Problèmes d'évolution

Virginie Ehrlacher et Frédéric Legoll

Mars 2020

Table des matières

Chapitre 1

Rappels

Ce chapitre a l'objectif de rappeler plusieurs notions élémentaires. Nous en profitons pour faire un certain nombre de remarques, illustrées par plusieurs exercices, et montrant la spécificité de la dimension infinie par rapport à la dimension finie.

La dernière section de ce chapitre est consacrée à l'étude du problème de l'élasticité linéaire. De même que l'équation de Poisson, ce modèle servira de motivation et d'illustration dans les chapitres suivants.

On rappelle tout d'abord la notation suivante pour un espace vectoriel normé E.

Définition 1.1. *La boule unité fermée de E est*

$$B_E = \{x \in E;\ \|x\|_E \leq 1\}.$$

1.1 Espaces de Hilbert

Dans cette section, on se place dans un espace de Hilbert V. On rappelle que V est donc un espace vectoriel muni d'un produit scalaire, qu'on note $\langle x, y\rangle$, que la norme induite par ce produit scalaire est $\|x\| = \sqrt{\langle x, x\rangle}$, et que V est complet pour cette norme.

1.1.1 Théorèmes fondamentaux

On rappelle maintenant quelques théorèmes fondamentaux pour les espaces de Hilbert.

Théorème 1.2 (Théorème de projection orthogonale)**.** *Soit V un espace de Hilbert et K un sous-espace vectoriel fermé de V. Pour tout $u \in V$, il existe un unique $v = P_K u \in K$, appelé projection orthogonale de u sur K, tel que*

$$\|P_K u - u\| = \inf_{w \in K} \|w - u\|.$$

De plus, $P_K u$ est caractérisé par

$$P_K u \in K \qquad et \qquad \forall w \in K,\ \langle u - P_K u, w\rangle = 0. \tag{1.1}$$

Démonstration. Cf. le cours de première année [14]. □

On peut faire un peu mieux, et simplement supposer que K est un sous-ensemble convexe et fermé de V.

Définition 1.3. *Soit E un espace vectoriel et C un sous-ensemble de E. L'ensemble C est convexe si, pour tout x et y dans C et tout $\lambda \in [0,1]$, on a $\lambda x + (1-\lambda)y \in C$.*

Théorème 1.4 (Théorème de projection sur un convexe)**.** *Soit V un espace de Hilbert et K un sous-ensemble fermé et convexe de V. Pour tout $u \in V$, il existe un unique $v = P_K u \in K$, appelé projection de u sur K, tel que*

$$\|P_K u - u\| = \inf_{w \in K} \|w - u\|.$$

De plus, $P_K u$ est caractérisé par

$$P_K u \in K \qquad et \qquad \forall w \in K,\ \langle u - P_K u, w - P_K u\rangle \leq 0. \tag{1.2}$$

Démonstration. La preuve est très similaire à celle du théorème de projection orthogonale donnée dans [14]. □

Le théorème suivant permet d'identifier un espace de Hilbert V avec son dual $V' = \mathcal{L}(V, \mathbb{R})$:

Théorème 1.5 (Théorème de Riesz)**.** *Soit V un espace de Hilbert. Etant donné $\varphi \in V'$, il existe un unique $u \in V$ tel que*

$$\forall w \in V, \quad \varphi(w) = \langle u, w\rangle.$$

De plus, on a $\|u\|_V = \|\varphi\|_{V'}$. En d'autres termes, l'application de V' dans V qui à φ associe u permet d'identifier l'espace de Hilbert V avec son dual.

Démonstration. Cf. le cours de première année [14]. □

La notion d'application bilinéaire coercive joue un rôle fondamental pour l'étude des équations aux dérivées partielles.

Définition 1.6. *Soit V un espace de Hilbert et soit a une forme bilinéaire sur V. On dit que a est coercive sur V s'il existe un réel $\alpha > 0$ tel que*

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

Théorème 1.7 (Théorème de Lax-Milgram)**.** *Soit V un espace de Hilbert et a une forme bilinéaire sur V, symétrique, continue et coercive. Soit b une forme linéaire continue sur V. Alors le problème*

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ \forall w \in V, \quad a(u,w) = b(w) \end{cases} \tag{1.3}$$

admet une unique solution. De plus, le problème (1.3) est équivalent au problème de minimisation

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ J(u) = \inf\limits_{w \in V} J(w) \end{cases} \tag{1.4}$$

où la fonctionnelle d'énergie $J(w)$ est définie par $J(w) = \dfrac{1}{2}a(w,w) - b(w)$.

Démonstration. Cf. les cours de première année [11, 14]. □

Remarque 1.8. *On peut supprimer l'hypothèse de symétrie sur la forme bilinéaire a. Alors le problème (1.3) admet encore une unique solution, mais il n'y a plus d'équivalence de (1.3) avec un problème de minimisation du type (1.4).*

1.1.2 Bases hilbertiennes

La notion de base hilbertienne généralise en dimension infinie la notion de base orthonormée.

Définition 1.9. *Soit V un espace de Hilbert. On appelle base hilbertienne de V une suite $(e_n)_{n\geq 1}$ d'éléments de V tels que*
- *pour tout n, $\|e_n\| = 1$ et pour tous $m \neq n$, $\langle e_n, e_m\rangle = 0$.*
- *l'espace vectoriel engendré par la famille $(e_n)_{n\geq 1}$ est dense dans V.*

Proposition 1.10. *Soit V un espace de Hilbert admettant une base hilbertienne $(e_n)_{n\geq 1}$. Soit $u \in V$ et posons $u_n = \langle u, e_n\rangle$ pour tout $n \geq 1$. Alors, les séries $\sum_{n\geq 1} u_n e_n$ et $\sum_{n\geq 1} |u_n|^2$ sont convergentes dans V et $\mathbb{R}$ respectivement, et on a*

$$u = \sum_{n\geq 1} u_n e_n \qquad \text{et} \qquad \|u\|^2 = \sum_{n\geq 1} |u_n|^2.$$

Démonstration. Cf. le cours de première année [14]. □

1.1.3 Orthogonal d'un sous-espace

Définition 1.11. *Soit V un espace de Hilbert, et $W \subset V$ un sous-espace vectoriel. On note*

$$W^\perp = \{v \in V; \quad \forall w \in W, \quad \langle v, w\rangle = 0\}.$$

Lemme 1.12. *Soit V un espace de Hilbert, et $W \subset V$ un sous-espace vectoriel. Alors $W^\perp$ est un sous-espace vectoriel fermé de V.*

Démonstration. Soit $(v_n)_{n\geq 1}$ une suite d'éléments de $W^\perp$ qui converge vers $v \in V$. Pour tout $w \in W$, et tout $n \geq 1$, on a $\langle v_n, w\rangle = 0$. En passant à la limite, on obtient donc $\langle v, w\rangle = 0$ et par conséquent $v \in W^\perp$. □

Lemme 1.13. *Soit V un espace de Hilbert, et $W \subset V$ un sous-espace vectoriel. Alors*

$$\left(W^\perp\right)^\perp = \overline{W}.$$

Démonstration. Par définition,

$$\left(W^\perp\right)^\perp = \left\{v \in V; \quad \forall w \in W^\perp,\ \langle v, w\rangle = 0\right\}.$$

On a immédiatement que $W \subset \left(W^\perp\right)^\perp$. D'après le lemme 1.12, $\left(W^\perp\right)^\perp$ est fermé, donc $\overline{W} \subset \left(W^\perp\right)^\perp$. Soit maintenant $x \in \left(W^\perp\right)^\perp$. Comme $\overline{W}$ est fermé, on peut appliquer le théorème de projection orthogonale de V sur $\overline{W}$ et décomposer x selon

$$x = P_{\overline{W}}x + y, \tag{1.5}$$

avec $y \in (\overline{W})^\perp$, et donc $\langle y, P_{\overline{W}}x\rangle = 0$. On a aussi $y \in W^\perp$, et comme $x \in \left(W^\perp\right)^\perp$, ceci implique $\langle x, y\rangle = 0$. Donc

$$0 = \langle x, y\rangle - \langle P_{\overline{W}}x, y\rangle = \langle x - P_{\overline{W}}x, y\rangle = \langle y, y\rangle,$$

ce qui conduit à $y = 0$. La relation (1.5) implique alors que $x \in \overline{W}$. On a donc montré que $\left(W^\perp\right)^\perp \subset \overline{W}$, ce qui termine la preuve. □

Théorème 1.14. *Si W est fermé dans V, et que $W^\perp = \{0\}$, alors $W = V$ tout entier.*

Démonstration. Soit $x \in V$. Comme W est fermé, on peut appliquer le théorème de projection orthogonale et décomposer x selon

$$u = P_W x + y. \tag{1.6}$$

La caractérisation (1.1) donne $\langle y, w\rangle = 0$ pour tout $w \in W$. Donc $y \in W^\perp$, et par conséquent $y = 0$. On déduit de (1.6) que $x = P_W x$, soit $x \in W$. Par conséquent, $W = V$. □

1.2 Espaces de Sobolev

Les espaces de Sobolev jouent un rôle central dans l'étude des équations aux dérivées partielles.

1.2.1 Définitions principales

Soit Ω un ouvert de $\mathbb{R}^d$. On rappelle que, pour tout $p \geq 1$, l'ensemble $L^p(\Omega)$ est l'ensemble des fonctions dont la puissance p-ième est intégrable sur Ω.

On rappelle qu'un multi-indice $\alpha = (\alpha_1, \ldots, \alpha_d)$ est un élément de $\mathbb{N}^d$. Sa longueur est $|\alpha| = \sum_{i=1}^d \alpha_i$ et on adopte la notation suivante : pour toute distribution $u \in \mathcal{D}'(\Omega)$,

$$\partial^\alpha u = \frac{\partial^{|\alpha|}\, u}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d} = \frac{\partial^{\alpha_1+\ldots+\alpha_d}\, u}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d}.$$

Définition 1.15. *Pour $k \geq 1$, l'espace de Sobolev $H^k(\Omega)$ est l'ensemble des fonctions $f \in L^2(\Omega)$ telles que les dérivées de f au sens des distributions, jusqu'à l'ordre k, s'identifient à des fonctions de $L^2(\Omega)$. Autrement dit,*

$$H^k(\Omega) = \left\{ f \in L^2(\Omega) \quad \textit{telles que} \quad \forall \alpha \in \mathbb{N}^d,\ |\alpha| \leq k,\ \partial_\alpha f \in L^2(\Omega) \right\}.$$

Comme l'espace $L^2(\Omega)$, les espaces $H^k(\Omega)$ sont des espaces de Hilbert.

Théorème 1.16. *Muni du produit scalaire*

$$(f, g)_{H^k} = \int_\Omega f(x)\, g(x)\, dx + \sum_{1 \leq |\alpha| \leq k} \int_\Omega \partial_\alpha f(x)\, \partial_\alpha g(x)\, dx,$$

l'espace $H^k(\Omega)$ est un espace de Hilbert. Sa norme est notée $\| \cdot \|_{H^k(\Omega)}$.

On rappelle maintenant un théorème de densité de l'ensemble des fonctions test.

Théorème 1.17. *Pour tout ouvert Ω de $\mathbb{R}^d$, l'ensemble $\mathcal{D}(\Omega)$ est dense dans $L^2(\Omega)$ pour la norme $L^2(\Omega)$.*

De plus, pour tout $k \geq 1$, l'ensemble $\mathcal{D}(\mathbb{R}^d)$ est dense dans $H^k(\mathbb{R}^d)$ pour la norme $H^k(\mathbb{R}^d)$.

Pour tout $k \geq 1$, si $\Omega \subset \mathbb{R}^d$ avec $\Omega \neq \mathbb{R}^d$, alors $\mathcal{D}(\Omega)$ n'est pas dense dans $H^k(\Omega)$.

Définition 1.18. *Pour $k \geq 1$, on définit $H_0^k(\Omega)$ comme la fermeture de $\mathcal{D}(\Omega)$ dans $H^k(\Omega)$ (pour la norme de $H^k(\Omega)$).*

On donne maintenant un résultat propre à la dimension 1.

Théorème 1.19. *Soit I un intervalle de $\mathbb{R}$ et $u \in H^1(I)$. Alors u s'identifie à une fonction continue et, pour tout x et y dans I,*

$$u(x) - u(y) = \int_y^x u'(s) ds.$$

On souligne que ce théorème est faux en dimension plus grande.

Démonstration. On esquisse ici la preuve, dont les détails sont laissés au lecteur. Soit $x_0 \in I$ fixé. Pour $u \in H^1(I)$, on définit

$$w(x) = \int_{x_0}^{x} u'(s)ds.$$

Grâce à l'inégalité de Cauchy-Schwarz, cette définition a bien un sens, et on montre que w est une fonction continue sur I. On calcule ensuite la dérivée de w au sens des distributions, en utilisant le théorème de Fubini. On montre ainsi que $w' = u'$ dans $\mathcal{D}'(I)$. Par conséquent, $w - u$ est une constante, et u s'identifie donc bien à une fonction continue. □

1.2.2 Trace

Pour une fonction définie dans un ouvert Ω, on souhaite définir sa valeur au bord de Ω. Pour les fonctions $u \in L^2(\Omega)$, cette notion n'a pas de sens. Par contre, si u est plus régulière, alors on peut définir rigoureusement cette notion.

Proposition 1.20. *Soit Ω un ouvert borné et régulier. On peut définir une application linéaire et continue*

$$\begin{array}{rccc} \gamma \ : & H^1(\Omega) & \longrightarrow & L^2(\partial\Omega) \\ & u & \mapsto & \gamma(u), \end{array}$$

et qui prolonge l'application trace pour les fonctions continues sur $\overline{\Omega}$: pour tout $u \in H^1(\Omega) \cap C^0(\overline{\Omega})$, $\gamma(u) = u|_{\partial\Omega}$.

L'application trace est continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$, ce qui signifie qu'il existe une constante C_Ω telle que

$$\forall u \in H^1(\Omega),\ \|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_\Omega\ \|u\|_{H^1(\Omega)}. \tag{1.7}$$

Remarque 1.21. *L'application trace n'est pas surjective sur $L^2(\partial\Omega)$, mais sur un espace plus petit, qui est $H^{1/2}(\partial\Omega)$. Elle est en fait continue de $H^1(\Omega)$ vers $H^{1/2}(\partial\Omega)$, si bien qu'il existe C_Ω tel que*

$$\forall u \in H^1(\Omega),\ \|\gamma(u)\|_{H^{1/2}(\partial\Omega)} \leq C_\Omega\ \|u\|_{H^1(\Omega)}.$$

Enfin, pour tout $u \in H^{1/2}(\partial\Omega)$, on a $\|u\|_{L^2(\partial\Omega)} \leq \|u\|_{H^{1/2}(\partial\Omega)}$.

L'espace $H_0^1(\Omega)$, défini comme la fermeture dans $H^1(\Omega)$ de $\mathcal{D}(\Omega)$, s'identifie à l'espace des fonctions à trace nulle :

Proposition 1.22. *Soit Ω un ouvert de $\mathbb{R}^d$. On a*

$$H_0^1(\Omega) = \left\{ u \in H^1(\Omega), \quad \gamma(u) = 0 \right\}.$$

1.2.3 Inégalité de Poincaré

On rappelle la notation suivante :

Définition 1.23. *Soit Ω un ouvert de $\mathbb{R}^d$. Pour une fonction u à valeur vectorielle $u = (u_1, \dots, u_d) \in L^2(\Omega)^d$, on note*

$$\|u\|_{L^2(\Omega)} = \sqrt{\sum_{i=1}^{d} \|u_i\|^2_{L^2(\Omega)}} \, .$$

Proposition 1.24 (Inégalité de Poincaré)**.** *Soit Ω un ouvert borné de $\mathbb{R}^d$. Alors il existe une constante C_Ω telle que*

$$\forall u \in H^1_0(\Omega), \quad \|u\|_{L^2(\Omega)} \leq C_\Omega \ \|\nabla u\|_{L^2(\Omega)} \, . \tag{1.8}$$

Démonstration. Cette inégalité est démontrée dans les cours [11, 14]. L'exercice 2.70 en propose une autre démonstration. L'exercice 3.7 donne une caractérisation de la meilleure constante C_Ω en terme de valeur propre du laplacien. □

1.2.4 Injections de Sobolev

On considère une fonction $u \in H^1(\Omega)$. Bien sûr, $u \in L^2(\Omega)$. On peut se demander si u n'est pas plus régulière que ceci, du fait que ∇u soit dans $L^2(\Omega)$. Le théorème suivant répond à cette question.

Théorème 1.25. *Soit Ω un ouvert régulier de $\mathbb{R}^d$, et soit k un entier. On a les injections continues suivantes :*

- *si $d > 2k$, alors $H^k(\Omega) \subset L^{p^*}(\Omega)$ avec $1/p^* = 1/2 - k/d$.*
- *si $d = 2k$, alors $H^k(\Omega) \subset L^q(\Omega)$ pour tout $q \in [2, +\infty[$.*
- *si $d < 2k$, alors $H^k(\Omega) \subset C^0(\overline{\Omega})$.*

On rappelle maintenant l'inégalité de Hölder.

Lemme 1.26 (Inégalité de Hölder)**.** *Soient p et q deux réels compris (au sens large) entre 1 et $+\infty$, avec $1/p + 1/q = 1$. Soient $f \in L^p(\Omega)$ et $g \in L^q(\Omega)$. Alors le produit $f\,g$ est dans $L^1(\Omega)$ et*

$$\|f\,g\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \ \|g\|_{L^q(\Omega)}.$$

On déduit de cette inégalité (le faire en exercice !) le résultat suivant :

Lemme 1.27. *Soient p et q deux réels compris (au sens large) entre 1 et $+\infty$, avec $p < q$. Soit $f \in L^p(\Omega) \cap L^q(\Omega)$. Alors, pour tout $r \in [p, q]$, on a $f \in L^r(\Omega)$, avec*

$$\|f\|_{L^r(\Omega)} \leq \|f\|^{\alpha}_{L^p(\Omega)} \ \|f\|^{1-\alpha}_{L^q(\Omega)},$$

où α est tel que $1/r = \alpha/p + (1-\alpha)/q$.

Ainsi, soit Ω un ouvert régulier de $\mathbb{R}^d$, et soit k un entier, avec par exemple $d > 2k$. On a vu que $H^k(\Omega) \subset L^{p^*}(\Omega)$ avec $1/p^* = 1/2 - k/d$. De plus, $H^k(\Omega) \subset L^2(\Omega)$. Donc $H^k(\Omega) \subset L^r(\Omega)$ pour tout $r \in [2, p^*]$.

1.3 Convergence faible

On rappelle qu'une suite d'éléments $(u_n)_{n\geq 0}$ d'un espace de Hilbert V converge vers $u \in V$ si $\lim_n \|u_n - u\| = 0$. On introduit ici une notion de convergence plus faible, la *convergence faible*. Pour éviter les confusions, on parlera alors de *convergence forte* pour la convergence usuelle.

Avant d'introduire cette nouvelle notion, on rappelle ici quelques notions liées à la compacité de sous-ensembles d'un espace vectoriel.

1.3.1 Compacité

On se place dans un espace vectoriel normé E. On rappelle la définition suivante :

Définition 1.28. *Un sous-ensemble $K \subset E$ est compact si, de toute suite $(u_n)_{n\geq 0}$ d'éléments de K, on peut extraire une sous-suite convergente dans K.*

Nous aurons besoin dans la suite de ce cours d'une notion plus fine que celle d'ensemble compact, et que nous introduisons maintenant :

Définition 1.29. *Un sous-ensemble $K \subset E$ est relativement compact si, de toute suite $(u_n)_{n\geq 0}$ d'éléments de K, on peut extraire une sous-suite convergente dans E.*

La différence avec la notion d'ensemble compact est donc que la limite de la suite n'appartient pas nécessairement à K.

La preuve de la proposition suivante est laissée en exercice :

Proposition 1.30. *Un sous-ensemble $K \subset E$ est relativement compact si et seulement si $\overline{K}$ est compact.*

On rappelle que les sous-ensembles compacts de E sont nécessairement des ensembles fermés et bornés. La réciproque n'est vraie que dans le cas où E est un espace de dimension finie. On a en effet le résultat suivant, caractéristique de la dimension infinie :

Théorème 1.31. *Soit V un espace de Hilbert de dimension infinie. Alors la boule unité fermée de V n'est pas compacte.*

Démonstration. Comme l'espace est de dimension infinie, on peut construire une suite orthonormée infinie $(e_n)_{n\geq 1}$ (en utilisant le procédé de Gram-Schmidt). Cette suite appartient bien à la boule unité fermée. Par ailleurs, pour $n \neq p$, on a

$$\|e_n - e_p\|^2 = \|e_n\|^2 + \|e_p\|^2 - 2\langle e_n, e_p\rangle = 2. \tag{1.9}$$

Supposons que la boule unité fermée est compacte. Alors on peut extraire de la suite $(e_n)_{n\geq 1}$ une sous-suite convergente, donc de Cauchy. Or ceci est contradictoire avec (1.9). □

1.3.2 Définition de la convergence faible

Avant de donner la définition de la notion de convergence faible, nous avons besoin de rappeler la définition de la limite inférieure d'une suite de réels.

Définition 1.32. *Soit u_n une suite de réels. On définit sa limite inférieure par*

$$\liminf u_n = \lim_{n\to\infty} \left(\inf_{k\geq n} u_k \right).$$

La suite $I_n = \inf_{k\geq n} u_k$ est une suite croissante de réels, qui admet donc bien une limite (éventuellement infinie).

Le lemme suivant montre que la notion de limite inférieure généralise la notion de limite.

Lemme 1.33. *Soit u_n une suite de réels qui converge vers λ. Alors $\lambda = \liminf u_n$.*

Dans le cas d'une suite quelconque, on a le résultat suivant :

Lemme 1.34. *Soit u_n une suite de réels, et soit $\lambda = \liminf u_n$. On peut extraire de u_n une sous-suite qui converge vers λ.*

Démonstration. On suppose $\lambda \in \mathbb{R}$ (le cas $\lambda = +\infty$ se traite de la même façon). On pose $I_n = \inf_{k\geq n} u_k$: par définition, $\lambda = \lim_n I_n$. Soit $\varepsilon > 0$ et $N > 0$. Il existe $n_0 > N$ tel que $\lambda \geq I_{n_0} \geq \lambda - \varepsilon$. De plus, il existe $k_0 \geq n_0$ tel que $\varepsilon + \inf_{k\geq n_0} u_k \geq u_{k_0} \geq \inf_{k\geq n_0} u_k$. Donc on a $\varepsilon + \lambda \geq u_{k_0} \geq \lambda - \varepsilon$, ce qui conclut la preuve. □

On introduit maintenant la notion de convergence faible.

Définition 1.35. *Soit V un espace de Hilbert. On dit qu'une suite u_n de V converge faiblement vers u dans V si $u \in V$ et*

$$\forall w \in V, \ \lim_{n\to+\infty} \langle u_n, w\rangle = \langle u, w\rangle.$$

On note $u_n \rightharpoonup u$.

Si V est de dimension finie, alors la convergence au sens faible est équivalente à la convergence au sens fort. En dimension infinie, les deux notions sont différentes.

On a également la caractérisation équivalente suivante de la convergence faible.

Proposition 1.36. *Soit V un espace de Hilbert, $u \in V$ et $(u_n)_{n\in N}$ une suite d'élements de V. Les deux propositions suivantes sont équivalentes :*

(i) $(u_n)_{n\in\mathbb{N}}$ converge faiblement vers u dans V ;

(ii) pour toute forme linéaire continue $\varphi \in V'$,

$$\varphi(u_n) \underset{n\to+\infty}{\longrightarrow} \varphi(u).$$

Démonstration. On montre que (ii) implique (i). Ceci découle du fait que, pour tout $w \in V$, l'application $\varphi : v \in V \mapsto \langle v, w\rangle \in \mathbb{R}$ est une forme linéaire continue. Montrons maintenant que (i) implique (ii). Ceci est une conséquence du théorème de Riesz. En effet, pour tout $\varphi \in V'$, il existe $w \in V$ tel que pour tout $v \in V$, $\varphi(v) = \langle w, v\rangle$. D'où le résultat. □

1.3.3 Propriétés de la convergence faible

Nous commençons par énoncer les liens entre convergence faible et convergence forte (au sens usuel).

Théorème 1.37. *Soit u_n une suite de V.*
- *si u_n converge fortement vers u dans V, alors u_n converge faiblement vers u dans V ;*
- *si u_n converge faiblement vers u dans V, alors la suite u_n est bornée dans V et $\|u\| \leq \liminf_{n\to\infty} \|u_n\|$.*
- *Si u_n converge vers u faiblement et w_n converge vers w fortement, alors on a $\lim_{n\to\infty}\langle u_n, w_n\rangle = \langle u, w\rangle$.*

Démonstration. La preuve de la première et de la troisième affirmation sont laissées au lecteur (utiliser l'inégalité de Cauchy-Schwarz). Le fait qu'une suite qui converge faiblement soit bornée est une propriété plus difficile à démontrer, et qui sera ici admise. Elle repose sur le théorème de Banach-Steinhaus (cf. par exemple [3, Théorème II.1 et Proposition III.5]). On prouve maintenant l'inégalité dans la deuxième affirmation. Supposons que u_n converge faiblement vers u. L'inégalité de Cauchy-Schwarz donne que

$$\left\langle \frac{u}{\|u\|}, u_n \right\rangle \leq \|u_n\|.$$

On passe à la limite inférieure et on utilise que le membre de gauche converge :

$$\lim_{n\to\infty} \left\langle \frac{u}{\|u\|}, u_n \right\rangle = \liminf_{n\to\infty} \left\langle \frac{u}{\|u\|}, u_n \right\rangle \leq \liminf_{n\to\infty} \|u_n\|,$$

d'où le fait que $\|u\| \leq \liminf_{n\to\infty} \|u_n\|$. □

L'intérêt de la convergence faible réside dans la proposition suivante, que nous admettrons.

Proposition 1.38. *Soit V un espace de Hilbert. La boule unité de V est faiblement compacte : de toute suite bornée de V, on peut extraire une sous-suite qui converge faiblement dans V.*

Dans un espace de Hilbert, pour montrer qu'une suite converge faiblement (à extraction près), il suffit donc de montrer qu'elle est bornée.

La définition d'ensemble fermé pour la topologie faible est naturelle :

Définition 1.39. *Soit V un espace de Hilbert, et C un sous-ensemble de V. On dit que C est faiblement fermé si, pour toute suite d'éléments $(u_n)_{n\geq 0}$ de C qui converge faiblement vers u dans V, on a $u \in C$.*

Comme la convergence forte implique la convergence faible, un ensemble faiblement fermé (i.e. fermé pour la topologie faible) est fortement fermé (i.e. fermé pour la topologie forte). La réciproque est fausse, sauf si l'ensemble est convexe, comme le montre le résultat suivant :

Proposition 1.40. *Soit V un espace de Hilbert, et C un sous-ensemble de V qui soit convexe et fortement fermé. Alors C est faiblement fermé.*

Démonstration. Soit u_n est une suite de points de C qui converge faiblement vers $u \in V$. Comme C est convexe et fortement fermé dans V, on peut considérer la projection de V sur C, qu'on note P_C. D'après le théorème 1.4, on a

$$\forall w \in C,\ \langle u - P_C u, w - P_C u\rangle \leq 0.$$

On écrit cette inégalité avec $w = u_n$ et on passe à la limite $n \to +\infty$ en utilisant la convergence faible de u_n vers u. Donc $\langle u - P_C u, u - P_C u\rangle \leq 0$, ce qui implique que $u = P_C u$ et donc $u \in C$. □

Proposition 1.41. *Soit V un espace de Hilbert et $J : V \to \mathbb{R}$ une fonction continue (pour la topologie forte de V) et convexe sur V. Pour toute suite u_n qui converge faiblement dans V vers u, on a*

$$J(u) \leq \liminf J(u_n).$$

Démonstration. Pour tout $\lambda \in \mathbb{R}$, l'ensemble $C(\lambda) = \{u \in V;\ J(u) \leq \lambda\}$ est convexe, car J est convexe. Comme J est continue, cet ensemble est fortement fermé. On utilise la proposition 1.40 : $C(\lambda)$ est faiblement fermé.

Soit $\lambda_0 = \liminf J(u_n)$. Le lemme 1.34 donne l'existence d'une sous-suite extraite $u_{\varphi(n)}$ telle que $\lim_n J(u_{\varphi(n)}) = \lambda_0$. Par conséquent, pour tout $\varepsilon > 0$, et pour tout $n \geq n_0(\varepsilon)$, on a $J(u_{\varphi(n)}) \leq \varepsilon + \lambda_0$, et donc $u_{\varphi(n)} \in C(\varepsilon + \lambda_0)$. Par ailleurs, la suite $u_{\varphi(n)}$ converge faiblement vers u. Donc $u \in C(\varepsilon + \lambda_0)$, soit $J(u) \leq \varepsilon + \lambda_0$, et ce pour tout ε. Donc $J(u) \leq \lambda_0$, ce qui conclut la preuve. □

On a donc vu que les notions de topologie faible et de convexité sont reliées.

En guise d'application de ces notions aux espaces de Sobolev, nous donnons la proposition suivante :

Proposition 1.42. *De toute suite bornée de $H_0^1(\Omega)$, on peut extraire une-suite qui converge faiblement vers u dans $H^1(\Omega)$. De plus, $u \in H_0^1(\Omega)$.*

Démonstration. La proposition 1.38 donne l'existence d'une sous-suite qui converge faiblement vers u dans $H^1(\Omega)$. L'espace $H_0^1(\Omega)$ est fermé dans $H^1(\Omega)$ et convexe, donc il est faiblement fermé en vertu de la proposition 1.40, et donc $u \in H_0^1(\Omega)$. □

1.4 Problème de l'élasticité linéarisée

Dans les cours de première année (cf. par exemple [11, 14]), on a étudié l'équation de Poisson $-\Delta u = f$ (avec $u \in H^1_0(\Omega)$ par exemple). Cette équation modélise par exemple le déplacement vertical d'une membrane soumise à des forces verticales $f(x)$ et dont les bords sont maintenus fixes (d'où la condition aux limites $u = 0$ sur $\partial\Omega$). L'équation de Poisson intervient aussi dans d'autres domaines, comme l'électrostatique (u représente alors un potentiel électrostatique), la thermique (u est alors la température locale dans un solide), ...

Nous nous intéressons dans cette section au problème de l'élasticité linéaire, qui est le modèle le plus simple apparaissant en mécanique des solides déformables. Une différence essentielle avec l'équation de Poisson est que l'inconnue est une fonction à valeur dans $\mathbb{R}^d$, et non pas à valeur scalaire comme dans l'équation de Poisson. Commençons par décrire plus précisément le modèle de l'élasticité linéarisée.

1.4.1 Le modèle

En mécanique, l'inconnue est le déplacement $u(x) \in \mathbb{R}^d$ d'un point matériel situé en x dans la configuration de référence. Soit donc Ω un ouvert de $\mathbb{R}^d$ et u une fonction définie sur Ω et à valeur dans $\mathbb{R}^d$. Une quantité importante est le tenseur des déformations, noté $e(u)$ et défini par

$$e(u) = \frac{1}{2}\left(\nabla u + (\nabla u)^t\right). \tag{1.10}$$

Donc $e(u)$ est une matrice symétrique de taille $d \times d$ dont les coefficients sont

$$e_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{1.11}$$

On s'intéresse à un solide déformable, et on fait l'hypothèse que les déplacements u et les déformations $e(u)$ sont petits. Cette hypothèse permet de linéariser les équations générales décrivant un solide élastique. On s'intéresse de plus ici aux équations stationnaires, c'est-à-dire indépendantes du temps, et qui décrivent l'équilibre d'un solide (leurs versions instationnaires, qui décrivent au contraire la dynamique du solide, seront étudiées plus loin, au chapitre 6).

En plus du tenseur des déformations, la modélisation fait intervenir le tenseur des contraintes σ. Comme $e(u)$, le tenseur σ est une fonction de Ω à valeur dans $\mathbb{R}^{d\times d}$. Le tenseur des contraintes est relié au tenseur des déformations par la loi constitutive du matériau, qui est ici linéaire. On s'intéresse à des matériaux homogènes et isotropes, si bien que cette relation s'écrit

$$\sigma(u) = 2\mu e(u) + \lambda(\text{tr } e(u))\,\text{Id}, \tag{1.12}$$

où λ et μ sont les coefficients de Lamé du matériau, qui varient d'un matériau à un autre, et où Id est la matrice identité de $\mathbb{R}^{d\times d}$. La relation (1.12) s'appelle loi de Hooke.

Pour des raisons thermodynamiques que nous ne détaillons pas ici, les coefficients de Lamé vérifient

$$\mu > 0 \quad \text{et} \quad 2\mu + d\lambda > 0, \tag{1.13}$$

où d est la dimension de l'espace dans lequel on travaille (en général, $d = 3$). Le tenseur $e(u)$ étant symétrique, le tenseur $\sigma(u)$ l'est aussi. On définit la divergence d'un tenseur symétrique σ comme le vecteur de composante

$$\forall 1 \leq i \leq d, \quad (\text{div } \sigma)_i = \sum_{j=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_j}.$$

On montre en mécanique que la relation d'équilibre pour un solide déformable soumis à des forces de volume f (fonction de Ω dans $\mathbb{R}^d$) s'écrit

$$-\text{div } \sigma(u) = f. \tag{1.14}$$

Nous préciserons le sens mathématique de (1.14) à la section 1.4.3 ci-dessous. Compte tenu de la loi de Hooke (1.12), on s'intéresse donc à l'équation aux dérivées partielles (d'inconnue u)

$$-\text{div } [2\mu e(u) + \lambda(\text{tr } e(u))\,\text{Id}] = f.$$

On décrit enfin les conditions aux limites. Très souvent, la frontière $\partial\Omega$ du solide peut être divisée en deux parties, $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ avec $\partial\Omega_D \cap \partial\Omega_N = \emptyset$, telles que

- sur $\partial\Omega_D$, on impose le déplacement. Par exemple, le solide est encastré, et on impose donc $u = 0$ sur $\partial\Omega_D$.
- sur le reste de la frontière $\partial\Omega_N$, on impose des forces de surface. Ces forces peuvent être nulles, ce qui correspond à un bord libre. Mathématiquement, cette condition aux limites s'écrit $\sigma \cdot n = g$, où n est le vecteur normal (sortant) au domaine, et g est la force surfacique imposée. En détaillant par composante, on a donc

$$\forall x \in \partial\Omega_N, \ \forall 1 \leq i \leq d, \quad \sum_{j=1}^{d} \sigma_{ij}(x) n_j(x) = g_i(x).$$

Le cas où $\partial\Omega_N = \emptyset$ est plus simple mathématiquement, mais moins réaliste du point de vue physique. Il est en effet rare d'imposer le déplacement sur l'ensemble de la frontière du solide.

On introduit ici quelques notations utiles pour la suite. On rappelle déjà la définition (1.23) : pour une fonction à valeur vectorielle $u = (u_1, \ldots, u_d) \in L^2(\Omega)^d$, où Ω un ouvert de $\mathbb{R}^d$, on note

$$\|u\|_{L^2(\Omega)} = \sqrt{\sum_{i=1}^{d} \|u_i\|^2_{L^2(\Omega)}}\ .$$

Définition 1.43. *Pour une fonction u à valeur matricielle $u = (u_{ij})_{1\leq i,j\leq d} \in L^2(\Omega)^{d\times d}$, on note*

$$\|u\|_{L^2(\Omega)} = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \|u_{ij}\|^2_{L^2(\Omega)}}\ .$$

On rappelle aussi le produit scalaire pour les matrices $d \times d$:

Définition 1.44. *Soit $A \in \mathbb{R}^{d\times d}$ et $B \in \mathbb{R}^{d\times d}$. On note*

$$A \cdot B = \sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}B_{ij}$$

le produit scalaire des deux matrices A et B.

1.4.2 Inégalité de Korn

L'inégalité de Poincaré joue un rôle fondamental dans l'étude mathématique de l'équation de Poisson, $-\Delta u = f$, dans un ouvert Ω borné. Dans cette équation, u est une fonction scalaire, représentant une température, un potentiel électrostatique, ... En mécanique, l'inconnue est une fonction à valeur vectorielle, et ce qui joue le rôle de l'inégalité de Poincaré est l'inégalité de Korn. Avant d'énoncer cette inégalité, nous donnons plusieurs lemmes qui permettent de mieux en saisir la portée.

Le lemme suivant permet de caractériser les déplacements u associés à un tenseur de déformation nul.

Lemme 1.45. *Soit Ω un ouvert connexe et borné de $\mathbb{R}^d$, avec $d = 2$ ou $d = 3$. Soit $\mathcal{R}$ l'ensemble des mouvements rigides de Ω définis par*

$$\mathcal{R} = \{u(x) = b + Mx,\ b \in \mathbb{R}^d \text{ et } M = -M^t \text{ matrice antisymétrique}\}.$$

Soit $u \in H^1(\Omega)^d$. Alors $u \in \mathcal{R}$ si et seulement si $e(u) = 0$.

Démonstration. Si $u \in \mathcal{R}$, il est clair que $e(u) = 0$. Réciproquement, si $e(u) = 0$, alors, pour tout $1 \leq i \leq d$, on a $\dfrac{\partial u_i}{\partial x_i} = 0$. Comme Ω est connexe, on en déduit que

$$u_i(x) = f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d),$$

où la fonction f_i ne dépend pas de x_i. On traite maintenant séparément le cas de la dimension 2 et celui de la dimension 3.

Si $d = 2$, on a donc $u_1 = f_1(x_2)$ et $u_2 = f_2(x_1)$. Comme $e_{12}(u) = 0$, on a $f_1'(x_2) + f_2'(x_1) = 0$, et donc il existe C tel que $f_1'(x_2) = -f_2'(x_1) = C$. Donc

$$u_1 = f_1(x_2) = Cx_2 + b_1 \quad \text{et} \quad u_2 = f_2(x_1) = -Cx_1 + b_2,$$

ce qui démontre le lemme.

Si $d = 3$, alors

$$u_1 = f_1(x_2, x_3), \quad u_2 = f_2(x_1, x_3), \quad u_3 = f_3(x_1, x_2).$$

En dérivant par rapport à x_2 la relation $e_{12}(u) = 0$, on obtient $\dfrac{\partial^2 f_1}{\partial x_2 \, \partial x_2} = 0$, ce qui donne $f_1(x_2, x_3) = x_2 g(x_3) + h(x_3)$. En utilisant que $\dfrac{\partial^2 f_1}{\partial x_3 \, \partial x_3} = 0$, on obtient qu'il existe a_1, b_1, c_1 et d_1 tels que

$$f_1(x_2, x_3) = a_1 x_2 x_3 + b_1 x_2 + c_1 x_3 + d_1.$$

Un raisonnement identique sur les autres composantes conduit à

$$f_2(x_1, x_3) = a_2 x_1 x_3 + b_2 x_1 + c_2 x_3 + d_2,$$
$$f_3(x_1, x_2) = a_3 x_1 x_2 + b_3 x_1 + c_3 x_2 + d_3.$$

Les relations $e_{12}(u) = 0$, $e_{13}(u) = 0$ et $e_{23}(u) = 0$ conduisent respectivement aux relations

$$\left\{ \begin{array}{l} a_1 + a_2 = 0 \\ b_1 + b_2 = 0 \end{array} \right. , \quad \left\{ \begin{array}{l} a_1 + a_3 = 0 \\ c_1 + b_3 = 0 \end{array} \right. , \quad \left\{ \begin{array}{l} a_2 + a_3 = 0 \\ c_2 + c_3 = 0 \end{array} \right. .$$

On obtient alors $a_1 = a_2 = a_3 = 0$ et donc $u = b + Mx$ avec

$$b = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} \quad \text{et} \quad M = \begin{pmatrix} 0 & b_1 & c_1 \\ -b_1 & 0 & c_2 \\ -c_1 & -c_2 & 0 \end{pmatrix}.$$

Ceci conclut la preuve du lemme. □

Avant d'énoncer l'inégalité de Korn, commençons par deux inégalités simples :

Lemme 1.46. *Soit Ω un ouvert de $\mathbb{R}^d$. Pour tout $u \in H^1(\Omega)^d$, on a*

$$\|e(u)\|_{L^2(\Omega)} \leq \|\nabla u\|_{L^2(\Omega)}, \tag{1.15}$$
$$\|div\ u\|_{L^2(\Omega)} \leq d \, \|\nabla u\|_{L^2(\Omega)}. \tag{1.16}$$

Démonstration. Par définition,

$$\begin{aligned} \|e(u)\|^2_{L^2(\Omega)} &= \sum_{1 \leq i,j \leq d} \int_\Omega e_{ij}^2(u) dx \\ &= \frac{1}{4} \sum_{1 \leq i,j \leq d} \int_\Omega \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)^2 dx \\ &\leq \frac{1}{2} \sum_{1 \leq i,j \leq d} \int_\Omega \left[\left(\frac{\partial u_i}{\partial x_j} \right)^2 + \left(\frac{\partial u_j}{\partial x_i} \right)^2 \right] dx \\ &\leq \sum_{1 \leq i,j \leq d} \int_\Omega \left(\frac{\partial u_i}{\partial x_j} \right)^2 dx = \|\nabla u\|^2_{L^2(\Omega)}, \end{aligned}$$

où on a utilisé à la troisième ligne la relation $(a+b)^2 \leq 2(a^2+b^2)$. La preuve de la seconde inégalité se fait de la même manière. □

L'inégalité inverse de (1.15) est fausse, il suffit de prendre un mouvement rigidifiant pour la mettre en défaut. Dans le cas de fonctions dans $H_0^1(\Omega)^d$, on a cependant le résultat ci-dessous.

Lemme 1.47. *Soit Ω un ouvert régulier de $\mathbb{R}^d$. Pour toute fonction $u \in H_0^1(\Omega)^d$, on a*

$$\|\nabla u\|_{L^2(\Omega)} \leq \sqrt{2}\ \|e(u)\|_{L^2(\Omega)}. \tag{1.17}$$

Démonstration. Soit $u \in C_0^\infty(\Omega)^d$. Suivant la définition 1.44 du produit scalaire de deux matrices, on a

$$\|e(u)\|^2_{L^2(\Omega)} = \sum_{i=1}^d \sum_{j=1}^d \int_\Omega e_{ij}(u)^2 dx = \int_\Omega e(u) \cdot e(u) dx.$$

Par définition de $e(u)$, on a $2e(u) \cdot e(u) = \nabla u \cdot \nabla u + \nabla u \cdot (\nabla u)^t$. On se concentre sur le deuxième terme :

$$\begin{aligned}
\int_\Omega \nabla u \cdot (\nabla u)^t dx &= \sum_{i=1}^d \sum_{j=1}^d \int_\Omega \frac{\partial u_i}{\partial x_j} \frac{\partial u_j}{\partial x_i} dx \\
&= -\sum_{i=1}^d \sum_{j=1}^d \int_\Omega \frac{\partial^2 u_i}{\partial x_j \partial x_i} u_j dx \\
&= -\sum_{j=1}^d \int_\Omega u_j \frac{\partial (\text{div } u)}{\partial x_j} dx \\
&= \int_\Omega (\text{div } u)^2 dx.
\end{aligned}$$

Par conséquent, pour tout $u \in C_0^\infty(\Omega)^d$, on a

$$2\|e(u)\|^2_{L^2(\Omega)} = \|\nabla u\|^2_{L^2(\Omega)} + \int_\Omega (\text{div } u)^2 dx.$$

On conclut en utilisant la densité de $C_0^\infty(\Omega)^d$ dans $H_0^1(\Omega)^d$. □

Dans le lemme précédent, la fonction u est nulle au bord. On énonce maintenant l'inégalité de Korn, qui généralise le lemme précédent à des fonctions non nulles au bord. Le prix à payer est que l'ouvert Ω doit être borné. La démonstration de la proposition ci-dessous est délicate, aussi nous l'admettrons.

Proposition 1.48 (Inégalité de Korn)**.** *Soit Ω un ouvert borné régulier de $\mathbb{R}^d$. Il existe une constante C_Ω telle que, pour toute fonction $u \in H^1(\Omega)^d$, on a*

$$\|u\|^2_{H^1(\Omega)} \leq C_\Omega \left(\|u\|^2_{L^2(\Omega)} + \|e(u)\|^2_{L^2(\Omega)} \right). \tag{1.18}$$

Cette inégalité est loin d'être triviale. En effet, au même titre que (1.17), le membre de gauche fait apparaitre toutes les dérivées partielles de u, alors que $e(u)$ ne fait intervenir que des combinaisons linéaires de ces dérivées partielles.

Une conséquence importante de l'inégalité de Korn, et dont nous aurons besoin dans l'étude de l'élasticité linéaire, est la proposition suivante.

Proposition 1.49. *Soit Ω un ouvert connexe, borné et régulier de $\mathbb{R}^d$, avec $d = 2$ ou $d = 3$. Soit $\Gamma_0 \subset \partial\Omega$ un sous-ensemble de la frontière de Ω de mesure superficielle non nulle, et soit*

$$V = \left\{ u \in H^1(\Omega)^d,\ u = 0 \text{ sur } \Gamma_0 \right\}.$$

Il existe une constante C_Ω telle que, pour toute fonction $u \in V$, on a

$$\|u\|_{H^1(\Omega)} \leq C_\Omega \ \|e(u)\|_{L^2(\Omega)}. \tag{1.19}$$

Supposons $d = 2$: alors $\partial\Omega$ est un ensemble de "dimension" 1. L'hypothèse que Γ_0 est de mesure superficielle non nulle implique que Γ_0 n'est pas réduit à un point. De même, si $d = 3$, alors Γ_0, comme objet bidimensionnel, n'est pas de mesure nulle. En particulier, Γ_0 n'est pas réduit à une droite.

Démonstration. On montre d'abord que, si $u \in V$ et $e(u) = 0$, alors $u = 0$. Si $u \in V$ et $e(u) = 0$, alors le lemme 1.45 indique que $u(x) = b + Mx$, où M est une matrice antisymétrique. On a de plus $u = 0$ sur Γ_0.

Si $d = 2$ et si $M \neq 0$, alors l'équation $u(x) = b + Mx = 0$ a une unique solution, ce qui est contradictoire avec le fait que Γ_0 ne soit pas réduit à un point. Donc $M = 0$, ce qui implique $u = 0$.

Si $d = 3$, et si $M \neq 0$, alors l'ensemble des solutions de l'équation $u(x) = b + Mx = 0$ est au plus une droite, ce qui à nouveau est contradictoire avec les hypothèses. Donc $u = 0$.

On prouve maintenant (1.19) par contradiction. Si (1.19) est faux, alors, pour tout n, il existe $u_n \in V$ telle que $\|u_n\|_{H^1(\Omega)} \geq n\|e(u_n)\|_{L^2(\Omega)}$. On peut choisir u_n tel que $\|u_n\|_{H^1(\Omega)} = 1$, et on a donc

$$\frac{1}{n} \geq \|e(u_n)\|_{L^2(\Omega)}. \tag{1.20}$$

Comme u_n est borné dans $H^1(\Omega)^d$, au vu du corollaire 2.69, on peut extraire une sous-suite $u_{\varphi(n)}$ qui converge fortement vers u dans $L^2(\Omega)^d$ et faiblement vers u dans $H^1(\Omega)^d$. On écrit l'inégalité de Korn (1.18) pour $u_{\varphi(n)} - u_{\varphi(p)}$:

$$\|u_{\varphi(n)} - u_{\varphi(p)}\|^2_{H^1(\Omega)} \leq C_\Omega \left(\|u_{\varphi(n)} - u_{\varphi(p)}\|^2_{L^2(\Omega)} + \|e(u_{\varphi(n)} - u_{\varphi(p)})\|^2_{L^2(\Omega)} \right).$$

Le premier terme du membre de droite peut être rendu petit pour n et p grands car $u_{\varphi(n)}$ converge fortement dans $L^2(\Omega)^d$. Il en est de même pour le second terme

grâce à (1.20). Donc la suite $u_{\varphi(n)}$ est de Cauchy dans $H^1(\Omega)^d$, et elle converge donc fortement vers u dans $H^1(\Omega)^d$. Comme V est fermé dans $H^1(\Omega)^d$, on a $u \in V$.

Avec (1.15), on voit donc que $e(u_{\varphi(n)})$ converge fortement dans $L^2(\Omega)^{d\times d}$ vers $e(u)$, ce qui, avec la majoration (1.20), implique que $e(u) = 0$.

Donc la fonction u est telle que $u \in V$ et $e(u) = 0$. La première partie de la preuve montre donc que $u = 0$. Ceci est contradictoire avec le fait que $\|u_{\varphi(n)}\|_{H^1(\Omega)} = 1$ et que $u_{\varphi(n)}$ converge fortement dans $H^1(\Omega)^d$ vers u. □

1.4.3 Problème aux limites

Nous reprenons maintenant le modèle présenté dans la section 1.4.1. On suppose que

$$\Omega \text{ est un ouvert connexe, borné et régulier de } \mathbb{R}^d, \text{ avec } d = 2 \text{ ou } d = 3, \tag{1.21}$$

$$\partial\Omega_D \text{ est un sous-ensemble de } \partial\Omega \text{ de mesure surfacique non nulle,} \tag{1.22}$$

$$f \in L^2(\Omega)^d \text{ et } g \text{ est la trace sur } \partial\Omega_N \text{ d'une fonction de } H^1(\Omega)^d. \tag{1.23}$$

Soit

$$V = \left\{u \in H^1(\Omega)^d,\ u = 0 \text{ sur } \partial\Omega_D\right\}. \tag{1.24}$$

On cherche $u \in V$ tel que

$$\left\{\begin{array}{rcl} -\text{div } [2\mu e(u) + \lambda(\text{tr } e(u))\,\text{Id}] &=& f \text{ dans } \mathcal{D}'(\Omega)^d, \\ \sigma(u)\cdot n &=& g \text{ sur } \partial\Omega_N, \end{array}\right. \tag{1.25}$$

où $\sigma(u)$ est relié à u par la loi de Hooke (1.12). On ne précise pour l'instant pas le sens mathématique exact de la condition aux limites sur $\partial\Omega_N$. Notons simplement que, puisque g est la trace d'une fonction de $H^1(\Omega)^d$, on a par la remarque 1.21 que $g \in H^{1/2}(\partial\Omega)^d \subset L^2(\partial\Omega)^d$.

1.4.4 Formulation variationnelle

On réalise maintenant la formulation variationnelle de (1.25). Soit $v \in V$. Formellement [1], on multiple chaque composante de l'équation aux dérivées partielles (1.14) (qui est la première ligne de (1.25)) par v_i, on somme sur les composantes et on intègre : pour tout $v \in V$,

$$\int_\Omega f\cdot v = -\sum_{i,j}\int_\Omega \frac{\partial\sigma_{ij}(u)}{\partial x_j}v_i = \sum_{i,j}\int_\Omega \sigma_{ij}(u)\frac{\partial v_i}{\partial x_j} - \sum_{i,j}\int_{\partial\Omega}\sigma_{ij}(u)v_i n_j. \tag{1.26}$$

En utilisant les conditions aux limites, on a

$$\sum_{i,j}\int_{\partial\Omega}\sigma_{ij}(u)v_i n_j = \sum_{i,j}\int_{\partial\Omega_N}\sigma_{ij}(u)v_i n_j = \int_{\partial\Omega_N} g\cdot v.$$

1. Remarquer que, si $u \in V \subset H^1(\Omega)^d$, la quantité div $\sigma(u)$ n'est pas une fonction, mais simplement une distribution.

