

1. **M. Abramowitz, I. A. Stegun**, *Handook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover (1964).
2. **Y. Achdou, O. Pironneau**, *Computational Methods for Option Pricing*, Frontiers in Applied Mathematics, SIAM (2008).
3. **C. Amrouche, C. Bernardi, M. Dauge, V. Girault**, *Vector potentials in three-dimensional non-smooth domains*, Math. Meth. Appl. Sci., 21, 823–864 (1998).
4. **F. Assous, P. Ciarlet, J. Segré**, *Numerical solution to the time-dependent Maxwell equations in two-dimensional singular domains : the Singular Complement Method*, J. Comput. Phys., 161, 218–249 (2000).
5. **I. Babuska, J. E. Osborn**, *Eigenvalue problems*, Dans Handbook of Numerical Analysis, Vol. II, Eds. P.G. Ciarlet and J.-L. Lions, North Holland, 641–787 (1991).
6. **E. Bécache**, *C7-2 : Schémas Numériques pour la Résolution de l'Equation des Ondes*, Cours ENSTA (2008-2009).
7. **A.-S. Bonnet-Ben Dhia, M. Lenoir**, *MA102 : Outils Elémentaires d'Analyse pour les EDP*, Cours ENSTA (2007).
8. **J. H. Bramble**, *A proof of the inf–sup condition for the Stokes equations on Lipschitz domains*, Math. Models Meth. App. Sci., 13, 361–371 (2003).
9. **H. Brezis**, *Analyse Fonctionnelle. Théorie et Applications*, Masson (1983).
10. **F. Brezzi, M. Fortin**, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, 15, Springer Verlag (1991).
11. **F. Chatelin**, *Valeurs Propres de Matrices*, Masson (1988).
12. **Z. Chen, Q. Du, J. Zou**, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37, 1542–1570 (2000).
13. **P. Ciarlet**, *MA261 : Introduction au Calcul Scientifique. Aspects Algorithmiques*, Cours ENSTA (2005).
14. **P. Ciarlet**, *Augmented formulations for solving Maxwell equations*, Comput. Methods Appl. Mech. Engrg., 194, 559–586 (2005).
15. **P. Ciarlet, E. Lunéville**, *La méthode des Eléments Finis : de la Théorie à la Pratique. I. Concepts Généraux*, Les Presses de l'ENSTA (2009).
16. **P. Ciarlet, J. Zou**, *Fully discrete finite element approaches for time-dependent Maxwell's equations*, Numer. Math., 82, 193–219 (1999).