La symétrie de $\sigma(u)$ permet d'écrire

$$\begin{aligned}\sum_{i,j}\int_\Omega \sigma_{ij}(u)\frac{\partial v_i}{\partial x_j} &= \frac{1}{2}\sum_{i,j}\int_\Omega \sigma_{ij}(u)\frac{\partial v_i}{\partial x_j} + \frac{1}{2}\sum_{i,j}\int_\Omega \sigma_{ij}(u)\frac{\partial v_j}{\partial x_i} \\ &= \sum_{i,j}\int_\Omega \sigma_{ij}(u)e_{ij}(v) \\ &= \lambda\sum_{i,j}\int_\Omega \text{tr } e(u)\ \delta_{ij}e_{ij}(v) + 2\mu\sum_{i,j}\int_\Omega e_{ij}(u)e_{ij}(v) \\ &= \lambda\int_\Omega \text{tr } e(u)\ \text{tr } e(v) + 2\mu\int_\Omega e(u)\cdot e(v).\end{aligned}$$

De plus, tr $e(u) = \text{div } u$. La relation (1.26) se récrit donc

$$\forall v \in V, \quad \lambda\int_\Omega \text{div } u\ \text{div } v + 2\mu\int_\Omega e(u)\cdot e(v) = \int_\Omega f\cdot v + \int_{\partial\Omega_N} g\cdot v. \tag{1.27}$$

Il est donc naturel d'introduire les formes a et b définies par

$$\forall u \in V,\ \forall v \in V, \quad a(u,v) = \lambda\int_\Omega \text{div } u\ \text{div } v + 2\mu\int_\Omega e(u)\cdot e(v), \tag{1.28}$$

$$\forall v \in V, \quad b(v) = \int_\Omega f\cdot v + \int_{\partial\Omega_N} g\cdot v. \tag{1.29}$$

La formulation variationnelle associée au problème aux limites (1.25) est donc

$$\text{Chercher } u \in V \ \text{ tel que } \quad \forall v \in V,\ a(u,v) = b(v). \tag{1.30}$$

On montre maintenant que ce problème variationnel est bien posé.

Grâce à l'inégalité de trace (1.7), la forme b est continue sur V. Avec (1.15) et (1.16), la forme a est continue sur V. Pour pouvoir appliquer le théorème de Lax-Milgram, il reste à montrer que a est coercive sur V, et c'est ici qu'on a besoin de l'inégalité de Korn. On a

$$a(u,u) = \lambda\int_\Omega (\text{div } u)^2 + 2\mu\int_\Omega e(u)\cdot e(u).$$

Pour minorer $a(u,u)$, on décompose la matrice $e(u)$ en sa partie diagonale et hors diagonale, suivant

$$e_1(u) = \frac{1}{d}(\text{tr } e(u))\ \text{Id}, \quad e_2(u) = e(u) - e_1(u).$$

Par construction, tr $e_1(u) =$ tr $e(u)$, donc tr $e_2(u) = 0$, et par conséquent $e_2(u) \cdot \mathrm{Id} =$ tr $e_2(u) = 0$. Donc

$$\begin{aligned} e(u) \cdot e(u) &= e_1(u) \cdot e_1(u) + e_2(u) \cdot e_2(u) + 2e_1(u) \cdot e_2(u) \\ &= e_1(u) \cdot e_1(u) + e_2(u) \cdot e_2(u) \\ &= \frac{1}{d}(\text{tr } e(u))^2 + e_2(u) \cdot e_2(u) \\ &= \frac{1}{d}(\text{div } u)^2 + e_2(u) \cdot e_2(u). \end{aligned}$$

Soit $\nu = \min(2\mu, 2\mu + d\lambda)$. Les hypothèses (1.13) donnent $\nu > 0$. On a

$$\begin{aligned} 2\mu e(u) \cdot e(u) + \lambda(\text{div } u)^2 &= \left(\frac{2\mu}{d} + \lambda\right)(\text{div } u)^2 + 2\mu e_2(u) \cdot e_2(u) \\ &\geq \nu \left(\frac{1}{d}(\text{div } u)^2 + e_2(u) \cdot e_2(u)\right) \\ &\geq \nu\, e(u) \cdot e(u). \end{aligned}$$

On en déduit donc que

$$a(u,u) \geq \nu \int_\Omega e(u) \cdot e(u) = \nu \|e(u)\|^2_{L^2(\Omega)}.$$

Avec les hypothèses (1.21) et (1.22), on peut appliquer la proposition 1.49 (conséquence de l'inégalité de Korn) pour minorer $\|e(u)\|_{L^2(\Omega)}$, et on obtient donc :

$$\exists C > 0,\ \forall u \in V,\ a(u,u) \geq C\|u\|^2_{H^1(\Omega)}, \tag{1.31}$$

ce qui donne la coercivité de a sur V. On a donc le résultat suivant :

Théorème 1.50. *On fait les hypothèses (1.21), (1.22) et (1.23). Soit V défini par (1.24), soit a la forme bilinéaire définie sur V par (1.28) et soit b la forme linéaire définie sur V par (1.29). Le problème de chercher $u \in V$ tel que*

$$\forall v \in V,\ a(u,v) = b(v)$$

admet une unique solution. De plus, cette solution est dans $H^2(\Omega)^d$.

Démonstration. La seule affirmation non démontrée est la régularité de la solution (le théorème de Lax-Milgram donne simplement $u \in H^1(\Omega)^d$). Remarquons que $g \in L^2(\partial\Omega_N)^d$ suffit pour donner un sens à la forme linéaire b et permet déjà d'appliquer le théorème de Lax-Milgram. La régularité supplémentaire de g induit la régularité supplémentaire de u. Nous l'admettrons ici. □

1.4.5 Interprétation des résultats

On fait ici le lien entre la formulation variationnelle (1.27) et le problème aux limites (1.25). Soit u l'unique solution de (1.27). Comme $u \in H^2(\Omega)^d$, on a $\nabla u \in H^1(\Omega)^{d\times d}$, et donc on peut définir la trace de ∇u sur $\partial\Omega_N$. On vérifie donc que u est solution de (1.25), la condition aux limites sur $\partial\Omega_N$ prenant le sens d'une équation sur la trace de ∇u.

Chapitre 2

Introduction à la théorie spectrale

Nous présentons dans ce chapitre les fondements de la théorie spectrale des opérateurs (définis en Section 2.1). Cette théorie est particulièrement utile et importante pour l'étude des équations aux dérivées partielles. En effet, un des buts premiers de l'étude d'un opérateur est la détermination de son spectre (Section 2.2), qui est la généralisation en dimension infinie de l'ensemble des valeurs propres d'une matrice. Dans les cas les plus simples, notamment pour les opérateurs dits compacts (Section 2.3), on peut déterminer complètement de manière qualitative le spectre d'un opérateur, et ensuite l'approcher numériquement. Ceci permet de résoudre des problèmes aux valeurs propres définis par une équation aux dérivées partielles (voir le Chapitre 3), ainsi que des problèmes d'évolution en mécanique, physique, etc, comme l'équation de la chaleur, l'équation des ondes, ou l'équation de Schrödinger (voir le Chapitre 6).

Nous verrons des applications concrètes de cette théorie dans le Chapitre 3.

2.1 Opérateurs linéaires

2.1.1 Domaine d'un opérateur

Définition 2.1. *Soient E et F deux espaces de Banach. Un opérateur linéaire est une application linéaire A d'un sous-espace vectoriel de E (noté $D(A)$, et appelé domaine de A) à valeurs dans F.*

Autrement dit, un opérateur linéaire est une application $A : D(A) \subset E \to F$ vérifiant

$$\forall (x,y) \in D(A) \times D(A), \qquad A(x+y) = Ax + Ay, \quad \forall \lambda \in \mathbb{C},\ A(\lambda x) = \lambda A x.$$

On a bien sûr $A(0) = 0$. Le domaine de l'opérateur est inclus dans l'ensemble des éléments de E pour lesquels Ax a un sens en tant qu'élément de F :

$$D(A) \subset \Big\{ x \in E \;\Big|\; Ax \in F \Big\}.$$

Se donner un opérateur linéaire, c'est se donner à la fois son domaine et son action sur les éléments de ce domaine.

Sauf mention du contraire, on supposera toujours dans ce cours que le domaine $D(A)$ est dense dans E, *i.e.* que tout élément $x \in E$ peut être approché par une suite $(x_n)_{n\geq 1}$ d'éléments de $D(A)$ tels que $\|x - x_n\|_E \to 0$ lorsque $n \to +\infty$.

Exemple 2.2 (Laplacien)**.** *Si $E = F = \mathrm{L}^2(\mathbb{R}^d)$, on peut définir l'opérateur $A = -\Delta$ de domaine $D(A) = \mathrm{H}^2(\mathbb{R}^d)$. On pourrait toutefois considérer un opérateur $B = -\Delta$ de domaine plus petit, par exemple restreint aux fonctions $\mathcal{C}^\infty$ et à support compact : $D(B) = \mathcal{C}_c^\infty(\mathbb{R}^d)$.*

On verra par la suite (cf. la Remarque 2.25) qu'il est important de définir à la fois l'action de l'opérateur (ici, appliquer $-\Delta$) et son domaine (sur quel ensemble de fonctions il agit). Deux opérateurs ayant la même action sont a priori différents si leurs domaines sont différents.

On ne peut pas toujours comparer les domaines de deux opérateurs en terme d'inclusion, mais lorsque cela est possible, on parle d'extension.

Définition 2.3 (Extension d'un opérateur)**.** *On dit qu'un opérateur A_2 est une extension de l'opérateur A_1, et on note $A_1 \subset A_2$, si $D(A_1) \subset D(A_2)$ et $A_1 x = A_2 x$ pour tout $x \in D(A_1)$.*

2.1.2 Opérateurs bornés

Définition 2.4. *On dit qu'un opérateur A est borné si*

$$\|A\| = \sup_{x\in D(A)\setminus\{0\}} \frac{\|Ax\|_F}{\|x\|_E} = \sup_{\|x\|_E\leq 1} \|Ax\|_F < +\infty. \tag{2.1}$$

Dans le cas où $D(A) = E$ (le domaine de l'opérateur est égal à l'espace tout entier), la définition ci-dessus est équivalente à la définition d'application linéaire continue vue en cours de première année. On rappelle la caractérisation suivante des applications linéaires continues de E dans F, lorsque E et F sont deux espaces vectoriels normés.

Proposition 2.5. *Soit A une application linéaire de E dans F, où E et F sont deux espaces vectoriels normés. Les 3 propositions suivantes sont équivalentes :*

- *A est continue.*
- *A est continue en 0.*
- *il existe une constante $c \geq 0$ telle que*

$$\forall u \in E, \quad \|Au\|_F \leq c\|u\|_E.$$

Démonstration. Cf. le cours de première année [14]. □

Remarque 2.6 (Extension d'un opérateur borné)**.** *Si $D(A)$ est dense dans E et*

$$\sup_{x\in D(A)\setminus\{0\}} \frac{\|Ax\|_F}{\|x\|_E} < +\infty,$$

alors on peut étendre de manière unique l'opérateur A de domaine $D(A)$ à un opérateur borné sur tout l'espace E. Il suffit pour cela que F soit un espace de Banach, E étant un espace vectoriel normé. Un opérateur borné défini sur l'espace E tout entier est exactement une application linéaire continue de E dans F.

Comme précisé ci-dessus, on supposera toujours dans ce cours (sauf mention du contraire) que $D(A)$ est dense dans E. Par conséquent, et sauf mention du contraire, on supposera toujours qu'un opérateur borné est défini sur l'espace E tout entier. Il s'agira donc d'une application linéaire continue.

L'intérêt de la notion d'opérateur borné est qu'il existe des opérateurs non bornés, comme le montre l'Exemple 2.7 et l'Exercice 2.8.

Exemple 2.7 (Laplacien)**.** *Les opérateurs A et B définis dans l'Exemple 2.2 ne sont pas des opérateurs bornés de E dans E. Il suffit de considérer la suite $f_n(x) = n^{d/2}\chi(nx)$ avec $\chi \in \mathcal{D}(\mathbb{R}^d)$. On a alors*

$$\frac{\|\Delta f_n\|_{\mathrm{L}^2}}{\|f_n\|_{\mathrm{L}^2}} = n^2 \frac{\|\Delta\chi\|_{\mathrm{L}^2}}{\|\chi\|_{\mathrm{L}^2}} \longrightarrow +\infty$$

lorsque $n \to +\infty$.

Exercice 2.8. *On considère les espaces de fonctions $C^0([0,1])$ et $C^1([0,1])$, qu'on munit de la norme*

$$\|f\| = \sup_{t\in[0,1]} |f(t)|.$$

L'application

$$\begin{array}{rcl} A : C^1([0,1]) & \longrightarrow & C^0([0,1]) \\ f & \longmapsto & f' \end{array}$$

est linéaire. Montrer qu'elle n'est pas continue.

Remarque 2.9. *On a cependant le résultat positif suivant. Soient E et F deux espaces de Banach, et A une application linéaire de E dans F qui est fermée, c'est-à-dire telle que l'ensemble $\cup_{u\in E}[u, Au]$ est fermé dans $E \times F$. Alors A est continue (cf. par exemple [3, Théorème II.21 p. 31]).*

Définition 2.10. *On note $\mathcal{L}(E,F)$ l'espace vectoriel des opérateurs bornés de E dans F. L'application $\|\cdot\|$ définie par*

$$\forall A \in \mathcal{L}(E,F), \quad \|A\| := \sup_{x\in E\setminus\{0\}} \frac{\|Ax\|_F}{\|x\|_E} = \sup_{x\in E,\ \|x\|_E=1} \|Ax\|_F, \tag{2.2}$$

est une norme sur cet espace.

Le seul point éventuellement délicat est de montrer l'inégalité triangulaire $\|A+B\| \leq \|A\| + \|B\|$. Pour ce faire, on fixe $f \in E \setminus \{0\}$ et on écrit

$$\|(A+B)f\|_F \leq \|Af\|_F + \|Bf\|_F \leq \Big(\|A\| + \|B\|\Big)\|f\|_E.$$

Ceci montre que

$$\frac{\|(A+B)f\|_F}{\|f\|_E} \leq \|A\| + \|B\|,$$

d'où le résultat en prenant le supremum sur $f \in E \setminus \{0\}$.

Exercice 2.11. *Soient E, F et G trois espaces de Banach, et $A \in \mathcal{L}(E,F)$ et $B \in \mathcal{L}(F,G)$. Montrer que $BA \in \mathcal{L}(E,G)$ et $\|BA\| \leq \|A\|\ \|B\|$.*

Un cas particulier important est lorsque l'espace d'arrivée est $\mathbb{R}$.

Définition 2.12. *L'ensemble $\mathcal{L}(E,\mathbb{R})$ des applications linéaires continues de E dans $\mathbb{R}$ est appelé espace dual de E et est noté E'. Un élément de E' est appelé forme linéaire continue et son action sur un élément $u \in E$ est notée à l'aide du crochet de dualité :*

$$\langle A, u\rangle_{E',E} = Au \in \mathbb{R}.$$

L'espace E' est équipé de la norme

$$\|A\|_{E'} = \sup_{u\in E, u\neq 0} \frac{|Au|}{\|u\|_E}.$$

Donnons quelques exemples d'opérateurs bornés.

Exemple 2.13 (Opérateurs de shift)**.** *On considère $E = F = \ell^p(\mathbb{N},\mathbb{C})$ (pour $1 \leq p \leq +\infty$ fixé), où*

$$\ell^p(\mathbb{N},\mathbb{C}) = \left\{ (x_1, x_2, \dots, x_n, \dots) \in \mathbb{C}^{\mathbb{N}} \;\middle|\; \sum_{n=1}^{+\infty} |x_i|^p < +\infty \right\}, \qquad 1 \leq p < +\infty,$$

et

$$\ell^\infty(\mathbb{N},\mathbb{C}) = \left\{ (x_1, x_2, \dots, x_n, \dots) \in \mathbb{C}^{\mathbb{N}} \;\middle|\; \sup_{i\in\mathbb{N}} |x_i| < +\infty \right\}.$$

On définit les opérateurs de shift à droite et de shift à gauche, de domaine $\ell^p(\mathbb{N},\mathbb{C})$, par

$$\tau_{\mathrm{d}}(x_1, x_2, \dots, x_n, \dots) = (0, x_1, x_2, \dots, x_n, \dots) \tag{2.3}$$

et

$$\tau_{\mathrm{g}}(x_1, x_2, \dots, x_n, \dots) = (x_2, x_3, \dots, x_n, \dots). \tag{2.4}$$

Ces deux opérateurs sont des opérateurs bornés. Il est immédiat que $\|\tau_{\mathrm{d}} x\| = \|x\|$ pour tout $x \in \ell^p(\mathbb{N},\mathbb{C})$ et donc $\|\tau_{\mathrm{d}}\| = 1$. Pour τ_{g}, on note tout d'abord que $\|\tau_{\mathrm{g}} x\| \leq \|x\|$, avec égalité par exemple pour $x = (0,1,0,\dots)$, ce qui donne $\|\tau_{\mathrm{g}}\| = 1$.

Exercice 2.14 (Opérateur de convolution)**.** *Soit $E = F = \mathrm{L}^2(\mathbb{R}^d)$ et $k \in \mathrm{L}^1(\mathbb{R}^d)$. Montrer que l'opérateur $T : E \to E$ d'action $Tf = k \star f$ est bien défini et est borné avec $\|T\| \leq \|k\|_{\mathrm{L}^1}$.*

Exercice 2.15 (Opérateur intégral)**.** *On considère $E = \mathrm{L}^1([0,1], \mathbb{R})$, $F = C^0([0,1], \mathbb{R})$, et $k \in C^0([0,1]^2, \mathbb{R})$. On rappelle que la norme sur l'espace de Banach F est $\|g\|_F = \sup_{x\in[0,1]} |g(x)|$. On considère l'opérateur K défini par*

$$Kf(x) = \int_0^1 k(x,y) f(y)\, dy.$$

Vérifier que $Kf \in F$ lorsque $f \in E$ puis que $K \in \mathcal{L}(E,F)$.

Exemple 2.16 (Opérateur de multiplication)**.** *Soit $E = F = \mathrm{L}^2(\mathbb{R}^d)$. Pour une fonction $V \in \mathrm{L}^\infty(\mathbb{R}^d, \mathbb{C})$ donnée, on définit l'opérateur A sur E par*

$$A\varphi = V\varphi.$$

On constate que, pour tout $\varphi \in E$, on a $A\varphi \in F$. Le domaine de A est donc $F = \mathrm{L}^2(\mathbb{R}^d)$. On vérifie de plus que $\|A\varphi\|_F \leq \|V\|_{\mathrm{L}^\infty} \|\varphi\|_E$, donc A est borné.

Pour une fonction mesurable V qui n'est pas bornée, il faudrait restreindre le domaine de l'opérateur aux fonctions $\varphi \in \mathrm{L}^2(\mathbb{R}^d)$ telles que

$$\int_{\mathbb{R}^d} |V\varphi|^2 < +\infty.$$

Il n'est toutefois pas clair que l'on obtiendrait ainsi un sous-ensemble dense de $\mathrm{L}^2(\mathbb{R}^d)$. C'est toutefois le cas si $V \in \mathrm{L}^2_{\mathrm{loc}}(\mathbb{R}^d)$ car alors $\mathcal{D}(\mathbb{R}^d) \subset D(A)$, et $\mathcal{D}(\mathbb{R}^d)$ est dense dans $\mathrm{L}^2(\mathbb{R}^d)$.

Exercice 2.17. *Montrer que si, dans l'Exemple 2.16, la fonction V est continue et bornée, alors $\|A\| = \sup_{x\in\mathbb{R}^d} |V(x)|$.*

Concluons cette section par un résultat important.

Proposition 2.18. *Si F est un espace de Banach et E un espace normé, alors $\mathcal{L}(E,F)$ est un espace de Banach.*

Démonstration. Considérons une suite de Cauchy $(A_n)_{n\geq 0}$ de $\mathcal{L}(E,F)$ pour la norme donnée par (2.1). Alors, pour tout $\varepsilon > 0$, il existe $N_\varepsilon \in \mathbb{N}$ tel que

$$\|A_n - A_m\| \leq \varepsilon \tag{2.5}$$

si $n, m \geq N_\varepsilon$. En particulier, la suite $(\|A_n\|)_{n\geq 0}$ est bornée, et il existe $C > 0$ tel que $0 \leq \|A_n\| \leq C < +\infty$ pour tout $n \in \mathbb{N}$. Pour $x \in E$ donné, on a

$$\|A_n x - A_m x\|_F \leq \varepsilon \|x\|_E \tag{2.6}$$

si $n, m \geq N_\varepsilon$. La suite $(A_n x)_{n\geq 0}$ est ainsi une suite de Cauchy dans l'espace de Banach F, et admet donc une limite $a_x \in F$. On peut construire un opérateur limite A en posant $Ax = a_x$. On vérifie facilement que A est linéaire (par unicité de la limite). Par ailleurs, en passant à la limite $m \to +\infty$ dans (2.6), on obtient

$$\|A_n x - Ax\|_F \leq \varepsilon \|x\|_E,$$

et donc, pour $n \geq N_\varepsilon$,

$$\|Ax\|_F \leq \|Ax - A_n x\|_F + \|A_n x\|_F \leq (\varepsilon + C)\|x\|_E.$$

Ainsi, A est dans $\mathcal{L}(E, F)$ et on peut passer à la limite dans (2.5) (ou prendre le supremum sur les $x \in E$ avec $\|x\|_E \leq 1$) et obtenir que, pour tout $\varepsilon > 0$, il existe $N_\varepsilon \in \mathbb{N}$ tel que

$$\|A_n - A\| \leq \varepsilon$$

pour tout $n \geq N_\varepsilon$. Ceci montre bien que $A_n \to A$ dans $\mathcal{L}(E, F)$. □

Finissons cette section en prouvant le résultat suivant :

Proposition 2.19. *Soient V et W deux espaces de Hilbert et $A \in \mathcal{L}(V, W)$ un opérateur borné de V dans W. Soit $(u_n)_{n\in\mathbb{N}}$ une suite d'éléments de V qui converge faiblement vers un élément $u \in V$. Alors la suite $(Au_n)_{n\in\mathbb{N}}$ converge faiblement vers Au dans W.*

Démonstration. Soit $w \in W$. Soit $\varphi : v \in V \mapsto \langle Av, w\rangle_W$. Comme $A \in \mathcal{L}(V, W)$, on vérifie facilement que $\varphi \in V'$. D'après la caractérisation équivalente de la convergence faible donnée par la Proposition 1.36, on a alors $\varphi(u_n) \underset{n\to+\infty}{\longrightarrow} \varphi(u)$, ce qui se réécrit

$$\langle Au_n, w\rangle_W \underset{n\to+\infty}{\longrightarrow} \langle Au, w\rangle_W.$$

Cette convergence a lieu pour tout $w \in W$, ce qui implique bien que la suite $(Au_n)_{n\in\mathbb{N}}$ converge faiblement vers Au dans W. □

2.1.3 Inverse d'un opérateur

Définition 2.20. *Pour un opérateur A de domaine $D(A)$, on définit son image*

$$\mathrm{Ran}(A) = A\Big(D(A)\Big) = \Big\{ y \in F \;\Big|\; \exists x \in D(A), \quad y = Ax \Big\},$$

et son noyau

$$\mathrm{Ker}(A) = \Big\{ x \in D(A) \;\Big|\; Ax = 0 \Big\}.$$

On dit que A est injectif si $\mathrm{Ker}(A) = \{0\}$, et que A est surjectif si $\mathrm{Ran}(A) = F$. L'opérateur est bijectif s'il est à la fois injectif et surjectif.

En dimension finie, on a le résultat classique suivant :

Proposition 2.21. *Soit E un espace vectoriel de dimension finie et A une application linéaire de E dans E. Alors A est continue, et de plus les 3 propositions suivantes sont équivalentes :*
- *A est injective sur E.*
- *A est surjective sur E.*
- *A est bijective de E dans E.*

Comme le montre l'exercice suivant, la situation en dimension infinie est plus complexe : une application linéaire continue peut être injective sans être surjective.

Exemple 2.22. *L'opérateur de shift à droite* (2.3) *est injectif, mais pas surjectif car* $(1, 0, \dots) \notin \mathrm{Ran}(\tau_{\mathrm{d}})$. *L'opérateur de shift à gauche* (2.4) *est surjectif, mais pas injectif.*

Si A est injectif, on peut définir l'opérateur inverse, de domaine $D(A^{-1}) = \mathrm{Ran}(A) \subset F$, à valeurs dans $D(A) \subset E$, par

$$x = A^{-1}y \Longleftrightarrow y = Ax.$$

Il n'y a aucune raison *a priori* que l'inverse soit borné. Ceci motive la définition suivante.

Définition 2.23 (Opérateur inversible)**.** *On dit qu'un opérateur A de domaine $D(A)$ est inversible si $A : D(A) \subset E \to F$ est bijectif et a un inverse $A^{-1} : F \to D(A) \subset E$ borné (comme opérateur de F dans E).*

Notons qu'on ne demande pas que A soit lui-même borné.

Enonçons une propriété qui nous sera utile par la suite, et qui permet de conclure à l'inversibilité d'un opérateur linéaire *borné* dès qu'il est bijectif (la preuve, omise, repose sur le lemme de Baire, voir par exemple [15]).

Proposition 2.24. *Si $A \in \mathcal{L}(E, F)$ et A est une bijection de E vers F, alors $A^{-1} \in \mathcal{L}(F, E)$.*

Soit A un opérateur inversible (qu'on suppose non borné) : il est donc bijectif de $D(A)$ sur F et son inverse B est borné. Donc $B \in \mathcal{L}(F, E)$. Cependant, B n'est pas nécessairement une bijection de F vers E (c'est seulement une bijection de F vers $D(A)$). Si B est une bijection de F vers E, alors on peut appliquer la proposition ci-dessus à B, ce qui donne le fait que $A \in \mathcal{L}(E, F)$.

A titre d'exemple, on peut prendre $E = \mathrm{H}^1_0(\Omega)$ pour un ouvert $\Omega \subset \mathbb{R}^d$ borné, $F = \mathrm{L}^2(\Omega)$, et l'opérateur A défini sur $D(A) = \mathrm{H}^2(\mathbb{R}^d) \cap E$ par $Au = -\Delta u$. On a déjà vu qu'un tel opérateur n'est pas borné (cf. l'exemple 2.7). L'opérateur A est bijectif de $D(A)$ sur F, et son inverse est borné : pour tout $g \in \mathrm{L}^2(\Omega)$, la solution $u \in D(A)$ de $-\Delta u = g$ satisfait $\|u\|_E \leq C\|g\|_F$. L'opérateur B est bijectif de F sur $D(A)$ qui est un sous-espace strict de E. En particulier, B n'est pas surjectif sur E.

Remarque 2.25 (De l'importance du domaine)**.** *Considérons l'espace de Banach des fonctions continues $E = F = \mathrm{C}^0([0,1],\mathbb{R})$, muni de la norme*

$$\|f\| = \sup_{t\in[0,1]} |f(t)|.$$

On peut définir un opérateur A_M de domaine "maximal" $D(A_M) = \mathrm{C}^1([0,1],\mathbb{R})$ par

$$A_M f = \frac{df}{dt}.$$

On peut également en définir plusieurs restrictions, qui ont la même action de dérivation, mais ont des domaines plus petits, en fonction des conditions de bord que l'on souhaite imposer (ou qui sont imposées par la physique du problème) :

— *l'opérateur A_k (pour $k \in \mathbb{R}$), de domaine*

$$D(A_k) = \Big\{ f \in \mathrm{C}^1([0,1],\mathbb{R}) \;\Big|\; f(0) = kf(1) \Big\},$$

avec les cas particuliers

$$D(A_0) = \Big\{ f \in \mathrm{C}^1([0,1],\mathbb{R}) \;\Big|\; f(0) = 0 \Big\}$$

et

$$D(A_\infty) = \Big\{ f \in \mathrm{C}^1([0,1],\mathbb{R}) \;\Big|\; f(1) = 0 \Big\};$$

— *l'opérateur A_{00} de domaine $D(A_{00}) = \Big\{ f \in \mathrm{C}^1([0,1],\mathbb{R}) \;\Big|\; f(0) = f(1) = 0 \Big\}$;*

— *l'opérateur A_m de domaine "minimal" $D(A_m) = \mathcal{D}(]0,1[,\mathbb{R})$.*

On a bien sûr $A_m \subset A_{00} \subset A_k \subset A_M$ pour tout $k \in \mathbb{R} \cup \{+\infty\}$. Ces différents opérateurs, bien qu'ayant la même action, ont des comportements très différents en ce qui concerne leur injectivité ou leur surjectivité. Ainsi,

— *A_M n'est pas injectif car toutes les fonctions $f + c$ pour $f \in D(A_M)$ fixé et $c \in \mathbb{R}$ quelconque ont la même image ;*

— *A_k est inversible si et seulement si $k \neq 1$, et*

$$A_k^{-1} f \;:\; t \mapsto \frac{1}{1-k}\left(\int_0^t f + k\int_t^1 f\right) = \int_0^t f + \frac{k}{1-k}\int_0^1 f.$$

On vérifie en particulier que $A_k^{-1} f(0) = k A_k^{-1} f(1)$;

— *A_{00} est inversible sur $D(A_{00}^{-1}) = \left\{ f \in \mathrm{C}^0([0,1],\mathbb{R}) \;\middle|\; \int_0^1 f = 0 \right\}$ et $A_{00}^{-1} f : t \mapsto \int_0^t f$ (et en particulier $A_{00}^{-1} f \in D(A_{00})$). Notons en effet que si $g \in D(A_{00})$, on a $\int_0^1 g' = g(1) - g(0) = 0$, ce qui motive la définition de $D(A_{00}^{-1}) = \mathrm{Ran}(A_{00})$;*

— *finalement, A_m n'est inversible que sur $\mathcal{D}(]0,1[,\mathbb{R}) \cap D(A_{00}^{-1})$.*

2.1.4 Adjoint d'un opérateur borné

Définition 2.26 (Adjoint d'un opérateur borné)**.** *Soit H un espace de Hilbert, muni d'un produit scalaire (complexe) noté $\langle\cdot,\cdot\rangle$, et $T \in \mathcal{L}(H)$. L'adjoint de T est l'opérateur T^* défini par*

$$\forall u \in H,\ \forall v \in H, \qquad \langle T^*u, v\rangle = \langle u, Tv\rangle.$$

On dit que T est auto-adjoint *si $T^* = T$.*

Exemple 2.27. *On vérifie facilement que l'adjoint sur $\ell^2(\mathbb{N},\mathbb{C})$ de l'opérateur $\tau_{\rm d}$ de shift à droite* (2.3) *est l'opérateur $\tau_{\rm g}$ de shift à gauche* (2.4) *(et réciproquement).*

Exercice 2.28. *Soit $V \in \mathrm{L}^\infty([a,b],\mathbb{R})$. Vérifier que l'opérateur $T : \mathrm{L}^2([a,b]) \to \mathrm{L}^2([a,b])$ défini par $Tf(x) = V(x)\,f(x)$ est autoadjoint.*

Exercice 2.29 (Opérateurs de Hilbert-Schmidt)**.** *Soit $H = \mathrm{L}^2(\mathbb{R}^d,\mathbb{C})$ et $K \in \mathrm{L}^2(\mathbb{R}^{2d},\mathbb{C})$. On considère l'opérateur intégral $\widehat{K} : H \to H$ défini par*

$$\widehat{K}f(x) = \int_{\mathbb{R}^d} K(x,y) f(y)\, dy.$$

On dit que K est le noyau *de $\widehat{K}$. Montrer que $\widehat{K} \in \mathcal{L}(H)$ et que*

$$\left\|\widehat{K}\right\| \leq \|K\|_{\mathrm{L}^2} = \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |K(x,y)|^2\, dx\, dy\right)^{1/2}.$$

Montrer également que $\widehat{K}^*$ est un opérateur intégral de noyau $\overline{K(y,x)}$.

On pourra vérifier en exercice la propriété suivante (voir [8, Section 4.2]).

Proposition 2.30. *Si $T \in \mathcal{L}(H)$ alors $T^* \in \mathcal{L}(H)$, $\|T^*\| = \|T\|$ et $T^{**} = T$. Si T_1 et T_2 sont dans $\mathcal{L}(H)$, alors $(T_1T_2)^* = T_2^*T_1^*$.*

Le résultat suivant sera utile dans la suite :

Proposition 2.31. *Soit $T \in \mathcal{L}(H)$ et $\lambda \in \mathbb{C}$. Alors*

$$\Big(\mathrm{Ran}(\lambda - T)\Big)^\perp = \mathrm{Ker}\left(\overline{\lambda} - T^*\right). \tag{2.7}$$

Démonstration. Par définition, on a, pour tout x et y dans H, que

$$\langle(\lambda - T)x, y\rangle = \langle x, (\overline{\lambda} - T^*)y\rangle.$$

Soit $\widetilde{x} \in \mathrm{Ran}(\lambda - T)$ et $y \in \mathrm{Ker}(\overline{\lambda} - T^*)$. Il existe x tel que $\widetilde{x} = (\lambda - T)x$ et ainsi $\langle\widetilde{x}, y\rangle = \langle x, (\overline{\lambda} - T^*)y\rangle = 0$. Ceci montre que $\mathrm{Ker}(\overline{\lambda} - T^*) \subset \Big(\mathrm{Ran}(\lambda - T)\Big)^\perp$. On montre l'inclusion inverse. Soit $y \in \Big(\mathrm{Ran}(\lambda - T)\Big)^\perp$. Pour tout $x \in H$, on a $\langle y, (\lambda - T)x\rangle = 0 = \langle(\overline{\lambda} - T^*)y, x\rangle$. Ceci étant vrai pour tout $x \in H$, on obtient $(\overline{\lambda} - T^*)y = 0$ et donc l'inclusion contraire $\Big(\mathrm{Ran}(\lambda - T)\Big)^\perp \subset \mathrm{Ker}(\overline{\lambda} - T^*)$. □

2.2 Théorie spectrale des opérateurs bornés

On va à présent étudier de plus près l'inversibilité d'opérateurs bornés d'un espace de Banach E dans lui-même. De telles considérations sont particulièrement intéressantes lorsqu'il s'agit de résoudre une équation du type

$$(\lambda \operatorname{Id} - A)u = f$$

avec $u, f \in E$ et $\lambda \in \mathbb{C}$. En effet, si l'inverse de l'opérateur $\lambda \operatorname{Id} - A$ est bien défini, alors $u = (\lambda \operatorname{Id} - A)^{-1} f$ est l'unique solution de cette équation.

2.2.1 Théorie générale

On peut définir aisément l'inverse d'un opérateur $\operatorname{Id} - A$ lorsque A est de norme suffisamment petite par le biais d'une série infinie. Plus précisément, la notion pertinente est le rayon spectral.

Lemme 2.32 (Rayon spectral)**.** *Soit $A \in \mathcal{L}(E)$. Alors la limite suivante existe :*

$$r(A) = \lim_{n \to +\infty} \|A^n\|^{1/n} = \inf_{n \geq 1} \|A^n\|^{1/n},$$

et est appelée rayon spectral. *On a en particulier $r(A) \leq \|A\|$.*

On peut avoir $r(A) < \|A\|$. Le cas le plus frappant est celui des opérateurs *nilpotents*, c'est-à-dire tels qu'il existe $N \in \mathbb{N}$ tel que $A^N = 0$. Dans ce cas, $r(A) = 0$. Par exemple, l'opérateur borné sur $E = \mathbb{R}^2$ dont la représentation matricielle dans la base canonique est

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

est tel que $\|A\| = 1$ mais $A^2 = 0$ et donc $r(A) = 0$.

Démonstration. On suit la preuve de [9, Section I.4.2]. Pour $n, m \in \mathbb{N}$, on a clairement

$$\|A^{n+m}\| \leq \|A^n\| \, \|A^m\|, \qquad \|A^n\| \leq \|A\|^n, \tag{2.8}$$

avec la convention $A^0 = \operatorname{Id}$. Ces inégalités proviennent de l'inégalité générale $\|AB\| \leq \|A\| \, \|B\|$ pour $A, B \in \mathcal{L}(E)$ (voir Exercice 2.11). Notons

$$a_n = \ln \|A^n\|.$$

Alors $a_n/n \leq \ln \|A\|$. Il s'agit de montrer que la suite $(a_n/n)_{n \geq 1}$ converge.

Les inégalités (2.8) montrent que $a_{n+m} \leq a_n + a_m$. Pour $m \in \mathbb{N}^*$ donné, considérons la division euclidienne de n par m : $n = qm + r$ avec $q, r \in \mathbb{N}$ et $r < m$. On montre alors que $a_n \leq q a_m + a_r$ et ainsi

$$\frac{a_n}{n} \leq \frac{q}{n} \, a_m + \frac{1}{n} \, a_r.$$

Lorsque $n \to +\infty$, $q/n \to 1/m$ alors que les valeurs de r sont limitées à $0, \dots, m-1$. Ainsi,

$$\sup_{r=0,\dots,m-1} \frac{1}{n} a_r \longrightarrow 0$$

lorsque $n \to +\infty$, et donc

$$\limsup_{n\to+\infty} \frac{a_n}{n} \leq \frac{a_m}{m}.$$

Comme m est arbitraire, on en déduit que

$$\limsup_{n\to+\infty} \frac{a_n}{n} \leq \inf_{m\geq 1} \frac{a_m}{m}.$$

Par ailleurs, on a trivialement

$$\liminf_{n\to+\infty} \frac{a_n}{n} \geq \inf_{m\geq 1} \frac{a_m}{m},$$

et on en déduit donc

$$\limsup_{n\to+\infty} \frac{a_n}{n} \leq \inf_{m\geq 1} \frac{a_m}{m} \leq \liminf_{n\to+\infty} \frac{a_n}{n}.$$

Les inégalités ci-dessus sont finalement des égalités, ce qui montre que la suite $(a_n/n)_{n\geq 1}$ est bien convergente, et qu'elle converge vers $\inf_{m\geq 1}(a_m/m)$. □

Exercice 2.33. *Soient $\tau_{\rm d}$ et $\tau_{\rm g}$ les opérateurs de shift définis par* (2.3) *et* (2.4). *Montrer que $r(\tau_{\rm d}) = r(\tau_{\rm g}) = 1$.*

Le lemme suivant, simple, va nous être utile dans la suite :

Lemme 2.34. *Soit $A \in \mathcal{L}(E)$ et soit $z \in \mathbb{C}$. La série $\sum_n z^n A^n$ est convergente dans $\mathcal{L}(E)$ si et seulement si $|z| < 1/r(A)$.*

Démonstration. Comme E est un Banach, l'espace $\mathcal{L}(E)$ est un espace de Banach (cf. la Proposition 2.18). D'après le cours de première année [14], on sait que, si la série est normalement convergente, i.e. si $\sum_n |z|^n \, \|A^n\|_E < \infty$, alors la série $\sum_n z^n A^n$ est convergente dans $\mathcal{L}(E)$.

Supposons $|z| < 1/r(A)$. Soit $\varepsilon > 0$. Par définition du rayon spectral, il existe N_ε tel que, pour tout $n > N_\varepsilon$, on a $\|A^n\|^{1/n} \leq r(A) + \varepsilon$, donc $|z|^n \, \|A^n\|_E \leq |z|^n (r(A)+\varepsilon)^n$. Grace à l'hypothèse sur z, on peut trouver ε tel que $|z| \, (r(A)+\varepsilon) < 1$. La série $\sum_n |z|^n \, \|A^n\|_E$ est donc convergente, donc la série $\sum_n z^n A^n$ est convergente dans $\mathcal{L}(E)$.

Supposons maintenant que la série $\sum_n z^n A^n$ est convergente dans $\mathcal{L}(E)$. Ceci implique que $z^n A^n$ converge vers 0 dans $\mathcal{L}(E)$: $\lim_n |z|^n \|A^n\|_E = 0$. Or $r(A) = \inf_n \|A^n\|_E^{1/n}$. On a donc $(|z|r(A))^n \leq |z|^n \|A^n\|_E$, et donc $\lim_n (|z|r(A))^n = 0$. Ceci implique que $|z|r(A) < 1$, d'où $|z| < 1/r(A)$. □

On peut à présent définir l'inverse de l'opérateur $\mathrm{Id} - A$ lorsque A a un rayon spectral strictement plus petit que 1.

Lemme 2.35 (Série de Neumann)**.** *Soit $A \in \mathcal{L}(E)$ tel que $r(A) < 1$. Alors l'opérateur $\mathrm{Id} - A$ a un inverse borné $(\mathrm{Id} - A)^{-1} \in \mathcal{L}(E)$ et*

$$(\mathrm{Id} - A)^{-1} = \sum_{n=0}^{+\infty} A^n. \tag{2.9}$$

Démonstration. Le lemme 2.34 montre que, pour tout z tel que $|z| < 1/r(A)$, la série $\displaystyle\sum_{n=0}^{+\infty} z^n A^n$ converge dans $\mathcal{L}(E)$. C'est donc en particulier le cas pour $z = 1$, ce qui indique que la série du membre de droite de (2.9) est une série convergente dans $\mathcal{L}(E)$.

On écrit ensuite que, pour tout N, on a

$$(\mathrm{Id} - A) \sum_{n=0}^{N} A^n = \mathrm{Id} - A^{N+1}. \tag{2.10}$$

On passe à la limite $N \to \infty$. Le membre de gauche converge vers $(\mathrm{Id} - A) \sum_{n=0}^{+\infty} A^n$. Pour étudier le membre de droite, on utilise le fait que $\|A^N\|^{1/N} \to r(A) < 1$. Il existe donc $\varepsilon > 0$ et N_ε tel que, pour tout $n > N_\varepsilon$, on a $\|A^N\|^{1/N} \leq 1 - \varepsilon$, si bien que $\|A^N\| \leq (1 - \varepsilon)^N$, et donc $\lim_{N\to\infty} \|A^N\| = 0$. On peut maintenant passer à la limite $N \to \infty$ dans (2.10), ce qui donne $(\mathrm{Id} - A) \sum_{n=0}^{\infty} A^n = \mathrm{Id}$, et donc le résultat escompté. □

Théorème-Définition 2.36. *Soit E un espace de Banach et $T \in \mathcal{L}(E)$. D'après la proposition 2.24, $\lambda - T$ est inversible si et seulement si $\lambda - T$ est bijectif.*

1. *On appelle* ensemble résolvant *de T l'ensemble*

$$\rho(T) = \Big\{\lambda \in \mathbb{C}, \quad \lambda - T \text{ inversible}\Big\}.$$

L'ensemble résolvant $\rho(T)$ est un ouvert de $\mathbb{C}$.

2. *Pour $\lambda \in \rho(T)$, on note $R(\lambda) = (\lambda - T)^{-1}$. La famille d'opérateurs linéaires bornés $(R(\lambda))_{\lambda\in\rho(T)}$ est appelée la résolvante de T. La fonction $\lambda \mapsto R(\lambda)$ est analytique de $\rho(T)$ dans $\mathcal{L}(E)$ et on a, pour tout $(\lambda, \mu) \in \rho(T) \times \rho(T)$, l'identité de la résolvante*

$$R(\lambda) - R(\mu) = (\mu - \lambda)R(\lambda)R(\mu).$$

3. *On appelle* spectre *de T l'ensemble*

$$\sigma(T) = \mathbb{C} \setminus \rho(T) = \Big\{\lambda \in \mathbb{C}, \quad \lambda - T \text{ non inversible}\Big\}.$$

L'ensemble $\sigma(T)$ est un compact de $\mathbb{C}$.

4. *On a*
$$\sigma(T) \subset \overline{D(0, r(T))},$$
où $\overline{D(0, r(T))}$ *est le disque fermé centré en 0 et de rayon* $r(T)$*. On a aussi que*
$$\sigma(T) \cap C(0, r(T)) \neq \emptyset$$
où $C(0, r(T))$ *est le cercle de centre 0 et de rayon* $r(T)$*. En particulier le spectre d'un opérateur borné n'est jamais vide.*

5. *L'ensemble* $\sigma(T)$ *se décompose en l'union disjointe*
$$\sigma(T) = \sigma_{\mathrm{p}}(T) \cup \sigma_{\mathrm{r}}(T) \cup \sigma_{\mathrm{c}}(T),$$
avec
$$\sigma_{\mathrm{p}}(T) = \Big\{\lambda \in \mathbb{C}, \quad \lambda - T \text{ non injectif}\Big\},$$
$$\sigma_{\mathrm{r}}(T) = \Big\{\lambda \in \mathbb{C}, \quad \lambda - T \text{ injectif et } \overline{(\lambda - T)E} \neq E\Big\},$$
et
$$\sigma_{\mathrm{c}}(T) = \Big\{\lambda \in \mathbb{C}, \quad \lambda - T \text{ injectif et } (\lambda - T)E \neq \overline{(\lambda - T)E} = E\Big\}.$$
L'ensemble $\sigma_{\mathrm{p}}(T)$ *est appelé le* spectre ponctuel *de* T*,* $\sigma_{\mathrm{c}}(T)$ *le* spectre continu *de* T*,* $\sigma_{\mathrm{r}}(T)$ *le* spectre résiduel *de* T*.*

Notons que les trois types de spectre définis ci-dessus ont été classés par ordre croissant de défaut d'inversibilité :
- pour le spectre ponctuel, on a un défaut d'injectivité ;
- pour le spectre résiduel, on a un défaut majeur de surjectivité : même en prenant l'adhérence de l'image de E, on ne retrouve pas E ;
- pour le spectre continu, l'inverse est bien défini sur un domaine dense, mais n'est pas borné. Montrons ceci par l'absurde.
  L'opérateur linéaire $\lambda - T$ est bijectif de E sur $(\lambda - T)E$. On introduit son inverse $B : (\lambda - T)E \to E$, qui est défini sur un sous-ensemble dense de E. Supposons B borné. On peut alors l'étendre par continuité comme un opérateur de E sur E. Soit $y \in E$ et $u = By$ (qui existe car B est maintenant défini sur tout E). Montrons que $y = (\lambda - T)u$:
  - Si $y \in (\lambda - T)E$, c'est évident.
  - Sinon, on sait qu'il existe une suite $y_n \in (\lambda - T)E$ telle que $y_n \to y$. Puisque $y_n \in (\lambda - T)E$, il existe $u_n \in E$ tel que $y_n = (\lambda - T)u_n$, et donc $u_n = By_n$. La suite y_n est convergente, donc de Cauchy. Puisque B est borné, on voit que u_n est aussi de Cauchy, donc convergente. Par définition, on a $u = By = \lim_n u_n$. On peut donc passer à la limite dans l'égalité $y_n = (\lambda - T)u_n$ (puisque $\lambda - T$ est continu), ce qui donne $y = (\lambda - T)u$.

On vient donc de démontrer que, pour tout $y \in E$, il existe $u \in E$ tel que $y = (\lambda - T)u$, ce qui donne $(\lambda - T)E = E$. On obtient donc une contradiction.

Démonstration. Soit $\lambda \in \mathbb{C}$ tel que $|\lambda| > r(T)$. On écrit

$$\lambda - T = \lambda \left(\mathrm{Id} - \frac{T}{\lambda} \right)$$

et $r(T/\lambda) = r(T)/|\lambda| < 1$. En utilisant le lemme 2.35, on voit que $\lambda - T$ est inversible, donc $\lambda \in \rho(T)$. Il en découle que

$$\sigma(T) \subset \overline{D(0, r(T))}.$$

Soit maintenant $\mu \in \rho(T)$. On écrit

$$\lambda - T = \mu - T + (\lambda - \mu)\mathrm{Id} = (\mu - T)\Big(\mathrm{Id} + (\lambda - \mu)(\mu - T)^{-1} \Big). \tag{2.11}$$

Donc, si $|\lambda - \mu|\, r\big((\mu - T)^{-1}\big) < 1$, alors $\lambda - T$ est inversible (en vertu du lemme 2.35). On en déduit que $\rho(T)$ est un ouvert de $\mathbb{C}$.

Comme $\sigma(T) = \mathbb{C} \setminus \rho(T)$, on obtient que $\sigma(T)$ est un fermé de $\mathbb{C}$. Comme $\sigma(T)$ est borné, c'est un compact de $\mathbb{C}$.

La relation (2.11) montre que $R(\lambda)$ est analytique dans $\rho(T)$.

En multipliant les deux membres de l'égalité

$$(\lambda - T) = (\mu - T) + (\lambda - \mu)\mathrm{Id}$$

à gauche par $R(\lambda)$ et à droite par $R(\mu)$, on obtient l'identité de la résolvante.

Supposons que $\sigma(T) \cap C(0, r(T)) = \emptyset$. Comme $\sigma(T)$ est compact, il existe $\varepsilon \in]0, r(T)[$ tel que

$$\mathbb{C} \setminus \overline{D(0, r(T) - \varepsilon)} \subset \rho(T).$$

Comme $R(\lambda)$ est analytique sur $\rho(T)$, il en résulte que $f(z) = R(1/z)$ est analytique sur $D(0, (r(T) - \varepsilon)^{-1})$. Or, un calcul explicite montre que le développement en série entière de $f(z)$ en 0 est donné par

$$f(z) = z \sum_{n \in \mathbb{N}} z^n T^n.$$

Sur l'ensemble $\mathcal{C} = \{z \in \mathbb{C};\ 1/r(T) < |z| < 1/(r(T) - \varepsilon)\}$, on obtient donc que $f(z)$ est analytique, alors que la série est divergente, d'après le lemme 2.34. On obtient donc une contradiction. □

Remarque 2.37. *Notons que $\sigma_{\mathrm{p}}(T)$ est l'ensemble des valeurs propres de T, i.e. l'ensemble des $\lambda \in \mathbb{C}$ tels qu'il existe $u \in E \setminus \{0\}$ tel que*

$$Tu = \lambda u.$$

En dimension finie, un opérateur linéaire injectif est bijectif. Ainsi,

$$\sigma(T) = \sigma_{\mathrm{p}}(T)$$

est simplement l'ensemble des valeurs propres de T dans ce cas.

Prouvons ici le lemme suivant qui sera utile par la suite.

Lemme 2.38. *Soit $T \in \mathcal{L}(E)$. Soit $(\lambda_k)_{k\geq 1}$ une suite de $\sigma_{\rm p}(T)$ de valeurs propres toutes distinctes, et soit $(u_k)_{k\geq 1}$ une suite de vecteurs propres associés. Alors les vecteurs $(u_k)_{k\geq 1}$ sont linéairement indépendants.*

Démonstration. On procède par récurrence. On suppose que les vecteurs $u_1, \dots, u_n$ sont indépendants. Si, au rang $n+1$, l'hypothèse de récurrence n'est pas vraie, alors il existe $(\alpha_k)_{1\leq k\leq n}$ tels que $u_{n+1} = \sum_{k=1}^n \alpha_k u_k$. Alors

$$Tu_{n+1} = \sum_{k=1}^n \alpha_k \lambda_k u_k = \lambda_{n+1} u_{n+1} = \lambda_{n+1} \sum_{k=1}^n \alpha_k u_k.$$

Par hypothèse de récurrence, la famille $(u_1, \dots, u_n)$ est libre, donc $\lambda_{n+1}\alpha_k = \alpha_k \lambda_k$ pour tout $1 \leq k \leq n$. Les valeurs propres étant distinctes deux à deux, on a ainsi $\alpha_k = 0$, ce qui donne $u_{n+1} = 0$, ce qui est contradictoire. On a donc démontré l'hypothèse de récurrence au rang $n+1$. □

Remarque 2.39 (Autre décomposition du spectre)**.** *Dans certains cas, il est plus commode de décomposer $\sigma(T)$ sous la forme $\sigma(T) = \sigma_{\rm d}(T) \cup \sigma_{\rm ess}(T)$, où $\sigma_{\rm d}(T) \subset \sigma_{\rm p}(T)$ est le* spectre discret*, qui est composé des valeurs propres isolées de multiplicité finie :*

$$\sigma_{\rm d}(T) = \Big\{ \lambda \in \mathbb{C} \;\Big|\; 0 < \dim(\mathrm{Ker}(\lambda - T)) < +\infty,\ \exists \varepsilon > 0,\ \ ]\lambda-\varepsilon, \lambda+\varepsilon[\cap \sigma(T) = \{\lambda\} \Big\}.$$

Donnons à présent quelques exemples de spectre résiduel et continu, afin de donner un début d'intuition sur ces notions.

Exercice 2.40 (Spectre résiduel)**.** *On considère l'opérateur de shift à droite $\tau_{\rm d}$ dans $\ell^2(\mathbb{N}, \mathbb{C})$ défini par* (2.3)*.*

1. *Vérifier que $\sigma_{\rm p}(\tau_{\rm d}) = \emptyset$ et que $\lambda - \tau_{\rm d}$ est injectif pour tout $\lambda \in \mathbb{C}$.*
2. *Montrer que $0 \in \sigma_{\rm r}(\tau_{\rm d})$.*
3. *Montrer que $\{\lambda \in \mathbb{C},\ |\lambda| < 1\} \subset \sigma_{\rm r}(\tau_{\rm d})$. Indication : considérer $x_\lambda = \left(1, \overline{\lambda}, \overline{\lambda}^2, \dots\right)$ et vérifier que $x_\lambda \in (\mathrm{Ran}(\lambda - \tau_{\rm d}))^\perp$.*

Exercice 2.41 (Spectre continu)**.** *Soit $a < b$ deux réels, $E = \mathrm{L}^2([a,b], \mathbb{C})$ et $T \in \mathcal{L}(E)$ défini par*

$$Tf(x) = x\, f(x).$$

Montrer que $\sigma(T) = \sigma_{\rm c}(T) = [a,b]$, en suivant les étapes ci-dessous :

1. *Montrer que $\sigma(T) \subset [a,b]$.*
2. *Montrer que $\sigma(T) = [a,b]$ (en supposant qu'il existe $\lambda \in [a,b]$ tel que $\lambda - T$ soit inversible, et en considérant $\varphi \in \mathbb{C}^\infty([a,b], \mathbb{C})$ valant 1 au voisinage de λ).*

3. *Montrer que* $\sigma(T) = \sigma_{\rm c}(T)$. *Pour cela, établir d'abord que* $\sigma_{\rm p}(T) = \emptyset$, *puis prouver que* $\overline{\mathrm{Ran}(\lambda - T)} = E$ *pour tout* $\lambda \in [a, b]$. *Pour ce dernier point, pour* $f \in E$ *donnée, considérer la suite* $(\varphi_n)_{n\geq 1}$ *de* E *définie par*

$$\varphi_n(x) = \begin{cases} \dfrac{f(x)}{\lambda - x} & \text{si } |x - \lambda| \geq \dfrac{1}{n} \text{ et } x \in [a, b], \\ 0 & \text{sinon.} \end{cases}$$

Exercice 2.42. *Le schéma de preuve ci-dessus montre qu'on peut en fait étendre l'argument à des opérateurs plus généraux, définis sur* $E = \mathrm{L}^2(\mathbb{R}^d, \mathbb{C})$, *de domaine*

$$D(T) = \left\{ f \in E \;\middle|\; \int_{\mathbb{R}^d} (1 + |V(x)|^2)|f(x)|^2 \, dx < +\infty \right\},$$

pour $V \in \mathrm{L}^\infty_{\rm loc}(\mathbb{R}, \mathbb{R})$, *et d'action*

$$Tf(x) = V(x)\, f(x).$$

Ainsi, pour une fonction V *continue, on montrera que* $\sigma(T) = [\min V, \max V]$, *et que* $\sigma(T) = \sigma_{\rm c}(T)$ *si* $V^{-1}(\{\lambda\})$ *est un ensemble au plus dénombrable sans point d'accumulation pour tout* $\lambda \in \mathbb{R}$. *Notons en revanche que si on considère une fonction* $V \in \mathbb{C}^\infty(\mathbb{R})$, *valant* $c \in \mathbb{R}$ *dans un voisinage* $]-\eta, \eta[$ *de l'origine, alors* $c \in \sigma_{\rm p}(T)$.

2.2.2 Cas des opérateurs bornés autoadjoints

Les opérateurs bornés auto-adjoints ont des propriétés intéressantes, qui se traduisent sur leur spectre.

Proposition 2.43. *Soit* H *un espace de Hilbert et* $T \in \mathcal{L}(H)$. *Si* T *est auto-adjoint, on a*

$$\sigma(T) \subset \mathbb{R}.$$

De plus, $r(T) = \|T\|$, $\sigma(T) \subset [-\|T\|, \|T\|]$ *et l'une au moins des deux extrémités du segment est dans* $\sigma(T)$. *Enfin,* $\sigma_{\rm r}(T) = \emptyset$ *et les vecteurs propres associés à des éléments différents de* $\sigma_{\rm p}(T)$ *sont orthogonaux.*

Démonstration. Pour prouver ce résultat, nous allons établir plusieurs résultats intermédiaires.

— Commençons par montrer que si $\lambda \in \mathbb{C}$ est tel que $\alpha = |\mathrm{Im}(\lambda)| \neq 0$, alors $\lambda - T$ est inversible.

Montrons tout d'abord que l'opérateur $\lambda - T$ est injectif. En effet, pour tout $x \in H$, on a

$$\langle (\lambda - T)x, x\rangle = -\langle Tx, x\rangle + \mathrm{Re}(\lambda)\, \langle x, x\rangle - \mathrm{i}\, \mathrm{Im}(\lambda)\, \langle x, x\rangle.$$

On voit que $\langle Tx, x\rangle = \overline{\langle x, Tx\rangle} = \overline{\langle T^*x, x\rangle} = \overline{\langle Tx, x\rangle}$. Donc $\langle Tx, x\rangle$ est réel. Il en résulte que

$$|\langle(\lambda - T)x, x\rangle| \geq \alpha\|x\|^2. \tag{2.12}$$

On en déduit par l'inégalité de Cauchy-Schwarz que

$$\|(\lambda - T)x\| \geq \alpha\|x\|. \tag{2.13}$$

Cette inégalité implique que l'opérateur $\lambda - T$ est injectif.

Montrons ensuite que l'opérateur $\lambda - T$ est surjectif. Soit $V = \mathrm{Ran}(\lambda - T)$. Nous allons montrer que $V = H$. Pour cela, montrons tout d'abord que V est fermé dans H. Soit $w_n = (\lambda - T)v_n$ une suite dans V qui converge vers $w \in H$. En utilisant (2.13), on obtient

$$\|w_p - w_q\| \geq \alpha\|v_p - v_q\|.$$

La suite $(w_n)_{n\geq 0}$ est de Cauchy, donc la suite $(v_n)_{n\geq 0}$ aussi. Elle converge donc vers un certain $v \in H$. Par continuité de l'application T,

$$w_n = (\lambda - T)v_n \longrightarrow (\lambda - T)v$$

dans H. Donc $w = (\lambda - T)v$, ce qui prouve que $w \in V$. Donc V est fermé dans H. Montrons enfin que V est dense. Une technique standard pour montrer cela est de prouver que $V^\perp = \{0\}$ (ce qui donne, grace au lemme 1.13, que $\overline{V} = (V^\perp)^\perp = H$). Soit donc $w \in V^\perp$. Pour tout $v \in H$, on a alors

$$\langle(\lambda - T)v, w\rangle = 0.$$

En particulier, pour $v = w$,

$$\langle(\lambda - T)w, w\rangle = 0.$$

En utilisant (2.12), on obtient $w = 0$, ce qui montre que $V^\perp = \{0\}$ d'où la densité de V dans H. Comme V est dense dans H et fermé dans H, on en déduit que $V = H$, et donc la surjectivité de $\lambda - T$.
Comme l'opérateur $\lambda - T \in \mathcal{L}(H)$ est bijectif, il est inversible (cf. la proposition 2.24). Noter également que l'inégalité (2.13) donne la borne suivante sur la résolvante :

$$\|(\lambda - T)^{-1}\| \leq \frac{1}{|\mathrm{Im}(\lambda)|}.$$

On a donc démontré que $\sigma(T) \subset \mathbb{R}$.

— Le théorème 2.36 implique alors que

$$\sigma(T) \subset \overline{D(0, r(T))} \cap \mathbb{R} = [-r(T), r(T)]$$

et que

$$\sigma(T) \cap C(0, r(T)) = \sigma(T) \cap C(0, r(T)) \cap \mathbb{R} = \sigma(T) \cap \{-r(T), r(T)\} \neq \emptyset.$$

— Nous allons maintenant prouver que $r(T) = \|T\|$. Tout d'abord, notons que $\|T^*T\| \leq \|T\| \, \|T^*\| = \|T\|^2$. Par ailleurs, comme $|\langle x, T^*Tx\rangle| \leq \|T^*T\| \, \|x\|^2$, on a

$$\|T^*T\| \geq \sup_{\|x\|=1} |\langle x, T^*Tx\rangle| = \sup_{\|x\|=1} \|Tx\|^2 = \left(\sup_{\|x\|=1} \|Tx\| \right)^2 = \|T\|^2,$$

ce qui montre que $\|T^2\| = \|T^*T\| = \|T\|^2$. Par récurrence, on a ensuite $\|T^{2^p}\| = \|T\|^{2^p}$. Pour $n \in \mathbb{N}$ quelconque, on considère p tel que $n \leq 2^p$ et on écrit

$$\|T\|^{2^p} = \|T^{2^p}\| \leq \|T^n\| \, \|T^{2^p-n}\| \leq \|T^n\| \, \|T\|^{2^p-n}.$$

Ceci montre que $\|T\|^n \leq \|T^n\|$. L'inégalité contraire étant par ailleurs toujours satisfaite, on en déduit que $\|T\|^n = \|T^n\|$, et donc $\|T^n\|^{1/n} = \|T\|$ pour tout $n \geq 1$. On a donc finalement $r(T) = \lim_{n\to\infty} \|T^n\|^{1/n} = \|T\|$.

— Montrons maintenant que $\sigma_{\mathrm{r}}(T) = \emptyset$. Pour ce faire, on considère $\lambda \in \sigma(T) \subset \mathbb{R}$ tel que $\mathrm{Ker}(\lambda - T) = \{0\}$. On a vu (cf. la proposition 2.31) que

$$\Big(\mathrm{Ran}(\lambda - T)\Big)^{\perp} = \mathrm{Ker}(\overline{\lambda} - T^*).$$

Dans le cas présent, ceci implique que $\Big(\mathrm{Ran}(\lambda - T)\Big)^{\perp} = \mathrm{Ker}(\lambda - T) = \{0\}$, ce qui implique (cf. le lemme 1.13) que signifie que $\overline{\mathrm{Ran}(\lambda - T)} = H$ et donc $\lambda \notin \sigma_{\mathrm{r}}(T)$.

— Enfin, soient u et v deux vecteurs propres associés respectivement à deux éléments $\lambda \neq \mu$ de $\sigma_{\mathrm{p}}(T)$. Alors,

$$\lambda\langle u, v\rangle = \langle Tu, v\rangle = \langle u, Tv\rangle = \mu\langle u, v\rangle.$$

Ceci montre que $\langle u, v\rangle = 0$.

□

Remarque 2.44. *On fait ici le lien entre le spectre résiduel d'un opérateur et le spectre ponctuel de son adjoint.*

La relation (2.7) *montre de manière générale que, pour un opérateur borné $T \in \mathcal{L}(E)$, si $\lambda \in \sigma_{\mathrm{r}}(T)$, alors $\overline{\lambda} \in \sigma_{\mathrm{p}}(T^*)$. Bien sûr, dans le cas des opérateurs autoadjoints, on a $T^* = T$ et donc $\lambda \in \sigma_{\mathrm{r}}(T) \cap \sigma_{\mathrm{p}}(T) = \emptyset$ par définition des différentes parties du spectre. Ceci montre bien que $\sigma_{\mathrm{r}}(T) = \emptyset$ pour des opérateurs autoadjoints.*

Par ailleurs, on peut montrer que, si $\lambda \in \sigma_{\mathrm{p}}(T)$, alors $\overline{\lambda} \in \sigma_{\mathrm{p}}(T^*) \cup \sigma_{\mathrm{r}}(T^*)$.

Exercice 2.45. *Donner un exemple d'opérateur borné tel que $\overline{\lambda} \in \sigma_{\mathrm{p}}(T^*)$ lorsque $\lambda \in \sigma_{\mathrm{p}}(T)$, et un exemple d'opérateur borné tel que $\overline{\lambda} \in \sigma_{\mathrm{r}}(T^*)$ lorsque $\lambda \in \sigma_{\mathrm{p}}(T)$.*

Exercice 2.46. *Soit V un espace de Hilbert et soit $T \in \mathcal{L}(V)$ un opérateur borné auto-adjoint. On suppose que $\langle Tu, u\rangle = 0$ pour tout $u \in V$. Montrer qu'alors $T = 0$.*

2.2.3 Invariance par transformation unitaire

Dans certains cas, il est plus facile d'étudier le spectre d'un opérateur UTU^{-1} que le spectre de l'opérateur T directement. Les deux opérateurs ci-dessus ont le même spectre sous certaines conditions sur la transformation U.

Définition 2.47. *Soit H un espace de Hilbert. Un opérateur $U \in \mathcal{L}(H)$ est une isométrie si $\|Ux\|_H = \|x\|_H$ pour tout $x \in H$, ou, de manière équivalente, si $\langle Ux, Uy\rangle = \langle x, y\rangle$ pour tout x et y dans H.*

L'équivalence est obtenue en développant la quantité $\|Ux + \lambda Uy\|_H^2$.

Notons qu'une isométrie est telle que $U^*U = \mathrm{Id}$, et est également une application injective. Par exemple, l'opérateur de shift à droite (2.3) est une isométrie. Cependant, une isométrie n'est pas nécessairement une bijection, ce qui motive la définition suivante.

Définition 2.48. *Soit H un espace de Hilbert. Un opérateur $U \in \mathcal{L}(H)$ est* unitaire *si U est une isométrie et* $\mathrm{Ran}(U) = H$.

Un opérateur unitaire est donc borné, et injectif et surjectif donc bijectif. Ceci implique que U^{-1} existe et est borné (cf. la proposition 2.24), et donc un opérateur unitaire est inversible. On a par ailleurs $U^{-1} = U^*$.

Exemple 2.49. *La transformée de Fourier est une transformation unitaire de $\mathrm{L}^2(\mathbb{R}^d)$ si on la normalise correctement. En effet, définissons $U : \mathrm{L}^1(\mathbb{R}^d) \to \mathrm{L}^\infty(\mathbb{R}^d)$ par*

$$Uf(k) = \widehat{f}(k) = \left(\frac{1}{2\pi}\right)^{d/2} \int_{\mathbb{R}^d} f(x)\, \mathrm{e}^{-\mathrm{i}k\cdot x}\, dx.$$

Il est possible (mais pas évident !) d'étendre la notion de transformée de Fourier à des fonctions plus générales, et en particulier à des fonctions dans $\mathrm{L}^2(\mathbb{R}^d)$. La construction de Uf pour $f \in \mathrm{L}^2(\mathbb{R}^d)$ montre que $\|Uf\|_{\mathrm{L}^2} = \|f\|_{\mathrm{L}^2}$ (c'est la formule de Parseval). Ceci donne que U est borné, et que U est une isométrie. On peut de plus montrer que U est surjectif sur $\mathrm{L}^2(\mathbb{R}^d)$. Donc U est unitaire.

Exercice 2.50. *Montrer que, pour tout $\eta > 0$, l'opérateur $U_\eta : \mathrm{L}^2(\mathbb{R}^d) \to \mathrm{L}^2(\mathbb{R}^d)$ défini par $U_\eta f(x) = \eta^{d/2} f(\eta x)$ est une isométrie. Est-il unitaire ?*

Exercice 2.51. *Montrer que, pour tout $a \in \mathbb{R}^n$, l'opérateur de translation $\tau_a : \mathrm{L}^2(\mathbb{R}^d) \to \mathrm{L}^2(\mathbb{R}^d)$ défini par $\tau_a f(x) = f(x - a)$ est un opérateur unitaire.*

Proposition 2.52. *Soit H un espace de Hilbert, T un opérateur borné et U un opérateur unitaire. On considère l'opérateur (borné)*

$$T_U = UTU^{-1} = UTU^*.$$

Alors $\sigma(T_U) = \sigma(T)$. Si T est autoadjoint, alors T_U l'est aussi.

La preuve de cette proposition est très simple, et repose sur le fait que si $\lambda \in \rho(T)$, alors $(\lambda - T_U)^{-1} = U(\lambda - T)^{-1}U^{-1}$, ce qui montre que $\lambda \in \rho(T_U)$. On montre de même l'implication contraire.

2.3 Opérateurs compacts

2.3.1 Définition et premières propriétés

Définition 2.53. *Soient E et F deux espaces de Banach et T un opérateur linéaire de E dans F. On dit que l'opérateur T est* compact *si, pour tout $B \subset E$,*

$$B \text{ borné dans } E \quad \Rightarrow \quad T(B) \text{ relativement compact dans } F.$$

On note $\mathcal{K}(E,F)$ l'ensemble des opérateurs compacts de E dans F.

Ainsi, un opérateur compact transforme une suite bornée en une suite convergente (à extraction près).

Proposition 2.54. *Tout opérateur linéaire compact est continu, i.e. $\mathcal{K}(E,F) \subset \mathcal{L}(E,F)$.*

Démonstration. Soit E et F deux espaces de Banach et T un opérateur linéaire compact de E dans F. Soit $\overline{B}_1 = \{x \in E,\ \|x\| \leq 1\}$ la boule unité fermée de E. L'ensemble $\overline{B}_1$ étant borné, son image par T est relativement compacte donc bornée : il existe une constante C telle que

$$\forall x \in \overline{B}_1, \qquad \|Tx\|_F \leq C.$$

On en déduit que

$$\forall x \in E \setminus \{0\}, \qquad \|Tx\|_F = \|x\|_E \left\| T\left(\frac{x}{\|x\|_E}\right)\right\|_F \leq C\|x\|_E.$$

L'opérateur linéaire T est donc continu. □

Nous avons la caractérisation équivalente suivante des opérateurs compacts dans le cas où les espaces E et F sont des opérateurs de Hilbert.

Proposition 2.55. *Soit E et F deux espaces de Hilbert. Soit $T \in \mathcal{L}(E,F)$. Alors les deux propositions suivantes sont équivalentes :*

(i) $T \in \mathcal{K}(E,F)$;

(ii) Pour toute suite $(u_n)_{n\in\mathbb{N}}$ qui converge faiblement vers u dans E, on peut extraire une sous-suite de la suite $(Tu_n)_{n\in\mathbb{N}}$ qui converge fortement vers Tu dans F.

Démonstration. On démontre l'implication (i) ⇒ (ii). Soit $T \in \mathcal{K}(E,F)$. Soit $(u_n)_{n\in\mathbb{N}}$ une suite d'éléments de E qui converge faiblement vers un élément $u \in E$. On utilise la Proposition 2.19 : comme T est un opérateur continu, la suite $(Tu_n)_{n\in\mathbb{N}}$ converge faiblement vers Tu dans F. Par ailleurs, la suite $(u_n)_{n\in\mathbb{N}}$ est bornée, et T est compact, donc on peut extraire une sous-suite de $(Tu_n)_{n\in\mathbb{N}}$ qui converge fortement vers un élément $w \in F$. Comme la convergence forte implique la convergence faible, par unicité de la limite, on a nécessairement $w = Tu$.

On prouve maintenant l'implication (ii) $\Rightarrow$ (i). Soit $T \in \mathcal{L}(E,F)$ qui vérifie la propriété (ii). Montrons que T est compact. Soit B un sous-ensemble borné de E. Montrons que $T(B)$ est un ensemble relativement compact dans F. Soit $(u_n)_{n\in\mathbb{N}}$ une suite d'éléments de B. Comme B est borné, la suite $(u_n)_{n\in\mathbb{N}}$ l'est aussi, et on peut donc en extraire une sous-suite qui converge faiblement dans E vers un élément $u \in E$. D'après la caractérisation (ii), il existe une extraction φ telle que la suite $(Tu_{\varphi(n)})_{n\in\mathbb{N}}$ converge fortement dans F vers Tu. Ceci montre qu'il existe une sous-suite de $(Tu_n)_{n\in\mathbb{N}}$ qui converge fortement dans F. L'ensemble $T(B)$ est donc relativement compact. □

Exercice 2.56. *Montrer que les opérateurs suivants sont compacts :*

1. *l'identité de E est compacte si et seulement si E est de dimension finie ;*
2. *si l'un des espaces E ou F est de dimension finie, alors tout opérateur linéaire continu T de E dans F est compact (en particulier, si $T \in \mathcal{L}(E,F)$ avec $\mathrm{Ran}(T)$ de dimension finie, alors $T \in \mathcal{K}(E,F)$) ;*
3. *si T_1 et T_2 sont deux opérateurs linéaires compacts de E dans F, alors T_1+T_2 est un opérateur compact ;*
4. *la restriction d'un opérateur compact $T \in \mathcal{K}(E,F)$ à un sous-espace vectoriel $\widetilde{E}$ de E est compacte.*

Exercice 2.57. *On considère l'opérateur de l'Exercice 2.15. Montrer que $K \in \mathcal{K}(E,F)$ en admettant le résultat de compacité suivant, connu sous le nom de lemme d'Ascoli :*

Soit $\mathcal{F}$ un sous-ensemble borné de $F = C^0([0,1],\mathbb{R})$ tel que la propriété d'équicontinuité suivante soit satisfaite : pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que

$$|x - x'| \leq \delta \;\Rightarrow\; \forall u \in \mathcal{F},\ |u(x) - u(x')| \leq \varepsilon.$$

Alors $\mathcal{F}$ est relativement compact dans F.

Théorème 2.58. *Soit E et F deux espaces de Banach. L'ensemble $\mathcal{K}(E,F)$ est un sous-espace vectoriel fermé de l'espace vectoriel $\mathcal{L}(E,F)$.*

Démonstration. Il est facile de montrer que $\mathcal{K}(E,F)$ est un espace vectoriel. Grace à la Proposition 2.54, on sait qu'il est inclus dans $\mathcal{L}(E,F)$. Il reste à prouver que c'est un sous-espace fermé de $\mathcal{L}(E,F)$. Considérons pour cela une suite d'opérateurs compacts $(T_k)_{k\in\mathbb{N}^*}$ qui converge dans $\mathcal{L}(E,F)$ vers un opérateur $T \in \mathcal{L}(E,F)$ et montrons que T est compact. Soit B un borné de E, soit $R > 0$ un réel tel que $B \subset \{x \in E,\ \|x\| \leq R\}$ et soit $(u_n)_{n\in\mathbb{N}}$ une suite de $T(B)$. Il faut montrer que on peut extraire de $(u_n)_{n\in\mathbb{N}}$ une sous-suite convergente (ceci prouvera que $T(B)$ est relativement compact et donc que T est compact).

Soit $(w_n)_{n\in\mathbb{N}}$ une suite d'éléments de B tels que pour tout $n \in \mathbb{N}$, $T(w_n) = u_n$. On va extraire de $(u_n)_{n\in\mathbb{N}}$ une sous-suite convergente en utilisant un procédé diagonal.

On pose $\{w_n^0\}_n = \{w_n\}_n$ et on construit, par récurrence sur k, la suite $\{w_n^k\}_n$, qui est une sous-suite de $(w_n^{k-1})_{n\in\mathbb{N}}$ telle que $(T_k(w_n^k))_{n\in\mathbb{N}}$ soit convergente. On utilise pour cela le fait que T_k est un opérateur compact, et que $\{w_n^{k-1}\}_n$, suite extraite de $(w_n)_n$, est bornée. On définit maintenant la suite $(v_n)_{n\in\mathbb{N}}$ par $v_n = w_n^n$. Pour tout $k \in \mathbb{N}^*$, $(v_n)_{n\geq k}$ est une sous-suite de $(w_n^k)_{n\in\mathbb{N}}$: la suite $(T_k(v_n))_{n\in\mathbb{N}}$ est donc convergente.

On pose $\widetilde{u}_n = T(v_n)$. La suite $(\widetilde{u}_n)_{n\in\mathbb{N}}$ est une sous-suite de $(u_n)_{n\in\mathbb{N}}$. On va montrer qu'elle est de Cauchy. Soit $\varepsilon > 0$ et $k \in \mathbb{N}^*$ tel que

$$\|T - T_k\|_{\mathcal{L}(E,F)} \leq \frac{\varepsilon}{3R}.$$

Soit ensuite $N \geq 0$ tel que $\forall q > p \geq N$,

$$\|T_k(v_p) - T_k(v_q)\|_F \leq \frac{\varepsilon}{3}.$$

Il vient que, pour tout $q > p \geq N$,

$$\begin{aligned}
\|\widetilde{u}_p - \widetilde{u}_q\| &= \|T(v_p) - T(v_q)\|_F \\
&\leq \|T(v_p) - T_k(v_p)\|_F + \|T_k(v_p) - T_k(v_q)\|_F + \|T_k(v_q) - T(v_q)\|_F \\
&\leq \|T - T_k\|_{\mathcal{L}(E,F)} \left(\|v_p\|_E + \|v_q\|_E\right) + \|T_k(v_p) - T_k(v_q)\|_F \\
&\leq \varepsilon.
\end{aligned}$$

La suite $(\widetilde{u}_n)_{n\in\mathbb{N}}$ est donc de Cauchy. Ceci conclut la preuve. □

Une des conséquences importantes de ce résultat est que, si T est la limite d'une suite d'opérateurs $(T_n)_{n\geq 0}$ de rang fini (*i.e.* tels que la dimension de $\mathrm{Ran}(T_n)$ est finie), au sens où

$$\|T_n - T\| \longrightarrow 0$$

où la norme est définie en (2.1), alors l'opérateur limite T est compact. En général, la réciproque est fausse : on ne peut pas approcher n'importe quel opérateur compact par une suite d'opérateurs de rang fini. Cette réciproque est cependant vraie si on considère $\mathcal{K}(E,F)$ avec F un espace de Hilbert (cf. [3, Section VI.1] ou la Remarque 2.73 pour le cas où $E = F$ est un espace de Hilbert).

Proposition 2.59. *Soient E, F et G trois espaces de Banach, et soient $T_1 \in \mathcal{L}(E,F)$ et $T_2 \in \mathcal{L}(F,G)$.*

Si T_1 est compact, ou bien si T_2 est compact, alors l'application $T_2 \circ T_1$ est compacte : $T_2 \circ T_1 \in \mathcal{K}(E,G)$.

Démonstration. On suppose que $T_1 \in \mathcal{L}(E,F)$ et $T_2 \in \mathcal{K}(F,G)$. Comme T_1 est continue, l'image par T_1 de la boule unité de E, qu'on note $T_1(B_E)$, est bornée. Comme T_2 est linéaire compacte, l'image par T_2 d'un ensemble borné est relativement compacte dans G. Donc $T_2 \circ T_1(B_E)$ est relativement compacte dans G, et $T_2 \circ T_1$ est une application compacte.

Supposons maintenant que $T_1 \in \mathcal{K}(E,F)$ et $T_2 \in \mathcal{L}(F,G)$. Soit $w_n = T_2 \circ T_1(u_n)$ une suite d'éléments de $T_2 \circ T_1(B_E)$, avec $u_n \in B_E$. On pose $v_n = T_1(u_n) \in F$. Comme T_1 est compacte, on peut extraire de v_n une sous-suite convergente dans F, qu'on note $v_{\varphi(n)}$, avec $\lim_{n\to\infty} v_{\varphi(n)} = v$. Par conséquent, comme T_2 est continue, on a

$$\lim_{n\to\infty} w_{\varphi(n)} = \lim_{n\to\infty} T_2(v_{\varphi(n)}) = T_2(v).$$

On peut donc extraire de toute suite de $T_2 \circ T_1(B_E)$ une sous-suite convergente : donc $T_2 \circ T_1$ est une application compacte. □

Concluons enfin avec quelques exercices d'application.

Exercice 2.60 (Opérateurs de Hilbert-Schmidt)**.** *Montrer que l'opérateur $\widehat{K}$ de l'Exercice 2.29 est compact.*

Exercice 2.61. *Soit V un espace de Hilbert de dimension infinie. Montrer que, si $A \in \mathcal{K}(V,V)$, alors A n'est pas bijectif.*

Exercice 2.62. *Soit $u = (u_i)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ une suite à valeur réelle. On considère l'ensemble $\ell_2 = \{u \in \mathbb{R}^{\mathbb{N}};\ \sum_{i\geq 0} u_i^2 < +\infty\}$ des suites de carré sommable, qu'on munit du produit scalaire $\langle u, v\rangle = \sum_{i\geq 0} u_i v_i$.*

Soit $(a_i)_{i\geq 0}$ une suite de réels bornés : $|a_i| \leq C < +\infty$ pour tout $i \geq 0$. On définit l'application linéaire A sur ℓ_2 par $Au = (a_i u_i)_{i\geq 0}$. Montrer que $Au \in \ell_2$ et que A est continue. Montrer que A est compacte si et seulement si $\lim_{i\to+\infty} a_i = 0$ (Indication : pour montrer que $\lim_{i\to+\infty} a_i = 0$ implique A est compacte, on pourra utiliser un principe d'extraction diagonale).

Proposition 2.63. *Soit V un espace de Hilbert et $A \in \mathcal{K}(V,V)$. Alors* $\mathrm{Ker}(\mathrm{Id} - A)$ *est de dimension finie.*

Démonstration. Soit $E_1 = \mathrm{Ker}(\mathrm{Id} - A)$. Montrons que la boule unité fermée de E_1 est compacte. Soit $v \in \mathrm{Ker}(\mathrm{Id} - A)$ avec $\|v\| \leq 1$: on a donc $v = Av$, donc $v \in A(B_V)$, et ainsi $B_{E_1} \subset A(B_V)$. Comme A est compacte, $A(B_V)$ est relativement compacte, et donc B_{E_1} est relativement compact. Comme B_{E_1} est fermée, on a donc que B_{E_1} est compacte. En application de la proposition 1.31, on a donc que E_1 est de dimension finie. □

2.3.2 Le théorème de Rellich

Définition 2.64. *Soient V et H deux espaces de Hilbert avec $V \subset H$. On note respectivement $\langle\cdot,\cdot\rangle_V$ et $\langle\cdot,\cdot\rangle_H$ leur produit scalaire. On dit que l'injection $V \subset H$ est compacte si l'application*

$$\begin{array}{rccc} \mathcal{I} : & V & \longrightarrow & H \\ & u & \longmapsto & u \end{array}$$

est continue et compacte, autrement dit :

— *il existe C tel que, pour tout $u \in V$, on a $\|u\|_H \leq C\ \|u\|_V$;*
— *de toute suite bornée de V (pour la norme $\|\cdot\|_V$), on peut extraire une sous-suite convergente dans H (pour la norme $\|\cdot\|_H$).*

On va à présent énoncer un résultat de compacité important (et très utile dans l'étude des équations aux dérivées partielles).

Théorème 2.65. *Soit Ω un ouvert* borné *de $\mathbb{R}^d$. L'injection canonique de $\mathrm{H}^1_0(\Omega)$ dans $\mathrm{L}^2(\Omega)$ est compacte.*

Un des intérêts de ce résultat est que, si on arrive à obtenir une borne (en norme $\mathrm{H}^1(\Omega)$) sur une suite de fonctions approchant la solution d'une équation (par exemple, en montrant qu'une énergie est uniformément bornée), alors on peut extraire de cette suite une sous-suite convergente (en norme $\mathrm{L}^2(\Omega)$). Cette limite est alors un candidat naturel pour être une solution de l'équation.

Dans ce chapitre, ce résultat va nous permettre de montrer que les inverses de certains opérateurs sont compacts, ce qui permettra de décrire complètement le spectre de l'opérateur en question.

Démonstration. La preuve comprend trois étapes.
— On commence par traiter le cas où $\Omega =]0, \pi[$. On note $e_k(x) = \sqrt{2/\pi}\,\sin(kx)$ le k-ième mode de Fourier valant 0 au bord de Ω. On note que $e_k \in \mathrm{H}^1_0(0,\pi)$, $\|e_k\|_{\mathrm{L}^2} = 1$ et $\|e_k\|^2_{\mathrm{H}^1} = 1 + k^2$.
En utilisant la transformée de Fourier, on peut montrer (et ce sera admis ici) qu'on peut caractériser les espaces $\mathrm{L}^2(0,\pi)$ et $\mathrm{H}^1_0(0,\pi)$ par

$$\mathrm{L}^2(0,\pi) = \left\{ u(x) = \sum_{k=1}^{+\infty} c_k\, e_k(x), \quad \sum_{k=1}^{+\infty} |c_k|^2 < +\infty \right\}$$

et

$$\mathrm{H}^1_0(0,\pi) = \left\{ u(x) = \sum_{k=1}^{+\infty} c_k\, e_k(x), \quad \sum_{k=1}^{+\infty} (1+k^2)|c_k|^2 < +\infty \right\}.$$

De plus,

$$\|u\|_{\mathrm{L}^2} = \left(\sum_{k=1}^{+\infty} |c_k|^2 \right)^{1/2}, \qquad \|u\|_{\mathrm{H}^1} = \left(\sum_{k=1}^{+\infty} (1+k^2)|c_k|^2 \right)^{1/2}.$$

On note que, pour montrer la complétude de la base des $\{e_k\}_{k\geq 1}$ dans $\mathrm{L}^2(0,\pi)$, il suffit de prendre une fonction de $\mathrm{L}^2(0,\pi)$, de l'antisymétriser pour en faire une fonction sur $]-\pi,\pi[$, d'étendre la fonction à tout $\mathbb{R}$ en la périodisant, et enfin de développer cette fonction sur la base des sinus et cosinus (en utilisant la théorie des séries de Fourier).

Soit

$$\begin{array}{rccc} I & : & \mathrm{H}^1_0(0,\pi) & \longrightarrow & \mathrm{L}^2(0,\pi) \\ & & u & \mapsto & u \end{array}$$

l'injection canonique de $\mathrm{H}^1_0(0,\pi)$ dans $\mathrm{L}^2(0,\pi)$. Pour tout $N \in \mathbb{N}^*$, soit I_N l'opérateur linéaire défini par

$$\begin{array}{rccc} I_N & : & \mathrm{H}^1_0(0,\pi) & \longrightarrow & \mathrm{L}^2(0,\pi) \\ & & u = \displaystyle\sum_{k=1}^{+\infty} c_k e_k & \mapsto & I_N(u) = \displaystyle\sum_{k=1}^{N} c_k e_k. \end{array}$$

Montrons que la suite $(I_N)_{N\in\mathbb{N}^*}$ converge vers I dans $\mathcal{L}(\mathrm{H}^1_0, \mathrm{L}^2)$. On calcule

$$\begin{aligned} \|I - I_N\|^2_{\mathcal{L}(\mathrm{H}^1_0,\mathrm{L}^2)} &= \sup_{u\in\mathrm{H}^1_0(\Omega),\ u\neq 0} \frac{\|(I-I_N)(u)\|^2_{\mathrm{L}^2}}{\|u\|^2_{\mathrm{H}^1_0}} \\ &= \sup_{(c_k)_{k\in\mathbb{N}^*}\neq 0,\ \sum(1+k^2)|c_k|^2<+\infty} \frac{\displaystyle\sum_{k=N+1}^{+\infty} |c_k|^2}{\displaystyle\sum_{k=1}^{+\infty}(1+k^2)|c_k|^2} \\ &\leq \sup_{(c_k)_{k\in\mathbb{N}^*}\neq 0,\ \sum(1+k^2)|c_k|^2<+\infty} \frac{\displaystyle\sum_{k=N+1}^{+\infty} |c_k|^2}{\displaystyle\sum_{k=N+1}^{+\infty}(1+k^2)|c_k|^2} \\ &\leq \frac{1}{1+(N+1)^2} \underset{N\to+\infty}{\longrightarrow} 0. \end{aligned}$$

Par ailleurs, pour tout $N \in \mathbb{N}^*$, l'opérateur I_N est de rang fini (égal à N). C'est donc un opérateur compact. Il en résulte que I est limite dans $\mathcal{L}(\mathrm{H}^1_0, \mathrm{L}^2)$ d'opérateurs compacts. C'est donc lui-même un opérateur compact d'après le Théorème 2.58.

— Pour $\Omega =]0,\pi[^d$, on montre de la même manière que l'injection canonique de $\mathrm{H}^1_0(\Omega)$ dans $\mathrm{L}^2(\Omega)$ est compacte. Il suffit de développer les fonctions $u \in \mathrm{H}^1_0(\Omega)$ dans la base tensorielle de Fourier :

$$u(x_1, x_2, \cdots, x_d) = \sum_{k_1,k_2,\cdots,k_d=1}^{+\infty} c_{k_1k_2\cdots k_d} \sin(k_1x_1)\sin(k_2x_2)\cdots\sin(k_dx_d).$$

— Enfin, si Ω est un ouvert borné quelconque de $\mathbb{R}^d$, on peut se ramener par homothétie et translation au cas où $\Omega \subset \omega =]0,\pi[^d$. Il suffit alors de remarquer que l'injection I_Ω de $\mathrm{H}^1_0(\Omega)$ dans $\mathrm{L}^2(\Omega)$ peut se décomposer en

$$I_\Omega : \mathrm{H}^1_0(\Omega) \xrightarrow{p} \mathrm{H}^1_0(\omega) \overset{I_\omega}{\hookrightarrow} \mathrm{L}^2(\omega) \xrightarrow{r} \mathrm{L}^2(\Omega)$$

où p désigne l'opérateur linéaire qui transforme une fonction de $\mathrm{H}^1_0(\Omega)$ en une fonction de $\mathrm{H}^1_0(\omega)$ en la prolongeant par 0 dans $\omega \setminus \Omega$, I_ω est l'injection canonique de $\mathrm{H}^1_0(\omega)$ dans $\mathrm{L}^2(\omega)$ et r est l'opérateur de restriction qui à $u \in \mathrm{L}^2(\omega)$ associe la fonction $u|_\Omega$ (qui est dans $\mathrm{L}^2(\Omega)$). Comme p et r sont des opérateurs continus et I_ω est un opérateur compact, il en résulte (cf. la proposition 2.59) que I_Ω est lui-même un opérateur compact.

Ceci conclut la preuve. □

Remarque 2.66 (Injection compacte de $\mathrm{H}^1(\Omega)$ dans $\mathrm{L}^2(\Omega)$)**.** *Une modification de la preuve ci-dessus permet de montrer facilement que l'injection de* $\mathrm{H}^1(\Omega)$ *dans* $\mathrm{L}^2(\Omega)$ *est compacte lorsque le domaine* Ω *est un parallélépipède* $\Omega = \prod_{i=1}^d]a_i, b_i[$*. Pour des domaines généraux, la question est plus difficile. Ce qui pose problème dans la preuve ci-dessus, c'est de montrer que l'opérateur d'extension (celui qui à une fonction* $f \in \mathrm{H}^1(\Omega)$ *associe une fonction* $\widetilde{f} \in \mathrm{H}^1(\omega)$ *où* ω *est un cube contenant* Ω *et* $\widetilde{f}\big|_\Omega = f$*) est bien défini et est borné. De tels résultats existent pour des domaines bornés* réguliers*, voir par exemple [6, Théorème 7.1.7] et [3, Théorème IX.7] et les résultats ci-dessous.*

On a le résultat suivant :

Théorème 2.67 (de Rellich-Kondrachov)**.** *Soit* Ω *ouvert* régulier *borné de* $\mathbb{R}^d$*. On a les injections compactes :*

- *si* $d > 2$*, alors* $H^1(\Omega) \subset L^q(\Omega)$ *pour tout* $q \in [1, p^*[$*, avec* $1/p^* = 1/2 - 1/d$*.*
- *si* $d = 2$*, alors* $H^1(\Omega) \subset L^q(\Omega)$ *pour tout* $q \in [1, +\infty[$*.*
- *si* $d = 1$*, alors* $H^1(\Omega) \subset C^0(\overline{\Omega})$*.*

On en déduit en particulier le résultat suivant.

Corollaire 2.68. *Soit* Ω *un ouvert régulier borné de* $\mathbb{R}^d$*. Alors l'injection* $H^1(\Omega) \subset L^2(\Omega)$ *est compacte.*

Donc, si Ω est un ouvert régulier borné, alors, de toute suite bornée de $H^1(\Omega)$, on peut extraire une sous-suite convergente dans $L^2(\Omega)$.

Démontration du Corollaire 2.68. Si $d \geq 2$, le résultat découle directement du théorème de Rellich-Kondrachov. Si $d = 1$, on remarque que l'injection $I : H^1(\Omega) \hookrightarrow L^2(\Omega)$ est la composition de deux injections

$$I_1 : H^1(\Omega) \hookrightarrow C^0(\overline{\Omega}) \quad \text{ et } I_2 : C^0(\overline{\Omega}) \hookrightarrow L^2(\Omega).$$

L'injection I_1 est compacte d'après le théorème de Rellich-Kondrachov, et l'injection I_2 est continue. L'injection $I = I_1 \circ I_2$ est donc compacte. □

Le corollaire suivant est alors une conséquence immédiate de la Proposition 2.55.

Corollaire 2.69. *Soit* Ω *un ouvert régulier borné de* $\mathbb{R}^d$*. Soit* u_n *une suite bornée de* $H^1(\Omega)$*. On peut extraire de la suite* u_n *une sous-suite qui converge faiblement vers* u *dans* $H^1(\Omega)$ *et qui converge fortement vers* u *dans* $L^2(\Omega)$*.*

Exercice 2.70. *En utilisant le corollaire ci-dessus, démontrer l'inégalité de Poincaré (1.8) par un raisonnement par l'absurde.*

2.3.3 Théorie spectrale des opérateurs autoadjoints compacts

Les opérateurs autoadjoints compacts ont une structure spectrale très particulière, qui ressemble beaucoup à celle des opérateurs linéaires en dimension finie.

Théorème 2.71 (Diagonalisation des opérateurs auto-adjoints compacts)**.** *Soit H un espace de Hilbert séparable de dimension infinie et $T \in \mathcal{L}(H)$ un opérateur autoadjoint compact. Alors il existe une suite (μ_n) de réels non nuls, finie ou tendant vers 0, et une base hilbertienne $(e_n) \cup (f_n)$ de H, telles que*

1. *$\sigma(T) = (\mu_n) \cup \{0\}$,*
2. *$Te_n = \mu_n e_n$ (et donc $\mu_n \in \sigma_{\mathrm{p}}(T)$),*
3. *(f_n) est une base de $\mathrm{Ker}(T)$.*

En outre, pour tout $\lambda \in \sigma(T) \setminus \{0\}$, l'espace propre $E_\lambda = \mathrm{Ker}(\lambda - T)$ est de dimension finie.

On note qu'on a toujours $0 \in \sigma(T)$. En effet :
- soit T n'est pas injectif, et alors $0 \in \sigma_{\mathrm{p}}(T)$;
- soit T n'est pas surjectif, et alors $0 \in \sigma_{\mathrm{r}}(T) \cup \sigma_{\mathrm{c}}(T)$ (en effet, si T est injectif et surjectif, alors il est bijectif, ce qui n'est pas possible en vertu de l'exercice 2.61) ; d'après la Proposition 2.43, on a que $\sigma_{\mathrm{r}}(T) = \emptyset$, donc $0 \in \sigma_{\mathrm{c}}(T)$.

Remarque 2.72. *La preuve ci-dessous montre que plusieurs cas (et uniquement ceux-là) peuvent se présenter :*

1. *on peut avoir $\sigma(T) = \sigma_{\mathrm{p}}(T)$, avec les cas suivants :*
   (a) *ou bien $\sigma(T) = \sigma_{\mathrm{p}}(T) = \{0\}$, auquel cas $T = 0$. Dans ce cas, la base (f_n) engendre tout l'espace, et la base (e_n) est vide ;*
   (b) *ou bien $\sigma(T) = \sigma_{\mathrm{p}}(T) = \{\mu_n\}_{n \in \{1,\dots,N\}} \cup \{0\}$, c'est-à-dire que T est de rang fini (et bien sur T n'est pas injectif). Dans ce cas, la base (e_n) est de cardinal fini N, et la base (f_n) est de cardinal infini ;*
   (c) *ou bien $\sigma(T) = \sigma_{\mathrm{p}}(T) = \{\mu_n\}_{n \geq 0} \cup \{0\}$, auquel cas T est non injectif. La base (e_n) est de cardinal infini, alors que la base (f_n) peut être de cardinal fini ou infini en fonction de la dégénerescence de la valeur propre 0 ;*
2. *si $\sigma_{\mathrm{p}}(T) \subsetneq \sigma(T)$, alors $\sigma(T)$ est l'union disjointe de $\sigma_{\mathrm{p}}(T)$ et de $\{0\}$. Dans ce cas, T est injectif (car $0 \notin \sigma_{\mathrm{p}}(T)$) et on a $\sigma_{\mathrm{p}}(T) = \{\mu_n\}_{n \geq 0}$ et $\sigma_{\mathrm{c}}(T) = \{0\}$ (en effet, $\{0\} = \sigma(T) \setminus \sigma_{\mathrm{p}}(T) = \sigma_{\mathrm{c}}(T) \cup \sigma_{\mathrm{r}}(T)$ et $\sigma_{\mathrm{r}}(T) = \emptyset$ d'après la Proposition 2.43).*

Démonstration. Nous décomposons cette (longue) preuve en plusieurs étapes.

1. Montrons pour commencer que

$$\sigma(T) \subset \sigma_{\mathrm{p}}(T) \cup \{0\}\,. \tag{2.14}$$

On rappelle que $\sigma(T) \subset \mathbb{R}$ par la Proposition 2.43. Pour montrer (2.14), considérons $\lambda \in \mathbb{R} \setminus \{0\}$ tel que $\lambda \notin \sigma_{\mathrm{p}}(T)$. Il s'agit de montrer que $\lambda \notin \sigma(T)$. Comme $\lambda \notin \sigma_{\mathrm{p}}(T)$, $(\lambda - T)$ est injectif. Etudions alors la surjectivité en nous intéressant à $V = \mathrm{Ran}(\lambda - T)$, et plus particulièrement, montrons que $V = H$, ce qui donnera le résultat escompté.

(a) On montre que V est fermé.

En effet, soit une suite $(w_n)_{n\in\mathbb{N}}$ d'éléments de V qui converge vers w dans H. Soit $(v_n)_{n\in\mathbb{N}}$ l'unique suite d'éléments de H définie par $w_n = (\lambda - T)v_n$ pour tout $n \in \mathbb{N}$. On a alors

$$v_n = \frac{1}{\lambda}[w_n + Tv_n].$$

Montrons d'abord que la suite (v_n) admet une sous-suite bornée. Par l'absurde, supposons que $\|v_n\| \longrightarrow +\infty$. En utilisant le fait que w_n converge, on aurait dans ce cas

$$\lambda \frac{v_n}{\|v_n\|} - T\frac{v_n}{\|v_n\|} = \frac{w_n}{\|v_n\|} \longrightarrow 0.$$

En utilisant la compacité de l'opérateur T, on extrait de (v_n) une sous-suite (v_{n_k}) telle que

$$T\frac{v_{n_k}}{\|v_{n_k}\|} \longrightarrow u \in H.$$

D'où

$$\frac{v_{n_k}}{\|v_{n_k}\|} \longrightarrow z = \frac{1}{\lambda}u$$

et z vérifie $(\lambda - T)z = 0$. Il en résulte que $z = 0$ puisque $\lambda - T$ est injectif. C'est impossible car z est la limite forte d'une suite de points de la sphère unité de H.

La suite $(v_n)_{n\in\mathbb{N}}$ admet donc une sous-suite bornée. L'opérateur T étant compact, (v_n) admet une sous-suite (v_{n_k}) bornée telle qu'on ait

$$Tv_{n_k} \longrightarrow w' \in H.$$

En utilisant à nouveau que w_n converge, il en résulte que

$$v_{n_k} \longrightarrow v = \frac{1}{\lambda}[w + w'] \in H,$$

ce qui indique que la suite $(v_n)_{n\in\mathbb{N}}$ admet une sous-suite convergente. Comme T est continu, on a finalement

$$w = \lim_{k\to+\infty} w_{n_k} = \lim_{k\to+\infty} (\lambda - T)v_{n_k} = (\lambda - T)v \in V,$$

ce qui montre bien que V est fermé.

(b) On montre que V est dense.
En effet, soit $w \in V^\perp$. Alors $\langle(\lambda - T)v, w\rangle = 0$ pour tout $v \in H$. Comme T est auto-adjoint et λ est réel, on en déduit que $\langle v, (\lambda - T)w\rangle = 0$ pour tout $v \in H$. Ceci implique que $(\lambda - T)w = 0$, et donc $w = 0$ puisque $(\lambda - T)$ est injectif. Donc $V^\perp = \{0\}$, et en utilisant le lemme 1.13, on en déduit que $\overline{V} = (V^\perp)^\perp = H$.

Ceci conclut la preuve de (2.14).

2. Montrons que $\sigma_{\mathrm{p}}(T)$ est ou bien une suite finie, ou bien une suite infinie qui converge vers 0.
Dans le cas contraire, on pourrait extraire de $\sigma_{\mathrm{p}}(T)$ une suite $(\lambda_n)_{n\in\mathbb{N}}$ de réels non nuls *tous distincts* qui converge vers un réel $\mu \neq 0$. Soit $e_n \in \mathrm{Ker}(\lambda_n - T)$ tel que $\|e_n\| = 1$. On a, pour tout $n \in \mathbb{N}$,

$$e_n = \frac{1}{\lambda_n} T e_n.$$

La suite $(e_n)_{n\in\mathbb{N}}$ étant bornée et T étant compact, on peut extraire une sous-suite (Te_{n_k}) qui converge dans H vers un certain u, d'où

$$e_{n_k} \longrightarrow \frac{1}{\mu} u.$$

Or la suite (e_n) est orthonormale par la Proposition 2.43, ce qui montre que la suite (e_{n_k}) n'est pas de Cauchy, donc ne peut pas converger. On a obtenu une contradiction, ce qui donne le résultat annoncé. En particulier, $\sigma_{\mathrm{p}}(T)$ est dénombrable.

3. A tout élément $\lambda_n \in \sigma_{\mathrm{p}}(T)$ tel que $\lambda_n \neq 0$, on associe $E_n = \mathrm{Ker}(\lambda_n - T)$. Montrons que les espaces E_n sont de dimension finie.
Soit en effet $T_n = T|_{E_n}$. Il est clair que $T_n = \lambda_n \,\mathrm{Id}_{E_n}$ (avec $\lambda_n \neq 0$) et que T_n est compact de E_n dans E_n (car c'est la restriction d'un opérateur compact à l'ensemble $E_n = \mathrm{Ker}(\lambda_n - T)$). L'opérateur T_n est donc compact et bijectif de E_n dans E_n. L'exercice 2.61 indique alors que E_n est de dimension finie.

4. Les espaces E_n sont deux à deux orthogonaux (par la Proposition 2.43) et sont orthogonaux à $F = \mathrm{Ker}(T)$.
Pour le second point, on procède comme dans la preuve de la Proposition 2.43. En effet, soit $\lambda_n \in \sigma_{\mathrm{p}}(T)$ avec $\lambda_n \neq 0$, $u \in E_n$ et $v \in F$. Alors

$$\lambda_n \langle u, v\rangle = \langle Tu, v\rangle = \langle u, Tv\rangle = 0,$$

d'où $\langle u, v\rangle = 0$ puisque $\lambda_n \neq 0$. Donc $F \subset E^\perp$.

5. Soit enfin $E = \bigoplus_n E_n$. Montrons que $H = E \oplus F$, où les sommes directes sont des sommes orthogonales dans les deux cas (selon le point précédent).

(a) Remarquons tout d'abord que E est stable par T. En effet, soit $x \in E$. On peut écrire

$$x = \sum_n x_n, \qquad x_n \in E_n, \qquad \sum_n \|x_n\|^2 < +\infty.$$

Comme par ailleurs (λ_n) est finie ou tend vers 0, la série $\sum_n \lambda_n x_n$ converge dans H. On a donc

$$Tx = \sum_n \lambda_n x_n, \qquad \lambda_n x_n \in E_n, \qquad \sum_n \|\lambda_n x_n\|^2 < +\infty,$$

ce qui montre que $Tx \in E$.

(b) Par ailleurs, $E^\perp$ est aussi stable par T. En effet, si $w \in E^\perp$, alors $\langle Tw, v\rangle = \langle w, Tv\rangle = 0$ pour tout $v \in E$ (on a utilisé que $Tv \in E$). Ceci montre que $Tw \in E^\perp$.