17. **P. Clément**, *Approximation by finite element functions using local regularization*, R.A.I.R.O. Anal. Numer., 9, 77–84 (1975).
18. **F. Collino**, *High order absorbing boundary conditions for wave propagation models : straight line boundary and corner cases*, Second International Conference on Mathematical and Numerical Aspects of Wave Propagation (Newark, DE, 1993), SIAM, Philadelphia, PA, 161–171 (1993).
19. **M. Costabel**, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal., 157, 527–541 (1991).
20. **M. Costabel, M. Dauge, S. Nicaise**, *Singularities of Maxwell interface problems*, Math. Mod. Num. Anal., 33, 627–649 (1999).
21. **R. Courant, D. Hilbert**, *Methods of Mathematical Physics*, Interscience Publishers (1961).
22. **M. Crouzeix, A. L. Mignot**, *Analyse Numérique des Equations Différentielles*, Masson (1983).
23. **R. Dautray, J. L. Lions**, *Analyse Mathématique et Calcul Numérique pour les Sciences et les Techniques*, Masson (1987).
24. **B. Engquist, A. Majda**, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (139), 629–651 (1977).
25. **P. Fernandes, G. Gilardi**, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions,* Math. Models Meth. App. Sci., 7, 957–991 (1997).
26. **J.-C. Gilbert**, *AO201 : Optimisation Différentiable, Théorie et Algorithmes*, Cours ENSTA (2009).
27. **V. Girault, P.-A. Raviart**, *Finite Element Methods for Navier-Stokes Equations*, Springer Series in Computational Mathematics, 5, Springer Verlag (1986).
28. **D. Givoli**, *Non reflecting boundary conditions*, J. Comput. Phys., 94 (1), 1–29 (1991).
29. **P. Grisvard**, *Singularities in Boundary Value Problems*, Coll. Recherche en Mathématiques Appliquées, 22, Masson (1992).
30. **T. Ha Duong and P. Joly**, *On the stability analysis of boundary conditions for the wave equation by energy methods; part I : The homogeneous case*, Math. Comp., 62, 539–563 (1994).
31. **C. Hazard, M. Lenoir**, *Formulations variationnelles pour la diffraction des ondes électromagnétiques*, Chapitre 5 de l'Ecole des Ondes INRIA : aspects récents en méthodes numériques pour les équations de Maxwell, INRIA Rocquencourt, mars 1998, Eds. G. Cohen et P. Joly, Collection Didactique INRIA (1998).
32. **C. Hazard, S. Lohrengel**, *A singular field method for Maxwell's equations : numerical aspects for 2D magnetostatics*, SIAM J. Numer. Anal., 40, 1021–1040 (2002).
33. **Ouvrage collectif POEMS**, *MA103 : Introduction à la Discrétisation des Equations aux Dérivées Partielles*, Cours ENSTA (2007).
34. **P. Joly**, *Analyse et Approximation de Modèles de Propagation d'Onde*, Cours de l'École Polytechnique, (2001-2002).
35. **V. Khoan**, *Distributions, Analyse de Fourier, Opérateurs aux Dérivées Partielles*, Vuibert (1972).
36. **H.-O. Kreiss and J. Lorenz**, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press (1989).
37. **D. Lamberton, B. Lapeyre**, *Stochastic Calculus Applied to Finance*, Chapman & Hall (1996).

38. **M. Lenoir**, *Approximation par Eléments Finis des Problèmes Elliptiques*, Publication ENSTA, 771 (1987).
39. **J.-L. Lions, E. Magenes**, *Problèmes aux Limites Non Homogènes et Applications, Vol. 1*, Dunod (1968).
40. **G. Meurant**, *Computer Solution of Large Linear Systems*, Elsevier (1999).
41. **J.-C. Nédélec**, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35, 315–341 (1980).
42. **B. N. Parlett**, *The Symmetric Eigenvalue Problem*, Series in Computational Mathematics, Prentice Hall (1980).
43. **O. Pironneau**, *Méthodes d'Elements Finis pour les Fluides*, Masson (1988).
44. **P.-A. Raviart, J.-M. Thomas**, *A Mixed Finite Element Method for Second Order Elliptic Problems*, dans Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics, 606, Springer, 292–315 (1977).
45. **P.-A. Raviart, J.-M. Thomas**, *Introduction à l'Analyse Numérique des Equations aux Dérivées Partielles*, Masson (1983).
46. **J. E. Roberts, J.-M. Thomas**, *Mixed and Hybrid Methods*, dans Handbook of Numerical Analysis, Vol. II, Eds. P.G. Ciarlet and J.-L. Lions, North Holland, 523–629 (1991).
47. **R. Seydel**, *Tools for Computational Finance*, Springer Verlag (2004).
48. **V. Thomée**, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Series in Computational Mathematics, 25, Springer Verlag (1997).
49. **L. Trefethen, L. Halpern**, *Well posedness of one way equations and absorbing boundary conditions*, Math. Comp., 47, 421–435 (1986).
50. **C. Weber**, *A local compactness theorem for Maxwell's equations*, Math. Meth. Appl. Sci., 2, 12-25 (1980).
51. **H. F. Weinberger**, *Variational Methods for Eigenvalue Approximation*, Regional Conference Series in Applied Mathematics (1974).
52. **J.-L. Yao Bi N'guessan**, *Méthode des éléments finis mixtes et conditions limites absorbantes pour la modélisation des phénomènes électromagnétiques hyperfréquences*, Thèse de l'Ecole Centrale de Lyon (1995).

Index