(c) Définissons maintenant $\widetilde{T}$, la restriction de T à l'ensemble fermé $E^\perp$:

$$\begin{array}{rccc} \widetilde{T} \ : & E^\perp & \to & E^\perp \\ & v & \mapsto & Tv. \end{array}$$

L'opérateur $\widetilde{T}$ est auto-adjoint et compact. En vertu de (2.14), on a $\sigma(\widetilde{T}) \subset \sigma_{\rm p}(\widetilde{T}) \cup \{0\}$. Supposons que $\sigma_{\rm p}(\widetilde{T}) \not\subset \{0\}$. Il existe alors $\lambda \in \sigma_{\rm p}(\widetilde{T})$ avec $\lambda \neq 0$, et il existe donc $v \in E^\perp \setminus \{0\}$ tel que

$$\widetilde{T}v = \lambda v,$$

d'où aussi $Tv = \lambda v$. Donc $\lambda \in \sigma_{\rm p}(T)$. Ceci signifie cependant que $\lambda = \lambda_n$ et que $v \in E_n$ pour un certain n. D'où

$$v \in E_n \cap E^\perp = \{0\},$$

ce qui contredit l'hypothèse $v \neq 0$. Donc $\sigma_{\rm p}(\widetilde{T}) \subset \{0\}$.

Il en résulte que $\sigma(\widetilde{T}) \subset \{0\}$, et comme le spectre n'est jamais vide, on obtient

$$\sigma(\widetilde{T}) = \{0\}.$$

D'après la proposition 2.43, la relation ci-dessus implique que $\|\widetilde{T}\| = 0$ et donc que $\widetilde{T} = 0$. Ainsi, $E^\perp \subset \mathrm{Ker}(T) = F$.

(d) On a $H = E \oplus E^\perp$ et on a vu ci-dessus que $F \subset E^\perp$. On vient de montrer que $E^\perp \subset \mathrm{Ker}(T) = F$. Donc $F = E^\perp$, ce qui donne bien que $H = E \oplus F$.

6. La base (e_n) et la suite (μ_n) sont construites de la manière suivante. Notons n_k la dimension de E_k. On prend $\mu_1 = \mu_2 = \cdots = \mu_{n_1} = \lambda_1$ et $(e_1, \cdots, e_{n_1})$ une base orthonormale de E_1. Puis on pose $\mu_{n_1+1} = \cdots = \mu_{n_1+n_2} = \lambda_2$ et $(e_{n_1+1}, \cdots, e_{n_1+n_2})$ une base orthonormale de E_2. On procède de même pour tous les espaces E_n.

Ceci conclut la preuve. □

Remarque 2.73. *Soit H est un espace de Hilbert. La preuve précédente montre qu'on peut écrire tout opérateur autoadjoint de $\mathcal{K}(H)$ (donc compact) comme une limite d'opérateurs de rang fini (voir [3]). En effet, comme $(e_n) \cup (f_n)$ forme une base hilbertienne de H, on peut écrire tout $u \in H$ sous la forme*

$$u = \sum_{n=1}^{+\infty} u_n,$$

et l'application T est diagonale dans cette base :

$$Tu = \sum_{n=1}^{+\infty} \lambda_n u_n, \tag{2.15}$$

avec $\lambda_n \to 0$ lorsque $n \to +\infty$ (éventuellement, il est possible que $\lambda_n = 0$ à partir d'un certain rang). Définissant les opérateurs de rang fini T_N par

$$T_N u = \sum_{n=1}^{N} \lambda_n u_n,$$

on voit facilement que $\|T - T_N\| \leq \sup_{m \geq N} |\lambda_m| \longrightarrow 0$ lorsque $N \to +\infty$.

Remarque 2.74 (Calcul fonctionnel)**.** *Notons également que la décomposition* (2.15) *permet de définir des opérateurs $f(T)$ par la formule*

$$f(T)u = \sum_{n=1}^{+\infty} f(\lambda_n) u_n.$$

Ceci généralise les opérations faites sur les matrices symétriques réelles.

2.3.4 Opérateurs autoadjoints compacts définis positifs

Dans la suite du cours, nous aurons besoin en particulier d'appliquer le théorème de décomposition spectrale à des opérateurs autoajoints compacts *définis positifs*. Donnons-en tout d'abord la définition.

Définition 2.75. *Soit V un espace de Hilbert, et soit A un opérateur borné de V dans V. On dit que A est défini positif si*

$$\forall u \in V \setminus \{0\}, \quad \langle Au, u \rangle > 0.$$

Remarque 2.76. *Soit V un espace de Hilbert, et soit A un opérateur borné de V dans V. On lui associe la forme bilinéaire a définie par*

$$a(u, w) = \langle Au, w \rangle.$$

En dimension finie, A est défini positif si et seulement si a est coercive. En dimension infinie, ce n'est plus le cas, comme le montre l'exercice 2.77 ci-dessous.

Exercice 2.77. *Soit Ω un ouvert borné de $\mathbb{R}^d$. On se place dans l'espace de Hilbert $L^2(\Omega)$. Pour tout $f \in L^2(\Omega)$, le problème*

$$\left\{ \begin{array}{l} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ -\Delta u = f \quad \text{dans } \mathcal{D}'(\Omega) \end{array} \right. \tag{2.16}$$

admet une unique solution. On considère l'opérateur

$$\begin{array}{rcl} A : L^2(\Omega) & \longrightarrow & L^2(\Omega) \\ f & \longmapsto & u \quad \text{solution du problème (2.16).} \end{array}$$

Montrer que A est un opérateur borné et que A est défini positif. Pour montrer que la forme bilinéaire associée a n'est pas coercive, on pourra supposer que Ω est la boule ouverte de centre 0 et de rayon 1, et considérer les fonctions $f_n(x) = n^{d/2}\chi(nx)$, où χ est une fonction fixée de $\mathcal{D}(\Omega)$.

Le théorème ci-dessous est alors un corollaire du Theorème 2.71 (on est dans le dernier cas évoqué dans la Remarque 2.72).

Théorème 2.78. *Soit V un espace de Hilbert de dimension infinie, et A un opérateur borné, défini positif, auto-adjoint et compact de V dans V. Alors les valeurs propres de A forment une suite $(\lambda_k)_{k\geq 1}$ de réels strictement positifs qui tend vers 0, et il existe une base hilbertienne $(u_k)_{k\geq 1}$ de V formée de vecteurs propres de A, avec*

$$\forall k \geq 1, \quad Au_k = \lambda_k u_k.$$

De plus, le sous-espace propre associé à chaque valeur propre est de dimension finie.

On remarque que le théorème ci-dessus ne caractérise que le spectre ponctuel de l'opérateur, alors que le Theorème 2.71 caractérise tout le spectre.

Remarque 2.79. *Comme $(u_k)_{k\geq 1}$ forme une base hilbertienne de V, on peut appliquer la proposition 1.10 et on a donc les relations suivantes pour tout $w \in V$:*

$$w = \sum_{k\geq 1} \langle w, u_k\rangle u_k \quad et \quad \|w\|^2 = \sum_{k\geq 1} |\langle w, u_k\rangle|^2.$$

Exercice 2.80. *On reprend les notations et hypothèses du théorème 2.78. Montrer que, pour $w \in V$, l'équation $Au = w$ admet une unique solution $u \in V$ si et seulement si w vérifie*

$$\sum_{k\geq 1} \frac{|\langle w, u_k\rangle|^2}{\lambda_k^2} < +\infty.$$

Exercice 2.81. *Soit $V = L^2(0,1)$ et A l'application linéaire de V dans V définie par $(Af)(x) = (x^2+1)f(x)$. Vérifier que A est continue, définie positive, auto-adjointe, mais pas compacte. Montrer que A n'a pas de valeurs propres. Montrer que $A - \lambda\mathrm{Id}$ est inversible si et seulement si $\lambda \notin [1,2]$.*

Nous présentons ici une démonstration directe du théorème 2.78. Dans ce but, nous aurons besoin des deux lemmes suivants.

Lemme 2.82. *Soit V un espace de Hilbert (non réduit au seul vecteur nul) et A une application linéaire continue auto-adjointe compacte de V dans V. On définit*

$$m = \inf_{u \in V \setminus \{0\}} \frac{\langle Au, u\rangle}{\langle u, u\rangle} \quad et \quad M = \sup_{u \in V \setminus \{0\}} \frac{\langle Au, u\rangle}{\langle u, u\rangle}.$$

Alors $\|A\|_{\mathcal{L}(V)} = \max(|m|, |M|)$ et soit m, soit M, est valeur propre de A.

Lemme 2.83. *Soit V un espace de Hilbert et A une application linéaire continue compacte de V dans V. Pour tout réel $\delta > 0$, il n'existe au plus qu'un nombre fini de valeurs propres de A en dehors de l'intervalle $]-\delta, \delta[$.*

Démontration du lemme 2.82. On voit que $|\langle Au, u\rangle| \leq \|A\|_{\mathcal{L}(V)}\|u\|^2$, par conséquent $\max(|m|, |M|) \leq \|A\|_{\mathcal{L}(V)}$. Comme A est auto-adjoint, on a, pour tout u et w dans V, que

$$\begin{aligned} 4\langle Au, w\rangle &= \langle A(u+w), u+w\rangle - \langle A(u-w), u-w\rangle \\ &\leq M\|u+w\|^2 - m\|u-w\|^2 \\ &\leq \max(|m|, |M|)\left(\|u+w\|^2 + \|u-w\|^2\right) \\ &\leq 2\max(|m|, |M|)\left(\|u\|^2 + \|w\|^2\right). \end{aligned}$$

Si $Au \neq 0$, on peut choisir $w = Au/\|Au\|$ dans l'inégalité précédente, et on obtient

$$2\|Au\| \leq \max(|m|, |M|)\left(\|u\|^2 + 1\right).$$

Cette dernière inégalité reste vraie si $Au = 0$. On prend maintenant le supremum sur les $u \in V$, $\|u\| = 1$, ce qui donne $2\|A\|_{\mathcal{L}(V)} \leq 2\max(|m|, |M|)$. En combinant cette inégalité avec l'inégalité inverse obtenue ci-dessus, on obtient que $\max(|m|, |M|) = \|A\|_{\mathcal{L}(V)}$.

On montre maintenant la deuxième partie du lemme. Si $m = M = 0$, alors, pour tout $u \in V$, on a $\langle Au, u\rangle = 0$. En utilisant l'exercice 2.46, on obtient que $A = 0$, ce qui termine la preuve du lemme. On suppose maintenant que soit m, soit M, est non nul, et donc $\max(|m|, |M|) > 0$. Par définition, on a $M \geq m$. Si $M \leq |m|$, alors on est dans un des deux cas suivants :

- soit $0 \geq M \geq m$: on change alors A en $-A$ ce qui permet de revenir au cas $M \geq m > 0$.
- soit $M \geq 0 \geq m$ et $M \leq |m|$: on change alors A en $-A$ ce qui permet de revenir au cas $M \geq |m| \geq 0$.

Sans perte de généralité, on peut donc supposer que $M \geq |m|$ et $M > 0$. Montrons que M est valeur propre de A. En utilisant la première partie du lemme et la définition de M, on a

$$\|A\|_{\mathcal{L}(V)} = M = \sup_{u \in V \setminus \{0\}} \frac{\langle Au, u\rangle}{\langle u, u\rangle}.$$

Soit $u_n \in V$ une suite maximisante, avec $\|u_n\| = 1$. On a donc $\lim_{n\to+\infty}\langle Au_n, u_n\rangle = M$. Comme u_n est bornée et A est compacte, on peut extraire de Au_n une sous-suite convergente : $\lim_{n\to+\infty} Au_{\varphi(n)} = v$. On a aussi

$$\langle Au_n, u_n\rangle \leq \|Au_n\|\ \|u_n\| \leq \|A\|_{\mathcal{L}(V)}\ \|u_n\|^2 = \|A\|_{\mathcal{L}(V)} = M.$$

Or $\lim_{n\to+\infty}\langle Au_n, u_n\rangle = M$, ce qui donne que $\lim_{n\to+\infty}\|Au_n\|\ \|u_n\| = M$. Comme $\|u_n\| = 1$, on en déduit que $\lim_{n\to+\infty}\|Au_n\| = M$. Sachant que Au_n converge à extraction près vers v, on obtient que $\|v\| = M$.

On voit aussi que

$$\|Au_n - Mu_n\|^2 = \|Au_n\|^2 + M^2 - 2M\langle Au_n, u_n\rangle \to_{n\to+\infty} 0,$$

ce qui implique $\lim_{n\to\infty} Au_n - Mu_n = 0$. Or $\lim\limits_{n\to+\infty} Au_{\varphi(n)} = v$, donc $\lim\limits_{n\to+\infty} Mu_{\varphi(n)} = v$. Comme A est continue, on a $\lim\limits_{n\to+\infty} MAu_{\varphi(n)} = Av$, et par unicité de la limite, on déduit que $Mv = Av$, avec $v \neq 0$. Donc M est bien valeur propre de A. □

Démontration du lemme 2.83. On procède par contradiction, et on suppose donc qu'il existe une suite infinie de valeurs propres $(\lambda_k)_{k\geq 1}$ distinctes telles que $|\lambda_k| \geq \delta$. Soient $(u_k)_{k\geq 1}$ les vecteurs propres associés, et E_k le sous-espace vectoriel engendré par $u_1, \ldots, u_k$.

Grâce au lemme 2.38, les vecteurs propres $(u_k)_{k\geq 1}$ sont linéairement indépendants, et donc E_{k-1} est strictement inclus dans E_k. Donc il existe w_k de norme 1, avec $w_k \in E_k$ et w_k orthogonal à E_{k-1}. Comme λ_k est isolé de 0, on voit que la suite de vecteurs w_k/λ_k est bornée. L'application A étant compacte, on en déduit que, à extraction près, la suite Aw_k/λ_k converge. Par ailleurs, pour $j < k$, on voit que

$$\begin{aligned}\frac{1}{\lambda_k}Aw_k - \frac{1}{\lambda_j}Aw_j &= \frac{1}{\lambda_k}\left(Aw_k - \lambda_k w_k\right) + w_k - \frac{1}{\lambda_j}Aw_j \\ &= \left(A - \lambda_k \mathrm{Id}\right)\frac{w_k}{\lambda_k} + w_k - \frac{1}{\lambda_j}Aw_j.\end{aligned}$$

Or, pour tout $w \in E_k$, on a $(A - \lambda_k\mathrm{Id})\, w \in E_{k-1}$. Par conséquent, les vecteurs $(A - \lambda_k\mathrm{Id})\,\dfrac{w_k}{\lambda_k}$ et $\dfrac{1}{\lambda_j}Aw_j$ sont dans E_{k-1}, tandis que w_k est orthogonal à E_{k-1}. Donc

$$\begin{aligned}\left\|\frac{1}{\lambda_k}Aw_k - \frac{1}{\lambda_j}Aw_j\right\|^2 &= \left\|\left(A - \lambda_k\mathrm{Id}\right)\frac{w_k}{\lambda_k} - \frac{1}{\lambda_j}Aw_j\right\|^2 + \|w_k\|^2 \\ &\geq \|w_k\|^2 = 1.\end{aligned}$$

Ceci est contradictoire avec le fait que la suite Aw_k/λ_k converge à extraction près. □

Démontration du théorème 2.78. Le lemme 2.82 montre que l'ensemble des valeurs propres n'est pas vide, tandis que le lemme 2.83 montre que cet ensemble est soit fini,

soit infini dénombrable avec 0 comme seul point d'accumulation. On note $(\lambda_k)_{k\geq 1}$ les valeurs propres de A et $V_k = \mathrm{Ker}(A - \lambda_k \mathrm{Id})$ les sous-espaces vectoriels propres associés. Comme A est défini positif, on voit que les valeurs propres sont toutes strictement positives.

Comme $\lambda_k \neq 0$, l'application $\dfrac{1}{\lambda_k}A$ est compacte, et la proposition 2.63 montre que $V_k = \mathrm{Ker}(\dfrac{1}{\lambda_k}A - \mathrm{Id})$ est de dimension finie.

Les sous-espaces propres sont orthogonaux deux à deux. En effet, si $v_k \in V_k$ et $v_j \in V_j$ avec $k \neq j$, alors, comme A est auto-adjoint,

$$\langle Av_j, v_k\rangle = \lambda_j\langle v_j, v_k\rangle = \langle v_j, Av_k\rangle = \lambda_k\langle v_j, v_k\rangle.$$

On déduit de $\lambda_k \neq \lambda_j$ que $\langle v_j, v_k\rangle = 0$.

Soit

$$W = \left\{ v \in V;\ \exists K \geq 1 \text{ tel que } v = \sum_{k=1}^{K} v_k,\ v_k \in V_k \right\}$$

l'espace vectoriel engendré par les $(v_k)_{k\geq 1}$. Montrons que W est dense dans V. Il est clair que W est stable par A, c'est-à-dire $A(W) \subset W$. L'application A étant auto-adjointe, ceci implique que $W^\perp$ est lui-aussi stable par A. On considère alors la restriction A_0 de A à $W^\perp$, qui est encore une application linéaire continue auto-adjointe compacte. Si $W^\perp \neq \{0\}$, on peut appliquer le lemme 2.82, et donc A_0 a une valeur propre λ. Soit u le vecteur propre associé : $u \in W^\perp$ et $Au = \lambda u$. Donc λ est une valeur propre de A, et par conséquent $u \in W$. Donc $u \in W \cap W^\perp$, ce qui est contradictoire avec le fait que $u \neq 0$. Donc $W^\perp = \{0\}$. Par conséquent, $V = \{0\}^\perp = (W^\perp)^\perp = \overline{W}$ (on a utilisé le lemme 1.13 pour obtenir la dernière égalité), ce qui montre que W est dense dans V.

On construit maintenant une base hilbertienne de V. Pour cela, on considère dans chacun des V_k (qui sont de dimension finie) une base orthonormée. Les réunions de ces bases forme une base hilbertienne de V, car les V_k sont orthogonaux deux à deux et W est dense dans V.

Comme V est de dimension infinie et que les V_k sont de dimension finie, on obtient aussi que A possède un nombre infini dénombrable de valeurs propres. □

Chapitre 3

Equations aux dérivées partielles et problèmes aux valeurs propres

3.1 Motivation

Ce chapitre est une introduction à l'étude mathématique et numérique des phénomènes vibratoires. Ces phénomènes ont une grande importance pour de nombreuses sciences de l'ingénieur : génie civil, acoustique (des instruments de musique mais aussi des véhicules), détection de fissure dans des matériaux (par contrôle non destructif), ...

D'un point de vue mathématique, il s'agit d'étudier les valeurs propres et vecteurs propres d'équations aux dérivées partielles. Illustrons notre propos sur un exemple concret. On considère une membrane élastique homogène et isotrope, dont le bord est maintenu fixe, initialement au repos, et on cherche à étudier sa réponse à une excitation dépendant du temps.

Lorsqu'on néglige les forces de gravitation devant les forces de tension superficielle, et qu'on se place dans le cadre de l'élasticité linéaire, le système vérifié par le déplacement vertical $u(t,x)$ d'un point de la membrane situé au repos à la position $x \in \Omega$ s'écrit :

$$\begin{cases} \dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2}(t,x) - \Delta u(t,x) = f(t,x) & \text{dans } \mathbb{R}^{+*} \times \Omega, \\ u(t,x) = 0 & \text{sur } \mathbb{R}^{+*} \times \partial\Omega, \\ u(0,x) = 0 & \text{sur } \Omega, \\ \dfrac{\partial u}{\partial t}(0,x) = 0 & \text{sur } \Omega, \end{cases} \tag{3.1}$$

où $c = \sqrt{S/\rho}$, S désignant la tension superficielle et ρ la masse surfacique de la membrane. On reconnaît dans l'EDP du système (3.1) une équation d'onde de célérité c comportant un terme source f.

L'analogue discret (en espace) de ce problème est le système dynamique d'incon-

nue $U(t) \in \mathbb{R}^N$ suivant :

$$\begin{cases} M \dfrac{d^2U}{dt^2}(t) + AU(t) = B(t), \\ U(0) = \dfrac{dU}{dt}(0) = 0, \end{cases} \tag{3.2}$$

où M et A sont deux matrices de taille $N \times N$ et $B(t)$ est un vecteur de $\mathbb{R}^N$ dépendant du temps.

Nous verrons plus loin dans le cours qu'on peut effectivement passer du système (3.1) au système (3.2) par une formulation variationnelle de (3.1), qui est ensuite approximée par une méthode de Galerkin (par exemple une méthode d'éléments finis ; cf. la section 8.2.1).

Supposons ici pour simplifier que M est la matrice identité, et que A est une matrice symétrique. Une méthode classique pour résoudre (3.2) est de diagonaliser la matrice A, ce qui consiste à chercher les couples $(\lambda_k, U_k)_{1 \leq k \leq N}$ de valeurs propres et de vecteurs propres de A, qui vérifient donc

$$\forall k, \quad AU_k = \lambda_k U_k. \tag{3.3}$$

Puisque A est symétrique, ses vecteurs propres forment une base orthonormée de $\mathbb{R}^N$. On cherche alors une solution de (3.2) comme une combinaison linéaire sur ces vecteurs propres :

$$U(t) = \sum_{k=1}^{N} \alpha_k(t) U_k \quad \text{avec} \quad \alpha_k(t) \in \mathbb{R}.$$

En insérant cette décomposition dans (3.2), on trouve que les α_k vérifient

$$\frac{d^2\alpha_k}{dt^2} + \lambda_k \alpha_k(t) = b_k(t) \tag{3.4}$$

avec $b_k(t) = \langle B(t), U_k \rangle$. On est donc ramené à la résolution d'une équation différentielle ordinaire scalaire.

L'argument clé qui a permis de ramener le système (3.2), posé en dimension N éventuellement grande, à la résolution des N équations scalaires *indépendantes* (3.4), est la diagonalisation de la matrice A et la recherche d'une solution comme combinaison linéaire de vecteurs propres. Essayons maintenant d'utiliser la même stratégie pour résoudre le problème (3.1). L'analogue de la matrice A, qui associe au vecteur U le vecteur $A\,U$, est l'opérateur $-\Delta$, qui à la distribution u associe la distribution $-\Delta u$. Il est donc naturel d'essayer de chercher des fonctions u_k, définies sur Ω, et des réels λ_k, tels que

$$-\Delta u_k = \lambda_k u_k \quad \text{dans} \quad \Omega. \tag{3.5}$$

Ce problème aux valeurs propres est l'équivalent en dimension infinie du problème (3.3). En fait, cette équation aux valeurs propres apparaît aussi naturellement si on

s'intéresse à l'équation sans second membre associée à (3.1), et qu'on en cherche une solution sous la forme $u(t,x) = \varphi(t)v(x)$. Oublions les conditions initiales : les fonctions φ et v doivent alors vérifier

$$\begin{cases} \dfrac{1}{c^2}\varphi''(t)v(x) - \varphi(t)\Delta v(x) = 0 & \text{pour tout } t > 0,\ x \in \Omega, \\ v(x) = 0 & \text{sur } \partial\Omega. \end{cases} \tag{3.6}$$

Formellement, on a donc

$$\forall t > 0,\ \forall x \in \Omega, \quad \frac{\varphi''(t)}{\varphi(t)} = \frac{\Delta v}{v} = -\lambda,$$

où $\lambda \in \mathbb{R}$ est une constante, et donc la fonction $v(x)$ est un vecteur propre du laplacien avec conditions de Dirichlet nulles au bord (on retrouve la relation (3.5)), tandis que φ suit l'équation suivante, similaire à (3.4) :

$$\varphi''(t) + \lambda\varphi(t) = 0.$$

Supposons $\lambda > 0$ (nous montrerons au théorème 3.3 ci-dessous que c'est effectivement le cas). Alors $\varphi(t) = a\cos(\sqrt{\lambda}t) + b\sin(\sqrt{\lambda}t)$, et la fonction

$$u(t,x) = av(x)\cos(\sqrt{\lambda}t) + bv(x)\sin(\sqrt{\lambda}t) \tag{3.7}$$

est solution de l'EDP apparaissant dans (3.1) avec $f = 0$. La fonction u s'interprète comme un mode propre de vibration de la membrane. La signification mécanique de λ se comprend sur la relation (3.7) : il s'agit du carré des pulsations propres de vibration.

La discussion ci-dessus permet donc de comprendre l'importance des valeurs propres et des vecteurs propres du laplacien, et de la signification du point de vue vibratoire de ces quantités.

La suite de ce chapitre est organisée ainsi. Les théorèmes abstraits qui ont été présentés au Chapitre 2 sont utilisés dans la section 3.2 pour étudier les modes propres du laplacien et de l'élasticité linéarisée. En pratique, on ne peut calculer qu'une approximation numérique des valeurs et vecteurs propres, et l'analyse d'erreur est discutée dans la section 3.3. Enfin, la mise en oeuvre numérique d'une méthode de discrétisation aboutit au bout du compte à un problème d'algèbre linéaire, qui consiste à diagonaliser une matrice. Quelques algorithmes pour la résolution d'un tel problème seront discutés dans la section 3.4.

3.2 Valeurs propres d'un problème elliptique

Pour commencer cette section, on se place dans un cadre assez général, qu'on pourra ensuite appliquer à différents modèles. Nous suivons en fait la même démarche

que dans les cours d'Analyse de première année [11, 14], dans lequel on a tout d'abord démontré, dans un cadre assez général, le théorème de Lax-Milgram, qu'on a ensuite appliqué à différentes équations. Nous appliquerons le résultat abstrait démontré à la section 3.2.1 dans les sections 3.2.2 (pour l'étude des valeurs propres du laplacien) et 3.2.3 (pour l'étude des modes propres de l'élasticité linéaire).

3.2.1 Problème variationnel abstrait

On se donne un espace de Hilbert V et une forme bilinéaire $a(\cdot,\cdot)$ sur V, qui est symétrique, continue et coercive. On se donne aussi un autre espace de Hilbert H, tel que

$$\begin{cases} V \subset H \text{ avec injection compacte au sens de la définition 2.64,} \\ V \text{ dense dans } H. \end{cases}$$

Pour ne pas confondre les produits scalaires sur H et sur V, nous les noterons respectivement $\langle\cdot,\cdot\rangle_H$ et $\langle\cdot,\cdot\rangle_V$. Les normes associées sont notées $\|\cdot\|_H$ et $\|\cdot\|_V$. Les hypothèses sur la forme a donnent donc l'existence de $M > 0$ et $\alpha > 0$ tels que

$$\begin{gathered} \forall u \in V,\ \forall w \in V, \quad |a(u,w)| \leq M\|u\|_V\|w\|_V, \\ \forall u \in V, \quad a(u,u) \geq \alpha\|u\|_V^2. \end{gathered}$$

Le problème qui nous intéresse ici est : trouver $\lambda \in \mathbb{R}$ et $u \in V \setminus \{0\}$ tels que

$$\forall w \in V, \quad a(u,w) = \lambda\langle u, w\rangle_H. \tag{3.8}$$

On dira alors que λ est valeur propre de la forme bilinéaire a (ou du probème variationnel (3.8)), et que u est le vecteur propre associé.

On donne dès à présent un cas typique d'application du cadre abstrait développé ici. Soit Ω un ouvert borné de $\mathbb{R}^d$. On pose $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, et

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

Nous montrerons à la section 3.2.2 que les hypothèses faites ci-dessus sont vérifiées, et que résoudre (3.8) est alors équivalent à chercher $\lambda \in \mathbb{R}$ et $u \in H_0^1(\Omega)$, $u \neq 0$, tels que

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

Ainsi, λ et u seront valeur propre et vecteur propre du laplacien dans Ω avec conditions aux limites de Dirichlet.

Théorème 3.1. *Soient V et H deux espaces de Hilbert de dimension infinie. On suppose $V \subset H$ avec injection compacte et V dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique, continue et coercive sur V. Alors les valeurs propres de (3.8)*

forment une suite croissante $(\lambda_k)_{k\geq 1}$ de réels strictement positifs qui tend vers l'infini, et il existe une base hilbertienne de H de vecteurs propres associés, c'est-à-dire :

$$u_k \in V \quad et \quad \forall w \in V,\ a(u_k, w) = \lambda_k \langle u_k, w\rangle_H. \tag{3.9}$$

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

Démonstration. L'injection $V \subset H$ étant continue, on sait qu'il existe $C > 0$ tel que

$$\forall w \in V, \quad \|w\|_H \leq C\|w\|_V. \tag{3.10}$$

Pour $f \in H$, on considère le problème variationnel

$$\left\{ \begin{array}{l} \text{Chercher } u \in V \text{ tel que} \\ \forall w \in V, \quad a(u,w) = \langle f, w\rangle_H. \end{array} \right. \tag{3.11}$$

Grâce au théorème de Lax-Milgram, ce problème admet une unique solution $u \in V$. On définit les applications linéaires

$$\begin{array}{rcl} \mathcal{A} : H & \longrightarrow & V \\ f & \longmapsto & u \text{ unique solution de (3.11)}, \end{array}$$

et

$$\begin{array}{rcl} A : H & \longrightarrow & H \\ f & \longmapsto & \mathcal{A}f. \end{array}$$

Comme a est coercive sur V, on a, pour u solution de (3.11),

$$\alpha\|u\|_V^2 \leq a(u,u) = \langle f, u\rangle_H \leq \|f\|_H\ \|u\|_H.$$

En utilisant (3.10), on obtient

$$\|\mathcal{A}f\|_V = \|u\|_V \leq \frac{C}{\alpha}\|f\|_H.$$

Donc $\mathcal{A}$ est linéaire continue de H dans V. En utilisant à nouveau (3.10), on obtient que A est linéaire continue de H dans H.

Montrons que A est définie positive, auto-adjointe et compacte sur H.

Comme A est la composition de $\mathcal{A} \in \mathcal{L}(H,V)$ et de l'injection de V dans H, qui est compacte, on a que A est compacte. Soient maintenant f et g dans H. On a

$$\langle f, Ag\rangle_H = \langle f, \mathcal{A}g\rangle_H = a(\mathcal{A}f, \mathcal{A}g) = a(\mathcal{A}g, \mathcal{A}f) = \langle g, \mathcal{A}f\rangle_H = \langle g, Af\rangle_H,$$

et donc A est auto-adjointe sur H. On montre enfin que A est définie positive sur H. En prenant $g = f$ dans l'égalité précédente, on voit que, pour tout $f \in H$,

$$\langle f, Af\rangle_H = a(\mathcal{A}f, \mathcal{A}f) \geq \alpha\|\mathcal{A}f\|_V^2 \geq 0.$$

Supposons que $\langle f, Af\rangle_H = 0$. Alors l'inégalité ci-dessus donne que $\mathcal{A}f = 0$. Par définition, on a

$$\forall w \in V, \quad a(\mathcal{A}f, w) = \langle f, w\rangle_H.$$

On déduit de $\mathcal{A}f = 0$ que $\langle f, w\rangle_H = 0$ pour tout $w \in V$. Or V est dense dans H, donc ceci implique que $\langle f, w\rangle_H = 0$ pour tout $w \in H$, et par conséquent $f = 0$. Finalement, pour tout $f \in H$, $f \neq 0$, on a $\langle f, Af\rangle_H > 0$ et donc A est définie positive sur H.

On peut donc appliquer le théorème 2.78. Il existe donc une base hilbertienne de H formée des vecteurs propres u_k de A, associés aux valeurs propres $(\mu_k)_{k\geq 1}$, qui forme une suite décroissante vers 0 :

$$\forall k \geq 1, \quad Au_k = \mu_k u_k.$$

Comme $\mu_k > 0$ et $Au_k \in V$, on voit que $u_k \in V$. On montre maintenant que les u_k sont vecteurs propres de la forme bilinéaire a. Par définition de A, on a

$$\forall w \in V,\ a(Au_k, w) = \langle u_k, w\rangle_H = \mu_k a(u_k, w),$$

et donc, en posant

$$\lambda_k = \frac{1}{\mu_k},$$

on obtient (3.9). Montrons que les v_k définis par

$$v_k = \frac{u_k}{\sqrt{\lambda_k}}$$

forment une base hilbertienne de V pour le produit scalaire $a(\cdot,\cdot)$. On a $v_k \in V$ et l'espace vectoriel engendré par les v_k est dense dans H, donc dense dans V. Enfin, les vecteurs v_k sont orthogonaux deux à deux, car

$$\begin{aligned} a(v_k, v_p) &= a\left(\frac{u_k}{\sqrt{\lambda_k}}, \frac{u_p}{\sqrt{\lambda_p}}\right) \\ &= \frac{1}{\sqrt{\lambda_k\lambda_p}} a(u_k, u_p) \\ &= \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_p}} \langle u_k, u_p\rangle_H = \delta_{kp}. \end{aligned}$$

Ceci conclut la preuve du théorème. □

On donne maintenant une caractérisation très utile des valeurs propres du problème (3.8), appelé principe du min-max ou de Courant-Fisher. Nous introduisons le quotient de Rayleigh défini, pour chaque $v \in V \setminus \{0\}$, par

$$R(v) = \frac{a(v, v)}{\|v\|_H^2}. \tag{3.12}$$

Proposition 3.2. *Soient V et H deux espaces de Hilbert de dimension infinie. On suppose $V \subset H$ avec injection compacte et V dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique, continue et coercive sur V. Pour $k \geq 0$, on note E_k l'ensemble des sous-espaces vectoriels de dimension k de V. On note $(\lambda_k)_{k\geq 1}$ la suite croissante des valeurs propres du problème variationnel (3.8). Alors, pour tout $k \geq 1$, la k-ième valeur propre est donnée par*

$$\lambda_k = \min_{W\in E_k} \left(\max_{v\in W\setminus\{0\}} R(v) \right) = \max_{W\in E_{k-1}} \left(\min_{v\in W^\perp\setminus\{0\}} R(v) \right). \tag{3.13}$$

En particulier, la première valeur propre vérifie

$$\lambda_1 = \min_{v\in V\setminus\{0\}} R(v), \tag{3.14}$$

et tout point de minimum dans (3.14) est un vecteur propre associé à λ_1.

Démonstration. Soit u_k une base hilbertienne de H formée des vecteurs propres de (3.8). On commence par caractériser H et V. On a

$$H = \left\{ v = \sum_{k\geq 1} \alpha_k u_k \text{ tel que } \sum_{k\geq 1} \alpha_k^2 < +\infty \right\}.$$

En effet, soit $v \in H$: comme u_k est une base hilbertienne de H, en utilisant la proposition 1.10, on a bien $v = \sum_{k\geq 1} \alpha_k u_k$ avec $\alpha_k = \langle v, u_k\rangle_H$. La série $\sum_{k\geq 1} \alpha_k^2$ est bien convergente car égale à $\|v\|_H^2$. Réciproquement, soit une suite α_k telle que $\sum_{k\geq 1} \alpha_k^2 < +\infty$. La suite $\sum_{k=1}^K \alpha_k u_k$ est bien dans H, et elle est de Cauchy, donc elle converge vers un élément de H.

On montre maintenant que

$$V = \left\{ v = \sum_{k\geq 1} \alpha_k u_k \text{ tel que } \sum_{k\geq 1} \lambda_k \alpha_k^2 < +\infty \right\}.$$

Soit $v \in V$: les $v_k = u_k/\sqrt{\lambda_k}$ forment une base hilbertienne de V pour $a(\cdot,\cdot)$, donc on peut décomposer v suivant ces v_k selon

$$v = \sum_{k\geq 1} \alpha_k v_k \text{ avec } a(v,v) = \sum_{k\geq 1} \alpha_k^2.$$

Posant $\beta_k = \alpha_k/\sqrt{\lambda_k}$, on obtient $v = \sum_{k\geq 1} \beta_k u_k$ avec $a(v,v) = \sum_{k\geq 1} \lambda_k \beta_k^2 < +\infty$. Réciproquement, supposons $v = \sum_{k\geq 1} \alpha_k u_k$ avec $\sum_{k\geq 1} \lambda_k \alpha_k^2 < +\infty$. Alors la suite $\sum_{k=1}^K \alpha_k u_k$ est une suite d'éléments de V qui est de Cauchy pour la norme induite par $a(\cdot,\cdot)$. Donc cette suite converge vers un élément de V.

Soit maintenant $v \in V \setminus \{0\}$. Alors on écrit $v = \sum_{k\geq 1} \alpha_k u_k$ et le quotient de Rayleigh s'écrit

$$R(v) = \frac{\sum_{k\geq 1} \lambda_k \alpha_k^2}{\sum_{k\geq 1} \alpha_k^2}.$$

L'égalité (3.14) est donc claire. Soit u un point de minimum : $R(u) = \lambda_1$. Soit $v \in V$ quelconque. La fonction $f(t) = R(u+tv)$ est minimale en $t = 0$, donc $f'(0) = 0$. Or

$$f'(0) = 2\frac{a(u,v)\|u\|_H^2 - \langle u, v\rangle_H a(u,u)}{\|u\|_H^4}.$$

Comme $f'(0) = 0$ et $a(u,u) = \lambda_1\|u\|_H^2$, on obtient $a(u,v) = \lambda_1\langle u, v\rangle_H$ pour tout $v \in V$, et donc u est vecteur propre associé à la valeur propre λ_1.

On démontre maintenant (3.13). Soit W_k l'espace vectoriel engendré par $(u_1, \ldots, u_k)$, qui est de dimension k. Soit $v \in W_k$: on a $R(v) = \dfrac{\sum_{j=1}^k \lambda_j \alpha_j^2}{\sum_{j=1}^k \alpha_j^2}$ donc

$$\lambda_k = \max_{v\in W_k, v\neq 0} R(v) \geq \min_{W\in E_k}\left(\max_{v\in W\setminus\{0\}} R(v)\right). \tag{3.15}$$

De même, pour $v \in W_{k-1}^\perp$, on a $R(v) = \dfrac{\sum_{j\geq k} \lambda_j \alpha_j^2}{\sum_{j\geq k} \alpha_j^2}$ et donc

$$\lambda_k = \min_{v\in W_{k-1}^\perp, v\neq 0} R(v) \leq \max_{W\in E_{k-1}}\left(\min_{v\in W^\perp\setminus\{0\}} R(v)\right).$$

Soit maintenant W un sous-espace vectoriel de V de dimension k. On a $V = W_{k-1} \oplus W_{k-1}^\perp$, donc $W = (W\cap W_{k-1})\oplus(W\cap W_{k-1}^\perp)$. Si $W\cap W_{k-1}^\perp = \{0\}$, alors $W = W\cap W_{k-1}$, ce qui n'est pas possible car W est de dimension k et $W \cap W_{k-1}$ est de dimension inférieure ou égale à $k-1$. Donc $(W \cap W_{k-1}^\perp) \setminus \{0\} \neq \emptyset$. On a

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

Par conséquent,

$$\min_{W\in E_k}\left(\max_{v\in W\setminus\{0\}} R(v)\right) \geq \lambda_k.$$

En rassemblant cette inégalité avec (3.15), on obtient la première égalité de (3.13). La seconde égalité de (3.13) s'obtient de manière analogue, en considérant $W \in E_{k-1}$ et en s'appuyant sur le fait que $W^\perp \cap W_k$ n'est pas réduit à $\{0\}$. □

3.2.2 Première application : valeurs propres du laplacien

Dans cette section, nous mettons en oeuvre le théorème 3.1, démontré dans un cadre abstrait, pour étudier les valeurs propres du laplacien.

Théorème 3.3. *Soit Ω un ouvert borné régulier de classe C^1 de $\mathbb{R}^d$. Il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de réels strictement positifs qui tend vers l'infini, et il existe une base hilbertienne de $L^2(\Omega)$, notée $(u_k)_{k\geq 1}$, telle que chaque u_k appartient à $H_0^1(\Omega)$ et vérifie*

$$\left\{ \begin{array}{rcll} -\Delta u_k & = & \lambda_k u_k & \text{dans } \mathcal{D}'(\Omega), \\ u_k & = & 0 & \text{sur } \partial\Omega. \end{array} \right. \tag{3.16}$$

Les $(\lambda_k)_{k\geq 1}$ et les $(u_k)_{k\geq 1}$ sont appelés les valeurs propres et vecteurs propres du laplacien avec conditions aux limites de Dirichlet sur l'ouvert Ω.

Démonstration. On va appliquer le théorème 3.1, avec les choix $V = H_0^1(\Omega)$ (muni du produit scalaire $(\cdot,\cdot)_{H^1}$), $H = L^2(\Omega)$ (muni du produit scalaire $(\cdot,\cdot)_{L^2}$), et

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

Comme $C_0^\infty(\Omega)$ est dense dans $L^2(\Omega)$ et inclus dans $H_0^1(\Omega)$, on a bien que V est dense dans H. Comme Ω est borné, on peut appliquer le théorème de Rellich 2.67, et l'injection $V \subset H$ est bien compacte. La forme a est bien bilinéaire, symétrique, continue et coercive sur V (ce dernier point résulte directement de l'inégalité de Poincaré (1.8)). Par conséquent, il existe une suite croissante $(\lambda_k)_{k\geq 1}$ de réels positifs et une base hilbertienne $(u_k)_{k\geq 1}$ de $L^2(\Omega)$ tels que $u_k \in H_0^1(\Omega)$ et

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

On obtient alors (3.16) par une simple intégration par partie. □

Remarque 3.4. *Supposons que Ω soit de classe C^∞. Alors les u_k solutions de (3.16) sont bien plus réguliers que $H_0^1(\Omega)$. On voit en effet que $-\Delta u_k = \lambda_k u_k$ avec $\lambda_k u_k$ de régularité H^1. Donc $\Delta u_k \in H^1(\Omega)$. Comme Ω est très régulier, ceci impose que $u_k \in H^3(\Omega)$, et donc $\Delta u_k \in H^3(\Omega)$, ce qui donne $u_k \in H^5(\Omega)$, ... On obtient finalement que $u_k \in C^\infty(\Omega)$.*

Remarque 3.5. *L'hypothèse que Ω est borné est fondamentale. Sans cette hypothèse, le théorème de Rellich est faux, et le théorème 3.3 est lui aussi faux.*

Exercice 3.6. *On se place en dimension 1 et on considère $\Omega =]0,1[$. Calculer explicitement toutes les valeurs propres et les fonctions propres du laplacien avec conditions aux limites de Dirichlet (3.16). En déduire que la série $\sum_{k\geq 1} a_k \sin(k\pi x)$ converge dans $L^2(0,1)$ si et seulement si $\sum_{k\geq 1} a_k^2 < +\infty$, et que la même série converge dans $H^1(0,1)$ si et seulement si $\sum_{k\geq 1} k^2 a_k^2 < +\infty$.*

En utilisant le principe de Courant-Fisher, on pourra résoudre l'exercice suivant.

Exercice 3.7. *On reprend les notations et hypothèses du théorème 3.3. Trouver une relation entre la plus petite constante C_Ω possible dans l'inégalité de Poincaré (1.8) et la première valeur propre λ_1 de (3.16).*

On donne enfin un résultat qualitatif très important à propos de la première valeur propre.

Théorème 3.8 (de Krein-Rutman)**.** *On reprend les notations et hypothèses du théorème 3.3. On suppose que l'ouvert Ω est connexe. Alors la première valeur propre λ_1 est simple (le sous-espace vectoriel associé est de dimension 1), et le premier vecteur propre peut être choisi positif presque partout dans Ω.*

Remarque 3.9. *Ce théorème est spécifique aux équations scalaires, c'est-à-dire pour lesquelles l'inconnue u est à valeurs dans $\mathbb{R}$. Dans le cas vectoriel, comme par exemple dans le cas de l'élasticité traitée dans la section 3.2.3, le résultat est faux.*

3.2.3 Seconde application : l'élasticité linéarisée

On s'intéresse maintenant au problème de l'élasticité linéarisée, et on va à nouveau utiliser le théorème 3.1 pour montrer l'existence de modes propres. On reprend les notations de la section 1.4. Pour éviter de confondre le coefficient de Lamé λ avec les valeurs propres, ces dernières sont notées ℓ_k.

Théorème 3.10. *Soit Ω un ouvert connexe borné régulier de classe C^1 de $\mathbb{R}^d$, avec $d = 2$ ou $d = 3$. Il existe une suite croissante $(\ell_k)_{k\geq 1}$ de réels strictement positifs qui tend vers l'infini, et il existe une base hilbertienne de $L^2(\Omega)^d$, notée $(u_k)_{k\geq 1}$, telle que chaque u_k appartient à $H^1_0(\Omega)^d$ et vérifie*

$$\left\{ \begin{array}{rcll} -div\ (2\mu e(u_k) + \lambda(tr\ e(u_k))\,\mathrm{Id}) & = & \ell_k u_k & \text{dans } \mathcal{D}'(\Omega)^d, \\ u_k & = & 0 & \text{sur } \partial\Omega. \end{array} \right. \tag{3.17}$$

Démonstration. On applique le théorème 3.1, avec les choix $V = H^1_0(\Omega)^d$, $H = L^2(\Omega)^d$, et

$$a(u,v) = \lambda \int_\Omega \mathrm{div}\ u\ \mathrm{div}\ v + 2\mu \int_\Omega e(u) \cdot e(v).$$

La preuve suit les mêmes étapes que la preuve du théorème 3.3. Le seul point délicat est la coercivité de la forme bilinéaire a, qui a été démontrée à la section 1.4.4 (cf. l'inégalité (1.31)). □

La quantité ℓ_k s'interprète comme le carré des pulsations propres de vibration, tandis que les fonctions u_k sont les modes propres de vibration du solide.

3.3 Méthodes numériques

Dans la section 3.2.1, nous nous sommes intéressés à la résolution du problème aux valeurs propres (3.8). Nous expliquons maintenant comment discrétiser ce problème pour aboutir à une méthode numérique permettant de calculer une approximation des valeurs propres (et éventuellement des vecteurs propres) de (3.8).

3.3.1 Discrétisation du problème

On réalise une approximation interne du problème (3.8). Soit donc $V_h \subset V$ un sous-espace de dimension finie de V. Typiquement, V_h est un espace d'éléments finis, tandis que H est l'espace $L^2(\Omega)$. Le problème discrétisé est : trouver $\lambda_h \in \mathbb{R}$ et $u_h \in V_h \setminus \{0\}$ tels que

$$\forall w_h \in V_h, \quad a(u_h, w_h) = \lambda_h \langle u_h, w_h \rangle_H. \tag{3.18}$$

Théorème 3.11. *On reprend les hypothèses du théorème 3.1 : soient V et H deux espaces de Hilbert de dimension infinie. On suppose $V \subset H$ avec injection compacte et V dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique, continue et coercive sur V, et soit $V_h \subset V$ un sous-espace de dimension finie J.*

Alors les valeurs propres de (3.18) forment une suite croissante finie

$$0 < \lambda_{1,h} \leq \ldots \leq \lambda_{J,h},$$

et il existe une base de V_h, orthonormale dans H, de vecteurs propres associés, c'est-à-dire : pour tout m, $1 \leq m \leq J$,

$$u_{m,h} \in V_h \quad et \quad \forall w_h \in V_h,\ a(u_{m,h}, w_h) = \lambda_{m,h} \langle u_{m,h}, w_h \rangle_H. \tag{3.19}$$

Pour démontrer ce théorème, nous aurons besoin du résultat d'algèbre linéaire suivant :

Proposition 3.12 (Factorisation de Cholesky)**.** *Soit A une matrice réelle symétrique définie positive. Il existe une unique matrice réelle B, triangulaire inférieure, telle que tous ses éléments diagonaux soient positifs, et qui vérifie*

$$A = BB^t.$$

Démonstration. Plutôt que de démontrer ce théorème en suivant le schéma de preuve du théorème 3.1, on suit ici une preuve plus algébrique. Soit $(\varphi_j)_{1 \leq j \leq J}$ une base de V_h (ce sont par exemple les fonctions de base d'une méthode d'éléments finis). On cherche u_h solution de (3.18) sous la forme

$$u_h(x) = \sum_{j=1}^{J} U_j \varphi_j(x).$$

On introduit les matrices de masse $\mathcal{M}_h$ et de rigidité $\mathcal{K}_h$ définies par, pour tout i et j, $1 \leq i, j \leq J$,

$$(\mathcal{M}_h)_{ij} = \langle \varphi_i, \varphi_j \rangle_H, \quad (\mathcal{K}_h)_{ij} = a(\varphi_i, \varphi_j).$$

Alors le problème (3.18) se récrit : trouver $\lambda_h \in \mathbb{R}$ et $U \in \mathbb{R}^J$, $U \neq 0$, tels que

$$\mathcal{K}_h U = \lambda_h \mathcal{M}_h U. \tag{3.20}$$

La terminologie matrice de masse et de rigidité est liée à la mécanique des solides. La matrice de rigidité $\mathcal{K}_h$ est la même que celle apparaissant dans la résolution par approximation interne du problème variationnel $a(u,w) = \langle f, w\rangle_H$. Les matrices $\mathcal{M}_h$ et $\mathcal{K}_h$ sont symétriques définies positives.

Pour résoudre le problème (3.20), on commence par calculer la factorisation de Cholesky de $\mathcal{M}_h$, c'est-à-dire calculer la matrice Q_h telle que $\mathcal{M}_h = Q_h Q_h^t$.

Une fois ceci fait, le problème (3.20) revient au problème classique

$$\tilde{\mathcal{K}}_h \tilde{U} = \lambda_h \tilde{U}, \tag{3.21}$$

avec $\tilde{U} = Q_h^t U$ et $\tilde{\mathcal{K}}_h = Q_h^{-1}\mathcal{K}_h(Q_h^t)^{-1}$. On note que la matrice $\tilde{\mathcal{K}}_h$ est symétrique et positive. Si ξ est tel que $\xi^t\tilde{\mathcal{K}}_h\xi = 0$, alors, puisque $\mathcal{K}_h$ est symétrique définie positive, on a $(Q_h^t)^{-1}\xi = 0$, donc $\xi = 0$. La matrice $\tilde{\mathcal{K}}_h$ est donc symétrique définie positive.

Pour le problème (3.21), on dispose d'algorithmes de calculs de valeurs propres et de vecteurs propres, dont certains seront décrits à la section 3.4.

On note $(\lambda_m, \tilde{U}_m)$ les éléments propres de $\tilde{\mathcal{K}}_h$: $\tilde{\mathcal{K}}_h\tilde{U}_m = \lambda_m\tilde{U}_m$. On définit $U_m = (Q_h^t)^{-1}\tilde{U}_m$ et on a donc $\mathcal{K}_h U_m = \lambda_m \mathcal{M}_h U_m$.

Soit U_m et U_n associés à des valeurs propres distinctes : $\lambda_m \neq \lambda_n$. Alors, en utilisant la symétrie de $\mathcal{K}_h$, on a

$$\lambda_m U_n^t \mathcal{M}_h U_m = U_n^t \mathcal{K}_h U_m = (U_n^t \mathcal{K}_h U_m)^t = U_m^t \mathcal{K}_h U_n = \lambda_n U_m^t \mathcal{M}_h U_n.$$

Puisque $\lambda_m \neq \lambda_n$, ceci implique que $U_m^t \mathcal{M}_h U_n = 0$. Les vecteurs propres solution de (3.20) sont donc orthogonaux pour $\mathcal{M}_h$ (et donc pour $\mathcal{K}_h$). □

Pour éviter d'avoir à calculer la factorisation de Cholesky de $\mathcal{M}_h$, on peut utiliser une formule de quadrature pour évaluer $\langle\varphi_i, \varphi_j\rangle_H$ qui rende la matrice de masse diagonale. Un tel procédé est appelé condensation de masse (ou mass lumping) et est souvent utilisé en pratique, par exemple dans l'esprit de l'exercice suivant.

Exercice 3.13. *On suppose que Ω est un ouvert borné de $\mathbb{R}^d$,*

$$V = H_0^1(\Omega), \quad H = L^2(\Omega), \quad a(u,v) = \int_\Omega \nabla u \cdot \nabla v.$$

On étudie donc $-\Delta u = \lambda u$ dans $H_0^1(\Omega)$. On suppose qu'on utilise une méthode d'éléments finis P_1 sur un maillage formé de triangles (en 2D) ou de tetraèdres (en 3D) de sommets $(a_i)_{1\leq i\leq d+1}$. On utilise la formule de quadrature

$$\int_K \psi(x)dx \approx \frac{\mathit{Volume(K)}}{d+1}\sum_{i=1}^{d+1}\psi(a_i), \tag{3.22}$$

où K est un triangle (ou un tétraèdre) du maillage. Ceci revient donc à choisir pour noeud d'intégration les sommets de K, qu'on affecte tous du même poids.

Vérifier que la formule de quadrature (3.22) conduit effectivement à une matrice de masse $\mathcal{M}_h$ diagonale.

3.3.2 Convergence et estimation d'erreur

Nous estimons ici la différence entre les valeurs propres du problème continu (3.8) et les valeurs propres du problème (3.19) (identique à (3.18)), qui est son approximation discrète. Cette estimation est fondée sur la caractérisation suivante des valeurs propres $(\lambda_{m,h})_{1\leq m\leq J}$ du problème discrétisé (3.19), analogue en dimension finie du principe de Courant-Fisher (cf. la proposition 3.2) :

$$\lambda_{m,h} = \min_{W\in E_{m,h}} \left(\max_{v\in W\setminus\{0\}} R(v) \right), \tag{3.23}$$

où $E_{m,h}$ est l'ensemble des sous-espaces vectoriels de dimension m de V_h, et $R(v)$ est le quotient de Rayleigh défini par (cf. (3.12))

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

La comparaison de (3.13) et de (3.23) donne déjà que, pour $1 \leq m \leq J$,

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

Pour obtenir une majoration de $\lambda_{m,h}$, on introduit l'opérateur de projection $\Pi_h \in \mathcal{L}(V, V_h)$ défini, pour tout $u \in V$, par

$$\forall w_h \in V_h, \quad a(\Pi_h u, w_h) = a(u, w_h). \tag{3.24}$$

Soient $(u_m)_{m\geq 1}$ les vecteurs propres de (3.8), et soit W_m le sous-espace vectoriel de V engendré par $(u_1, \ldots, u_m)$, qui est de dimension m.

Lemme 3.14. *Pour tout $1 \leq m \leq J$, on pose*

$$\sigma_{m,h} = \inf_{v\in W_m, \|v\|_H=1} \|\Pi_h v\|_H.$$

Si $\sigma_{m,h} > 0$, on a

$$\lambda_{m,h} \leq \frac{\lambda_m}{\sigma_{m,h}^2}.$$

Démonstration. On utilise le principe de Courant-Fisher (caractérisation (3.23)) avec le choix $W_{m,h} = \text{Vect}\ \{\Pi_h u_1, \ldots, \Pi_h u_m\}$. On a bien $W_{m,h} \subset V_h$ et $\dim W_{m,h} \leq m$. Montrons que $W_{m,h}$ est de dimension m. Si ce n'est pas le cas, alors il existe $(\alpha_i)_{1\leq i\leq m}$ non tous nuls tels que

$$0 = \sum_{i=1}^m \alpha_i \Pi_h u_i = \Pi_h \left(\sum_{i=1}^m \alpha_i u_i \right),$$

ce qui contredit l'hypothèse $\sigma_{m,h} > 0$. Donc $\dim W_{m,h} = m$ et (3.23) implique que

$$\lambda_{m,h} \leq \max_{v \in W_{m,h} \setminus \{0\}} R(v) = \max_{v \in W_m, \|v\|_H = 1} \frac{a(\Pi_h v, \Pi_h v)}{\|\Pi_h v\|_H^2}.$$

Pour tout $v \in V$, on a

$$a(v,v) = a\left(\Pi_h v, \Pi_h v\right) + a\left(v - \Pi_h v, v - \Pi_h v\right) + 2a\left(v - \Pi_h v, \Pi_h v\right).$$

Par définition de $\Pi_h v$, le dernier terme est nul. Par coercivité de a, le second terme est positif. Donc $a(v,v) \geq a\left(\Pi_h v, \Pi_h v\right)$ et donc

$$\lambda_{m,h} \leq \max_{v \in W_m, \|v\|_H = 1} \frac{a(v,v)}{\|\Pi_h v\|_H^2}.$$

Pour $v \in W_m$ tel que $\|v\|_H = 1$, on a $v = \sum_{i=1}^m \alpha_i u_i$ avec $\sum_{i=1}^m \alpha_i^2 = 1$, donc $a(v,v) \leq \lambda_m$, d'où

$$\lambda_{m,h} \leq \lambda_m \max_{v \in W_m, \|v\|_H = 1} \frac{1}{\|\Pi_h v\|_H^2} = \frac{\lambda_m}{\sigma_{m,h}^2}.$$

Ceci conclut la preuve. □

On a donc l'estimation

$$\lambda_m \leq \lambda_{m,h} \leq \lambda_m / (\sigma_{m,h}^2). \tag{3.25}$$

On voit donc que la différence entre $\lambda_{m,h}$ et λ_m est reliée aux propriétés d'approximation de V par V_h. Plus V_h est "proche" de V, plus on s'attend à ce que la solution $\Pi_h u \in V_h$ du problème (3.24) soit proche de u, donc en particulier que $\|\Pi_h u\|_H$ soit proche de $\|u\|_H$. Ceci implique alors que $\sigma_{m,h}$ est proche de 1 (puisqu'on minimise $\|\Pi_h v\|_H$ sur des vecteurs v de norme 1). On remarque donc que, pour aller plus loin dans l'estimation de $\lambda_{m,h}$, il n'est plus nécessaire de faire appel à la spécificité du problème (c'est un problème aux valeurs propres). Disposer de propriétés d'approximation de V par V_h suffit.

Mentionnons enfin que ces propriétés d'approximation sont souvent reliées à l'existence d'une application r_h de V dans V_h telle que, pour tout $v \in V$, on a $\lim_{h \to 0} \|v - r_h(v)\|_V = 0$. Dans le cas d'une approximation par éléments finis P_1, l'application r_h est par exemple l'interpolation de v sur les noeuds du maillage.

Précisons tout ceci dans un cas particulier. On revient à la définition (3.24) de l'opérateur Π_h. En utilisant le fait que la forme bilinéaire a est coercive et continue, on a, pour tout $u \in V$,

$$\begin{aligned} \alpha \|u - \Pi_h u\|_V^2 &\leq a(u - \Pi_h u, u - \Pi_h u) \\ &\leq a(u - \Pi_h u, u - \Pi_h u + w_h) \\ &\leq M \|u - \Pi_h u\|_V \ \|u - \Pi_h u + w_h\|_V \end{aligned}$$

pour tout $w_h \in V_h$. Donc

$$\|u - \Pi_h u\|_V \leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|u - w_h\|_V. \tag{3.26}$$

On suppose maintenant que $V = H_0^1(\Omega)$ pour un ouvert Ω borné de $\mathbb{R}^n$, et que V_h est le sous-espace de V correspondant à la méthode des éléments finis P_k, avec $k+1 > n/2$. On considère alors l'interpolée $r_h v$ d'une fonction v. C'est un résultat classique [1] que cette application r_h est bien définie sur $H^{k+1}(\Omega)$ et qu'il existe une constante C vérifiant

$$\forall v \in H^{k+1}(\Omega), \quad \|v - r_h v\|_{H^1(\Omega)} \leq C h^k \|v\|_{H^{k+1}(\Omega)}. \tag{3.27}$$

Supposons maintenant que W_m, l'espace vectoriel engendré par les m premiers vecteurs propres de la forme bilinéaire a, soit inclus dans $H^{k+1}(\Omega)$. Alors, il existe C_m tel que, pour tout $v \in W_m$ de norme 1, on a

$$\|v - r_h v\|_{H^1(\Omega)} \leq C_m h^k. \tag{3.28}$$

Détaillons ceci. On peut toujours supposer que les m premiers vecteurs propres de a, notés u_j, $1 \leq j \leq m$, sont orthogonaux deux à deux pour le produit scalaire de H^1, et sont de norme 1 : $\|u_j\|_{H^1} = 1$. On a supposé que $W_m \subset H^{k+1}(\Omega)$, donc $u_j \in H^{k+1}(\Omega)$ vérifie la majoration (3.27). En posant $\overline{C}_m = C \sup_{1 \leq j \leq m} \|u_j\|_{H^{k+1}(\Omega)}$, on a donc

$$\forall j,\ 1 \leq j \leq m, \quad \|u_j - r_h u_j\|_{H^1(\Omega)} \leq \overline{C}_m h^k. \tag{3.29}$$

Soit maintenant $v \in W_m$, avec $\|v\|_{H^1} = 1$. On décompose v sur la base des u_j :

$$v = \sum_{j=1}^{m} \alpha_j u_j \quad \text{avec} \quad \|v\|_{H^1}^2 = \sum_j \alpha_j^2 = 1.$$

La dernière relation implique que $|\alpha_j| \leq 1$ pour tout j. On calcule maintenant

$$\|v - r_h v\|_{H^1(\Omega)} = \left\| \sum_j \alpha_j (u_j - r_h u_j) \right\|_{H^1(\Omega)} \leq \sum_j |\alpha_j| \, \|u_j - r_h u_j\|_{H^1(\Omega)}.$$

En utilisant $|\alpha_j| \leq 1$ et la majoration (3.29), on arrive à

$$\|v - r_h v\|_{H^1(\Omega)} \leq \sum_{j=1}^{m} \overline{C}_m h^k = C_m h^k,$$

ce qui est exactement (3.28).

En rassemblant (3.26) et (3.28), on a donc, pour tout $v \in W_m$ de norme 1, que

$$\|v - \Pi_h v\|_{H^1(\Omega)} \leq \frac{M}{\alpha} C_m h^k,$$

soit $\|\Pi_h v\|_{H^1(\Omega)} \geq 1 - \tilde{C}_m h^k$. Ceci implique $\sigma_{m,h} \geq 1 - \tilde{C}_m h^k$. L'estimation (3.25) donne donc, pour une constante C_m, l'encadrement $\lambda_m \leq \lambda_{m,h} \leq \lambda_m(1+C_m h^k)$, soit

$$0 \leq \lambda_{m,h} - \lambda_m \leq C_m h^k. \tag{3.30}$$

Nous finissons cette section en énonçant un résultat précis de convergence pour les valeurs propres et les vecteurs propres du laplacien, définis par (3.16), approximés par une méthode d'éléments finis triangulaires P_k. Un tel résultat se généralise à d'autres problèmes et d'autres types d'éléments finis.

Théorème 3.15. *Soit Ω un ouvert borné et régulier de $\mathbb{R}^d$. Soit $(\mathcal{T}_h)_{h\geq 0}$ une suite de maillages triangulaires réguliers de Ω. Soit V_{0h} le sous-espace de $H^1_0(\Omega)$ défini par la méthode des éléments finis P_k, de dimension J.*

Soient $(\lambda_m, u_m)_{m\geq 1}$ les valeurs propres et vecteurs propres du problème (3.16), et soit $(\lambda_{m,h})_{1\leq m\leq J}$ les valeurs propres de l'approximation variationnelle (3.18) correspondante sur l'espace de dimension finie V_{0h}. Pour tout $m \geq 1$ fixé, on a

$$\lim_{h\to 0} |\lambda_m - \lambda_{m,h}| = 0.$$

Il existe une famille de vecteurs propres $(u_{m,h})_{1\leq m\leq J}$ de (3.18) dans V_{0h} telle que, si λ_m est valeur propre simple, alors

$$\lim_{h\to 0} \|u_m - u_{m,h}\|_{H^1(\Omega)} = 0.$$

Si le sous-espace engendré par $(u_1, \ldots, u_m)$ est inclus dans $H^{k+1}(\Omega)$ avec $k+1 > d/2$, alors il existe C_m indépendant de h tel que

$$|\lambda_m - \lambda_{m,h}| \leq C_m h^{2k}. \tag{3.31}$$

Si λ_m est valeur propre simple, alors

$$\|u_m - u_{m,h}\|_{H^1(\Omega)} \leq C_m h^k. \tag{3.32}$$

Il est important à ce stade de faire plusieurs remarques :

— la constante C_m dans (3.31) et (3.32) tend vers $+\infty$ lorsque m tend vers $+\infty$. Donc, à h fixé, les plus grandes valeurs propres discrètes (par exemple, $\lambda_{J,h}$) ne sont pas nécessairement une bonne approximation des valeurs propres exactes. Pour avoir une bonne approximation de λ_J, il peut donc être nécessaire de travailler avec un espace d'approximation V_{0h} de dimension bien plus grande que J.

— la convergence des vecteurs propres ne peut s'obtenir que si la valeur propre est simple. Si λ_m est multiple, alors il se peut que la suite $u_{m,h}$ ne converge pas, mais admette plusieurs points d'accumulation, qui sont des combinaisons linéaires de vecteurs propres associés à λ_m.

— l'ordre de convergence des valeurs propres est le double de celui pour les vecteurs propres [1]. On retrouvera ce phénomène (lié au caractère auto-adjoint de l'opérateur) dans les algorithmes de calcul des valeurs propres et vecteurs propres d'une matrice (cf. par exemple la proposition 3.17).

3.4 Algorithmes pour le calcul de valeurs et de vecteurs propres

Les valeurs propres d'une matrice sont les racines de son polynôme caractéristique $P(\lambda) = \det(A - \lambda \mathrm{Id})$. Cependant, il n'existe pas de méthodes directes (c'est-à-dire qui donnent le résultat en un nombre fini d'opérations) pour calculer les racines d'un polynôme quelconque, dès que son ordre est supérieur ou égal à 5. De plus, tout polynôme est le polynôme caractéristique d'une matrice, donc le calcul des valeurs propres d'une matrice est un problème aussi difficile que celui du calcul des racines d'un polynôme quelconque.

Calculer les valeurs propres d'une matrice est en fait un problème beaucoup plus difficile que la résolution d'un système linéaire. Il n'existe que des méthodes itératives. Nous nous concentrons dans cette section sur le cas des matrices réelles symétriques, pour lesquelles le problème est plus simple.

Nous mentionnons ici trois méthodes typiques pour une matrice symétrique :

— la méthode de la puissance, analysée dans la section 3.4.1. C'est la méthode la plus simple, mais elle ne permet (au mieux) que de calculer les valeurs propres de plus grande et de plus petite valeur absolue.

— la méthode de Given-Householder, qui permet de calculer une ou plusieurs valeurs propres de rang quelconque sans avoir à calculer toutes les valeurs propres. Cette méthode est en fait la concaténation de deux algorithmes, l'algorithme de Householder qui permet de transformer une matrice symétrique en une matrice tridiagonale de mêmes valeurs propres, et l'algorithme de Givens qui permet le calcul des valeurs propres d'une matrice tridiagonale. Nous n'en dirons pas plus et renvoyons à la bibliographie pour plus de détails.

— la méthode de Lanczos, analysée dans la section 3.4.2. Comme l'algorithme de gradient conjugué, cette méthode fait appel aux espaces de Krylov. Nous en décrirons ci-dessous l'esprit. Cette méthode est à la base de nombreux développements récents qui conduisent aux méthodes les plus efficaces pour de grandes matrices creuses.

1. On voit aussi que l'estimation (3.30) sur les valeurs propres n'est pas optimale, si la forme bilinéaire a correspond au laplacien.

3.4.1 Méthode de la puissance

Il s'agit de la méthode la plus simple pour calculer la valeur propre de plus grande (ou de plus petite) valeur absolue. Une limitation de la méthode est que cette valeur propre doit être simple.

Algorithme 3.16 (Méthode de la puissance)**.** *Soit A une matrice symétrique réelle d'ordre n, et ε une précision souhaitée.*

1. *Initialisation : soit $x_0 \in \mathbb{R}^n$ avec $\|x_0\| = 1$.*
2. *Itération : pour $k \geq 1$,*
   (a) *on calcule $y_k = Ax_{k-1}$.*
   (b) *on pose $x_k = y_k/\|y_k\|$.*
   (c) *test de convergence : si $\|x_k - x_{k-1}\| \leq \varepsilon$, on s'arrête.*

La proposition suivante indique sous quelles conditions et à quelle vitesse cet algorithme converge.

Proposition 3.17. *On suppose que A est une matrice réelle symétrique de taille n, de valeurs propres $(\lambda_1, \ldots, \lambda_n)$ rangées par ordre de valeur absolue croissante, et que λ_n est positive et simple : $|\lambda_1| \leq \ldots \leq |\lambda_{n-1}| < \lambda_n$. Soit $(e_1, \ldots, e_n)$ une base de vecteurs propres orthonormés. On suppose que x_0 n'est pas orthogonal à e_n. Alors la méthode de la puissance converge, au sens où*

$$\lim_{k\to+\infty} \|y_k\| = \lambda_n, \quad \lim_{k\to+\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_n.$$

La convergence est géométrique, avec une vitesse proportionnelle à $|\lambda_{n-1}|/|\lambda_n|$:

$$|\|y_k\| - \lambda_n| \leq C \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{k}.$$

Remarque 3.18. *Comme on l'a remarqué dans le théorème 3.15, la convergence de la valeur propre se fait à un ordre deux fois plus grand que la convergence du vecteur propre.*

Démonstration. On décompose le vecteur initial sur les vecteurs propres de A : $x_0 = \sum_{i=1}^n \beta_i e_i$, avec $\beta_n \neq 0$ par hypothèse. Le vecteur x_k est proportionnel à $A^k x_0 = \sum_{i=1}^n \beta_i \lambda_i^k e_i$ et de norme 1, donc

$$x_k = \frac{\beta_n e_n + \displaystyle\sum_{i=1}^{n-1} \beta_i (\lambda_i/\lambda_n)^k e_i}{\left(\beta_n^2 + \displaystyle\sum_{i=1}^{n-1} \beta_i^2 (\lambda_i/\lambda_n)^{2k}\right)^{1/2}}. \tag{3.33}$$

Comme $|\lambda_i| < \lambda_n$, on voit que x_k converge vers $x_\infty = \text{signe}\,(\beta_n)e_n$. On déduit de (3.33) que

$$y_{k+1} = \frac{\beta_n \lambda_n e_n + \sum_{i=1}^{n-1} \beta_i (\lambda_i/\lambda_n)^k \lambda_i e_i}{\left(\beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 (\lambda_i/\lambda_n)^{2k}\right)^{1/2}},$$

ce qui donne la convergence de $\|y_{k+1}\|$ vers λ_n au rythme $|\lambda_{n-1}/\lambda_n|^{2k}$. □

On est souvent intéressé par le calcul des valeurs propres petites. L'algorithme suivant, très inspiré de la méthode de la puissance, permet de calculer la valeur propre de valeur absolue la plus petite.

Algorithme 3.19 (Méthode de la puissance inverse)**.** *Soit A une matrice symétrique réelle inversible d'ordre n, et ε une précision souhaitée.*

1. *Initialisation : soit $x_0 \in \mathbb{R}^n$ avec $\|x_0\| = 1$.*
2. *Itération : pour $k \geq 1$,*
   (a) *résoudre $Ay_k = x_{k-1}$.*
   (b) *on pose $x_k = y_k/\|y_k\|$.*
   (c) *test de convergence : si $\|x_k - x_{k-1}\| \leq \varepsilon$, on s'arrête.*

La proposition suivante indique sous quelles conditions et à quelle vitesse cet algorithme converge.

Proposition 3.20. *On suppose que A est une matrice réelle symétrique inversible de taille n, de valeurs propres $(\lambda_1, \ldots, \lambda_n)$ rangées par ordre de valeur absolue croissante, et que λ_1 est positive et simple : $0 < \lambda_1 < |\lambda_2| \leq \ldots \leq |\lambda_n|$. Soit $(e_1, \ldots, e_n)$ une base de vecteurs propres orthonormés. On suppose que x_0 n'est pas orthogonal à e_1. Alors la méthode de la puissance inverse converge, au sens où*

$$\lim_{k\to+\infty} \frac{1}{\|y_k\|} = \lambda_1, \quad \lim_{k\to+\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_1.$$

La convergence est géométrique, avec une vitesse proportionnelle à $|\lambda_1|/|\lambda_2|$:

$$\left|\|y_k\|^{-1} - \lambda_1\right| \leq C \left|\frac{\lambda_1}{\lambda_2}\right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left|\frac{\lambda_1}{\lambda_2}\right|^{k}.$$

Démonstration. La preuve de cette proposition est similaire à celle de la proposition 3.17. □

3.4.2 Méthode de Lanczos

Cette méthode utilise la notion d'espace de Krylov, qui apparait aussi dans l'algorithme de gradient conjugué, et qu'on rappelle ci-dessous. Comme nous l'avons précisé ci-dessus, cette méthode (et ses généralisations) est très efficace pour les matrices de grande taille. On donne ici l'esprit de la méthode plutôt qu'une description précise d'une implémentation numérique efficace.

Dans toute la suite, A est une matrice symétrique réelle d'ordre n, $r_0 \neq 0$ est un vecteur de $\mathbb{R}^n$ donné, et K_k est l'espace de Krylov associé :

Théorème-Définition 3.21. *Soit $r_0 \neq 0$ un vecteur de $\mathbb{R}^n$ donné. Pour tout $k \geq 1$, l'espace de Krylov K_k associé est*

$$K_k = \mathit{Vect}\ \left\{r_0, Ar_0, \dots, A^k r_0\right\}.$$

Il existe un entier $k_0 \leq n-1$, appelé dimension critique de Krylov, tel que :
- *si $k \leq k_0$, alors la famille $(r_0, \dots, A^k r_0)$ est libre et $\dim K_k = k+1$;*
- *si $k > k_0$, alors $K_k = K_{k_0}$.*

L'algorithme de Lanczos consiste à construire une suite de vecteurs v_j par la formule de récurrence

$$\forall j \geq 2, \quad \hat{v}_j = Av_{j-1} - \langle Av_{j-1}, v_{j-1}\rangle v_{j-1} - \|\hat{v}_{j-1}\| v_{j-2} \ \text{ et } \ v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}, \tag{3.34}$$

avec les initialisations $v_0 = 0$ et $v_1 = r_0/\|r_0\|$. On montrera ci-dessous que, tant que $j \leq k_0 + 1$, on a $\hat{v}_j \neq 0$ et donc v_j est bien défini, tandis que $\hat{v}_{k_0+2} = 0$. La relation entre les v_j et les espaces de Krylov sera explicitée dans le lemme ci-dessous.

Pour tout entier $k \leq k_0 + 1$, on définit la matrice V_k de taille $n \times k$ dont les colonnes sont les vecteurs $v_1, \dots, v_k$, ainsi que la matrice symétrique tridiagonale de taille $k \times k$ définie par

$$(T_k)_{i,i} = \langle Av_i, v_i\rangle, \quad (T_k)_{i,i+1} = (T_k)_{i+1,i} = \|\hat{v}_{i+1}\|, \quad (T_k)_{i,j} = 0 \text{ sinon.}$$

Lemme 3.22. *Pour tout $j \leq k_0 + 1$, on a $\hat{v}_j \neq 0$ et donc v_j est bien défini, tandis que $\hat{v}_{k_0+2} = 0$.*

Pour $1 \leq k \leq 1 + k_0$, la famille $(v_1, \dots, v_k)$ coïncide avec la base orthonormée de l'espace de Krylov K_{k-1} construite par le procédé de Gram-Schmidt appliqué à la famille $(r_0, \dots, A^{k-1} r_0)$.

Soit e_k le k-ième vecteur de la base canonique de $\mathbb{R}^k$, et Id_k la matrice identité de taille $k \times k$. Alors, pour $1 \leq k \leq 1 + k_0$, on a

$$AV_k = V_k T_k + \hat{v}_{k+1} e_k^t \tag{3.35}$$

et

$$V_k^t A V_k = T_k \quad \textit{et} \quad V_k^t V_k = \mathrm{Id}_k. \tag{3.36}$$

Démonstration. On introduit la suite de vecteurs w_j définie par $w_0 = 0$, $w_1 = r_0/\|r_0\|$ et, pour $j \geq 2$,

$$\hat{w}_j = Aw_{j-1} - \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i \rangle w_i \quad \text{et} \quad w_j = \frac{\hat{w}_j}{\|\hat{w}_j\|}. \tag{3.37}$$

On montrera ci-dessous que $w_j = v_j$. On montre par récurrence que les vecteurs w_j (tant qu'ils existent) sont orthonormés. Supposons que ce soit vrai jusqu'au rang $j-1$: pour tout $p, q \leq j-1$, on suppose que $\langle w_q, w_p \rangle = \delta_{qp}$. On prouve maintenant l'hypothèse de récurrence au rang j. Soit $p \leq j-1$: alors

$$\begin{aligned} \langle \hat{w}_j, w_p \rangle &= \langle Aw_{j-1}, w_p \rangle - \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i \rangle \langle w_i, w_p \rangle \\ &= \langle Aw_{j-1}, w_p \rangle - \langle Aw_{j-1}, w_p \rangle = 0, \end{aligned}$$

donc $\langle w_j, w_p \rangle = \delta_{pj}$ pour tout $p \leq j$, ce qui donne l'hypothèse de récurrence au rang j.

Par récurrence, on montre aussi que $w_j \in K_{j-1}$, tant que les vecteurs w_j existent.

Supposons maintenant que l'algorithme stoppe à l'indice j (c'est-à-dire que j est le premier indice tel que $\hat{w}_j = 0$), avec $j \leq k_0 + 1$. Alors

$$Aw_{j-1} = \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i \rangle w_i. \tag{3.38}$$

Or $w_i \in K_{i-1}$ pour tout $i \leq j-1$, donc on a $w_i = \sum_{p=0}^{i-1} \beta_i^p A^p r_0$. On insère cette décomposition dans (3.38), ce qui donne

$$\sum_{p=0}^{j-2} \beta_{j-1}^p A^{p+1} r_0 = \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i \rangle \sum_{p=0}^{i-1} \beta_i^p A^p r_0,$$

soit, en isolant le terme de plus haut degré à gauche,

$$\beta_{j-1}^{j-2} A^{j-1} r_0 = \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i \rangle \sum_{p=0}^{i-1} \beta_i^p A^p r_0 - \sum_{p=0}^{j-3} \beta_{j-1}^p A^{p+1} r_0.$$

Le vecteur du membre de droite est dans K_{j-2}. Comme $j-1 \leq k_0$, la famille $(r_0, \ldots, A^{j-1} r_0)$ est libre, donc $A^{j-1} r_0 \notin K_{j-2}$. Donc $\beta_{j-1}^{j-2} = 0$. Par conséquent, la décomposition de w_{j-1} s'écrit

$$w_{j-1} = \sum_{p=0}^{j-3} \beta_i^p A^p r_0 \in K_{j-3}.$$

Donc la famille $(w_1, \ldots, w_{j-1})$ est une famille de $j-1$ vecteurs orthogonaux deux à deux et qui appartiennent tous à K_{j-3}, qui est de dimension $j-2$. Ceci est contradictoire : donc l'algorithme stoppe à un indice $j > k_0 + 1$.

Supposons maintenant que $\hat{w}_{k_0+2} \neq 0$. Alors la famille $(w_1, \ldots, w_{k_0+2})$ est une famille de k_0+2 vecteurs orthogonaux deux à deux et qui appartiennent tous à $K_{k_0+1} = K_{k_0}$, qui est de dimension k_0+1. Ceci est à nouveau contradictoire. Donc l'algorithme stoppe exactement à l'indice k_0+2.

Pour tout $j \leq k_0+1$, la famille $(w_1, \ldots, w_j)$ est une famille de j vecteurs orthonormés et qui appartiennent tous à K_{j-1}, qui est de dimension j : donc cette famille constitue une base orthonormée de K_{j-1}, qui coïncide avec la base orthonormée construite par le procédé de Gram-Schmidt appliqué à la famille $(r_0, \ldots, A^{j-1}r_0)$.

On montre maintenant que $w_j = v_j$ pour tout $j \leq k_0+1$. Comme A est symétrique, on a

$$\begin{aligned} \langle Aw_p, w_{j-1}\rangle &= \langle w_p, Aw_{j-1}\rangle \\ &= \langle w_p, \hat{w}_j\rangle + \sum_{i=1}^{j-1} \langle Aw_{j-1}, w_i\rangle\langle w_p, w_i\rangle. \end{aligned}$$

Supposons $j \leq p-1$: alors, pour les i tels que $1 \leq i \leq j-1$, on a $i \leq p-2 < p$ et $\langle w_p, w_i\rangle = 0$. Donc, pour $j \leq p-1$, on a $\langle Aw_p, w_{j-1}\rangle = 0$. On voit aussi que

$$\langle Aw_p, w_{p-1}\rangle = \langle w_p, \hat{w}_p\rangle = \|\hat{w}_p\|.$$

Donc la récurrence (3.37) définissant $\hat{w}_j$ se récrit

$$\begin{aligned} \hat{w}_j &= Aw_{j-1} - \langle Aw_{j-1}, w_{j-1}\rangle w_{j-1} - \langle Aw_{j-1}, w_{j-2}\rangle w_{j-2} \\ &= Aw_{j-1} - \langle Aw_{j-1}, w_{j-1}\rangle w_{j-1} - \|\hat{w}_{j-1}\| w_{j-2}, \end{aligned}$$

ce qui est exactement la récurrence (3.34). Par conséquent, on a bien $w_j = v_j$ pour tout $j \leq k_0+1$.

On montre maintenant (3.35). La colonne p de la matrice AV_k est exactement, pour $1 \leq p \leq k$, égale à

$$\mathrm{Col}_p(AV_k) = Av_p = \hat{v}_{p+1} + \langle Av_p, v_p\rangle v_p + \|\hat{v}_p\| v_{p-1}.$$

Un simple calcul montre que les colonnes de V_kT_k sont

$$\begin{aligned} \forall p, \quad 2 \leq p \leq k-1, \quad \mathrm{Col}_p(V_kT_k) &= \hat{v}_{p+1} + \langle Av_p, v_p\rangle v_p + \|\hat{v}_p\| v_{p-1}, \\ \mathrm{Col}_1(V_kT_k) &= \hat{v}_2 + \langle Av_1, v_1\rangle v_1, \\ \mathrm{Col}_k(V_kT_k) &= \langle Av_k, v_k\rangle v_k + \|\hat{v}_k\| v_{k-1}. \end{aligned}$$

Enfin, la colonne p de $\hat{v}_{k+1}e_k^t$ est nulle si $p < k$, tandis que la colonne k vaut exactement $\hat{v}_{k+1}$. On a donc bien la relation (3.35).

Les vecteurs v_k étant orthogonaux deux à deux et de norme 1, on a $V_k^tV_k = \mathrm{Id}_k$. On multiplie enfin à gauche la relation (3.35) par V_k^t : du fait que $\hat{v}_{k+1}$ est orthogonal aux v_j pour $j \leq k$, on a $V_k^t\hat{v}_{k+1} = 0$ et on obtient finalement la relation (3.36). $\square$

Nous comparons maintenant les valeurs propres de A et celle de la matrice T_{k_0+1}. Notons que ces deux matrices ne sont pas en général de même taille. On note $\lambda_1 < \lambda_2 < \ldots < \lambda_m$ les valeurs propres distinctes de la matrice A qui est de taille $n \times n$ (donc $1 \leq m \leq n$), et soient P_i les matrices de projection orthogonale sur les sous-espaces propres correspondants de A. Par construction,

$$A = \sum_{i=1}^{m} \lambda_i P_i, \quad \mathrm{Id}_n = \sum_{i=1}^{m} P_i, \quad P_i P_j = 0 \text{ si } i \neq j, \quad P_i^2 = P_i \text{ pour tout } i.$$

Lemme 3.23. *Les valeurs propres de T_{k_0+1} sont aussi valeurs propres de A.*

Réciproquement, si on suppose que $P_i r_0 \neq 0$ pour tout i, alors toutes les valeurs propres de A sont aussi valeurs propres de T_{k_0+1} et $k_0 + 1 = m$. Les valeurs propres de T_{k_0+1} sont simples.

Dans le cas où $P_i r_0 \neq 0$ pour tout i, la récurrence de Lanczos permet donc de construire une matrice T_{k_0+1} qui est tridiagonale et dont les valeurs propres sont exactement les valeurs de A. On pourrait alors penser calculer les valeurs propres de A de la façon suivante :

- on applique la récurrence de Lanczos jusqu'à l'ordre $k_0 + 1$, ce qui permet de construire la matrice T_{k_0+1}.
- on calcule les valeurs propres de la matrice T_{k_0+1}. Le problème sur T_{k_0+1} est plus simple que le problème initial sur A, car T_{k_0+1} est tridiagonale et il existe des algorithmes pour le calcul des valeurs propres qui sont spécifiques aux matrices tridiagonales, comme l'algorithme de Givens.
- comme (dans les bons cas) T_{k_0+1} a exactement les mêmes valeurs propres que A, on a ainsi calculé les valeurs propres de A.

Une telle approche n'est cependant pas la meilleure façon d'exploiter la récurrence de Lanczos, à cause d'instabilités numériques liées à des erreurs d'arrondi. Une bonne façon d'exploiter la récurrence de Lanczos sera donnée par le lemme 3.24 ci-dessous. On démontre maintenant le lemme 3.23.

Démontration du lemme 3.23. Soit λ valeur propre de T_{k_0+1}, et soit $y \neq 0$ un vecteur propre associé : $T_{k_0+1} y = \lambda y$. Comme $\hat{v}_{k_0+2} = 0$, on déduit de (3.35) que $A V_{k_0+1} = V_{k_0+1} T_{k_0+1}$, et donc que

$$A V_{k_0+1} y = \lambda V_{k_0+1} y.$$

Si $V_{k_0+1} y = 0$, alors les colonnes de V_{k_0+1} sont liées (puisque $y \neq 0$), ce qui est contradictoire avec le fait que la famille $(v_1, \ldots, v_{k_0+1})$ forme une base orthonormée de K_{k_0}. Donc $V_{k_0+1} y \neq 0$, et λ est valeur propre de A.

Réciproquement, on suppose que $P_i r_0 \neq 0$ pour tout $1 \leq i \leq m$. Supposons la famille $(P_1 r_0, \ldots, P_m r_0)$ liée : alors, par exemple, il existe $\alpha_1, \ldots, \alpha_{m-1}$ tels que

$$P_m r_0 = \sum_{i=1}^{m-1} \alpha_i P_i r_0.$$

Comme $P_m P_i = 0$ pour tout $i < m$, on obtient $0 = P_m^2 r_0 = P_m r_0$, ce qui est contradictoire avec les hypothèses. Donc la famille $(P_1 r_0, \ldots, P_m r_0)$ est libre et $E_m =$ Vect $\{P_1 r_0, \ldots, P_m r_0\}$ est de dimension m.

Montrons que $m = k_0 + 1$. On voit que $A^k r_0 = \sum_{i=1}^m \lambda_i^k P_i r_0$ donc $A^k r_0 \in E_m$, et par conséquent $K_k \subset E_m$ pour tout k. Donc $k_0 + 1 = \dim K_{k_0} \leq \dim E_m = m$.

On montre l'inégalité inverse. La famille $(P_1 r_0, \ldots, P_m r_0)$ est libre. On se place dans cette base. La famille $(r_0, \ldots, A^{m-1} r_0)$ est représentée dans cette base par la matrice

$$M = \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \ldots & \lambda_1^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \ldots & \lambda_m^{m-1} \end{pmatrix},$$

qui est une matrice de Van Der Monde inversible car les λ_j sont distincts deux à deux. Donc la famille $(r_0, \ldots, A^{m-1} r_0)$ est libre, ce qui implique $m - 1 \leq k_0$. On a donc bien $m = k_0 + 1$ et $E_m = K_{k_0}$.

Soit λ_i une valeur propre de A : le vecteur $P_i r_0$ est vecteur propre associé. Or $P_i r_0 \in E_m = K_{k_0}$, et les colonnes de V_{k_0+1} forment une base orthonormée de K_{k_0}. Donc il existe $y \neq 0$ tel que $V_{k_0+1} y = P_i r_0$. La relation (3.36) donne

$$\begin{aligned} T_{k_0+1} y &= V_{k_0+1}^t A V_{k_0+1} y \\ &= V_{k_0+1}^t A P_i r_0 \\ &= \lambda_i V_{k_0+1}^t P_i r_0 \\ &= \lambda_i V_{k_0+1}^t V_{k_0+1} y = \lambda_i y, \end{aligned}$$

donc λ_i est aussi valeur propre de T_{k_0+1}. La matrice A possède $m = k_0 + 1$ valeurs propres distinctes, et toutes ces valeurs propres sont aussi valeurs propres de T_{k_0+1}, qui est de dimension $k_0 + 1$. Donc les valeurs propres de T_{k_0+1} sont simples. □

Comme nous l'avons précisé plus haut, la bonne façon d'exploiter la récurrence de Lanczos n'est pas de calculer la matrice T_{k_0+1} pour ensuite la diagonaliser. Il est plus intéressant d'exploiter le lemme que nous donnons maintenant :

Lemme 3.24. *Soit un entier k, $1 \leq k \leq k_0 + 1$. Soit λ valeur propre de T_k et soit $y \in \mathbb{R}^k$ un vecteur propre associé. Alors il existe une valeur propre λ_i de la matrice A telle que*

$$|\lambda - \lambda_i| \leq \sqrt{m}\ \|\hat{v}_{k+1}\|\ \frac{|\langle e_k, y \rangle|}{\|y\|},$$

où e_k est le k-ième vecteur de la base canonique de $\mathbb{R}^k$.

Ce lemme vient compléter la discussion qui fait suite au lemme 3.23. Une façon efficace d'utiliser la récurrence de Lanczos est en effet la suivante : si la dernière composante d'un vecteur propre de T_k est petite, i.e. $|\langle e_k, y \rangle| \ll \|y\|$, alors la valeur propre correspondante est une bonne approximation d'une valeur propre de A. Ainsi, le calcul (d'une approximation) des valeurs propres de A passe toujours

par la diagonalisation de la matrice T_k. Cependant, le lemme ci-dessus donne une estimation d'erreur qu'il est possible d'évaluer en pratique.

Démonstration. Soit λ valeur propre de T_k et y vecteur propre associé : $T_k y = \lambda y$. La relation (3.35) donne

$$AV_k y = \lambda V_k y + \langle y, e_k \rangle \hat{v}_{k+1}.$$

En utilisant les projections P_i, on a donc

$$\sum_{i=1}^{m} (\lambda_i - \lambda) P_i V_k y = \langle y, e_k \rangle \hat{v}_{k+1}.$$

Soit $\varepsilon_j = \text{signe}\,(\lambda_j - \lambda)$, on prend le produit scalaire de l'égalité ci-dessus avec $\varepsilon_j P_j V_k y$:

$$\varepsilon_j (\lambda_j - \lambda) \|P_j V_k y\|^2 = \langle y, e_k \rangle \varepsilon_j \langle \hat{v}_{k+1}, P_j V_k y \rangle.$$

On somme sur les j, avec $\varepsilon_j(\lambda_j - \lambda) = |\lambda_j - \lambda| \geq \min_i |\lambda_i - \lambda|$:

$$\min_i |\lambda_i - \lambda| \sum_{j=1}^{m} \|P_j V_k y\|^2 \leq \langle y, e_k \rangle \sum_{j=1}^{m} \varepsilon_j \langle \hat{v}_{k+1}, P_j V_k y \rangle.$$

Or $\sum_{j=1}^{m} \|P_j V_k y\|^2 = \|V_k y\|^2 = \|y\|^2$, donc

$$\begin{aligned}
\min_i |\lambda_i - \lambda| &\leq \frac{|\langle e_k, y \rangle|}{\|y\|^2} \left| \sum_{j=1}^{m} \varepsilon_j \langle \hat{v}_{k+1}, P_j V_k y \rangle \right| \\
&\leq \frac{|\langle e_k, y \rangle|}{\|y\|^2} \sum_{j=1}^{m} \|\hat{v}_{k+1}\| \, \|P_j V_k y\| \\
&\leq \frac{|\langle e_k, y \rangle|}{\|y\|^2} \|\hat{v}_{k+1}\| \sqrt{m} \sqrt{\sum_{j=1}^{m} \|P_j V_k y\|^2}.
\end{aligned}$$

En utilisant à nouveau $\sum_{j=1}^{m} \|P_j V_k y\|^2 = \|y\|^2$, on obtient le résultat annoncé. □

Chapitre 4

Introduction aux problèmes d'évolution

Cette deuxième partie du cours constitue en une brève introduction à l'étude mathématique et numérique d'équations aux dérivées partielles dépendant du temps. En fait, nous étudierons surtout l'équation de la chaleur dans le cadre de ce cours mais nous évoquerons brièvement d'autres types d'équations d'évolution. Nous renvoyons à [7, 1, 5, 13] pour une présentation plus détaillée.

4.1 Exemples d'équations d'évolution

Avant toute chose, commen_cons par donner une liste de plusieurs équations d'évolution courantes (arbitrairement sélectionnées !).

Bien sûr, pour que ce que nous écrivons ait un sens mathématique précis, il faudra ajouter une (des) condition(s) initiales et préciser quel sens on donne aux dérivées. Pour simplifier, le lecteur pourra supposer que les fonctions apparaissant dans les équations suivantes sont suffisamment dérivables par rapport à toutes leurs variables. Nous verrons plus loin que définir des solutions en un sens plus faible pourra être très utile. C'est même indispensable lorsque l'on cherche à décrire certains phénomènes physiques possédant des singularités.

Equation de transport linéaire

Il s'agit de l'équation

$$\frac{\partial u}{\partial t} + b \cdot \nabla u = f$$

qui décrit des phénomènes comme le transfert de chaleur, de masse, etc (u est la densité). La fonction f est le terme source et b est la direction d'écoulement.

L'équation de Boltzmann (ou de Liouville/Vlasov si $f = 0$)

$$\frac{\partial u}{\partial t} + p \cdot \nabla_x u + F \cdot \nabla_p u = f$$

décrit elle l'évolution de la distribution statistique $u(t,x,p)$ des particules d'un fluide dans l'espace des phases $\mathbb{R}^3\times\mathbb{R}^3 \ni (x,p)$. La fonction F modélise les forces appliquées à chaque particule et la fonction f décrit les collisions entre les particules.

Equation de la chaleur

L'équation de la chaleur est le prototype de toute une famille d'équations appelées *paraboliques* :

$$\frac{\partial u}{\partial t} - \Delta u = f. \tag{4.1}$$

Typiquement, $u(t,x)$ représente la température au temps t et au point $x \in \Omega$ d'un matériau homogène situé dans un domaine $\Omega \subset \mathbb{R}^3$. La fonction f s'interprète alors comme une source de chaleur. Si Ω est borné, il faut ajouter des conditions aux bords (de type Dirichlet ou Neumann par exemple).

En fait la même équation intervient dans de très nombreuses situations : par exemple u peut aussi modéliser la diffusion d'une concentration dans le domaine Ω, ou l'évolution du champ de pression d'un fluide s'écoulant en milieu poreux, ou encore la loi d'un mouvement brownien dans Ω. Des généralisations de l'équation (4.1) peuvent permettre de décrire des matériaux non homogènes ou en présence d'un effet convectif. Elles peuvent devenir non linéaires (de type réaction-diffusion par exemple) :

$$\frac{\partial u}{\partial t} - \Delta u = F(u). \tag{4.2}$$

L'équation de la chaleur (4.1) est également la base de plusieurs problèmes de frontières libres, comme le problème de l'obstacle parabolique

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = -1_{u>0}, \\ u \geq 0 \end{cases}$$

qui décrit un système composé de deux phases (par exemple un gla_con plongé dans de l'eau). Il y a ici deux inconnues : la solution, et le domaine dans lequel l'équation est vérifiée. On appelle frontière libre le bord de l'ensemble $\{u = 0\}$, qui évolue au cours du temps. Des modèles similaires interviennent en finance mathématique (modèles de type Black-Scholes).

Equation des ondes

Il s'agit d'une équation du deuxième ordre en temps :

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \tag{4.3}$$

qui est le prototype d'une famille d'équations appelées *hyperboliques*. Lorsqu'elle est posée dans un domaine borné $\Omega \subset \mathbb{R}^2$, u peut représenter le déplacement vertical

d'une membrane élastique (des conditions de Dirichlet au bord de Ω signifient alors que la membrane est attachée). De manière générale, c'est l'équation adaptée à la description de phénomènes vibratoires comme la propagation d'ondes sonores, lumineuses ou à la surface de l'eau. Elle intervient en acoustique, électromagnétisme, dynamique des fluides...

Tout comme l'équation de la chaleur, l'équation des ondes (4.3) est bien sûr la base de nombreux modèles plus complexes, comme par exemple l'équation des ondes non linéaire

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = F(u). \tag{4.4}$$

Equation de Schrödinger

L'équation de Schrödinger

$$-i\frac{\partial u}{\partial t} - \Delta u + Vu = 0 \tag{4.5}$$

ressemble beaucoup à l'équation de la chaleur (si $V = 0$) puisque c'est comme si on avait remplacé t par it. Elle intervient abondamment en mécanique quantique pour la description de la matière Ã l'échelle microscopique. Si on suppose $\int_\Omega |u(t = 0, x)|^2 dx = 1$, $|u(t, x)|^2 dx$ peut représenter la probabilité de présence au temps t d'une particule quantique dans le vide ($V = 0$) ou en présence d'un champ électrique extérieur V. Grâce Ã la présence du complexe i, on aura $\int_\Omega |u(t, x)|^2 dx = 1$ pour tout $t \in \mathbb{R}$.

Alors que les équations précédentes étaient posées dans l'espace physique ($\mathbb{R}^2$ ou $\mathbb{R}^3$...), l'équation de Schrödinger a la particularité d'être fréquemment étudiée dans des espaces $\mathbb{R}^n$ avec n très grand. Par exemple si on désire décrire l'évolution d'un système comportant N particules quantiques dans l'espace $\mathbb{R}^3$, on aura $u = u(x_1, ..., x_N)$ où chaque $x_i \in \mathbb{R}^3$: $|u(x_1, ..., x_N)|^2$ représente alors la probabilité de présence de trouver la particule numéro k en $x_k \in \mathbb{R}^3$. Ceci rend la description de la matière à l'échelle microscopique très complexe et justifie la nécessité de trouver des modèles valides à des échelles supérieures. La dérivation de tels modèles à partir de l'échelle microscopique est alors d'un grand intérêt.

Autres équations

L'équation de Burgers

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}$$

apparaît en mécanique des fluides ou en acoustique (elle peut modéliser la dynamique d'un gaz par exemple). Lorsque $\nu = 0$,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

on obtient le prototype d'une équation pour laquelle il peut apparaître des discontinuités (ondes de choc).

L'équation d'Airy

$$\frac{\partial u}{\partial t} + \frac{\partial^3 u}{\partial x^3} = 0$$

est elle-même la base de l'équation (non linéaire) de Korteweg-de Vries (KdV)

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0$$

qui est un modèle prototype pour la description d'ondes de type "solitons" comme on peut parfois en observer à la surface de l'eau.

L'équation d'Euler-Bernoulli

$$\frac{\partial u}{\partial t} + \frac{\partial^4 u}{\partial x^4} = 0$$

peut quant à elle modéliser la torsion d'une poutre unidimensionnelle.

4.2 Préliminaires

On se pose généralement différentes questions lors de l'étude d'une équation d'évolution.

La première est bien sûr *l'existence et l'unicité de solutions* dans un espace fonctionnel bien choisi. Pour cela, il pourra être très utile de commencer par choisir des espaces fonctionnels assez "gros", c'est-à-dire contenant beaucoup plus de fonctions que celles qui sont régulières par rapport à toutes leurs variables. On parle alors de *solutions faibles*. Intuitivement, plus l'espace fonctionnel est grand et plus il sera facile de démontrer l'existence de la solution. Mais, on peut aussi se demander quelle est la régularité de la solution lorsque la condition initiale est elle-même régulière ainsi que les autres paramètres de l'équation. On peut ainsi obtenir l'existence et l'unicité de solutions régulières *a posteriori*, qu'on appelle des *solutions fortes*. L'utilisation de solutions faibles peut donc être soit un intermédiaire utile pour démontrer l'existence de solutions plus régulières, soit une nécessité lors de l'étude d'équations pour lesquelles on se s'attend pas à ce que la solution soit ou reste régulière au cours du temps.

Le fait qu'il existe une unique solution à un modèle mathématique n'implique pas automatiquement que le modèle considéré soit un "bon" modèle. Une autre question importante est celle de la dépendance de la solution en fonction des conditions initiales et des divers paramètres apparaissant dans l'équation. D'une part on peut se demander comment la solution varie (dans l'espace fonctionnel choisi) si on change un peu ces paramètres et reste robuste par rapport à de légers changements de ces paramètres. C'est une question d'apparence anodine mais qui est **fondamentale**

lorsque l'on envisage d'utiliser des méthodes de simulation pour pouvoir approcher numériquement les solutions de telles équations. Le mathématicien Jacques Hadamard a donné une définition de ce qu'est un "bon" modèle, en parlant de **problème bien posé**. En notant f les données du modèles (le second membre, les données initiales, le domaine, etc.), u la solution recherchée, et $\mathcal{A}$ l'opérateur qui agit sur u, supposons que le problème considéré soit de trouver u solution d'un problème du type

$$\mathcal{A}(u) = f. \tag{4.6}$$

Définition 4.1. *On dit que le problème (4.6) est bien posé si pour toute donnée f il admet une solution u unique, et si cette solution u dépend continûment de la donnée f.*

La troisième condition, la moins évidente, est pourtant cruciale dans une perspective d'approximation numérique. En effet, faire un calcul numérique d'une solution approchée de (4.6) revient à perturber les données (qui de continues deviennent discrètes) et à résoudre (4.6) pour ces données perturbées. Si de petites perturbations des données conduisent à de grandes perturbations de la solution, il n'y a aucune chance pour que la simulation numérique soit proche de la réalité (ou du moins de la solution exacte). Par conséquent, cette dépendance continue de la solution par rapport aux données est une condition absolument nécessaire pour envisager des simulations numériques précises.

Une question reliée à la question ci-dessus est alors le choix d'un schéma numérique pour approcher les solutions de l'équation dont on peut prouver qu'elle converge en un certain sens vers la solution lorsque les paramètres de discrétisation tendent dans une certaine limite (par exemple un pas de temps ou un pas de maillage aura vocation à tendre vers 0), et de quantifier l'erreur faite par rapport à la solution exacte de l'équation par rapport aux paramètres de la discrétisation.

Enfin, on peut chercher ensuite à décrire un peu plus précisément le comportement de la solution au cours du temps, en particulier en relation avec des motivations physiques (signe de la solution, vitesse de propagation, comportement en temps grand, etc). Le comportement qualitatif peut alors être très différent suivant le type d'équation considérée.

Dans le cadre de ce cours, nous aborderons brièvement quelques-unes de ces questions sur quelques exemples types d'équations d'évolution. Nous étudierons tout particulièrement l'équation de la chaleur.

Nous avons pris le partie dans ce polycopié de commencer par vous présenter tout d'abord une méthode numérique, qui est une des plus anciennes et des plus simples méthodes pour approcher numériquement les solutions de problèmes d'évolution, à savoir la méthode des différences finies. La présentation de cette méthode fera l'objet du Chapitre 5.

Dans le Chapitre 6, nous montrerons les propriétés mathématiques théoriques des solutions de l'équation de la chaleur. Le chapitre 7 sera consacré à l'étude mathématique d'autres types de problèmes, à savoir l'équation de transport et l'équation des ondes. Les notions de théorie spectrale que vous avez vues lors de la première partie du cours seront tout particulièrement utiles pour ces deux chapitres.

Enfin, le Chapitre 8 sera consacré à l'étude d'une autre méthode numérique utilisée pour discrétiser les problèmes d'évolution, à savoir la méthode des éléments finis.

Chapitre 5

Méthode des différences finies

A part dans quelques cas très particuliers, il est impossible de calculer explicitement les solutions des différents modèles présentés ci-dessus. Il est donc nécessaire d'avoir recours au calcul numérique pour estimer qualitativement et quantitativement ces solutions. Le principe de toutes les méthodes de résolution numérique des équations aux dérivées partielles est d'obtenir des valeurs numériques *discrètes* (c'est-à-dire en nombre fini) qui approchent (en un sens convenable à préciser) la solution exacte.

Il existe de nombreuses méthodes d'approximation numérique des solutions d'équations aux dérivées partielles. Nous présentons dans ce chapitre une des plus anciennes et des plus simples, appelée méthode des différences finies (nous verrons plus loin une autre méthode, dite méthode des éléments finis).

Nous nous contenterons ici d'illustrer la méthode des différences finies sur le cas de l'équation de la chaleur, mais il faut savoir que celle-ci est utilisé pour de nombreux autres problèmes.

5.1 Principe de la méthode des différences finies

Nous nous contentons ici de présenter la méthode en dimension 1 pour simplifier l'exposé. Nous considérons l'équation de la chaleur dans le domaine borné $\Omega = (0,1) \subset \mathbb{R}$:

$$\begin{cases} \partial_t u(t,x) - D\partial_{xx} u(t,x) = 0, & \text{pour } (x,t) \in (0,1) \times \mathbb{R}_+^*, \\ u(0,x) = u_0(x), & \text{pour } x \in (0,1), \end{cases} \tag{5.1}$$

avec une condition initiale $u_0 : (0,1) \to \mathbb{R}$ régulière ($\mathcal{C}^\infty([0,1])$ par exemple) et un coefficient de diffusion $D > 0$.

Nous considèrerons deux types de conditions aux bords pour ce problème : soit des conditions aux bords de Dirichlet homogènes, i.e.

$$u(t,0) = u(t,1) = 0, \quad \text{pour } t \in \mathbb{R}_+^*; \tag{5.2}$$

soit des conditions aux bords périodiques, i.e.

$$u(t,0) = u(t,1), \quad \text{pour } t \in \mathbb{R}_+^*. \tag{5.3}$$

Nous supposerons dans toute la suite du chapitre qu'il existe bien une et une seule solution $u(t,x)$ à ce problème, et que celle-ci est une fonction régulière en temps et en espace. Nous verrons dans le chapitre suivant les résultats qui permettent d'affirmer que tel est bien le cas.

Pour définir un schéma numérique basé sur une méthode de différences finies, il est nécessaire d'introduire les différentes *discrétisations* du problème, à savoir la discrétisation en espace et la discrétisation en temps.

Commençons par la discrétisation en espace. Soit $N \in \mathbb{N}^*$ et soit $\Delta x := \frac{1}{N+1}$ un pas d'espace. On définit un maillage régulier de l'intervalle $[0,1]$ comme suit :

$$\forall 0 \leq j \leq N+1, \quad x_j := j\Delta x,$$

de telle sorte que

$$x_0 = 0 < x_1 < x_2 < \cdots < x_N < x_{N+1} = 1.$$

Considérons maintenant la discrétisation en espace. On introduit un pas de temps $\Delta t > 0$ et on définit

$$\forall n \in \mathbb{N}, \quad t_n := n\Delta t.$$

Le but d'une méthode de différences finies est d'approcher les valeurs de la fonction u prises au temps t_n et au point x_j par des quantités u_j^n qui seront calculées par un schéma numérique. Autrement dit,

$$u(t_n, x_j) \approx u_j^n, \quad \forall 0 \leq j \leq N+1, \ \forall n \in \mathbb{N},$$

où il reste à déterminer les valeurs $\left(u_j^n\right)_{0\leq j\leq N+1, n\in\mathbb{N}}$.

Différents schémas numériques de différences finies correspondent à différentes manières de calculer les quantités approchées $\left(u_j^n\right)_{0\leq j\leq N+1, n\in\mathbb{N}}$. Nous en présentons ici quelques-uns.

Les valeurs de $(u_j^0)_{0\leq j\leq N+1}$ sont fixées par la condition initiale. Celle-ci est discrétisée comme suit :

$$\forall 0 \leq j \leq N+1, \quad u_j^0 := u_0(x_j).$$

Comme mentionné ci-dessus, les conditions aux limites de (5.1) peuvent être de plusieurs types, mais leur choix n'intervient pas dans la définition des schémas. Les conditions limites de Dirichlet homogènes (5.2) se discrétisent de la manière suivante :

$$\forall n \in \mathbb{N}^*, \quad u_0^n = u_{N+1}^n = 0.$$

Dans le cas de conditions limites périodiques (5.3), on impose les relations suivantes :

$$\forall n \in \mathbb{N}^*, \quad u_0^n = u_{N+1}^n.$$

Il nous reste à discrétiser l'équation aux dérivées partielles (5.1) en tant que telle. Le principe d'une méthode de différences finies est de remplacer les dérivées intervenant dans l'équation considérée par des *différences finies* en utilisant des formules de Taylor dans lesquelles on néglige les restes. Par exemple, on approche la dérivée seconde en espace (le laplacien en dimension 1) par la formule suivante :

$$-\partial_{xx} u(t_n, x_j) \approx \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2}. \tag{5.4}$$

En effet, la formule (5.4) vient de la formule de Taylor suivante :

$$-u(t_n, x_j - \Delta x) + 2u(t_n, x_j) - u(t_n, x_j - \Delta x) = -(\Delta x)^2 \partial_{xx} u(t_n, x_j) - \frac{(\Delta x)^4}{12} \frac{\partial^4 u}{\partial x^4}(t_n, x_j) + \mathcal{O}\left((\Delta x)^6\right).$$

Si Δx est petit, la formule (5.4) est une "bonne" approximation (elle est naturelle, mais pas unique). La formule (5.4) est dit *centrée* car elle est symmétrique en j.

Remarque 5.1. *La formule (5.4) nécessite de définir les valeurs de u_j^n pour $j \leq 0$ ou $j \geq N+1$. Dans le cas de conditions aux bords de Dirichlet homogènes, on impose*

$$u_j^n = 0 \quad \textit{pour tout } j \leq 0 \textit{ ou } j \geq N+1.$$

Dans le cas de conditions limites périodiques, on impose

$$u_j^n = u_{N+1+j}^n \textit{ pour } j \leq 0 \quad \textit{et} \quad u_j^n = u_{j-(N+1)}^n \textit{ pour } j \geq N+1.$$

Il ne nous reste plus qu'à appprocher la dérivée en temps $\partial_t u$ intervenant dans l'équation. Plusieurs choix sont possibles, et en fonction de ce choix, on obtient différents schémas numériques ayant des propriétés mathématiques différentes. Examinons ici trois choix naturels possibles.

— Une première possibilité est de considérer une approximation aux différences finies par la formule centrée

$$\partial_t u(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$$

ce qui aboutit à un schéma complètement symmétrique par rapport à n et j (appelé schéma centré ou *schéma de Richardson*) :

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + D\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0.$$

Aussi "naturel" et évident soit-il, **ce schéma est incapable de calculer des solutions approchées** de l'équation de la chaleur (5.1) (voir TP !). Pour l'instant, indiquons simplement que la difficulté provient du caractère centré de la différence finie qui approche la dérivée en temps. Dans le reste du chapitre, nous nous concentrerons essentiellement sur l'analyse des deux autres méthodes décentrées mentionnées ci-dessous.

— Un deuxième choix consiste à utiliser un schéma de différences finies *décentré amont*(on remonte le temps ; on parle aussi de schéma d'Euler rétrograde)

$$\partial_t u(t_n, x_j) \approx \frac{u_j^n - u_j^{n-1}}{\Delta t}$$

qui conduit au schéma

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + D\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{5.5}$$

— Un troisième choix consiste à utiliser un schéma de différences finies *décentré aval*(on avance dans le temps ; on parle aussi de schéma d'Euler progressif)

$$\partial_t u(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

qui conduit au schéma

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + D\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{5.6}$$

Notez que, quitte à décaler de 1 l'indice en temps n, le schéma (5.6) est équivalent au schéma

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + D\frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0. \tag{5.7}$$

Les schémas numériques de différences finies peuvent se réécrire de manière équivalente en termes de systèmes matriciels. Détaillons ces systèmes linéaires dans le cas de conditions aux bords de Dirichlet homogènes et de conditions aux bords périodiques.

Dans le cas de conditions aux bords de Dirichlet homogènes, comme $u_0^n = u_{N+1}^n = 0$ pour tout $n \in \mathbb{N}$, il suffit de connaître la formule qui relie les valeurs du vecteur $U^n := (u_j^n)_{1\leq j\leq N} \in \mathbb{R}^N$ aux valeurs du vecteur $U^{n-1} := (u_j^{n-1})_{1\leq j\leq N} \in \mathbb{R}^N$. Dans ce cas, le schéma (5.7) se réécrit de manière équivalente

$$\frac{U^n - U^{n-1}}{\Delta t} + A^{\mathrm{Dir}} U^{n-1} = 0, \tag{5.8}$$

et le schéma (5.5) se réécrit de manière équivalente

$$\frac{U^n - U^{n-1}}{\Delta t} + A^{\mathrm{Dir}} U^n = 0, \tag{5.9}$$

avec $A^{\mathrm{Dir}} \in \mathbb{R}^{N\times N}$ la matrice définie par

$$A^{\mathrm{Dir}} := D \begin{pmatrix} \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & \cdots & 0 & 0 \\ \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 \\ 0 & \cdots & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} \\ 0 & 0 & \cdots & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} \end{pmatrix}.$$

Dans le cas de conditions limites périodiques, comme $u_0^n = u_{N+1}^n$ pour tout $n \in \mathbb{N}$, il suffit de connaître la formule qui relie les valeurs du vecteur $U^n := (u_j^n)_{1\leq j\leq N+1} \in \mathbb{R}^{N+1}$ aux valeurs du vecteur $U^{n-1} := (u_j^{n-1})_{1\leq j\leq N+1} \in \mathbb{R}^{N+1}$. Dans ce cas, le schéma (5.7) se réécrit de manière équivalente

$$\frac{U^n - U^{n-1}}{\Delta t} + A^{\mathrm{Per}} U^{n-1} = 0, \tag{5.10}$$

et le schéma (5.5) se réécrit de manière équivalente

$$\frac{U^n - U^{n-1}}{\Delta t} + A^{\mathrm{Per}} U^n = 0, \tag{5.11}$$

avec $A^{\mathrm{Per}} \in \mathbb{R}^{(N+1)\times(N+1)}$ la matrice définie par

$$A^{\mathrm{Per}} := D \begin{pmatrix} \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & \cdots & 0 & \frac{-1}{\Delta x^2} \\ \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} & 0 \\ 0 & \cdots & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} & \frac{-1}{\Delta x^2} \\ \frac{-1}{\Delta x^2} & 0 & \cdots & \cdots & 0 & \frac{-1}{\Delta x^2} & \frac{2}{(\Delta x)^2} \end{pmatrix}.$$

Le schéma (5.5) est appelé schéma d'Euler *implicite* et le schéma (5.7) est appelé schéma d'Euler explicite. Cette dénomination vient de la remarque suivante : la formule (5.7) (ou de manière équivalente les formules 5.8 et (5.10)) donne une expression explicite des valeurs de $(u_j^n)_{0\leq j\leq N+1}$ en fonction des valeurs précédentes de $(u_j^{n-1})_{0\leq j\leq N+1}$ (ou de manière équivalente du vecteur U^n en fonction du vecteur U^{n-1}). En effet, pour le schéma d'Euler explicite, on a alors

$$U^n = U^{n-1} - \Delta t A U^{n-1} = (I - \Delta t A) U^{n-1},$$

avec $A = A^{\mathrm{Dir}}$ ou $A = A^{\mathrm{Per}}$ en fonction du type de conditions aux limites considérées et I la matrice identité.

A contrario, la formule (5.5) (ou de manière équivalente les formules 5.9 et (5.11)) indique qu'il est nécessaire de résoudre un système d'équations linéaires pour calculer les valeurs $(u_j^n)_{0\leq j\leq N-1}$ en fonction des valeurs précédentes $(u_j^{n-1})_{0\leq j\leq N+1}$. En effet, pour le schéma d'Euler implicite, on a alors

$$U^n = (I + \Delta t A)^{-1} U^{n-1}.$$

Il existe également beaucoup d'autres schémas ! Un des buts de l'analyse numérique va être de comparer et de sélectionner les meilleurs schémas suivant des critères de précision, de coût ou de robustesse.

Remarque 5.2. *S'il y a un second membre $f(t,x)$ dans l'équation de la chaleur (5.1), c'est-à-dire si l'équation aux dérivées partielles à résoudre s'écrit*

$$\partial_t u(t,x) - \partial_{xx} u(t,x) = f(t,x),$$

alors les schémas se modifient en remplaçant 0 au second membre par une approximation de $f(t,x)$ au point (t_n, x_j). Par exemple, si on choisit l'approximation $f(t_n, x_j)$, le schéma explicite (5.6) devient

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + D\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = f(t_n, x_j),$$

ou de manière équivalente

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + D\frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = f(t_{n-1}, x_j),$$

Dans le cas de conditions aux limites de Dirichlet, en notant $F^{n-1} := (f(t_{n-1}, x_j))_{1\leq j\leq N}$, on obtient alors l'expression du vecteur U^n en fonction du vecteur U^{n-1} comme suit

$$\frac{U^n - U^{n-1}}{\Delta t} + A^{\mathrm{Dir}} U^{n-1} = F^{n-1}.$$

Exercice 5.3. *Ecrire l'expression de U^n en fonction du vecteur U^{n-1} et du vecteur $F^n := (f(t_n, x_j))_{1\leq j\leq N}$ dans le cas d'un problème de la chaleur avec second membre et d'un schéma d'Euler implicite avec conditions limites de Dirichlet homogènes.*

5.2 Consistance et précision

Bien sûr les formules des schémas présentées ci-dessus ne sont pas choisies au hasard : elles résultent d'une approximation de l'équation par développement de Taylor comme nous l'avons expliqué plus haut. Pour formaliser cette approximation de l'équation aux dérivées partielles par des différences finies, on introduit la

notion de *consistance* et de *précision*. Bien que pour l'instant nous ne considérons que l'équation de la chaleur (5.1), nous allons donner une définition de consistance valable pour n'importe quelle équations aux dérivées partielles que nous notons

$$F(u) = 0.$$

Remarquons que $F(u)$ est une notation pour une fonction de u et de ses dérivées partielles en tout point (t, x). De manière générale, un schéma aux différences finies est défini, pour tous les indices possibles n, j par la formule

$$F_{\Delta t,\Delta x}\left(\{u_{j+k}^{n+m}\}_{m^-\le m\le m^+,k^-\le k\le k^+}\right) = 0 \tag{5.12}$$

où les entiers m^-, m^+, k^-, k^+ définissent ce qu'on appelle la *largeur du stencil* du schéma. On appelle schéma à deux niveaux un schéma tel que $m^- = -1$ et $m^+ = 0$ (ou tel que $m^- = 0$ et $m^+ = 1$).

Exemple 5.4. *Pour le schéma d'Euler explicite (5.7), $m^- = -1$, $m^+ = 0$, $k^- = -1$, $k^+ = +1$ et*

$$F_{\Delta t,\Delta x}\left(\{u_{j+k}^{n+m}\}_{m^-\le m\le m^+,k^-\le k\le k^+}\right) = \frac{u_j^n - u_j^{n-1}}{\Delta t} + D\frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2}.$$

Le schéma d'Euler explicite est donc un schéma à deux niveaux.

Exercice 5.5. *Ecrire la valeur de m^-, m^+, k^-, k^+ et de $F_{\Delta t,\Delta x}\left(\{u_{j+k}^{n+m}\}_{m^-\le m\le m^+,k^-\le k\le k^+}\right)$ pour le schéma d'Euler implicite (5.5). Est-ce un schéma à deux niveaux ?*

Définition 5.6. *Un schéma aux différences finies (5.12) est dit consistant avec l'équation aux dérivées partielles $F(u) = 0$, si, pour toute solution $u(t, x)$ suffisamment régulière de cette équation, l'erreur de troncature du schéma, définie par*

$$F_{\Delta t,\Delta x}\left(\{u\left(t + m\Delta t, x + k\Delta x\right)\}_{m^-\le m\le m^+,k^-\le k\le k^+}\right), \tag{5.13}$$

tend vers 0, uniformément par rapport à $(t, x) \in \mathbb{R}_+ \times \Omega$, lorsque Δt et Δx tendent vers 0 indépendamment. De plus, on dit que le schéma est précis à l'ordre p en espace et à l'ordre q en temps si l'erreur de troncature (5.13) tend vers 0 comme $\mathcal{O}\left((\Delta x)^p + (\Delta t)^q\right)$ lorsque Δt et Δx tendent vers 0.

Remarque 5.7. *Il faut prendre garde dans la formule (5.12) à une petite ambiguïté quant à la définition du schéma. En effet, on peut toujours multiplier n'importe quelle formule par une puissance suffisamment élevée de Δt et Δx de manière à ce que l'erreur de troncature tende vers 0. Cela rendrait consistant n'importe quel schéma ! Pour éviter cet inconvénient, on supposera toujours que la formule*

$$F_{\Delta t,\Delta x}\left(\{u_{j+k}^{n+m}\}_{m^-\le m\le m^+,k^-\le k\le k^+}\right) = 0$$

a été écrite de telle manière que, pour une fonction régulière $u(t, x)$ qui n'est pas une solution de l'équation $F(u) = 0$, la limite de l'erreur de troncature n'est pas nulle.

Concrètement, on calcule l'erreur de troncature d'un schéma en remplaçant u_{j+k}^{n+m} dans la formule (5.12) par $u(t+m\Delta t, x+k\Delta x)$. Comme application de la Définition 5.6, nous allons montrer le lemme suivant.

Lemme 5.8. *Le schéma explicite (5.7) est consistant, précis à l'ordre* 1 *en temps et* 2 *en espace pour l'équation de la chaleur (5.1). De plus, si on choisit de garder constant le rapport* $D\frac{\Delta t}{\Delta x^2} = \frac{1}{6}$*, alors ce schéma est précis à l'ordre* 2 *en temps et* 4 *en espace.*

Démonstration. Soit $v(t,x)$ une fonction de classe $\mathcal{C}^6$. Par développement de Taylor autour du point (t,x), on calcule l'erreur de troncature du schéma (5.7)

$$\begin{aligned}
&\frac{v(t+\Delta t,x)-v(t,x)}{\Delta t}\\
&+D\frac{-v(t,x-\Delta x)+2v(t,x)-v(t,x+\Delta x)}{(\Delta x)^2}\\
&=(\partial_t v(t,x)-D\partial_{xx}v(t,x))\\
&+\frac{\Delta t}{2}\partial_{tt}v(t,x)-D\frac{(\Delta x)^2}{12}\partial_{xxxx}v(t,x)\\
&+\mathcal{O}\left((\Delta t)^2+(\Delta x)^4\right).
\end{aligned}$$

Si v est une solution de l'équation de la chaleur (5.1), on obtient ainsi aisément la consistance ainsi que la précision à l'ordre 1 en temps et 2 en espace. Si on suppose de plus que $D\frac{\Delta t}{(\Delta x)^2} = \frac{1}{6}$, alors les termes en Δt et en $(\Delta x)^2$ se simplifient car $\partial_{tt}v = D\partial_{txx}v = D^2\partial_{xxxx}v$. □

Exercice 5.9. *Montrer que le schéma implicite (5.5) est consistant, précis à l'ordre* 1 *en temps et* 2 *en espace.*

5.3 Stabilité et analyse de Fourier

Pour simplifier les idées, nous supposons dans cette section (sauf mention du contraire) que des conditions aux bords de Dirichlet homogènes sont imposées. Mais les notions présentées ci-dessous s'étendent bien évidemment sans difficulté à tous types ed conditions aux limites.

Pour tout $U := (u_j)_{1\le j\le N} \in \mathbb{R}^N$, nous définissons pour tout $1 \le p < +\infty$ la norme :

$$\|U\|_p := \left(\sum_{j=1}^{N} \Delta x |u_j|^p\right)^{1/p}, \tag{5.14}$$

et

$$\|U\|_\infty := \max_{1\le j\le N} |u_j|.$$

Nous introduisons ici la notion de schéma *stable* pour une certaine norme.

Définition 5.10. *Soit $1 \leq p \leq +\infty$. Un schéma aux différences est dit inconditionnellement stable pour la norme L^p s'il existe une constante $K > 0$ indépendante de Δt et de Δx (lorsque ces valeurs tendent vers 0) telle que*

$$\|U^n\|_p \leq K\|U^0\| \text{ pour tout } n \in \mathbb{N}, \tag{5.15}$$

quelle que soit la donnée initiale U^0. Si (5.15) n'a lieu que pour des pas Δt et Δx astreints à certaines inégalités, on dit que le schéma est conditionnellement stable.

Remarque 5.11. *Puisque toutes les normes sont équivalentes dans $\mathbb{R}^N$, le lecteur trop rapide pourrait croire que la stabilité par rapport à une norme implique la stabilité par rapport à toutes les normes. Malheureusement, il n'en est rien et il existe des schémas qui sont stables par rapport à une norme mais qui ne le sont pas par rapport à une autre. En effet, le point crucial de la Définition 5.10 est que la majoration est uniforme par raport à Δx alors même que les normes définies par (5.14) dépendent de Δx.*

Définition 5.12. *Un schéma aux différences finies est dit linéaire si la formule $F_{\Delta t,\Delta x}(\{u_{j+k}^{n+m}\}) = 0$ qui le définit est linéaire par rapport à ses arguments u_{j+k}^{n+m}.*

La stabilité d'un schéma linéaire à deux niveaux est très facile à interpréter. En effet, par linéarité tout schéma linéaire à deux niveaux peut s'écrire sous la forme condensée

$$U^n = MU^{n-1} \tag{5.16}$$

où M est un opérateur linéaire (une matrice, dite d'itération) de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Par exemple, pour le schéma explicite (5.7), la matrice M vaut

$$M = (I - \Delta t A^{\mathrm{Dir}}) = \begin{pmatrix} 1-2c & c & 0 & \cdots & \cdots & 0 & 0 \\ c & 1-2c & c & 0 & \cdots & \cdots & 0 \\ 0 & c & 1-2c & c & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & c & 1-2c & c & 0 \\ 0 & \cdots & \cdots & 0 & c & 1-2c & c \\ 0 & 0 & \cdots & \cdots & 0 & c & -2c \end{pmatrix}.$$

avec $c = D\frac{\Delta t}{(\Delta x)^2}$.

Exercice 5.13. *Ecrire la matrice M d'itération du schéma (5.5) en fonction de A^{Dir} et Δt.*

A l'aide de cette matrice d'itération, on a

$$U^n = M^n U^0$$

et par conséquent la stabilité du schéma est équivalente à

$$\|A^n U^0\|_p \leq K\|U^0\|_p, \quad \forall n \in \mathbb{N},\ \forall U^0 \in \mathbb{R}^N.$$

On introduit la norme matricielle subordonnée :

$$\forall P \in \mathbb{R}^{N\times N}, \quad \|P\|_{\mathcal{L}^p} := \sup_{U\in\mathbb{R}^N,\, U\neq 0} \frac{\|PU\|_p}{\|U\|_p}.$$

La stabilité du schéma en norme L^p est alors équivalente à

$$\|M^n\|_{\mathcal{L}^p} \leq K,\ \forall n \in \mathbb{N}.$$

Deux notions de stabilité sont particulièrement importantes pour l'analyse de schémas numériques aux différences finies pour l'équation de la chaleur : la stabilité en norme L^∞ et la stabilité en norme L^2. Nous détaillons les techniques de preuve usuelles pour ces deux types de stabilité dans les sections suivantes.

5.3.1 Stabilité en norme L^∞

La stabilité en norme L^∞ est très liée au principe du maximum discret dont nous donnons la définition ci-dessous.

Définition 5.14. *Un schéma aux différences finies vérifie le principe du maximum discret si pour tout $n \in \mathbb{N}$ et pour tout $1 \leq j \leq N$ on a*

$$\min\left(0, \min_{0\leq j\leq N+1} u_j^0\right) \leq u_j^n \leq \max\left(0, \max_{0\leq j\leq N+1} u_j^0\right)$$

quelle que soit la donnée initiale U^0.

Remarque 5.15. *Dans le Définition 5.14, les inégalités tiennent compte non seulement du minimum et du maximum de U^0 mais aussi de 0 qui est la valeur imposée au bord par les conditions de Dirichlet. Cela est nécessaire si la donnée initiale U^0 ne vérifie pas les conditions aux limites de Dirichlet (nous verrons dans un prochain chapitre que cela peut effectivement être le cas), et inutile dans le cas contraire.*

Le principe du maximum discret permet de démontrer le résultat suivant.

Lemme 5.16. *Le schéma explicite (5.7) est stable en norme L^∞ si et seulement si la condition*

$$2D\Delta t \leq (\Delta x)^2 \tag{5.17}$$

est satisfaite. On appelle la condition (5.17) la condition de Courant-Friedrich-Lewy ou condition CFL.

Pour la petite histoire, la condition de stabilité (5.17) fut découverte en 1928 (avant l'apparition des premiers ordinateurs !). C'est une des remarques les plus profondes de l'analyse numérique.

Démonstration. Le schéma d'Euler explicite peut se réécrire sous la forme

$$u_j^n = D\frac{\Delta t}{(\Delta x)^2}u_{j-1}^{n-1} + \left(1 - 2D\frac{\Delta t}{(\Delta x)^2}\right)u_j^{n-1} + D\frac{\Delta t}{(\Delta x)^2}u_{j+1}^{n-1}. \tag{5.18}$$

Si la condition CFL est vérifiée, alors (5.18) montre que u_j^n est une combinaison convexe des valeurs au temps précédent u_{j-1}^{n-1}, u_j^{n-1}, u_{j+1}^{n-1}. En effet, tous les coefficients dans le membre de droite de (5.18) sont positifs et leur somme vaut 1. En particulier si la donnée initiale U^0 est bornée par deux constants m et M telles que

$$m \leq u_j^0 \leq M, \quad \forall 0 \leq j \leq N+1,$$

alors une récurrence facile montre que les mêmes inégalités restent vraies pour tous les temps ultérieurs

$$\min(0, m) \leq u_j^n \leq \max(0, M), \quad \forall 0 \leq j \leq N+1,$$

en prenant en compte les conditions aux limites de Dirichlet. Le schéma (5.7) vérifie donc le principe du maximum discret et est donc stable en norme L^∞.

Supposons maintenant au contraire que la condition CFL ne soit pas vérifiée, c'est-à-dire que

$$2D\Delta t > (\Delta x)^2.$$

Alors pour certaines données initiales, le schéma n'est pas stable (il peut être stable pour certaines conditions initiales exceptionnelles, par exemple pour $U^0 = 0$!). Prenons la donnée initiale définie par

$$u_j^0 = (-1)^j, \ \forall 0 \leq j \leq N+1,$$

qui est bien uniformément bornée.

Une récurrence facile montre alors que pour tout $0 \leq j \leq N+1$,

$$u_j^n \leq 0 \text{ si } n+j \text{ est impair} \qquad \text{et } u_j^n \geq 0 \text{ si } n+j \text{ est pair.}$$

En conséquence, en utilisant (5.18), pour tout $n \in \mathbb{N}^*$, on obtient que

$$|u_j^n| \geq \left|D\frac{\Delta t}{(\Delta x)^2}\right| \left(\left|u_{j-1}^{n-1}\right| + \left|u_{j+1}^{n-1}\right|\right),$$

et une récurrence facile montre alors que

$$|u_j^n| \geq \left|2D\frac{\Delta t}{(\Delta x)^2}\right|^n.$$

On obtient donc que $|u_j^n| \longrightarrow_{n\to+\infty} +\infty$ et le shéma n'est donc pas stable en norme L^∞. □

Exercice 5.17. *Montrer que le schéma implicite (5.5) est stable en norme L^∞ quels que soient les pas de temps Δt et d'espace Δx. On dit que le schéma d'Euler implicite est inconditionnellement stable en norme L^∞.*

Indication : On supposera que la condition initiale U^0 est telle qu'il existe deux constantes $m \leq 0 \leq M$ telles que

$$m \leq u_j^0 \leq M, \quad \forall 1 \leq j \leq N,$$

et on cherchera à prouver (par récurrence) que pour tout $n \in \mathbb{N}^*$,

$$m \leq u_j^n \leq M, \quad \forall 1 \leq j \leq N,$$

et ceci sans condition sur Δt et Δx.

5.3.2 Stabilité en norme L^2

De nombreux schémas vérifient le principe du maximum discret mais sont néanmoins de "bons" schémas. Pour ceux-là, il faut vérifier la stabilité dans une autre norme que la norme L^∞. La norme L^2 se prête très bien à l'étude de la stabilité grâce à l'outil très puissant de l'analyse de Fourier que nous présentons maintenant. Pour ce faire, nous supposons désormais que les conditions aux limites pour l'équation de la chaleur sont des conditions aux limites de périodicité. Rappelons que dans ce cas, à chaque itération $n \in \mathbb{N}^*$ du schéma numérique, nous devons déterminer le vecteur $U^n := (u_j)_{1 \leq j \leq N+1} \in \mathbb{R}^{N+1}$.

A chaque vecteur $U^n := (u_j^n)_{1 \leq j \leq N+1}$, on associe une fonction définie sur $\mathbb{R}$, constante par morceaux, périodique de période 1 et définie sur $[0, 1]$ par

$$u^n(x) = u_j^n \quad \text{ si } x_{j-1/2} \leq x < x_{j+1/2},$$

avec $x_{j+1/2} = (j + 1/2)\Delta x$ pour $0 \leq j \leq N$, $x_{-1/2} = 0$ et $x_{N+1+1/2} = 1$. Ainsi définie, la fonction $u^n(x)$ est périodique de période 1 et appartient à $L^2(0, 1)$. Or, d'après l'analyse de Fourier, on peut écrire une telle fonction $u^n(x)$ en utilisant sa décomposition en séries de Fourier :

$$u^n(x) = \sum_{k \in \mathbb{Z}} \widehat{u}^n(k) e^{2i\pi kx},$$

avec

$$\widehat{u}^n(k) = \int_0^1 u^n(x) e^{-2i\pi kx} \, dx.$$

On a de plus la formule de Plancherel

$$\int_0^1 |u^n(x)|^2 \, dx = \sum_{k \in \mathbb{Z}} |\widehat{u}^n(k)|^2.$$

Rappelons qu'une propriété (importante pour la suite) de la transformée de Fourier des fonctions périodiques est la suivante : si on note $v^n(x) := u^n(x + \Delta x)$, alors $\widehat{v}^n(k) = \widehat{u}^n(k)e^{2i\pi k\Delta x}$.

Expliquons maintenant la méthode sur l'exemple du schéma explicite (5.7). Avec les notations introduites ci-dessus, on peut réécrire ce schéma, pour $0 \leq x \leq 1$,

$$\frac{u^n(x) - u^{n-1}(x)}{\Delta t} + D\frac{-u^{n-1}(x - \Delta x) + 2u^{n-1}(x) - u^{n-1}(x + \Delta x)}{(\Delta x)^2} = 0.$$

Par application de la transformée de Fourier, il vient

$$\widehat{u}^n(k) = \left(1 - D\frac{\Delta t}{(\Delta x)^2}\left(-e^{2i\pi k\Delta x} + 2 - e^{2i\pi k\Delta x}\right)\right)\widehat{u}^{n-1}(k).$$

Autrement dit,

$$\widehat{u}^n(k) = M(k)\widehat{u}^{n-1}(k) = M(k)^n\widehat{u}^0(k),$$

avec

$$M(k) := 1 - 2D\frac{\Delta t}{(\Delta x)^2}\left(1 - \cos(2\pi k\Delta x)\right) = 1 - 4D\frac{\Delta t}{(\Delta x)^2}\left(\sin(\pi k\Delta x)\right)^2.$$

Pour $k \in \mathbb{Z}$, le coefficient de Fourier $\widehat{u}^n(k)$ est borné lorsque n tend vers l'infini si et seulement si le facteur d'amplification vérifie $|M(k)| \leq 1$, c'est-à-dire

$$4D\frac{\Delta t}{(\Delta x)^2}\left(\sin(\pi k\Delta x)\right)^2 \leq 2,$$

soit

$$2D\Delta t\left(\sin(\pi k\Delta x)\right)^2 \leq (\Delta x)^2. \tag{5.19}$$

Si la condition CFL (5.17) , i.e. $2D\Delta t \leq (\Delta x)^2$ est satisfaite, alors l'inégalité (5.19) est vraie quel que soit le mode de Fourier $k \in \mathbb{Z}$, et par la formule de Plancherel, on en déduit

$$\|U^n\|_2^2 = \int_0^1 |u^n(x)|^2\,dx = \sum_{k\in\mathbb{Z}} |\widehat{u}^n(k)|^2 \leq \sum_{k\in\mathbb{Z}} |\widehat{u}^0(k)|^2 = \int_0^1 |u^0(x)|^2\,dx = \|U^0\|_2^2,$$

ce qui n'est rien d'autre que la stabilité L^2 du schéma explicite. Si la condition CFL n'est pas satisfaite, le schéma est instable. En effet, il suffit de choisir Δx (éventuellement suffisammenet petit) et k_0 (suffisamment grand) et une donnée initiale ayant une seule composante de Fourier non nulle $\widehat{u}^0(k) \neq 0$ avec $\pi k_0\Delta x \approx \pi/2$ (modulo π) de telle manière que $|M(k_0)| > 1$. On a donc démontré le lemme suivant :

Lemme 5.18. *Le schéma explicite (5.7) est stable en norme L^2 si et seulement si la condition CFL $2D\Delta t \leq (\Delta x)^2$ est satisfaite.*

Exercice 5.19. *Montrer que, pour des conditions aux limites périodiques, le schéma implicite (5.5) est stable en norme L^2.*

Remarque 5.20. *Traduisons sous forme de "recette" la méthode de l'analyse de Fourier pour prouver la stabilité L^2 d'un schéma. En utilisant les notations ci-dessus, on obtient une relation*

$$\widehat{u}^n(k) = M(k)\widehat{u}^{n-1}(k)$$

et on en déduit la valeur du coefficient d'amplification $M(k)$. On appelle condition de stabilité de von Neumann l'inégalité

$$|M(k)| \leq 1, \quad \forall k \in \mathbb{Z}.$$

Si la condition de stabilité de von Neumann est satisfaite (avec éventuellement des relations sur Δt et Δx), alors le schéma est stable pour la norme L^2, sinon il est instable.

Remarque 5.21. *On peut également montrer que le schéma explicite (5.7) avec conditions limites de Dirichlet est stable en norme L^2 si et seulement la condition CFL $2D\Delta t \leq (\Delta x)^2$. La preuve est juste un peu plus pénible que dans le cas de conditions de bords périodiques. De même, le schéma implicite (5.5) avec conditions limites de Dirichlet est inconditionnellement stable en norme L^2.*

5.4 Convergence

Nous avons maintenant tous les outils pour démontrer la convergence des schémas de différences finies. Le principal résultat en ce sens est le Théorème de Lax qui affirme que, pour un schéma linéaire à deux niveaux, consistance et stabilité implique convergence.

Théorème 5.22 (Théorème de Lax)**.** *Soit $u(t,x)$ la solution suffisamment régulière de l'équation de la chaleur (5.1) (avec des conditions aux limites appropriées). Soit u_j^n la solution unmérique discrète obtenue par un schéma de différences finies avec la donnée initiale $u_j^0 = u_0(x_j)$. On suppose que le schéma est linéaire, à deux niveaux, consistant et stable pour une norme $\|\cdot\|_p$ pour $1 \leq p \leq +\infty$. Alors, le schéma est convergent au sens où*

$$\forall T > 0, \quad \lim_{\Delta t, \Delta x \to 0} \left(\sup_{t_n \leq T} \|e^n\|_p \right) = 0, \tag{5.20}$$

avec e^n le vecteur "erreur" défini par ses composantes $e_j^n = u_j^n - u(t_n, x_j)$.

De plus, si le schéma est précis à l'ordre q en espace et à l'ordre r en temps, alors pour tout $T > 0$ il existe une constante $C_T > 0$ telle que

$$\sup_{0 \leq t_n \leq T} \|e^n\|_p \leq C_T \left((\Delta x)^q + (\Delta t)^r \right). \tag{5.21}$$

Démonstration. Pour simplifier, on suppose que le schéma est discrétisé avec des conditions aux limites de Dirichlet. La même démonstration est aussi valable pour des conditions aux limites de périodicité. Un schéma linéaire à deux niveaux peut s'écrire sous la forme condensée (5.16), i.e.

$$U^{n+1} = MU^n, \tag{5.22}$$

où M est la matrice d'itération (carrée de taille N). Soit u la solution (supposée suffisamment régulière) de l'équation de la chaleur (5.1). On note $\widetilde{U}^n := \left(\widetilde{u}_j^n\right)_{1\leq j\leq N} \in \mathbb{R}^N$ avec $\widetilde{u}_j^n := u(t_n, x_j)$. Comme le schéma est consistant, il existe un vecteur ϵ^n tel que

$$\widetilde{U}^{n+1} = M\widetilde{U}^n + \Delta t\epsilon^n, \quad \text{avec } \lim_{\Delta t,\Delta x\to 0} \|\epsilon^n\|_p = 0, \tag{5.23}$$

et la convergence de ϵ^n est uniforme pour tous les temps $0 \leq t_n \leq T$. Si le schéma est précis à l'ordre q en espace et à l'ordre r en temps, alors

$$\|\epsilon^n\|_p \leq C\left((\Delta x)^q + (\Delta t)^r\right).$$

En posant $e_j^n = u_j^n - u(t_n, x_j)$, on obtient par soustraction de (5.23) à (5.22)

$$e^{n+1} = Me^n - \Delta t\epsilon^n,$$

d'où par récurrence

$$e^n = M^n e^0 - \Delta t \sum_{k=1}^{n} M^{n-k}\epsilon^{k-1}. \tag{5.24}$$

Or la stabilité du schéma veut dire que $\|U^n\|_p = \|M^nU^0\|_p \leq K\|U^0\|_p$ pour toute donnée initiale, c'est-à-dire que $\|M^n\|_{\mathcal{L}^p} \leq K$ où la constante K ne dépend pas de n. D'autre part, $e^0 = 0$, donc (5.24) donne

$$\|e^n\|_p \leq \Delta t \sum_{k=1}^{n} \|M^{n-k}\|_{\mathcal{L}^p}\|\epsilon^{k-1}\|_p \leq \Delta t n K C\left((\Delta x)^q + (\Delta t)^r\right),$$

ce qui donne l'inégalité (5.21) avec la constante $C_T = TKC$. La démonstration de (5.20) est similaire. □

Remarque 5.23. *Le Théorème de Lax est en fait valable pour toute équation aux dérivées partielles linéaire. Il admet une réciproque au sens où un schéma linéaire consistant à deux niveaux qui converge est nécessairement stable. Remarquer que la vitesse de convergence dans (5.21) est exactement la précision du schéma. Enfin, il est bon de noter que cette estimation (5.21) n'est valable que sur un intervalle borné de temps* $[0, T]$ *mais qu'elle est indépendant du nombre de points de discrétisation* N.

Chapitre 6

Problèmes d'évolution paraboliques

Ce chapitre est une brève introduction à l'étude mathématique et numérique d'équations aux dérivées partielles dépendant du temps. Ici, nous étudierons surtout l'équation de la chaleur. Nous renvoyons à [7, 1, 5, 13] pour une présentation plus détaillée.

6.1 Préliminaires

6.1.1 Lemme de Gronwall

Avant de rappeler le lemme de Gronwall, nous donnons ici la définition d'une fonction absolument continue à valeurs dans un espace de Banach.

Définition 6.1. *Soit $I \subset \mathbb{R}$ un intervalle de $\mathbb{R}$ et soit X un espace de Banach. On dit qu'une fonction continue $u : I \to X$ est une fonction absolument continue si et seulement si pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que pour toute suite finie $(\alpha_n)_{n\leq N}, (\beta_n)_{n\leq N} \subset I$ tels que*

$$(\alpha_n, \beta_n) \cap (\alpha_m, \beta_m) = \emptyset \quad \forall n \neq m$$

et

$$\sum_{n\in\mathbb{N}} |\beta_n - \alpha_n| \leq \delta,$$

alors

$$\sum_{n\in\mathbb{N}} \|u(\beta_n) - u(\alpha_n)\|_X \leq \epsilon.$$

Proposition 6.2. *Une fonction f est absolument continue sur un intervalle compact $I = [a, b] \subset \mathbb{R}$ si et seulement si il existe une fonction $g \in L^1([a, b])$ telle que*

$$\forall x \in [a, b], \quad f(x) - f(a) = \int_a^x g(t)\, dt.$$

Remarque : Les fonctions absolument continues sont uniformément continues et différentiables presque partout. Les fonctions Lipschitz sont absolument continues.

Le lemme suivant est très classique et très utile :

Lemme 6.3 (Gronwall)**.** *Soit $T > 0$. Soit η une fonction positive absolument continue sur $[0, T]$ et vérifiant :*

$$\eta'(t) \leq \varphi(t)\eta(t) + \psi(t)$$

pour tout $t \in [0; T]$, où φ et ψ sont des fonctions positives de $L^1(0, T)$. Alors

$$\forall t \in [0; T], \qquad \eta(t) \leq e^{\int_0^t \varphi(s)ds} \left(\eta(0) + \int_0^t \psi(s)ds \right).$$

6.1.2 Rappels sur l'espace $H^{-1}(\Omega)$

Définition 6.4. *Soit Ω un ouvert de $\mathbb{R}^d$. On note $H^{-1}(\Omega)$ l'espace vectoriel des distributions $T \in \mathcal{D}'(\Omega)$ telles qu'il existe une constante C telle que*

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad |\langle T, \varphi \rangle_{\mathcal{D}',\mathcal{D}}| \leq C \|\varphi\|_{H^1}.$$

Remarque : Il est clair que $L^2(\Omega) \subset H^{-1}(\Omega)$. En effet, si $f \in L^2(\Omega)$

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad |\langle f, \varphi \rangle_{\mathcal{D}',\mathcal{D}}| = \left| \int_\Omega f\varphi \right| \leq \|f\|_{L^2} \|\varphi\|_{L^2} \leq \|f\|_{L^2} \|\varphi\|_{H^1}.$$

Théorème 6.5. *On peut identifier $H^{-1}(\Omega)$ au dual topologique de $H_0^1(\Omega)$.*

Démonstration. Soit $T \in H^{-1}(\Omega)$. L'application linéaire

$$\mathcal{D}(\Omega) \ni \varphi \mapsto \langle T, \varphi \rangle_{\mathcal{D}',\mathcal{D}}$$

est continue sur $\mathcal{D}(\Omega)$ *muni de la norme* H^1. Comme $\mathcal{D}(\Omega)$ est dense dans $H_0^1(\Omega)$ pour cette norme, cette application se prolonge (de manière unique) en une application linéaire continue sur $H_0^1(\Omega)$, notée

$$H_0^1(\Omega) \ni \varphi \mapsto \langle T, \varphi \rangle_{H^{-1},H_0^1}$$

qui vérifie en particulier

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad \langle T, \varphi \rangle_{H^{-1},H_0^1} = \langle T, \varphi \rangle_{\mathcal{D}',\mathcal{D}}.$$

On peut donc associer à tout $T \in H^{-1}(\Omega)$ un élément du dual topologique de $H_0^1(\Omega)$ (i.e. de l'espace vectoriel des formes linéaires continues sur $H_0^1(\Omega)$). On définit ainsi

$$\alpha : \left\{ \begin{array}{ccc} H^{-1}(\Omega) & \longrightarrow & (H_0^1(\Omega))' \\ T & \mapsto & \langle T, \cdot \rangle_{H^{-1},H_0^1}. \end{array} \right.$$

Réciproquement, soit $L \in (H_0^1(\Omega))'$. Il existe une constante C telle que

$$\forall \varphi \in H_0^1(\Omega), \qquad |L(\varphi)| \leq C\|\varphi\|_{H^1}.$$

Si on restreint L à $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, on obtient une forme linéaire sur $\mathcal{D}(\Omega)$ qui vérifie

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad |L(\varphi)| \leq C\|\varphi\|_{H^1}.$$

Il reste à vérifier que L est une distribution (c'est-à-dire qu'elle est continue sur $\mathcal{D}(\Omega)$ pour la topologie de $\mathcal{D}(\Omega)$). Soit donc K compact inclus dans Ω et $\varphi \in \mathcal{D}_K(\Omega)$. Il vient

$$|L(\varphi)| \leq C\|\varphi\|_{H^1} \leq C\left(\|\varphi\|_{L^2}^2 + \|\nabla\varphi\|_{L^2}^2\right)^{1/2}.$$

Or $\|\varphi\|_{L^2} \leq \sqrt{|K|}\,\sup|\varphi|$ et $\|\nabla\varphi\|_{L^2} \leq \sqrt{|K|}\,\sup|\nabla\varphi|$. Donc

$$|L(\varphi)| \leq C' \sup_{|\alpha|\leq 1,\, x\in K} |\partial^\alpha \varphi|.$$

Donc L définit une distribution (d'ordre ≤ 1). On définit ainsi

$$\beta : \left\{ \begin{array}{ccc} (H_0^1(\Omega))' & \longrightarrow & H^{-1}(\Omega) \\ L & \mapsto & L|_{\mathcal{D}(\Omega)}. \end{array} \right.$$

On vérifie sans difficulté que $\alpha \circ \beta = I_{(H_0^1(\Omega))'}$ et que $\beta \circ \alpha = I_{H^{-1}(\Omega)}$. □

Proposition 6.6. *(Caractérisation des éléments de H^{-1}). Soit Ω un ouvert de $\mathbb{R}^d$. Une distribution T appartient à $H^{-1}(\Omega)$ si et seulement si il existe, pour tout $|\alpha| \leq 1$ une fonction $g_\alpha \in L^2(\Omega)$ telle que*

$$T = \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha.$$

Démonstration. Il est clair que si T est de la forme

$$T = \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha$$

avec $g_\alpha \in L^2(\Omega)$, on a

$$\begin{aligned}\forall \varphi \in \mathcal{D}(\Omega), \qquad |\langle T, \varphi\rangle_{\mathcal{D}',\mathcal{D}}| &= |\langle \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha, \varphi\rangle_{\mathcal{D}',\mathcal{D}}| \\ &\leq |\sum_{|\alpha|\leq 1} (-1)^{|\alpha|} \langle g_\alpha, \partial^\alpha \varphi\rangle_{\mathcal{D}',\mathcal{D}}| \\ &= |\sum_{|\alpha|\leq 1} (-1)^{|\alpha|} \langle g_\alpha, \partial^\alpha \varphi\rangle_{L^2}| \\ &\leq \sum_{|\alpha|\leq 1} \|g_\alpha\|_{L^2} \|\partial^\alpha \varphi\|_{L^2} \\ &\leq \left(\sum_{|\alpha|\leq 1} \|g_\alpha\|_{L^2}\right) \|\varphi\|_{H^1}.\end{aligned}$$

Donc $T \in H^{-1}(\Omega)$. La réciproque est plus délicate et admise ici. □

Conséquences : on a les inclusions

$$\mathcal{D}(\Omega) \subset H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) \subset \mathcal{D}'(\Omega)$$

et on a par ailleurs pour $T \in L^2(\Omega)$ et $\varphi \in \mathcal{D}(\Omega)$,

$$\langle T, \varphi\rangle = \langle T, \varphi\rangle_{H^{-1},H_0^1} = \langle T, \varphi\rangle_{L^2} = \int_\Omega T\,\varphi.$$

L'espace $L^2(\Omega)$ est appelé l'"espace pivot" de ces dualités.

6.2 Les espaces de Bochner

6.2.1 Intégrale de Bochner

Dans cette section, nous introduisons la notion d'intégrale de Bochner, qui permet de généraliser la notion d'intégrale de Lebesgue, à des fonctions à valeurs dans un espace de Banach.

Dans toute cette section, on considère $a, b \in \mathbb{R} \cup \{\pm\infty\}$ et X un espace de Banach.

Définition 6.7. *Une fonction $f : [a,b] \to X$ est dite mesurable si et seulement si pour tout ensemble ouvert $B \subset X$, l'ensemble $f^{-1}(B)$ est un ensemble borélien de $[a,b]$.*

On voit aisément que cette définition est une extension directe de la notion de mesurabilité pour des fonctions à valeurs scalaires (ou à valeurs dans $\mathbb{R}^n$ avec $n \in \mathbb{N}^*$). Le théorème suivant est également une extension directe d'un résultat que vous connaissez bien pour des fonctions à valeurs scalaires.

Théorème 6.8. *Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions mesurables définies sur $[a,b]$ à valeurs dans X. Si $(f_n)_{n\in\mathbb{N}}$ converge simplement (dans X) vers une fonction $f : [a,b] \to X$, alors f est une fonction mesurable.*

Comme pour l'intégrale de Lebesgue, nous allons définir l'intégrale de Bochner comme la limite d'intégrales d'une suite de fonctions étagées. Dans notre cas, on appellera une fonction étagée toute fonction $s : [a,b] \to X$ telle qu'il existe $M \in \mathbb{N}^*$, $u_1, \cdots, u_M \in X$ et $B_1, \cdots, B_M$ des sous-ensembles boréliens de $[a,b]$ de mesure de Lebesgue finie tels que

$$\forall t \in [a,b], \quad s(t) = \sum_{m=1}^{M} u_m \chi_{B_m}(t),$$

où $\chi_B : [a,b] \to \{0,1\}$ désigne la fonction caractéristique du sous-ensemble $B \subset [a,b]$.

Pour une telle fonction, on peut définir son intégrale de Bochner comme l'élément de X suivant :

$$\int_{[a,b]} s(t)\, dt = \sum_{m=1}^{M} u_m \lambda(B_m) \in X,$$

où λ désigne la mesure de Lebesgue sur l'intervalle $[a,b]$.

Exercice 6.9. *Montrer que si $s : [a,b] \to X$ est une fonction étagée, alors*

$$\left\| \int_{[a,b]} s(t)\, dt \right\|_X \leq \int_{[a,b]} \|s(t)\|_X \, dt.$$

Pour définir l'intégrale de Lebesgue, nous utilisions dans le cas scalaire le résultat crucial suivant : toute fonction mesurable (à valeurs scalaires) peut être vue comme la limite simple d'une suite de fonctions étagées. Il se trouve que ce résultat n'est pas toujours valide dans le cas d'espaces de Banach généraux, ce qui justifie la définition suivante :

Définition 6.10. *On dit qu'une fonction $f : [a,b] \to X$ est Lebesgue-mesurable s'il existe une suite de fonctions étagées $(s_n)_{n\in\mathbb{N}}$ qui converge simplement vers f presque partout sur $[a,b]$ (au sens de la mesure de Lebesgue).*

Exercice 6.11. *Montrer que si une suite $(f_n)_{n\in\mathbb{N}}$ de fonctions Lebesgue-mesurables converge simplement presque partout vers une fonction f, alors f est une fonction Lebesgue-mesurable.*

Les notions de mesurabilité et de Lebesgue-mesurabilité ne sont pas équivalentes en général. La Lebesgue-mesurabilité implique la mesurabilité comme l'indique la proposition suivante.

Proposition 6.12. *Soit une fonction $f : [a, b] \to X$ Lebesgue-mesurable. Alors f est mesurable.*

La réciproque est fausse en général. Elle est cependant vraie dans le cas où l'espace X est un espace *séparable* au sens de la définition suivante.

Définition 6.13. *Un espace de Banach X est dit séparable s'il existe un sous-ensemble dense de X au plus dénombrable.*

Exercice 6.14. *Un espace de Hilbert séparable (i.e. muni d'une base hilbertienne) est un espace de Banach séparable au sens de la Définition 6.13.*

En pratique, tous les espaces de Banach que vous connaissez (espaces de Lebesgue L^p, de Sobolev H^k...) sont des espaces séparables. Si X est un espace de Banach séparable, on a alors le résultat suivant :

Théorème 6.15. *Soit X un espace de Banach séparable. Alors, pour toute fonction $f : [a, b] \to X$ mesurable, il existe une suite $(s_n)_{n\in\mathbb{N}}$ de fonctions étagées définies sur $[a, b]$ à valeurs dans X telle que $(s_n)_{n\in\mathbb{N}}$ converge simplement vers f presque partout (au sens de la mesure de Lebesgue) sur $[a, b]$.*

Autrement dit, si X est un espace de Banach séparable, toute fonction $f : [a, b] \to X$ est mesurable si et seulement si elle est Lebesgue-mesurable.

Pour pouvoir définir l'intégrale de Bochner, nous avons besoin de définir la notion de fonction intégrable dans notre contexte. C'est le but de la définition suivante.

Définition 6.16. *Une fonction $f : [a, b] \to X$ est dite intégrable si et seulement si il existe une suite $(s_n)_{n\in\mathbb{N}}$ de fonctions étagées telles que*
(i) $(s_n)_{n\in\mathbb{N}}$ converge simplement vers f presque partout sur $[a, b]$;
(ii) $\displaystyle\int_{[a,b]} \|f - s_n\|_X \underset{n\to+\infty}{\longrightarrow} 0.$

Le théorème suivant énonce une formulation équivalente de la notion d'intégrabilité, qui vous sera probablement plus familière.

Théorème 6.17 (Critère d'intégrabilité de Bochner)**.** *Une fonction $f : [a, b] \to X$ est intégrable si et seulement si*

$$\int_{[a,b]} \|f(t)\|_X \, dt < +\infty.$$

Nous sommes armés à présent pour pouvoir définir l'intégrale de Bochner d'une fonction intégrable.

Théorème-Définition 6.18. *Soit $f : [a,b] \to X$ une fonction intégrable et $(s_n)_{n\in\mathbb{N}}$ une suite de fonctions étagées vérifiant les propriétés (i) et (ii) de la Définition 6.16. Alors, l'intégrale de Bochner de f sur $[a,b]$ est définie par*

$$\int_{[a,b]} f = \lim_{n\to+\infty} \int_{[a,b]} s_n.$$

On a de plus la propriété suivante :

$$\left\| \int_{[a,b]} f \right\|_X \leq \int_{[a,b]} \|f\|_X . \tag{6.1}$$

Une propriété très importante de l'intégrale de Bochner est donnée dans la proposition suivante.

Proposition 6.19. *Soit Y un espace de Banach et $T : X \to Y$ une application linéaire continue. Si $f : [a,b] \to X$ est intégrable, alors $T(f) : [a,b] \to Y$ est intégrable et*

$$T\left(\int_{[a,b]} f\right) = \int_{[a,b]} T(f).$$

Exercice 6.20. *Soit H un espace de Hilbert, $v \in H$ et $f : [a,b] \to H$ une fonction intégrable. Montrer que*

$$\left\langle \int_{[a,b]} f(t)\,dt, v \right\rangle_H = \int_{[a,b]} \langle f(t), v\rangle_H \,dt.$$

Le théorème de convergence dominée de Lebesgueque est toujours valide pour l'intégrale de Bochner. Plus précisément, on a le résultat suivant.

Théorème 6.21 (Théorème de convergence dominée de Lebesgue pour l'intégrale de Bochner)**.** *Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions définies sur $[a,b]$ et à valeurs dans X, et $f : [a,b] \to X$ telles que*

- *$f_n(t) \longrightarrow_{n\to+\infty} f(t)$ dans X pour presque tout $t \in [a,b]$;*
- *Il existe une fonction $g \in L^1([a,b],\mathbb{R})$ telle que pour tout $n \in \mathbb{N}$, $\|f_n(t)\|_X \leq g(t)$ pour presque tout $t \in [a;b]$.*

Alors, f est intégrable et

$$\int_{[a,b]} \|f_n(t) - f(t)\|_X \,dt \longrightarrow 0,$$

ce qui implique que

$$\int_{[a,b]} f_n(t)\,dt \underset{n\to+\infty}{\longrightarrow} \int_{[a,b]} f(t)\,dt \quad \text{dans } X.$$

Exercice 6.22. *Démontrer le théorème 6.21 en utilisant le Théorème de Convergence Dominée de Lebesgue pour les fonctions à valeurs réelles. On commencera tout d'abord par montrer que* $[a,b] \ni t \mapsto \|f(t)\|_X$ *est une fonction intégrable sur* $[a,b]$.

Une extension immédiate des notions vues dans cette section permet de définir les fonctions $f : [a,b] \times [c,d] \to X$ intégrables sur $[a,b] \times [c,d]$ à valeurs dans un espace de Banach X ainsi que leur intégrale de Bochner. Pour de telles fonctions, il existe également un théorème de Fubini pour des fonctions intégrables pour l'intégrale de Bochner, similaire à celui que vous connaissez pour l'intégrale de Lebesgue. Plus précisément, on a

Théorème 6.23 (Fubini)**.**

Exercice 6.24. *Une extension immédiate permet de définir les fonctions* $f : [a,b] \times [c,d] \to X$ *intégrables sur* $[a,b] \times [c,d]$ *ainsi que leur intégrale de Bochner. Montrer que le théorème de Fubini est toujours valide dans ce contexte.*

Nous terminons cette section par un dernier théorème, très utile, qui s'appelle le théorème de différentiabilité de Lebesgue.

Théorème 6.25. *Soit* X *un espace de Banach*

Démonstration. Commençons par prouver le résultat dans le cas où $X = \mathbb{R}$, i.e. dans le cas □

6.2.2 Espaces dépendant du temps

Dans cette section, nous introduisons plusieurs espaces fonctionnels adaptés à l'étude des équations d'évolution, et qui seront à la base de la définition des *solutions faibles*.

Le contenu de ce chapitre peut être trouvé en détails dans [7, Section 5.9.2 et Appendice E.5].

L'idée générale est de séparer la variable temporelle en voyant $u(t,x)$ non pas comme une fonction des deux variables t et x, mais plutôt comme une fonction de t à valeurs dans un espace de fonctions de la variable x :

$$u : t \mapsto \{x \mapsto u(t,x)\}.$$

Soit X un espace de Banach et I un intervalle de $\mathbb{R}$. Nous noterons $C^k(I,X)$, $k \geq 0$, l'espace des fonctions k fois continuement dérivables sur I à valeurs dans X. De même on peut définir l'espace $L^p(I,X)$ contenant les fonctions $u : I \to X$ (définies presque partout et mesurables en un sens approprié, voir l'appendice E.5 de [7]) telles que la fonction $t \mapsto \|u(t)\|_X$ appartient à l'espace usuel $L^p(I,\mathbb{R})$:

$$\int_I \|u(t)\|_X^p \, dt < \infty.$$

Tous ces espaces sont eux-mêmes des espaces de Banach lorsqu'ils sont munis des normes associées :

$$\|u\|_{\mathcal{C}^k(I,X)} = \sum_{m=0}^{k} \sup_{t\in I} \left\|u^{(m)}(t)\right\|_X ,$$

$$\|u\|_{L^p(I,X)} = \left(\int_I \|u(t)\|_X^p \, dt\right)^{1/p}, \quad 1 \le p < \infty,$$

$$\|u\|_{L^\infty(I,X)} = \sup_{t\in I} \|u(t)\|_X .$$

De façon similaire, on dit que $u \in L^1_{\text{loc}}(I,X)$ si $u \in L^1([a;b],X)$ pour tout $[a;b] \subset I$, $a, b \in \mathbb{R}$. Notons que si X est un espace de Hilbert, alors $L^2(I,X)$ est aussi un espace de Hilbert muni du produit scalaire

$$\langle u, v\rangle_{L^2(I,X)} = \int_I \langle u(t), v(t)\rangle_X dt.$$

Nous utiliserons souvent des espaces du type $L^2(I, H_0^p(\Omega))$ ou $L^2(I, H^{-r}(\Omega))$. Ce sont tous des espaces de Hilbert.

Si $u \in L^p(I, H^k(\Omega))$ pour un ouvert régulier $\Omega \subset \mathbb{R}^d$ avec $p \ge 1$ et $k \ge 1$, alors on peut évidemment définir ∇u par

$$\nabla u : t \mapsto \{x \mapsto \nabla_x u(t,x)\}.$$

Bien sûr dans ce cas $\nabla u \in L^p(I, (H^{k-1}(\Omega))^d)$.

Nous aurons besoin dans la suite de définir la notion de *dérivée faible* temporelle pour des fonctions appartenant à de tels espaces. Cette notion est donnée dans le Théorème-Définition 6.26.

Théorème-Définition 6.26 (Dérivée faible temporelle)**.** *Soit I un intervalle ouvert de $\mathbb{R}$, X un espace de Banach. On dit que $v \in L^1_{\text{loc}}(I,X)$ est la* dérivée faible *de $u \in L^1_{\text{loc}}(I,X)$ (et on note $v = u'$) si et seulement si*

$$\forall \varphi \in \mathcal{C}_c^\infty(I,\mathbb{R}), \quad \int_I \varphi(t)v(t)dt = -\int_I \varphi'(t)u(t)dt \ \text{ dans } X. \tag{6.2}$$

Exercice 6.27. *Montrer que si X est un espace de Hilbert et si $w_1, w_2 \in L^1_{\text{loc}}(I,X)$ vérifient $\int_I w_1(t)\varphi(t)\,dt = \int_I w_2(t)\,\varphi(t)\,dt$ pour toute fonction $\varphi \in \mathcal{C}_c^\infty(I)$, alors nécessairement $u(t) = w(t)$ pour presque tout $t \in I$. En déduire que la dérivée faible de u définie par (6.2) est définie de manière unique.*

Le résultat de l'Exercice 6.27 reste valable pour un espace de Banach X général.

Remarque 6.28. *La dérivée faible est juste la dérivée au sens des distributions, mais pour la distribution u à valeurs vectorielles, c'est-à-dire dans l'espace X.*

Nous avons le lemme suivant, qui se montre de manière analogue que dans le cas de fonctions à valeurs scalaires.

Lemme 6.29. *Soit $u \in L^1_{\mathrm{loc}}(I,X)$ telle que $u'(t)=0$ pour tout $t \in I$. Alors, il existe $u_0 \in X$ tel que $u(t)=u_0$ pour presque tout $t \in I$.*

Preuve : Soit $\eta \in \mathcal{C}^\infty_c(I)$ telle que $\int_I \eta = 1$ et soit $a \in I$. Pour toute fonction $\varphi \in \mathcal{C}^\infty_c(I)$, on a

$$\varphi(t) = A\eta(t) + \psi'(t),$$

où $A = \int_I \varphi(t)\,dt$ et $\psi(t) = \int_a^t [\varphi(s) - A\eta(s)]\,ds$. On a alors

$$\begin{aligned}\int_I u(t)\varphi(t)\,dt &= A\int_I u(t)\eta(t)\,dt + \int_I u(t)\psi'(t)\,dt,\\ &= \left(\int_I \varphi(t)\,dt\right)u_0 - \int_I u'(t)\psi(t)\,dt = u_0\left(\int_I \varphi(t)\,dt\right),\end{aligned}$$

où $u_0 := \int_I A\eta(t)u(t)\,dt$. En utilisant des arguments similaires à ceux de l'Exercice 6.27, ceci implique bien que $u(t)=u_0$ pour presque tout $t \in I$. ♢

La dérivée faible possède des propriétés intéressantes qui nous seront utiles par la suite, que nous donnons dans la Proposition 6.30.

Proposition 6.30. *Soit X un espace de Banach. Soit $u \in L^1_{\mathrm{loc}}(]a,b[,X)$ tel que $u' \in L^1_{\mathrm{loc}}(]a,b[,X)$. Alors,*

$$u(t) - u(s) = \int_s^t u'(\tau)\,d\tau, \qquad \textit{pour presque tout } t,s \in I. \tag{6.3}$$

Preuve : Montrons d'abord qu'il existe $u_0 \in X$ tel que $u(t) - \int_s^t u'(\tau)\,d\tau = u_0$ pour presque tout $t \in I$. Notons $v(t) := \int_s^t u'(\tau)\,d\tau$ pour tout $t \in I$. Montrons tout d'abord que

$$v'(t) = u'(t).$$

Soit $\varphi \in \mathcal{C}^\infty_c(I,\mathbb{R})$ et soit $c,d \in]a,b[$ tel que $[c,d] \subset]a,b[$ et $\{s\} \cup \mathrm{Supp}\varphi \subset [c,d]$.

$$\int_{]c,d[} v'(t)\varphi(t)\,dt = -\int_{]a,b[} \varphi'(t)\left(\int_s^t u'(\tau)\,d\tau\right)dt.$$

La fonction $(t,\tau) \in [c,d]\times[c,d] \mapsto \varphi'(t)u'(\tau)$ est une fonction intégrable sur $[c,d]\times$

$[c,d]$. On peut donc appliquer le théorème de Fubini (voir Exercice 6.24) pour obtenir

$$\begin{aligned}-\int_{]c,d[}\varphi'(t)\left(\int_s^t u'(\tau)\,d\tau\right)dt &= -\int_s^d\left(\int_\tau^d \varphi'(t)u'(\tau)\,dt\right)d\tau+\int_c^s\left(\int_c^\tau \varphi'(t)u'(\tau)\,dt\right)d\tau\\ &=\int_s^d\varphi(\tau)u'(\tau)\,d\tau+\int_c^s\varphi(\tau)u'(\tau)\,d\tau\\ &=\int_{[c,d]}\varphi'(\tau)u'(\tau)\,d\tau.\end{aligned}$$

Cette dernière égalité prouve bien que $u'(t)=v'(t)$ pour tout $t\in]a,b[$. Donc il existe $u_0\in X$ tel que

$$u(t)=u_0+\int_s^t u'(\tau)\,d\tau. \tag{6.4}$$

Il nous reste à prouver que $u_0=u(s)$. En utilisant le théorème de convergence dominée, on peut montrer aisément que la fonction v est continue (voir Exercice 6.22). Ceci implique, en utilisant la formule (6.4) que la fonction u est continue. De plus, toujours en utilisant le théorème de convergence dominée, on montre que $v(t)\underset{t\to s}{\longrightarrow}0$, ce qui montre que nécessairement $u_0=u(s)$. ♢

Exercice 6.31. *Soit X un espace de Banach. Soit $u\in L^1_{\rm loc}(]a,b[,X)$ tel que $u'\in L^1_{\rm loc}(]a,b[,X)$. Montrer que*

$$\lim_{h\to0}\frac{u(t+h)-u(t)}{h}=u'(t)\ \textit{dans } X,\qquad \textit{pour presque tout } t\in I, \tag{6.5}$$

6.2.3 Théorème de Aubin-Lions

S'il n'est pas très difficile de définir ce qu'est une solution faible pour une équation aux dérivées partielles (linéaire), il n'est en revanche *a priori* pas du tout évident de donner un sens précis aux conditions initiales : que peut bien vouloir dire $u(t=0)=u_0$ si u n'est définie que presque partout en t ?

Voici maintenant un résultat fournissant une meilleure régularité pour u lorsque l'on sait dans quel espace fontionnel vit u', et qui va être crucial dans la suite pour définir correctement les conditions initiales. Ceci est à comparer avec les injections de Sobolev usuelles en dimension un.

On considère un espace de Hilbert H séparable que l'on identifie avec son dual, et un autre espace de Hilbert V tel que $V\hookrightarrow H$ (injection continue), avec V dense dans H. On a donc

$$V\hookrightarrow H=H'\hookrightarrow V'.$$

L'espace H est alors l'espace pivot dans les inclusions ci-dessus.

Exemple : L'exemple que l'on utilisera le plus dans ce cours est celui où $V=H^1_0(\Omega)$, $H=L^2(\Omega)$ et $V'=H^{-1}(\Omega)$.

Théorème 6.32 (Théorème de Aubin-Lions). *Soient $a, b \in \mathbb{R}$. Si $u \in L^2(]a;b[,V)$ est tel que $u' \in L^2(]a;b[,V')$, alors on a :*

1. *$u \in C^0([a;b],H)$;*
2. *$\sup_{t\in[a;b]} \|u(t)\|_H \leq C\left(\|u\|_{L^2(]a;b[,V)} + \|u'\|_{L^2(]a;b[,V')}\right)$ pour une constante C ne dépendant pas de u ;*
3. *Soient $u, v \in L^2(]a;b[,V)$ tels que $u', v' \in L^2(]a;b[,V')$. Alors la fonction $t \mapsto \langle u(t), v(t)\rangle_H$ est absolument continue et on a*

$$\frac{d}{dt}\langle u(t), v(t)\rangle_H = {}_{V'}\langle u'(t), v(t)\rangle_V + {}_{V'}\langle v'(t), u(t)\rangle_V.$$

Nous donnerons la preuve uniquement dans le cas où $V = H = V'$. La preuve dans le cas général est plus longue : nous renvoyons par exemple à [5, Chap. XVIII § 1] pour le cas général ou à [7] pour le cas où $V = H_0^1(\Omega)$ et $H = L^2(\Omega)$.

Preuve : [Théorème 6.32]

L'idée lorsque $V = H$ est d'utiliser la formule (6.3). D'après la Proposition 6.30, on a

$$u(t) - u(s) = \int_s^t u'(\tau)\, d\tau,$$

pour presque tout $s, t \in]a, b[$.

Notons que l'intégrale du terme à droite a un sens puisque $u' \in L^2(]a;b[,H)$ par hypothèse et que $1_{[s,t]} \in L^2(]a;b[)$. Les points *1.* et *2.* du théorème sont des conséquences faciles de la formule 6.3. En effet, on a par l'inégalité de Cauchy-Schwarz

$$\|u(t) - u(s)\|_H \leq |t-s|^{1/2} \|u'\|_{L^2(]a;b[,H)}$$

qui démontre la continuité (en fait $t \mapsto u(t)$ est même Hölder). En utilisant l'inégalité triangulaire et en intégrant par rapport à s, on trouve aussi

$$(b-a)\|u(t)\|_H \leq (b-a)^{1/2}\|u\|_{L^2(]a;b[,H)} + (b-a)^{3/2}\|u'\|_{L^2(]a;b[,H)}$$

donc

$$\sup_{t\in[a;b]} \|u(t)\|_H \leq (b-a)^{-1/2}\|u\|_{L^2(]a;b[,H)} + (b-a)^{1/2}\|u'\|_{L^2(]a;b[,H)}.$$

♢

Remarque 6.33. *Si $u \in L^2(]a;b[,V)$, alors on a également $u \in L^2(]a;b[,V')$ car $V \hookrightarrow V'$. L'hypothèse du théorème signifie simplement que u est différentiable dans V' (au sens de la définition 6.26 avec $X = V'$), et que sa dérivée u' appartient en plus à $L^2(]a;b[,V')$.*

Remarque 6.34. *Introduisons l'espace fonctionnel (de Banach)*

$$W(]a;b[,V,V') := \{u \in L^2(]a;b[,V) \mid u' \in L^2(]a;b[,V')\}.$$

Alors les points 1. *et* 2. *du Théorème 6.32 signifient que l'on a une injection continue*

$$W(]a;b[,V,V') \hookrightarrow C^0([a;b],H)$$

où $C^0([a;b],H)$ *est muni de la norme uniforme* $\|\cdot\|_{L^\infty([a;b],H)}$.

Ceci est très important car cela permet en particulier de donner un sens à $u(a)$ *et* $u(b)$ *dans* H*, donc aux conditions aux limites.*

Remarque 6.35. *En prenant* $u = v$ *dans* 3.*, on trouve que*

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|_H^2 = {}_{V'}\langle u'(t), u(t)\rangle_V.$$

De même on trouve que si $v \in V$*, alors* $t \mapsto \langle u(t), v\rangle_V$ *est absolument continue et on a*

$${}_{V'}\langle u'(t), v\rangle_V = \frac{d}{dt}\langle u(t), v\rangle_H.$$

Remarque 6.36. *Dans la pratique, nous utiliserons souvent le théorème précédent avec*

$$V = H_0^1(\Omega) \subset H = L^2(\Omega) \subset V' = H^{-1}(\Omega)$$

où Ω *est un ouvert borné de* $\mathbb{R}^n$ *(ou* $\Omega = \mathbb{R}^n$*). Le choix de* $V = H_0^1(\Omega)$ *correspond aux conditions au bord de Dirichlet.*

On obtient donc que si $u \in L^2(]0;T[,H_0^1(\Omega))$ *est tel que* $u' \in L^2(]0;T[,H^{-1}(\Omega))$*, alors* $u \in C^0([0;T],L^2(\Omega))$*, donc* $u(0)$ *et* $u(T)$ *ont un sens dans* $L^2(\Omega)$*.*

Nous étudions dans le reste de ce chapitre avec plus de détails l'équation de la chaleur, qui est l'exemple prototype par excellence d'une équation parabolique.

Nous commencerons par le cas simple de tout l'espace avant de traiter celui d'un domaine borné. Nous n'aborderons pas le cas d'un domaine non borné différent de $\mathbb{R}^n$ dont l'approche classique est basée sur la théorie des semi-groupes et le théorème de Hille-Yosida.

6.3 L'équation de la chaleur dans tout l'espace

Commençons par chercher une solution particulière $G(t,x)$ régulière (pour $t > 0$) de l'équation de la chaleur

$$\frac{\partial}{\partial t}G - \Delta G = 0.$$

Pour cela, on utilise la tranformée de Fourier définie par

$$\widehat{f}(k) = (2\pi)^{-n/2}\int_{\mathbb{R}^n} f(x)e^{-ik\cdot x}dx.$$

On trouve donc que $\widehat{G}$ doit résoudre l'équation suivante

$$\frac{\partial}{\partial t}\widehat{G}(t,k) + |k|^2\widehat{G}(t,k) = 0, \tag{6.6}$$

c'est-à-dire

$$\widehat{G}(t,k) = Ce^{-t|k|^2}.$$

Remarquons que G n'est bien définie que lorsque $t > 0$. Si $t = 0$, $\widehat{G} = C$ donc G est égal à une constante multipliée par la distribution δ. Si $t < 0$, $\widehat{G}$ n'est pas dans $\mathcal{S}'$ et on ne peut pas définir G. On voit donc apparaître dès maintenant une propriété importante de l'équation de la chaleur : la *non-réversibilité*. La solution ne sera définie que pour les temps futurs, c'est-à-dire $t \geq 0$ si la condition initiale est donnée en $t = 0$.

Revenant dans les variables d'espace et choisissant la constante C de façon adéquate, on obtient une solution de l'équation de la chaleur :

$$G(t,x) = (4\pi t)^{-n/2}e^{\frac{-|x|^2}{4t}}$$

avec

$$\forall t > 0, \qquad \int G(t,x)dx = 1.$$

Comme on a $G(t,\cdot) \to \delta$ quand $t \to 0$ au sens des distributions, on dit que G est la solution fondamentale de l'équation de la chaleur, c'est-à-dire formellement celle de

$$\begin{cases} \dfrac{\partial}{\partial t}G - \Delta G = 0, & t > 0 \\ G(0) = \delta. \end{cases} \tag{6.7}$$

Remarque 6.37. *Une autre façon de trouver la fonction G est de remarquer que si $u(t,x)$ est une solution de l'équation de la chaleur, alors $u(\lambda^2 t, \lambda x)$ l'est également. Il est donc naturel de chercher une fonction solution sous la forme $u(t,x) = v(|x|^2/t)$. On tombe alors sur la même fonction G.*

On peut maintenant utiliser la fonction G pour construire une solution de l'équation de la chaleur avec condition initiale g. En fait on remarque que (6.6) reste vérifiée si on multiplie $\widehat{G}$ par une fonction ne dépendant que de k, ce qui revient à faire une convolution dans l'espace de départ. On introduit donc pour $x \in \mathbb{R}^n$ et $t > 0$

$$u(t)(x) = (G(t,\cdot) * g)(x) = \int_{\mathbb{R}^n} G(t,x-y)g(y)dy = (4\pi t)^{-n/2}\int_{\mathbb{R}^n} e^{\frac{-|x-y|^2}{4t}}g(y)dy. \tag{6.8}$$

Comme $G(t,\cdot) \in L^1(\mathbb{R}^n)$ pour tout $t > 0$, on déduit que si $g \in L^p(\mathbb{R}^n)$, alors $u(t) \in L^p(\mathbb{R}^n)$ pour tout $t > 0$.

Théorème 6.38 (Solution de l'équation de la chaleur dans $\mathbb{R}^n$)**.** *Soit $g \in L^2(\mathbb{R}^n)$. Le problème*

$$\begin{cases} \dfrac{\partial}{\partial t}u - \Delta u = 0, & t > 0 \\ u(0) = g, \end{cases} \tag{6.9}$$

a une solution unique $u \in C^0([0;\infty), L^2(\mathbb{R}^n)) \cap C^1((0;\infty), H^2(\mathbb{R}^n))$, donnée par la formule (6.8)*.*

Preuve : Il est clair que la définition (6.8) fournit une solution

$$u \in C^0([0;\infty), L^2(\mathbb{R}^n)) \cap C^1((0;\infty), H^2(\mathbb{R}^n))$$

de l'équation (6.9), le vérifier en exercice.

Si maintenant $v \in C^0([0;\infty), L^2(\mathbb{R}^n)) \cap C^1((0;\infty), H^2(\mathbb{R}^n))$ résout (6.9) avec $g \equiv 0$, on peut prendre le produit scalaire avec la fonction $x \mapsto v(t,x) \in L^2(\mathbb{R}^n)$ et on intègre sur $[0;t_0]$ avec $t_0 \leq T$. On obtient

$$\|v(t_0,\cdot)\|^2_{L^2(\mathbb{R}^n)} + \int_0^{t_0} dt \int_{\mathbb{R}^n} dx |\nabla v(t,x)|^2 = \frac{1}{2}\|v(0,\cdot)\|^2_{L^2(\mathbb{R}^n)} = 0$$

qui démontre l'unicité. ♢

Proposition 6.39. *Si $g \in L^2(\mathbb{R}^n)$, la solution* (6.8) *de* (6.9) *est dans $C^\infty((0;\infty)\times\mathbb{R}^n)$.*

Preuve : Il s'agit juste de remarquer que G est de classe C^∞ sur $(0;\infty)\times\mathbb{R}^n$ et que toutes ses dérivées sont dans $C^0((0;\infty), L^2(\mathbb{R}^n))$, puis d'appliquer les résultats classiques de régularité d'intégrales dépendant d'un paramètre. ♢

Ainsi, bien que nous ayons seulement supposé $g \in L^2(\mathbb{R}^n)$, on obtient que la solution $u(t,x)$ est de classe $\mathcal{C}^\infty$ par rapport à x pour tout $t > 0$. On dit que l'équation de la chaleur a un *effet régularisant.*

De même, notons que si $g \geq 0$ alors $u(t,x) > 0$ pour tout $x \in \mathbb{R}^n$ et $t > 0$, puisque $G > 0$. Même si g s'annule par endroits au temps initial, la solution sera strictement positive sur tout l'espace quand $t > 0$. On parle de *propagation à vitesse infinie.*

Ces propriétés de l'équation de la chaleur sont très spécifiques aux équations de type parabolique et ne seront plus vraies pour l'équation des ondes, par exemple. Démontrer ces propriétés dans le cas d'un ouvert borné nous prendra un peu plus de temps mais tout restera vrai.

Remarque 6.40. *Introduisons l'opérateur agissant sur $L^2(\mathbb{R}^n)$ défini par $U(t)g = G(t,\cdot) * g$. C'est en fait juste l'opérateur de multiplication par $\widehat{G}(t,k)$ en Fourier (à une constante multiplicative près). Il est facile de montrer que c'est un semi-groupe, c'est-à-dire qu'il vérifie les propriétés suivantes :*

$$U(0) = I, \qquad U(t)U(s) = U(t+s)$$

et la solution de l'équation de la chaleur avec condition initiale g s'écrit justement $U(t)g$. En fait on a "formellement" (on peut donner un sens mathématique précis) $U(t) = e^{t\Delta}$.

Exercice 6.41 (Principe du maximum)**.** *On suppose que $g \in L^2(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$. Montrer que la solution u donnée par* (6.8) *est dans $L^\infty((0;\infty), L^\infty(\mathbb{R}^n))$ et que*

$$\sup_{t>0} \|u(t)\|_{L^\infty(\mathbb{R}^n)} \leq \|g\|_{L^\infty(\mathbb{R}^n)}.$$

Exercice 6.42 (Comportement asymptotique)**.** *Montrer que*

$$\forall t > 0, \qquad \lim_{|x|\to\infty} u(t,x) = 0,$$

$$\forall x \in \mathbb{R}^n, \qquad \lim_{t\to\infty} u(t,x) = 0.$$

Exercice 6.43 (Équation de la chaleur dans tout l'espace avec second membre)**.** *Soient $g \in L^2(\mathbb{R}^n)$ et $f \in C^1([0;\infty), L^2(\mathbb{R}^n))$. Montrer que le problème*

$$\begin{cases} \dfrac{\partial}{\partial t}u - \Delta u = f, \quad t > 0 \\ u(0) = g, \end{cases} \tag{6.10}$$

admet une solution unique $u \in C^0([0;\infty), L^2(\mathbb{R}^n)) \cap C^1((0;\infty), L^2(\mathbb{R}^n))$, donnée par la formule de Duhamel

$$u(t) = U(t)g + \int_0^t U(t-s)f(s)\,ds. \tag{6.11}$$

6.4 L'équation de la chaleur sur un ouvert borné Ω

Pour étudier l'équation de la chaleur sur un ouvert borné, on ne peut utiliser la transformée de Fourier comme nous l'avons fait dans tout l'espace.

Considérons un ouvert borné $\Omega \subset \mathbb{R}^n$ et un réel $T > 0$. On désire résoudre

$$\begin{cases} \dfrac{\partial}{\partial t}u - \Delta u = f, & \text{dans } (0;T)\times\Omega \\ u_{|(0;T)\times\partial\Omega} = 0 & \text{(conditions au bord de Dirichlet)} \\ u(0,x) = g(x). \end{cases} \tag{6.12}$$

6.4.1 Théorème d'existence de solutions faibles

On considère $g \in L^2(\Omega)$ et $f \in L^2(]0;T[, L^2(\Omega))$.

Définition 6.44 (Solutions faibles)**.** *Soit* $u \in L^2(]0;T[, H_0^1(\Omega))$ *telle que* $u' \in L^2(]0;T[, H^{-1}(\Omega))$*. On dit que* u *est une* solution faible *de* (6.12) *si on a*

$(C1)$ ${}_{H^{-1}(\Omega)}\langle u', v\rangle_{H_0^1(\Omega)} + \displaystyle\int_\Omega \nabla u(t)\cdot\nabla v = \int_\Omega f(t)v$ *pour tout* $v \in H_0^1(\Omega)$ *et presque partout en* $t \in]0;T[$.

$(C2)$ $u(0) = g$.

Rappelons que d'après le Théorème 6.32, on a $u \in C^0([0;T], L^2(\Omega))$ qui permet de donner un sens à $(C2)$. Rappelons aussi que pour tout $v \in H_0^1(\Omega)$,

$$\frac{d}{dt}\langle u, v\rangle_{L^2(\Omega)} = {}_{H^{-1}(\Omega)}\langle u', v\rangle_{H_0^1(\Omega)}.$$

Dans $(C1)$, la fonction v ne dépend pas du temps.

Le but de cette section est principalement de démontrer le résultat suivant :

Théorème 6.45 (Existence et unicité de solutions faibles)**.** *Soit* $\Omega \subset \mathbb{R}^n$ *un ouvert régulier. On suppose que* $g \in L^2(\Omega)$ *et* $f \in L^2(]0;T[, L^2(\Omega))$*. Alors le problème* (6.12) *admet une unique solution faible* u*. De plus, il existe une constante* $C > 0$ *qui ne dépend que de* Ω*, indépendante de* f *et* g*, telle que*

$$\max_{0\le t\le T}\|u(t)\|_{L^2(\Omega)} + \|u\|_{L^2(]0;T[,H_0^1(\Omega))} + \|u'\|_{L^2(]0;T[,H^{-1}(\Omega))} \le C\left(\|f\|_{L^2(]0;T[,L^2(\Omega))} + \|g\|_{L^2(\Omega)}\right). \tag{6.13}$$

Remarque 6.46. *L'estimée (6.13) montre que le problème de la chaleur est bien posé au sens de Hadamard. En effet, soit* $f_1, f_2 \in L^2(]0;T[, L^2(\Omega))$*,* $g_1, g_2 \in L^2(\Omega)$ *et notons* u_1, u_2 *les solutions de l'équation de la chaleur associées respectivement à* (f_1, g_1) *et* (f_2, g_2)*. Par linéarité de l'équation de la chaleur, la fonction* $u_1 - u_2$ *est alors solution de l'équation de la chaleur associée aux données* $(f_1 - f_2, g_1 - g_2)$*. L'estimée (6.13) montre alors que*

$$\begin{aligned}&\max_{0\le t\le T}\|u_1(t) - u_2(t)\|_{L^2(\Omega)} + \|u_1 - u_2\|_{L^2(]0;T[,H_0^1(\Omega))} + \|u_1' - u_2'\|_{L^2(]0;T[,H^{-1}(\Omega))}\\ &\le C\left(\|f_1 - f_2\|_{L^2(]0;T[,L^2(\Omega))} + \|g_1 - g_2\|_{L^2(\Omega)}\right).\end{aligned}$$

Autrement dit, la solution u *de l'équation de la chaleur varie continûment par rapport aux données de l'équation, à savoir* f *et* g*.*

Il existe deux méthodes de preuve de ce théorème, que nous allons voir dans ce cours. La première méthode présentée est appelée méthode par approximation de Galerkin. Cette méthode est utile car elle est à l'origine de la méthode d'approximation numérique la plus couramment utilisée pour discrétiser ce type d'équations paraboliques, à savoir la méthode des éléments finis. La deuxième méthode, dite

méthode spectrale, utilise la décomposition de la solution sur les fonctions propres de l'opérateur Laplacien. Cette dernière permet de prouver très facilement des propriétés qualitatives fines sur le comportement de la solution, qui ne pourraient pas être facilement accessibles via une méthode d'approximation de Galerkin.

C'est pour cette raison, nous vous présentons ces deux approches dans le détail dans le cadre de ce cours.

Preuve :

Dans cette méthode, on va chercher à approcher une solution faible par des solutions de problèmes approchés définis dans des espaces de dimension finie (approximations de Galerkin).

Étape 1 : Approximations de Galerkin.

Considérons une famille de fonctions $(w_k)_{k\geq 1} \subset H_0^1(\Omega)$, telle que
— $(w_k)_{k\geq 1}$ est une base *orthogonale* de $H_0^1(\Omega)$;
— $(w_k)_{k\geq 1}$ est une base *orthonormée* de $L^2(\Omega)$.
Par exemple, on peut prendre les fonctions propres du Laplacien avec conditions de Dirichlet au bord de Ω, qui vérifient $-\Delta w_k = \lambda_k w_k$ où $\mathrm{Sp}_{H_0^1(\Omega)}(-\Delta) = \{\lambda_k\}$. Voir le théorème 3.3.

On pose alors

$$V_m := \mathrm{Vect}(w_1, ..., w_m)$$

et on cherche une solution $u_m \in C^0([0;T], V_m)$ faible dans V_m, c'est-à-dire vérifiant

$(i)_m$ $\langle u_m'(t), w_k\rangle_{L^2} + \displaystyle\int_\Omega \nabla u_m(t)\cdot\nabla w_k = \langle f(t), w_k\rangle_{L^2}$ pour tout $k = 1...m$ et presque partout en $t \in [0;T]$;

$(ii)_m$ $\langle u_m(0), w_k\rangle_{L^2(\Omega)} = \langle g, w_k\rangle_{L^2(\Omega)}$ pour tout $k = 1...m$.

Si on écrit

$$u_m(t,x) = \sum_{k=1}^m d_k^m(t) w_k(x),$$

alors $(i)_m$ et $(ii)_m$ équivalent à

$(i)'_m$ $\dfrac{d}{dt} d_k^m(t) + d_k^m(t)\,\|\nabla w_k\|^2_{L^2(\Omega)} = \langle f(t), w_k\rangle_{L^2(\Omega)}$, $\forall k = 1...m$, p.p. $t \in [0;T]$;

$(ii)'_m$ $d_k^m(0) = \langle g, w_k\rangle_{L^2(\Omega)}$, $\forall k = 1...m$.

Il s'agit d'un système (diagonal) d'équations différentielles ordinaires, dont on admettra qu'il admet une unique solution absolument continue $(d_k^m(t))_{k=1}^m$, définie sur tout $[0;T]$. La preuve découle d'une extension du théorème de Cauchy-Lipschitz que vous connaissez.

Exercice 6.47. *Prouver ce résultat dans le cas où $f \in C^1([0,T]; L^2(\Omega))$.*

Étape 2 : estimées d'énergie.

On désire maintenant passer à la limite quand $m \to \infty$. Pour cela, nous commençons par démontrer le lemme suivant.

Lemme 6.48 (Estimées d'énergie)**.** *Il existe une constante C qui ne dépend que de Ω et $T > 0$ telle que pour tout $m \geq 1$,*

$$\max_{0\leq t\leq T} \|u_m(t)\|_{L^2(\Omega)} + \|u_m\|_{L^2(]0;T[,H_0^1(\Omega))} + \|u'_m\|_{L^2(]0;T[,H^{-1}(\Omega))} \leq C\left(\|f\|_{L^2(]0;T[,L^2(\Omega))} + \|g\|_{L^2(\Omega)}\right). \quad (6.14)$$

Remarque 6.49. *Il est clair que les solutions dépendent linéairement de g et f : si $u_{m,1}$ et $u_{m,2}$ sont des solutions faibles associées aux problèmes avec respectivement (f_1, g_1) et (f_2, g_2), alors $u_{m,1} + u_{m,2}$ est solution du problème associé au couple $(f_1 + f_2, g_1 + g_2)$. On parle de* principe de superposition. *Une fois que nous aurons démontré l'unicité de la solution, on obtient donc une application linéaire qui à tout (f, g) associe la solution u_m. Alors la formule* (6.22) *signifie que cette application linéaire est* continue *dans les bons espaces fonctionnels.*

Preuve : (du Lemme 6.48). Par linéarité, on peut prendre $w = u_m$ dans $(i)_m$. On obtient :

$$\langle u'_m, u_m\rangle_{L^2} + \int_\Omega |\nabla u_m(t)|^2 = \langle f(t), u_m(t)\rangle_{L^2}. \quad (6.15)$$

On a

$$\langle f(t), u_m(t)\rangle_{L^2} \leq \frac{1}{2}\left(\|f(t)\|^2_{L^2(\Omega)} + \|u_m(t)\|^2_{L^2(\Omega)}\right)$$

donc,

$$\langle u'_m, u_m\rangle_{L^2} + \int_\Omega |\nabla u_m(t)|^2 \leq \frac{1}{2}\left(\|f(t)\|^2_{L^2(\Omega)} + \|u_m(t)\|^2_{L^2(\Omega)}\right). \quad (6.16)$$

En posant $\eta(t) = \|u_m(t)\|^2_{L^2(\Omega)}$, on obtient

$$\eta'(t) \leq \frac{1}{2}\eta(t) + \frac{1}{2}\|f(t)\|^2_{L^2(\Omega)}.$$

D'après le lemme de Gronwall, on déduit

$$\begin{aligned}\|u_m(t)\|^2_{L^2(\Omega)} &\leq e^{t/2}\left(\|g\|^2_{L^2(\Omega)} + \frac{1}{2}\int_0^t \|f(s)\|^2_{L^2(\Omega)}\, ds\right)\\ &\leq e^{T/2}\left(\|g\|^2_{L^2(\Omega)} + \frac{1}{2}\|f\|^2_{L^2(]0;T[,L^2(\Omega))}\right).\end{aligned}$$

Ceci fournit bien l'estimée sur $\max_{0\leq t\leq T}\|u_m(t)\|_{L^2(\Omega)}$. Intégrons maintenant l'inégalité (6.24) sur $[0; T]$. Nous obtenons

$$\|u_m(T)\|^2_{L^2(\Omega)} + \|u_m\|^2_{L^2(]0;T[,H_0^1(\Omega))} \leq \|g\|^2_{L^2(\Omega)} + \frac{1}{2}\|f\|^2_{L^2(]0;T[,L^2(\Omega))} + T\max_{0\leq t\leq T}\|u_m(t)\|_{L^2(\Omega)}.$$

On déduit alors l'estimée sur $\|u_m\|_{L^2(]0;T[,H_0^1(\Omega))}$ en utilisant ce que nous avons déjà démontré pour majorer le dernier terme ci-dessus.

On estime maintenant $\|u'_m\|_{L^2(]0;T[,H^{-1}(\Omega))}$ par dualité. On considère une fonction fixée $v \in H_0^1(\Omega)$, telle que $\|v\|_{H_0^1(\Omega)} \leq 1$. On peut alors écrire $v = v^1 + v^2$ avec $v^1 \in V_m$ et $v^2 \in V_m^\perp = \text{Vect}(w_k,\ k \geq m+1)$. Bien sûr $\|v^1\|_{H_0^1(\Omega)} \leq 1$. On a alors

$$\begin{aligned} {}_{H^{-1}(\Omega)}\langle u'_m, v\rangle_{H_0^1(\Omega)} &= {}_{L^2(\Omega)}\langle u'_m, v^1\rangle_{L^2(\Omega)} \\ &= \langle f(t), v^1\rangle_{L^2} - \int_\Omega \nabla u_m(t)\cdot\nabla v^1. \end{aligned}$$

Ceci démontre que

$$\|u'_m\|_{H^{-1}(\Omega)} \leq \|f(t)\|_{L^2(\Omega)} + \|\nabla u_m\|_{L^2(\Omega)}.$$

On obtient alors l'estimée voulue en passant au carré, en intégrant sur $[0;T]$ et en utilisant les résultats précédents. ♢

Étape 3 : existence.

Nous pouvons maintenant démontrer l'existence d'au moins une solution en passant à la limite faible.

Comme u_m et u'_m sont des suites bornées, respectivement dans les espaces de Hilbert $L^2(]0;T[,H_0^1(\Omega))$ et $L^2(]0;T[,H^{-1}(\Omega))$, on peut extraire des sous-suites $u_{\varphi(m)}$ et $u'_{\varphi(m)}$ telles que $u_{\varphi(m)} \rightharpoonup u$ dans $L^2(]0;T[,H_0^1(\Omega))$ et $u'_{\varphi(m)} \rightharpoonup v$ dans $L^2(]0;T[,H^{-1}(\Omega))$. Il est facile de voir que $v = u'$, donc que $u \in C^0([0;T],L^2(\Omega))$.

Soit maintenant une fonction test de la forme

$$v(t) = \sum_{k=1}^{M} d_k(t) w_k, \tag{6.17}$$

avec $d_k(t)$ des fonctions régulières de t. Comme $v(t) \in V_m$ pour tout $m \geq M$ et tout $t \in [0;T]$, on a d'après (6.23)

$$\int_0^T \langle u'_m(t), v(t)\rangle_{L^2} dt + \int_0^T \int_\Omega \nabla u_m(t)\cdot\nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2} dt.$$

Par convergence faible pour la sous-suite $u_{\varphi(m)}$, on a donc

$$\int_0^T \langle u'(t), v(t)\rangle_{L^2} dt + \int_0^T \int_\Omega \nabla u(t)\cdot\nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2} dt$$

pour tout $v(t)$ de la forme (6.17) ci-dessus, donc pour tout $v \in L^2(]0;T[,H_0^1(\Omega))$ par densité. Ainsi, u est une solution faible de l'équation.

Il reste à vérifier que $u(0) = g$. Soit pour cela une fonction v de la forme (6.17) qui est régulière et satisfait de plus $v(T) \equiv 0$. En intégrant l'égalité ci-dessus par parties, on obtient

$$-\int_0^T \langle u(t), v'(t)\rangle_{L^2} dt + \int_0^T \int_\Omega \nabla u(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2} dt + \langle u(0), v(0)\rangle.$$

En intégrant par parties l'équation pour u_m, on trouve de même :

$$-\int_0^T \langle u_m(t), v'(t)\rangle_{L^2} dt + \int_0^T \int_\Omega \nabla u_m(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2} dt + \langle g, v(0)\rangle.$$

Par passage à la limite faible, on trouve donc

$$\langle u(0), v(0)\rangle = \langle g, v(0)\rangle,$$

c'est-à-dire $u(0) = g$ puisque $v(0)$ était quelconque.

♢

Preuve : [Preuve 2 : Décomposition sur les fonctions propres du Laplacien] On utilise la méthode de décomposition sur les fonctions propres du Laplacien.

Étape 1 : forme de la solution.

Soit $u \in L^2(]0;T[, H_0^1(\Omega))$ telle que $u' \in L^2(]0;T[, H^{-1}(\Omega))$, une solution faible de (6.12). D'après le Théorème 6.32, on a $u \in C^0([0;T], L^2(\Omega))$.

Considérons maintenant la famille $(w_k)_{k\geq 1} \subset H_0^1(\Omega)$ des fonctions propres du Laplacien avec conditions de Dirichlet au bord de Ω :

$$-\Delta w_k = \lambda_k w_k$$

où les λ_k sont les valeurs propres du Laplacien de Dirichlet, voir le Théorème 3.3. Rappelons que l'on a $\langle w_k, w_\ell\rangle = \delta_{k\ell}$ et $\langle \nabla w_k, \nabla w_\ell\rangle = \lambda_k \delta_{k\ell}$. Comme $u \in C^0([0;T], L^2(\Omega))$, on peut écrire pour tout t

$$u(t) = \sum_{k\geq 1} \alpha_k(t) w_k$$

où chaque $\alpha_k(t) = \langle u(t), w_k\rangle_{L^2(\Omega)}$ est une fonction absolument continue sur $[0;T]$ d'après le Théorème 6.32. En choisissant $v = w_k$ dans $(C1)$, on obtient que chaque α_k est une solution du problème

$$\begin{cases} \alpha_k'(t) + \lambda_k \alpha_k(t) = \beta_k(t) \quad \text{dans }]0;T[\\ \alpha_k(0) = \alpha_k^0 \end{cases}$$

où

$$\beta_k(t) = \langle f(t), w_k\rangle, \qquad \alpha_k^0 = \langle g, w_k\rangle.$$

Il s'agit pour chaque k d'une équation différentielle ordinaire dont l'unique solution est

$$\alpha_k(t) = \alpha_k^0 e^{-\lambda_k t} + \int_0^t \beta_k(s) e^{-\lambda_k(t-s)} ds, \qquad t > 0.$$

Ainsi, on trouve que u doit vérifier

$$u(t) = \sum_{k\geq 1} e^{-\lambda_k t} \langle g, w_k\rangle w_k + \int_0^t \sum_{k\geq 1} e^{-\lambda_k(t-s)} \langle f(s), w_k\rangle w_k \tag{6.18}$$

si cette formule a un sens. L'unicité est donc automatique si nous pouvons montrer que cette formule a un sens dans les espaces fonctionnels adaptés. Introduisons l'opérateur

$$U(t) = \sum_{k\geq 1} e^{-\lambda_k t} |w_k\rangle\langle w_k| \tag{6.19}$$

où la notation $|w_k\rangle\langle w_k|$ désigne le projecteur orthogonal dans $L^2(\Omega)$ sur $\mathrm{Vect}(w_k)$. Comme $-\Delta \geq 0$, on obtient que $\lambda_k \geq 0$ pour tout k, donc que pour chaque $t > 0$ $U(t)$ définit un opérateur auto-adjoint borné tel que

$$0 \leq U(t) \leq 1. \tag{6.20}$$

La formule (6.18) s'écrit alors

$$u(t) = U(t)g + \int_0^t U(t-s) f(s)\, ds, \tag{6.21}$$

expression à rapprocher de (6.11). Si nous pouvons montrer que cette formule fournit bien une fonction $u \in L^2(]0;T[, H_0^1(\Omega))$ telle que $u' \in L^2(]0;T[, H^{-1}(\Omega))$, nous aurons démontré à la fois l'existence et l'unicité.

Étape 2 : propriétés du propagateur $U(t)$.

Nous démontrons ici certaines propriétés utiles de l'opérateur $U(t)$. Nous noterons $B(X,Y)$ l'espace de Banach des opérateurs auto-adjoints bornés entre les espaces de Hilbert X et Y, muni de la norme usuelle

$$\|U\|_{X\to Y} = \sup_{x\in X,\ \|x\|_X=1} \|Ux\|_Y .$$

Remarquons que

$$U \in L^\infty([0;T], B(L^2(\Omega), L^2(\Omega)))$$

d'après (6.20).

Prouvons maintenant le

Lemme 6.50. *On a*

$$U \in B(L^2(\Omega), L^2([0,T[; H_0^1(\Omega)))$$

et

$$U' \in B(L^2(\Omega), L^2(]0,T[, H^{-1}(\Omega))).$$

Preuve : Commençons par estimer $\|U(t)\|_{L^2(\Omega)\to H_0^1(\Omega)}$. Pour cela, nous prenons une fonction $\psi \in L^2(\Omega)$ et calculons

$$\begin{aligned}\|U(t)\psi\|^2_{H_0^1(\Omega)} &= \|\nabla U(t)\psi\|^2_{L^2(\Omega)} \\ &= \sum_{k\geq 1} e^{-2t\lambda_k}\lambda_k \langle w_k,\psi\rangle^2\end{aligned}$$

d'après la formule de Parseval. Ainsi en utilisant le théorème de Fubini pour les fonctions positives, on a

$$\begin{aligned}\|U(t)\psi\|^2_{L^2(]0,T[,H_0^1(\Omega)} &= \int_{]0,T[}\sum_{k\geq 1} e^{-2t\lambda_k}\lambda_k\langle w_k,\psi\rangle^2 \\ &= \sum_{k\geq 1}\int_{]0,T[} e^{-2t\lambda_k}\lambda_k\, dt\langle w_k,\psi\rangle^2 \\ &= \sum_{k\geq 1}\frac{1}{2}\left(1-e^{-T\lambda_k}\right)\langle w_k,\psi\rangle^2 \\ &\leq \frac{1}{2}\left(1-e^{-T\lambda_1}\right)\sum_{k\geq 1}\langle w_k,\psi\rangle^2 \\ &= \frac{1}{2}\left(1-e^{-T\lambda_1}\right)\|\psi\|^2_{L^2(\Omega)}.\end{aligned}$$

On a donc bien que $U \in B(L^2(\Omega); L^2(]0;T[, H_0^1(\Omega)))$.

Ensuite, on remarque que

$$U'(t) = \sum_{k\geq 1}\lambda_k e^{-\lambda_k t}|w_k\rangle\langle w_k| = (-\Delta)U(t) = U(t)(-\Delta).$$

On a donc pour tout $(\varphi,\psi)\in L^2(\Omega)\times H_0^1(\Omega)$, et pour presque tout $t\in]0,T[$,

$$\begin{aligned}{}_{H^{-1}(\Omega)}\langle U'(t)\varphi,\psi\rangle_{H_0^1(\Omega)} &= {}_{H^{-1}(\Omega)}\langle(-\Delta)U(t)\varphi,\psi\rangle_{H_0^1(\Omega)} \\ &= {}_{L^2(\Omega)}\langle\nabla U(t)\varphi,\nabla\psi\rangle_{L^2(\Omega)} \\ &\leq \|\psi\|_{H_0^1(\Omega)}\|U(t)\varphi\|_{H_0^1(\Omega)}.\end{aligned}$$

Ainsi,

$$\|U'(t)\varphi\|_{H^{-1}(\Omega)} \leq \|U(t)\varphi\|_{H_0^1(\Omega)}$$

et

$$\|U'(t)\varphi\|_{L^2(]0,T[,H^{-1}(\Omega))} \leq \|U(t)\varphi\|_{L^2(]0,T[,H_0^1(\Omega))}\|U(t)\|_{L^2(\Omega)\to L^2(]0,T[,H_0^1(\Omega))}\|\varphi\|_{L^2(\Omega)}.$$

En conséquence, on obtient que

$$\|U'(t)\|_{L^2(\Omega)\to L^2(]0,T[,H^{-1}(\Omega))} \leq \|U(t)\|_{L^2(\Omega)\to L^2(]0,T[,H_0^1(\Omega))}.$$

♢

Étape 3 : conclusion.

Montrons maintenant que (6.21) fournit bien une fonction de $L^2(]0;T[,H_0^1(\Omega))$ telle que $u' \in L^2(]0;T[,H^{-1}(\Omega))$ en utilisant le Lemme 6.50. Posons

$$u_1(t) = U(t)g \quad \text{et} \quad u_2(t) = \int_0^t U(t-s)f(s)\,ds.$$

D'après le Lemme 6.50, on a d'abord que $u_1(t) \in L^2(]0,T[,H_0^1(\Omega))$. De plus, comme $u_1'(t) = U'(t)g$, on a, toujours d'après le Lemme 6.50, que $u_1' \in L^2(]0,T[,H^{-1}(\Omega)$.

On a aussi

$$\begin{aligned}
\|u_2(t)\|^2_{L^2(]0,T[,H_0^1(\Omega))} &= \int_{]0,T[} \|u_2(t)\|^2_{H_0^1(\Omega)}\,dt,\\
&= \int_{]0,T[} \|\int_0^t U(t-s)f(s)\,ds\|^2_{H_0^1(\Omega)}\,dt,\\
&\le \int_{]0,T[}\int_0^t \|U(t-s)f(s)\|^2_{H_0^1(\Omega)}\,ds\,t\,dt,\\
&\le T\int_{]0,T[}\int_0^t \|U(t-s)f(s)\|^2_{H_0^1(\Omega)}\,ds\,t\,dt,\\
&= T\int_{]0,T[}\int_0^t \|U(t-s)f(s)\|^2_{H_0^1(\Omega)}\,dt\,ds,\\
&= T\int_{]0,T[}\int_0^t \|U(t')f(s)\|^2_{H_0^1(\Omega)}\,dt'\,ds,\\
&= T\int_{]0,T[}\left(\int_{]0,T[} \|U(t')f(s)\|^2_{H_0^1(\Omega)}\,dt'\right)\,ds,\\
&= T\int_{]0,T[} \|U(\cdot)f(s)\|^2_{L^2(]0,T[,H_0^1(\Omega))}\,dt'\,ds,\\
&\le T\int_{]0,T[} \|U\|^2_{L^2(\Omega)\to L^2(]0,T[,H_0^1(\Omega))}\|f(s)\|^2_{L^2(\Omega)}\,ds,\\
&= T\|U\|^2_{L^2(\Omega)\to L^2(]0,T[,H_0^1(\Omega))}\|f\|^2_{L^2(]0,T[,L^2(\Omega))}.
\end{aligned}$$

Ceci montre bien que $u_2 \in L^2(]0,T[,H_0^1(\Omega))$.

Finalement, on a

$$u_2'(t) = f(t) + \int_0^t U'(t-s)f(s)\,ds.$$

Or $f \in L^2(]0;T[,L^2(\Omega)) \subset L^2(]0;T[,H^{-1}(\Omega))$ et le second terme est traité comme ci-dessus et on obtient bien que $u_2' \in L^2(]0,T[,H^{-1}(\Omega))$.

Ainsi on obtient en particulier que u appartient à $C^0([0;T],L^2(\Omega))$. Notons que u satisfait par construction la formulation faible (i) pour $v=w_k$ pour tout $k\geq 1$, donc pour tout $v\in H_0^1(\Omega)$. Il faut encore vérifier que l'on a bien

$$\lim_{t\to 0}\|u(t)-g\|_{L^2(\Omega)}=0.$$

Notons d'abord que

$$\begin{aligned}\left\|\int_0^t U(t-s)f(s)\,ds\right\|_{L^2(\Omega)} &\leq \int_0^t \|U(t-s)\|_{L^2(\Omega)\to L^2(\Omega)}\,\|f(s)\|_{L^2(\Omega)}\\ &\leq \sqrt{t}\,\|f\|_{L^2(]0;T[,L^2(\Omega))}\end{aligned}$$

où nous avons utilisé que $\|U(t)\|_{L^2(\Omega)\to L^2(\Omega)}\leq 1$ et l'inégalité de Cauchy-Schwarz. Ceci démontre que le dernier terme de (6.21) tend vers 0 dans $L^2(\Omega)$ quand $t\to 0$. Ainsi, nous devons juste prouver le

Lemme 6.51. *Soit $g\in L^2(\Omega)$. Alors on a*

$$\lim_{t\to 0}\|U(t)g-g\|_{L^2(\Omega)}=0.$$

Preuve : On a, comme $(w_k)_{k\geq 1}$ est une base orthonormée de $L^2(\Omega)$,

$$U(t)g-g=\sum_{k\geq 1}(e^{-\lambda_k t}-1)\langle g,w_k\rangle w_k$$

donc

$$\|U(t)g-g\|^2_{L^2(\Omega)}=\sum_{k\geq 1}(e^{-\lambda_k t}-1)^2\langle g,w_k\rangle^2$$

qui tend vers 0 par convergence dominée (ou en coupant la série en deux). ♢

Ceci termine la preuve du Théorème 6.45. ♢

Remarque 6.52. *Si $f\in L^2(]0;T[,L^2(\Omega))$ pour tout $T>0$, alors on obtient une unique solution définie pour tout $t>0$, mais les estimées sur cette solution dépendent du temps final considéré.*

Voici maintenant un résultat fournissant la régularité par rapport aux conditions initiales :

Théorème 6.53 (Régularité par rapport aux conditions initiales)**.** *Il existe une constante C (dépendant de T et Ω) telle que pour tous $g\in L^2(\Omega)$ et $f\in L^2(]0;T[,L^2(\Omega))$, l'unique solution u de* (6.12) *vérifie :*

$$\begin{aligned}\max_{0\leq t\leq T}\|u(t)\|_{L^2(\Omega)}+\|u\|_{L^2(]0;T[,H_0^1(\Omega))}+\|u'\|_{L^2(]0;T[,H^{-1}(\Omega))}&\\ \leq C\left(\|f\|_{L^2(]0;T[,L^2(\Omega))}+\|g\|_{L^2(\Omega)}\right).&\quad(6.22)\end{aligned}$$

Preuve : On peut prendre $v = u$ dans la formulation faible $(C1)$. On obtient, pour presque tout $t > 0$,

$$ {}_{H^{-1}(\Omega)}\langle u', u\rangle_{H_0^1(\Omega)} + \int_\Omega |\nabla u(t)|^2 = \langle f(t), u(t)\rangle_{L^2(\Omega)}. \tag{6.23}$$

On a

$$\langle f(t), u(t)\rangle_{L^2(\Omega)} \leq \frac{1}{2}\left(\|f(t)\|^2_{L^2(\Omega)} + \|u(t)\|^2_{L^2(\Omega)}\right)$$

donc,

$$ {}_{H^{-1}(\Omega)}\langle u', u\rangle_{H_0^1(\Omega)} + \int_\Omega |\nabla u(t)|^2 \leq \frac{1}{2}\left(\|f(t)\|^2_{L^2(\Omega)} + \|u(t)\|^2_{L^2(\Omega)}\right). \tag{6.24}$$

En posant $\eta(t) = \|u(t)\|^2_{L^2(\Omega)}$ et en utilisant le Théorème 6.32, on obtient

$$\eta'(t) \leq \eta(t) + \|f(t)\|^2_{L^2(\Omega)}.$$

D'après le lemme de Gronwall, on déduit

$$\|u(t)\|^2_{L^2(\Omega)} \leq e^t\left(\|g\|^2_{L^2(\Omega)} + \int_0^t \|f(s)\|^2_{L^2(\Omega)}\, ds\right)$$

donc

$$\max_{0\leq t\leq T}\|u(t)\|_{L^2(\Omega)} \leq e^T\left(\|g\|^2_{L^2(\Omega)} + \|f\|^2_{L^2(]0;T[,L^2(\Omega))}\right).$$

D'après l'inégalité (6.24), on a aussi, en intégrant sur $[0;T]$,

$$\|u\|^2_{L^2(]0;T[,H_0^1(\Omega))} \leq \frac{1}{2}\|g\|^2_{L^2(\Omega)} + \frac{1}{2}\|f\|^2_{L^2(]0;T[,L^2(\Omega))} + \frac{T}{2}\max_{0\leq t\leq T}\|u(t)\|^2_{L^2(\Omega)}.$$

On déduit alors l'estimée sur $\|u\|_{L^2(]0;T[,H_0^1(\Omega))}$ en utilisant ce que nous avons déjà démontré pour majorer le dernier terme ci-dessus.

On estime maintenant $\|u'\|_{L^2(]0;T[,H^{-1}(\Omega))}$ par dualité. Soit une fonction fixée $v \in H_0^1(\Omega)$, telle que $\|v\|_{H_0^1(\Omega)} \leq 1$. On a alors par définition des solutions faibles

$$ {}_{H^{-1}(\Omega)}\langle u', v\rangle_{H_0^1(\Omega)} = \langle f(t), v\rangle_{L^2} - \int_\Omega \nabla u(t)\cdot\nabla v. \tag{6.25}$$

Ceci démontre que

$$\|u'(t)\|_{H^{-1}(\Omega)} \leq \|f(t)\|_{L^2(\Omega)} + \|\nabla u(t)\|_{L^2(\Omega)}.$$

On obtient alors l'estimée voulue en passant au carré, en intégrant sur $[0;T]$ et en utilisant les résultats précédents. ♢

Exercice 6.54. (Un théorème général)

1. *En s'inspirant de l'une des deux démonstrations du Théorème 6.45, démontrer le résultat général suivant :*

   Théorème 6.55. *Soient H et V deux espaces de Hilbert tels que $V \hookrightarrow H$ avec injection compacte et V est dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique continue et coercive dans V. Soit un temps final $T > 0$, une condition initiale $g \in H$ et un terme source $f \in L^2(]0;T[,H)$. Il existe une unique solution faible $u \in L^2(]0;T[,V)$ telle que $u' \in L^2(]0;T[,V')$ au problème*

$$\begin{cases} \frac{d}{dt}\langle u(t),v\rangle_H + a(u(t),v) = \langle f(t),v\rangle_H \quad \forall v \in V,\ t \in]0;T[\\ u(0) = g. \end{cases}$$

   De plus il existe une constante C telle que

$$\|u\|_{L^2(]0;T[,V)} + \|u\|_{C^0([0;T],H)} \leq C(\|f\|_{L^2(]0;T[,H)} + \|g\|_H).$$

2. (Équation de la chaleur en milieu inhomogène). *Soit $\Omega \subset \mathbb{R}^n$ un ouvert borné régulier et A une fonction définie sur Ω à valeurs dans les matrices symétriques réelles définies positives de taille n, telle que*

$$\alpha I_n \leq A(x) \leq \beta I_n$$

   p.p. $x \in \Omega$, où $\alpha, \beta > 0$ et I_n est l'identité de $\mathbb{R}^n$. En déduire l'existence d'une unique solution faible au problème

$$\begin{cases} \frac{\partial}{\partial t}u(t,x) - \operatorname{div}(A(x)\nabla u(t,x)) = f, & (t,x) \in]0;T[\times\Omega, \\ u(t,x) = 0, & (t,x) \in]0;T[\times\partial\Omega \\ u(0,x) = g(x), \end{cases}$$

   où $g \in L^2(\Omega)$ et $f \in L^2(]0;T[,L^2(\Omega))$.

Exercice 6.56. *Soit u l'unique solution faible de (6.12). On suppose $f \in L^2((0,T)\times\Omega)$ et $g \in H_0^1(\Omega)$. Montrer alors que solution $u \in L^\infty((0,T),H_0^1(\Omega)) \cap H^1((0,T),L^2(\Omega))$ et satisfait l'estimée d'énergie : $\forall t \in [0,T]$,*

$$\int_\Omega |\nabla u|^2(t) + \int_0^t \int_\Omega |\partial_t u|^2 \leq \int_\Omega |\nabla g|^2 + \int_0^T \|f\|^2_{L^2(\Omega)}.$$

En déduire que $u \in L^2((0,T),H^2(\Omega))$.

6.4.2 Propriétés qualitatives des solutions faibles

Théorème 6.57 (Comportement asymptotique)**.** *Soit Ω un ouvert borné régulier, $g \in L^2(\Omega)$ et $u \in C^0([0;T],L^2(\Omega))$ l'unique solution faible obtenue avec le Théorème 6.45, avec $f \equiv 0$. Alors on a :*

$$\lim_{t\to+\infty} \|u(t)\|_{L^2(\Omega)} = 0.$$

Preuve : D'après l'inégalité de Poincaré (cf. la Proposition 1.24 et l'Exercice 3.7), on sait que la première valeur propre du Laplacien sur Ω avec conditions de Dirichlet est strictement positive. On a donc $-\Delta \geq \epsilon > 0$ au sens des formes quadratiques, ce qui signifie aussi que $\lambda_k \geq \epsilon > 0$ où les λ_k sont les valeurs propres introduites dans la preuve du Théorème 6.45. Ceci démontre en particulier que $0 \leq U(t) \leq e^{-\epsilon t}$ et donc que

$$\|U(t)\|_{L^2(\Omega)\to L^2(\Omega)} \leq e^{-\epsilon t}.$$

Or si $f \equiv 0$, on a $u(t) = U(t)g$ d'après (6.21), donc

$$\|u(t)\|_{L^2(\Omega)} \leq e^{-\epsilon t} \|g\|_{L^2(\Omega)} \to_{t\to+\infty} 0.$$

$\diamondsuit$

Exercice 6.58. *Montrer que si $f \in L^2(\Omega)$ ne dépend pas du temps, l'unique solution faible obtenue par le Théorème 6.45 vérifie*

$$\lim_{t\to+\infty} \|u(t) - v\|_{L^2(\Omega)} = 0$$

où v est l'unique solution de l'équation de Laplace

$$-\Delta v = f$$

dans $H^1_0(\Omega)$.

Examinons maintenant la régularité de la solution lorsque les données initiales sont plus ou moins régulières.

Théorème 6.59 (Effet régularisant avec $f \equiv 0$)**.** *On suppose que Ω est un ouvert borné de $\mathbb{R}^n$, de classe C^∞. Soit $g \in L^2(\Omega)$ une condition initiale et u l'unique solution faible obtenue par le Théorème 6.45. Alors pour tout $0 < \epsilon < T$, on a*

$$u \in C^\infty([\epsilon; T] \times \overline{\Omega}).$$

Preuve : La preuve est plus difficile que dans le cas de l'espace tout entier et nous ne donnons que les idées générales. Fixons $0 < \epsilon < T$. Nous voulons montrer que $(t, x) \mapsto (U(t)g)(x)$ est une fonction régulière par rapport au couple (t, x) lorsque $g \in L^2(\Omega)$. L'idée est de prouver que pour tous $\ell \geq 0$ et $m \geq 0$, il existe une constante $C_{\ell,m}$ telle que

$$\left\|\partial_t^\ell (-\Delta)^m U(t)g\right\|_{L^2(]\epsilon;T[,L^2(\Omega))} \leq C_{\ell,m} \|g\|_{L^2(\Omega)}. \tag{6.26}$$

Ceci signifie par régularité elliptique que

$$u = U(t)g \in H^r(]\epsilon; T[\times\Omega)$$

pour tout $r \geq 0$. D'après les injections de Sobolev, on obtient bien que $u \in C^\infty([\epsilon; T] \times \overline{\Omega})$.

Pour démontrer (6.26), on peut par densité prendre $g \in \text{Vect}(w_1, ..., w_m)$ et se rendre compte suivant un argument précédent que

$$\left\|\partial_t^\ell (-\Delta)^m U(t) g\right\|_{L^2(\Omega)} \leq \sup_{k \geq 1} \left(\lambda_k^{\ell+m} e^{-\lambda_k t}\right) \|g\|_{L^2(\Omega)}.$$

On obtient bien (6.26) avec

$$C_{\ell,m} := (T-\epsilon)^{1/2} \sup_{x>0} x^{\ell+m} e^{-x\epsilon}.$$

♢

Obtenir la régularité jusqu'à $t = 0$ ou avec un terme source $f \neq 0$ est plus difficile. Schématiquement, on peut démontrer

g	f	u
$H^{2m+1}(\Omega)$	$\frac{d^k}{dt^k} f \in L^2(]0;T[, H^{2m-2k}(\Omega))$ $k = 0, ..., m$	$\frac{d^k}{dt^k} u \in L^2(]0;T[, H^{2m+2-2k}(\Omega))$ $k = 0, ..., m+1$

mais il faut ajouter des *conditions de compatibilité*. Par exemple la dérivée u' vérifie aussi l'équation de la chaleur mais avec condition initiale $u'(0) = \Delta g + f(0, \cdot)$. Pour pouvoir utiliser les résultats précédents, il faut donc que cette fonction soit au moins dans $L^2(\Omega)$, ce qui impose des conditions sur f et g. Voir par exemple [7] pour plus de détails.

Nous nous contenterons du résultat partiel suivant :

Théorème 6.60 (Régularité)**.** *On suppose que Ω est un ouvert de bord C^∞ et que $g \in C_0^\infty(\Omega)$, $f \in C^\infty([0;T], C_0^\infty(\Omega))$. Alors*

$$u \in C^\infty([0;T] \times \overline{\Omega}).$$

Preuve : Comme précédemment, on démontre que $\partial_t^\ell (-\Delta)^m u \in L^2(]0;T[\times\Omega)$ pour tous $\ell, m \geq 0$. On utilise la propriété fondamentale vue plus haut

$$U'(t) = (-\Delta)U(t) = U(t)(-\Delta).$$

Ainsi

$$(-\Delta)^m u = U(t)(-\Delta)^m g + \int_0^t U(t-s)(-\Delta)^m f(s) ds.$$

Or $(-\Delta)^m g \in L^2(\Omega)$ et $(-\Delta)^m f \in L^2(]0;T[, L^2(\Omega))$ par hypothèse. Donc toute l'étude précédente implique que

$$(-\Delta)^m u \in L^2(]0;T[, H_0^1(\Omega)) \cap C^0([0;T], L^2(\Omega)), \quad \partial_t (-\Delta)^m u \in L^2(]0;T[, H^{-1}(\Omega))$$

pour tout $m \geq 1$. Rappelons que $\partial_t u = \Delta u + f$ donc

$$\partial_t(-\Delta)^{m-1}u = -(-\Delta)^m u + (-\Delta)^{m-1}f$$

au moins au sens des distributions. Or $(-\Delta)^m u \in C^0([0;T];L^2(\Omega))$ et bien sûr $(-\Delta)^{m-1}f \in C^0([0;T];L^2(\Omega))$ donc finalement

$$\partial_t(-\Delta)^{m-1}u \in C^0([0;T];L^2(\Omega))$$

pour tout $m \geq 1$.

Ensuite on a

$$\partial_t^2(-\Delta)^{m-1}u = -\partial_t(-\Delta)^m u + \partial_t(-\Delta)^{m-1}f$$

au moins au sens des distributions, donc

$$\partial_t^2(-\Delta)^{m-1}u \in C^0([0;T],L^2(\Omega)).$$

La démonstration suit en itérant l'argument précédent. ♢

Théorème 6.61 (Principe du maximum faible)**.** *Soient Ω un ouvert borné de $\mathbb{R}^n$, $T > 0$, $g \in L^2(\Omega)$, $f \in L^2(]0;T[,L^2(\Omega))$, et u l'unique solution faible obtenue grâce au Théorème 6.45. Si $f \geq 0$ presque partout dans $[0;T]\times\Omega$ et $g \geq 0$ presque partout dans Ω, alors $u \geq 0$ presque partout dans $[0;T]\times\Omega$.*

Preuve : Nous commençons par démontrer ce résultat en supposant que f et g sont respectivement dans $C^\infty([0;T],C_0^\infty(\Omega))$ et $C_0^\infty(\Omega)$ et que $f(t) > 0$ sur Ω. D'après le théorème précédent, u est très régulière (en fait on a seulement besoin que u soit de classe C^2 par rapport à (t,x)).

Soit $(t_0,x_0) \in [0;T]\times\overline{\Omega}$ un point où u atteint son minimum. Si $t_0 = 0$ ou $x_0 \in \partial\Omega$, on a clairement $u(t_0,x_0) = 0$ par positivité de g et la condition de Dirichlet au bord. On peut donc supposer que $x_0 \in \Omega$ et $t_0 \in]0;T]$. Supposons pour commencer que $t_0 < T$. Alors comme le minimum de u est atteint dans l'ouvert $]0;T[\times\Omega$, on a

$$\partial_t u(t_0,x_0) = 0 \quad \text{ et } \nabla u(t_0,x_0) = 0.$$

Comme il s'agit d'un minimum, la Hessienne de u est nécessairement positive en (t_0,x_0), donc on obtient

$$-\Delta u(t_0,x_0) = -\text{tr }(\text{Hess}(u)(t_0,x_0)) \leq 0.$$

Or d'après l'équation,

$$-\Delta u(t_0,x_0) = f(t_0,x_0) > 0$$

donc c'est absurde.

Si maintenant le minimum est atteint en (T,x_0), on a seulement

$$\frac{\partial}{\partial t}u(T,x_0) \leq 0$$

mais on a toujours

$$-\Delta u(t_0, x_0) = -\text{tr } (\text{Hess}(u)(t_0, x_0)) \leq 0$$

car $x \mapsto u(T, x)$ admet un minimum local en x_0 dans l'ouvert Ω. L'équation donne alors

$$0 < f(T, x_0) = \frac{\partial}{\partial t} u(T, x_0) - \Delta u(T, x_0) \leq 0$$

qui est aussi absurde. Nous avons prouvé que soit $t_0 = 0$, soit $x_0 \in \partial\Omega$. On a donc bien $\min_{[0;T]\times\overline{\Omega}} u(t, x) = 0$.

Nous venons donc de démontrer que si f et g sont des fonctions régulières strictement positives sur Ω, alors $u \geq 0$. Le cas général s'obtient par densité des fonctions régulières positives dans $L^2(\Omega)$ et $L^2(]0; T[, L^2(\Omega))$, et en utilisant la continuité de u par rapport aux données f et g, prouvée au Théorème 6.53. ♢

Remarque 6.62. *Dans le cas où $g \in H^{1/2}(\Omega)$, il existe une autre preuve de ce résultat qu'il est utile de connaître, et dont on donne ici les grandes idées sans rentrer dans les détails. On prend $v = u^-$ dans la formulation variationnelle $(C1)$ et on obtient donc (on admet que l'on peut prendre $v = u^-$ comme fonction test, même si u^-, qui est bien une fonction de $H_0^1(\Omega)$ pour presque tout temps, dépend du temps)*

$$\int_\Omega \frac{\partial u}{\partial t} u^- + \int_\Omega \nabla u \cdot \nabla u^- = \int_\Omega f u^-.$$

On en déduit :

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |u^-|^2 + \int_\Omega |\nabla u^-|^2 \leq 0,$$

et donc, comme $\int_\Omega |u_0^-|^2 = 0$, pour tout $t \geq 0$,

$$\frac{1}{2}\int_\Omega |u^-|^2 + \int_0^t \int_\Omega |\nabla u^-|^2 \leq 0.$$

Ceci permet de conclure que $u^- = 0$ et donc $u \geq 0$ presque partout.

Pour être tout à fait rigoureux et justifier l'égalité $\int_\Omega \frac{\partial u}{\partial t} u^- = \frac{1}{2}\frac{d}{dt}\int_\Omega |u^-|^2$, on peut utiliser la méthode des troncatures de Stampacchia. On renvoie à [3, Théorème X.3]. Ainsi, si u et v désignent deux solutions de l'équation de la chaleur, avec même second membre f et mêmes conditions aux limites g, et si les conditions initiales satisfont $u_0 \leq v_0$ alors $u \leq v$.

Remarque 6.63. *Le fait que u reste ≥ 0 lorsque les données sont ≥ 0 est important physiquement, par exemple si u représente une température.*

Voici maintenant un résultat plus précis quand $f \equiv 0$ et qui traduit l'existence d'une *propagation à vitesse infinie* : même si la condition initiale s'annule à l'intérieur de Ω, la solution u vérifie $u(t, x) > 0$ pour tout $t > 0$ et $x \in \Omega$.

Théorème 6.64 (Propagation à vitesse infinie)**.** *Soit Ω un ouvert borné régulier de $\mathbb{R}^n$, un temps final $T > 0$ et une fonction $g \in L^2(\Omega)$ telle que $g \neq 0$ et $g \geq 0$ presque partout. Alors la solution u obtenue par le Théorème 6.45 avec $f \equiv 0$ vérifie*

$$u(t,x) > 0 \qquad \forall x \in \Omega$$

pour tout temps $t > 0$.

La démonstration, complexe, repose sur une inégalité de type Harnack parabolique, ou une formule de la moyenne parabolique. Voir [7] pour plus de détails.

Chapitre 7

Autres problèmes d'évolution

Ce chapitre est une brève introduction à l'étude mathématique d'autres types d'équations aux dérivées partielles dépendant du temps, à savoir l'équation de transport et l'équation des ondes. Nous renvoyons à [7, 1, 5, 13] pour une présentation plus détaillée.

7.1 L'équation de transport

Nous commençons par une étude rapide de l'équation de transport dans tout l'espace $\mathbb{R}^n$

$$\frac{\partial}{\partial t}u + b \cdot \nabla_x u = 0 \tag{7.1}$$

où b est un vecteur fixe de $\mathbb{R}^n$ (indépendant de x et t). Supposons tout d'abord que u est une fonction régulière. On remarque alors que (7.1) signifie qu'une certaine dérivée de u s'annule. Soit $(t,x) \in \mathbb{R} \times \mathbb{R}^n$ fixé. Introduisons la fonction auxilliaire $z(s) = u(t+s, x+sb)$. Alors (7.1) signifie que $z'(s) = 0$, donc que $s \mapsto u(t+s, x+sb)$ est une fonction constante sur tout $\mathbb{R}$. Ainsi, pour chaque point $(t,x) \in \mathbb{R} \times \mathbb{R}^n$, u est constante sur la droite de direction $(1,b) \in \mathbb{R}^{n+1}$ passant par (t,x). La fonction régulière u est donc connue partout pourvu que l'on connaisse u sur au moins un point de chacune de ces droites (c'est la *méthode des caractéristiques*).

Considérons alors le problème avec condition initiale (régulière) $g \in C^1(\mathbb{R}^n)$:

$$\begin{cases} \dfrac{\partial}{\partial t}u(t,x) + b \cdot \nabla_x u(t,x) = 0, & (t,x) \in (0,\infty) \times \mathbb{R}^n, \\ u(0,x) = g(x), & x \in \mathbb{R}^n. \end{cases} \tag{7.2}$$

Les arguments précédents montrent que la fonction u définie sur $[0,\infty) \times \mathbb{R}^n$ par

$$u(t,x) := g(x - bt) \tag{7.3}$$

est l'unique solution de (7.2) dans $C^1([0,\infty) \times \mathbb{R}^n)$. Notons que la formule (7.3) décrit une *onde progressive* avançant dans la direction b à la vitesse $\|b\|_{\mathbb{R}^n}$.

Maintenant si g n'est pas une fonction C^1, on ne peut bien sûr chercher une solution C^1 à (7.2). Pourtant la définition (7.3) a toujours un sens avec des hypothèses très faibles sur g et on peut décider arbitrairement que ceci définit une *solution faible* de (7.2). Par exemple si $g \in L^p(\mathbb{R}^n)$, on aura $u \in C^0([0,\infty), L^p(\mathbb{R}^n))$.

Remarque 7.1. *Introduisons l'opérateur de translation $\tau_b(t)$ défini par*

$$(\tau_b(t)f)(x) = f(x - bt).$$

Pour tout $t \geq 0$ fixé, $\tau_b(t)$ est un opérateur borné de $W^{m,p}(\mathbb{R}^n)$ dans lui même pour tous $m \geq 0$, $p \geq 1$. Notons que la famille $(\tau_b(t))_{t\geq 0}$ vérifie les deux propriétés importantes

$$\tau_b(0) = Id$$

$$\tau_b(t+s) = \tau_b(t)\tau_b(s).$$

On parle de semi-groupe. *Si $g \in H^k(\mathbb{R}^n)$, la "solution faible" (7.3) s'écrit alors $u(t) = \tau_b(t)g \in \mathcal{C}^0([0,\infty), H^k(\mathbb{R}^n))$. Toutes les équations d'évolution linéaires d'ordre un en temps et sans second membre vont s'écrire sous cette forme pour un semi-groupe bien choisi.*

Reste à savoir en quel sens une telle solution résout (7.2). Voici un résultat facile dont le but principal est d'habituer le lecteur à la manipulation des espaces introduits à la section précédente.

Proposition 7.2. *Si $g \in H^1(\mathbb{R}^n)$, l'expression (7.3) fournit une fonction u qui satisfait*

$$u \in C^0([0,\infty), H^1(\mathbb{R}^n)) \cap C^1([0,\infty), L^2(\mathbb{R}^n)), \quad \nabla u \in C^0([0,\infty), L^2(\mathbb{R}^n)), \quad (7.4)$$

et l'égalité

$$u' + b \cdot \nabla u = 0 \quad (7.5)$$

a lieu dans $C^0([0,\infty), L^2(\mathbb{R}^n))$.

D'autre part, on a

$$\lim_{t\to 0} \|u(t,\cdot) - g\|_{H^1(\mathbb{R}^n)} = 0 \quad (7.6)$$

qui donne un sens à la condition initiale $u(0,x) = g(x)$.

Enfin, la solution (7.3) est l'unique fonction satisfaisant (7.4), (7.5) et (7.6).

Preuve : Soit $g \in H^1(\mathbb{R}^n)$ et $u = \{t \mapsto \tau_b(t)g\}$. Il est clair que $u \in C^0([0;\infty), H^1(\mathbb{R}^n))$. Définissons alors $v := \{t \mapsto -b \cdot \tau_b(t)\nabla g\}$ qui est une fonction de $C^0([0;\infty), L^2(\mathbb{R}^n))$ car $\nabla g \in L^2(\mathbb{R}^n)$. Notons que $v = -b \cdot \nabla u$. Il suffit de montrer que $u' = v$. On utilise la définition (**??**) : soit $w \in C_c^\infty(\mathbb{R}^n)$ et $\varphi \in C_c^\infty([0;\infty))$ deux fonctions test fixées. On a

$$\left\langle \int_0^\infty v(t)\varphi(t)dt, w \right\rangle_{L^2} = \int_0^\infty \langle v(t), w\rangle_{L^2} \varphi(t)dt.$$

Comme w est régulière, on a

$$\langle v(t), w\rangle_{L^2} = \frac{d}{dt}\langle g, \tau_b(-t)v\rangle_{L^2} = \frac{d}{dt}\langle \tau_b(t)g, v\rangle_{L^2}$$

(pour le voir, faire une intégration par partie). On obtient donc

$$\left\langle \int_0^\infty v(t)\varphi(t)dt, w \right\rangle_{L^2} = -\left\langle \int_0^\infty u(t)\varphi'(t)dt, w \right\rangle_{L^2}$$

pour tout $w \in C_c^\infty(\mathbb{R}^n)$, d'où l'égalité

$$\int_0^\infty v(t)\varphi(t)dt = -\int_0^\infty u(t)\varphi'(t)dt$$

dans $L^2(\mathbb{R}^n)$, par densité.

Pour l'unicité, il suffit de montrer que si u satisfait (7.4), (7.5) et (7.6) avec $g = 0$, alors nécessairement $u = 0$. On utilise une technique d'énergie : on prend le produit scalaire L^2 de (7.5) contre $u(t) \in L^2(\mathbb{R}^n)$ à t fixé (tout a un sens d'après (7.4)). On obtient pour presque tout t

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\mathbb{R}^n)} = \langle u'(t), u(t)\rangle_{L^2(\mathbb{R}^n)} = 0$$

car pour toute fonction $\varphi \in H^1(\mathbb{R}^n)$, $\int \varphi\nabla\varphi = 0$ et où nous avons utilisé le résultat du Théorème 6.32. Ainsi $\|u(t)\|_{L^2}$ est constant et il s'annule par (7.6) avec $g = 0$, c'est-à-dire $u \equiv 0$. ♢

Problème non homogène

Regardons rapidement l'équation de transport non homogène

$$\begin{cases} \dfrac{\partial}{\partial t}u(t,x) + b\cdot\nabla_x u(t,x) = f(t,x), & (t,x) \in (0,\infty)\times\mathbb{R}^n, \\ u(0,x) = g(x), & x \in \mathbb{R}^n. \end{cases} \tag{7.7}$$

On suppose comme précédemment que f et g sont de classe C^1. En posant comme avant $z(s) = u(t+s, x+sb)$, on trouve que

$$z'(s) = f(t+s, x+sb)$$

et donc que

$$u(t,x) - g(x-tb) = \int_{-t}^0 z'(s)ds = \int_0^t f(s, x+(s-t)b)ds,$$

c'est-à-dire

$$u(t,x) = g(x-tb) + \int_0^t f(s, x+(s-t)b)ds$$

résout (7.7) dans $\mathcal{C}^1([0,\infty)\times\mathbb{R}^n)$. Nous utiliserons plus tard cette formule pour l'étude de l'équation des ondes.

7.2 L'équation des ondes

7.2.1 L'équation des ondes 1D

Nous commençons par le cas simple de l'équation des ondes posée sur tout $\mathbb{R}$:

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}u(t,x) - c^2\dfrac{\partial^2}{\partial x^2}u(t,x) = 0, \quad (t,x) \in]0;\infty[\times\mathbb{R} \\ u(0,x) = u_0(x), \\ \frac{\partial}{\partial t}u(0,x) = u_1(x), \end{cases} \tag{7.8}$$

Nous supposons pour commencer que u_0 et u_1 sont des fonctions suffisamment régulières. On a :

$$\frac{\partial^2}{\partial t^2} - c^2\frac{\partial^2}{\partial x^2} = \left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x}\right)$$

donc une solution de l'équation aux dérivées partielles s'écrit (comparer avec l'équation de transport)

$$u(t,x) = f(x-ct) + g(x+ct).$$

La fonction $(t,x) \mapsto f(x-ct)$ représente une onde progressive avançant à la vitesse c vers la droite, alors que $(t,x) \mapsto g(x+ct)$ est une onde progressive avançant à la vitesse c vers la gauche. On calcule

$$u(0,x) = u_0(x) = f(x) + g(x)$$

et

$$\frac{\partial}{\partial t}u(0,x) = u_1(x) = -cf'(x) + cg'(x).$$

On trouve donc la formule de d'Alembert :

$$u(t,x) = \frac{1}{2}(u_0(x-ct) + u_0(x+ct)) + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)\,ds. \tag{7.9}$$

Cette formule a un sens dès que u_0 et u_1 sont dans L^1_{loc}, u étant alors solution de l'équation des ondes au sens des distributions. Pour donner un sens précis aux conditions aux limites, on peut démontrer un équivalent de la Proposition 7.2 :

Proposition 7.3. *Si $u_0 \in H^1(\mathbb{R})$ et $u_1 \in L^2(\mathbb{R})$, alors la formule* (7.9) *fournit une solution*

$$u \in C^0(\mathbb{R}, H^1(\mathbb{R})) \cap C^1(\mathbb{R}, L^2(\mathbb{R})) \cap C^2(\mathbb{R}, H^{-1}(\mathbb{R})) \tag{7.10}$$

de l'équation des ondes

$$\frac{\partial^2}{\partial t^2}u - c^2\frac{\partial^2}{\partial x^2}u = 0 \tag{7.11}$$

où cette égalité a lieu dans $C^0(\mathbb{R}, H^{-1}(\mathbb{R}))$ et telle que

$$\lim_{t\to 0}\|u(t) - u_0\|_{H^1(\mathbb{R})} = 0, \qquad \lim_{t\to 0}\|u'(t) - u_0\|_{L^2(\mathbb{R})} = 0 \tag{7.12}$$

C'est l'unique solution vérifiant (7.10), (7.11) *et* (7.12).

Preuve : La faire en exercice ! ♢

Sur le cas de l'équation 1D, on voit déjà de grandes différences de comportement par rapport à l'équation de la chaleur. Par exemple, si

$$\mathrm{Supp}(u_0) \cup \mathrm{Supp}(u_1) \subseteq [a; b],$$

alors on aura pour $t > 0$

$$\mathrm{Supp}(u(t, \cdot)) \subseteq [a - ct; b + ct].$$

Ainsi, la propagation a lieu à la vitesse c, il n'y a pas de propagation à vitesse infinie comme pour l'équation de la chaleur.

De même, on voit qu'il n'y a aucun gain ou aucune perte de régularité de la solution comme c'est le cas pour l'équation de la chaleur (effet régularisant) ou l'équation de Burgers (apparition de singularités). Par exemple si $u_1 \equiv 0$ et $u_0 \in H^1(\mathbb{R})$, alors $u(t, \cdot) \in H^1(\mathbb{R})$ pour tout $t > 0$, tout comme pour l'équation de transport.

Notons que si on prend $u_0 \equiv 0$ et $u_1 = \epsilon^{-1}\varphi(x/\epsilon)$, alors on trouve formellement que

$$u(t, x) \to_{\epsilon \to 0} G(t, x)$$

où

$$G(t, x) = \frac{1}{2c} 1_{[-ct;ct]}(x)$$

est donc la solution (formelle) sur $\mathbb{R}^2$, du problème

$$\begin{cases} \dfrac{\partial^2}{\partial t^2} G - c^2 \dfrac{\partial^2}{\partial x^2} G = 0, \\ G(0, \cdot) = 0, \\ G'(0, \cdot) = \delta. \end{cases} \tag{7.13}$$

On peut alors remarquer que la solution générale (7.9) s'écrit

$$\begin{aligned} u(t, x) &= \frac{1}{2}(u_0(x - ct) + u_0(x + ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} u_1(s)\, ds \\ &= \frac{d}{dt} \left(\frac{1}{2c} \int_{x-ct}^{x+ct} u_0(s)\, ds \right) + \frac{1}{2c} \int_{x-ct}^{x+ct} u_1(s)\, ds \\ &= \frac{d}{dt} \left(\int_{\mathbb{R}} G(t, x - y) u_0(y)\, dy \right) + \int_{\mathbb{R}} G(t, x - y) u_1(y)\, dy. \end{aligned}$$

La solution de l'équation des ondes dans tout $\mathbb{R}^n$ avec $n > 1$ s'écrit de façon similaire avec une fonction G adaptée voir l'exercice 7.4.

Exercice 7.4. (L'équation des ondes dans $\mathbb{R}^n$). *On suppose que g et h sont régulières, et que $f \equiv 0$. Montrer que la solution de l'équation des ondes sans second membre sur tout $\mathbb{R}^n$ avec conditions initiales $u(0) = u_0$ et $u'(0) = u_1$ s'écrit*

$$u(t) = \frac{d}{dt}G(t)u_0 + G(t)u_1$$

où

$$(G(t)\psi)(x) = \frac{1}{2\pi}\int_{\mathbb{R}^2} 1_{B(0;t)}(y)\frac{\psi(x-y)}{\sqrt{t^2-|y|^2}}\,dy \qquad \text{si } n = 2, \tag{7.14}$$

$$(G(t)\psi)(x) = \frac{1}{4\pi t}\int_{S(0;t)} \psi(x-y)\,d\sigma(y) \qquad \text{si } n = 3, \tag{7.15}$$

où $B(0;t)$ est la boule ouverte de centre 0 et de rayon t et $S(0;t)$ est la sphère de centre 0 et de rayon t, $d\sigma(y)$ est la mesure surfacique de cette sphère.

Vérifier que la solution se propage à vitesse finie : si $\mathrm{Supp}(u_0) \cup \mathrm{Supp}(u_1) \subset B(0,r)$, *alors*

$$\mathrm{Supp}(u(t,\cdot)) \subset B(0, r+t).$$

Si n est pair quelconque, $G(t)$ s'obtient par une formule similaire à (7.14) *tandis que si n est impair quelconque, $G(t)$ s'obtient par une formule similaire à* (7.15), *voir [7].*

7.2.2 L'équation des ondes dans un ouvert borné Ω

Tout comme pour l'équation de la chaleur, nous étudions maintenant l'équation des ondes dans un ouvert borné $\Omega \subset \mathbb{R}^n$, avec des conditions de Dirichlet au bord et terme source f :

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}u(t,x) - \Delta u(t,x) = f(t,x), \quad (t,x) \in]0;T[\times\Omega, \\ u(t,x) = 0 \text{ si } x \in \partial\Omega, \\ u(0,x) = g(x), \\ \frac{\partial}{\partial t}u(0,x) = h(x), \end{cases} \tag{7.16}$$

Cette fois, nous avons pris $c = 1$ pour simplifier.

7.2.2.1 Solutions faibles

Comme pour l'équation de la chaleur, nous commençons par introduire une notion de solution faible. On considère $f \in L^2(]0;T[, L^2(\Omega))$, $g \in H_0^1(\Omega)$ et $h \in L^2(\Omega)$.

Définition 7.5 (Solutions faibles)**.** *Soit $u \in L^2(]0;T[, H_0^1(\Omega))$ telle que $u' \in L^2(]0;T[, L^2(\Omega))$ et $u'' \in L^2(]0;T[, H^{-1}(\Omega))$. On dit que u est une* solution faible *de* (7.16) *si on a*

(O1) ${}_{H^{-1}(\Omega)}\langle u''(t), v\rangle_{H_0^1(\Omega)} + \displaystyle\int_\Omega \nabla u(t)\cdot\nabla v = \int_\Omega f(t)v$ *pour tout $v \in H_0^1(\Omega)$ et presque tout en $t \in]0;T[$.*

$(O2)$ $u(0) = g$.

$(O3)$ $u'(0) = h$.

Rappelons que d'après le Théorème 6.32, on a $u \in C^0([0;T], L^2(\Omega))$ et $u' \in C^0([0;T], H^{-1}(\Omega))$ qui permet de donner un sens à $(O2)$ et $(O3)$.

Le but de cette section est principalement de démontrer le résultat suivant :

Théorème 7.6 (Existence et unicité de solutions faibles)**.** *Soit $\Omega \subset \mathbb{R}^n$ un ouvert borné régulier. On suppose que $f \in L^2(]0;T[, L^2(\Omega))$, $g \in H_0^1(\Omega)$ et $h \in L^2(\Omega)$. Alors le problème* (7.16) *admet une unique solution faible u.*

De plus, on a

$$u \in L^\infty(]0;T[, H_0^1(\Omega)), \qquad u' \in L^\infty(]0;T[, L^2(\Omega))$$

et

$$\sup_{0\le t\le T} \left(\|u(t)\|_{H_0^1(\Omega)} + \|u'(t)\|_{L^2(\Omega)} \right) + \|u''\|_{L^2(]0;T[,H^{-1}(\Omega))} \\ \le C \left(\|f\|_{L^2(]0;T[,L^2(\Omega))} + \|g\|_{H_0^1(\Omega)} + \|h\|_{L^2(\Omega)} \right) \quad (7.17)$$

pour une constante C indépendante de u.

Remarque 7.7. *Bien noter que nous avons pris $g \in H_0^1(\Omega)$ alors qu'*a priori *on pourrait donner un sens à une solution faible avec seulement $g \in L^2(\Omega)$, puisque $u \in C^0([0;T], L^2(\Omega))$. Toutefois, contrairement à l'équation de la chaleur, il n'y a pas d'effet régularisant avec l'équation des ondes. Pour avoir $u \in L^2(]0;T[, H_0^1(\Omega))$ on est donc obligé de supposer $g \in H_0^1(\Omega)$. De même, on doit prendre $h \in L^2(\Omega)$.*

Remarque 7.8. *On peut en fait montrer que $u \in C^0([0;T], H_0^1(\Omega))$ et $u' \in C^0([0;T], L^2(\Omega))$. Nous verrons cela plus tard, au Corollaire 7.16.*

Preuve : Pour l'équation de la chaleur nous avons donné la preuve par décomposition sur les modes propres du Laplacien et proposé la méthode de Galerkin en exercice. Nous faisons l'inverse cette fois.

Étape 1 : approximations de Galerkin.

Considérons la famille des fonctions propres du Laplacien $(w_k)_{k\ge1} \subset H_0^1(\Omega)$, introduite dans la preuve pour l'équation de la chaleur. Rappelons que

— $(w_k)_{k\ge1}$ est une base *orthogonale* de $H_0^1(\Omega)$;

— $(w_k)_{k\ge1}$ est une base *orthonormée* de $L^2(\Omega)$.

On pose alors

$$V_m := \mathrm{vect}(w_1, ..., w_m)$$

et on cherche une solution $u_m \in C^2([0;T], V_m)$ faible dans V_m, c'est-à-dire vérifiant

$(O1)_m$ $\langle u_m'', w_k\rangle + \displaystyle\int_\Omega \nabla u_m(t)\cdot\nabla w_k = \langle f(t), w_k\rangle_{L^2}$ pour tout $k = 1...m$ et presque partout en $t \in [0;T]$;

$(O2)_m$ $\langle u_m(0), w_k\rangle = \langle g, w_k\rangle$ pour tout $k = 1...m$.

$(O3)_m$ $\langle u_m'(0), w_k\rangle = \langle h, w_k\rangle$ pour tout $k = 1...m$.

On pourrait *a priori* prendre une base quelconque de $H_0^1(\Omega)$ (c'est en quelque sorte ce que nous ferons à la section **??**). Mais les expressions seront plus simples en choisissant la base des fonctions propres du Laplacien. Si on écrit

$$u_m(t,x) = \sum_{k=1}^m d_k^m(t) w_k(x),$$

alors $(O1)_m$ et $(O2)_m$ équivalent à (nous utilisons que $\langle u_m'', w\rangle = \frac{d^2}{dt^2}\langle u_m, w\rangle$ d'après le Théorème 6.32)

$(O1)'_m$ $\dfrac{d^2}{dt^2} d_k^m(t) + d_k^m(t)\,\|\nabla w_k\|_{L^2}^2 = \langle f(t), w_k\rangle_{L^2}$, $\forall k = 1...m$, p.p. $t \in [0;T]$;

$(O2)'_m$ $d_k^m(0) = \langle g, w_k\rangle$, $\forall k = 1...m$.

$(O3)'_m$ $\frac{d}{dt} d_k^m(0) = \langle h, w_k\rangle$, $\forall k = 1...m$.

Il s'agit d'un système d'équations différentielles ordinaires du second ordre, qui admet une unique solution $(d_k^m(t))_{k=1}^m$, définie sur tout $[0;T]$. Bien noter cependant que l'on a seulement $t \mapsto \langle f(t), w_k\rangle_{L^2} \in L^2(]0;T[)$, donc l'unique solution $d_k^m(t)$ est en fait elle-même seulement dans $H^2(]0;T[) \hookrightarrow C^1([0;T])$. On obtient donc que $u_m \in C^1([0;T], H^2(\Omega)\cap H_0^1(\Omega))$, $u_m' \in C^0([0;T], H^2(\Omega)\cap H_0^1(\Omega))$ et $u_m'' \in L^2(]0;T[, H^2(\Omega)\cap H_0^1(\Omega))$ puisque les w_k sont au moins dans $H^2(\Omega)$ (elles sont bien plus régulières si Ω est lui même très régulier).

Étape 2 : estimées d'énergie.

On désire maintenant passer à la limite quand $m \to \infty$. Pour cela, nous commençons par démontrer le lemme suivant.

Lemme 7.9 (Estimées d'énergie)**.** *Il existe une constante C qui ne dépend que de Ω et $T > 0$ telle que pour tout $m \geq 1$,*

$$\max_{0\leq t\leq T}\left(\|u_m(t)\|_{H_0^1(\Omega)} + \|u_m'(t)\|_{L^2(\Omega)}\right) + \|u_m''(t)\|_{L^2(]0;T[,H^{-1}(\Omega))} \leq C\left(\|f\|_{L^2(]0;T[,L^2(\Omega))} + \|g\|_{H_0^1(\Omega)} + \|h\|_{L^2(\Omega)}\right). \quad (7.18)$$

Preuve : Par linéarité dans $(O1)_m$, on obtient presque partout en $t \in]0;T[$

$$\langle u_m'', u_m'\rangle_{L^2(\Omega)} + \int_\Omega \nabla u_m \cdot \nabla u_m' = \langle f(t), u_m'\rangle_{L^2(\Omega)}.$$

Notons que comme u_m est une somme finie de w_k's, on a que $u_m'', u_m' \in L^2(]0;T[, H_0^1(\Omega))$. L'équation précédente équivaut à

$$\frac{1}{2}\frac{d}{dt}\left(\|u_m'\|^2_{L^2(\Omega)} + \int_\Omega |\nabla u_m|^2\right) = \langle f(t), u_m'\rangle_{L^2(\Omega)} \leq \frac{1}{2}\left(\|f(t)\|^2_{L^2(\Omega)} + \|u_m'\|^2_{L^2(\Omega)}\right). \tag{7.19}$$

Remarque 7.10. *Si $f \equiv 0$, on trouve que l'énergie est conservée au cours du temps :*

$$\forall t \in [0,T], \qquad \|u_m'(t)\|^2_{L^2(\Omega)} + \|\nabla u_m(t)\|^2_{L^2(\Omega)} = \|h_m\|^2_{L^2(\Omega)} + \|\nabla g_m\|^2_{L^2(\Omega)}$$

où $g_m = \sum_{k=1}^m \langle w_k, g\rangle w_k$ et $h_m = \sum_{k=1}^m \langle w_k, h\rangle w_k$. Ceci donne donc automatiquement une borne sur u_m dans $L^\infty(]0;T[, H_0^1(\Omega))$ et sur u_m' dans $L^\infty(]0;T[, L^2(\Omega))$, à condition bien sûr que les termes de droite soient eux-mêmes bornés, c'est-à-dire que $g \in H_0^1(\Omega)$ et $h \in L^2(\Omega)$.

Revenons maintenant à (7.19) et posons

$$\eta(t) = \|u_m'\|^2_{L^2(\Omega)} + \int_\Omega |\nabla u_m|^2$$

de sorte que

$$\eta'(t) \leq \|f(t)\|^2_{L^2(\Omega)} + \eta(t).$$

D'après le Lemme de Gronwall, on obtient donc

$$\begin{aligned}\eta(t) = \|u_m'(t)\|^2_{L^2(\Omega)} &+ \int_\Omega |\nabla u_m(t)|^2 \\ &\leq e^t\left(\|u_m'(0)\|^2_{L^2(\Omega)} + \int_\Omega |\nabla u_m(0)|^2 + \int_0^t \|f(s)\|^2_{L^2(\Omega)}\right). \end{aligned} \tag{7.20}$$

Or on a

$$\|u_m'(0)\|^2_{L^2(\Omega)} = \sum_{k=1}^m (d_k^m)'(0)^2 = \sum_{k=1}^m \langle h, w_k\rangle^2 \leq \|h\|^2_{L^2(\Omega)}$$

où nous avons utilisé que les w_k forment une base orthonormée de $L^2(\Omega)$. De même

$$\int_\Omega |\nabla u_m(0)|^2 = \sum_{k=1}^m (d_k^m(0))^2 \int_\Omega |\nabla w_k|^2 = \sum_{k=1}^m (d_k^m(0))^2 \lambda_k \leq \int_\Omega |\nabla g|^2.$$

Ainsi on obtient l'estimée

$$\begin{aligned}\max_{t\in[0;T]} &\left(\|u_m'(t)\|^2_{L^2(\Omega)} + \|u_m(t)\|^2_{H_0^1(\Omega)}\right) \\ &\leq e^T\left(\|g\|^2_{H_0^1(\Omega)} + \|h\|^2_{L^2(\Omega)} + \|f\|^2_{L^2(]0;T[,L^2(\Omega))}\right). \end{aligned} \tag{7.21}$$

Pour obtenir l'estimée sur u''_m, nous raisonnons par dualité et considérons une fonction fixe $v \in H^1_0(\Omega)$. Nous pouvons écrire $v = v_1 + v_2$ avec $v_1 \in V_m$ et $v_2 \in (V_m)^\perp$. Comme $u''_m(t) \in V_m \subseteq L^2(\Omega)$ pour p.p. $t \in]0;T[$, on a

$$ {}_{H^{-1}(\Omega)}\langle u''_m, v\rangle_{H^1_0(\Omega)} = \langle u''_m, v\rangle_{L^2(\Omega)} = \langle u''_m, v_1\rangle_{L^2(\Omega)}. $$

On peut alors utiliser $(O1)_m$ pour obtenir

$$ {}_{H^{-1}(\Omega)}\langle u''_m, v\rangle_{H^1_0(\Omega)} = \langle f(t), v_1\rangle - \int_\Omega \nabla u_m \cdot \nabla v_1 $$

donc

$$ |{}_{H^{-1}(\Omega)}\langle u''_m, v\rangle_{H^1_0(\Omega)}| \leq \left(C \left\| f(t) \right\|_{L^2(\Omega)} + \left\| u_m(t) \right\|_{H^1_0(\Omega)} \right) \left\| v_1 \right\|_{H^1_0(\Omega)}. $$

Ceci montre que

$$ \left\| u''_m(t) \right\|_{H^{-1}(\Omega)} \leq C \left\| f(t) \right\|_{L^2(\Omega)} + \left\| u_m(t) \right\|_{H^1_0(\Omega)}. $$

Pour en déduire l'estimée sur $\|u''_m\|_{L^2(]0;T[,H^{-1}(\Omega))}$, on utilise que $f \in L^2(]0;T[,L^2(\Omega))$ et (7.21). ♢

Étape 3 : existence d'une solution faible.

D'après le Lemme 7.9, nous savons que (u_m) est bornée dans $L^2(]0;T[,H^1_0(\Omega)) \cap L^\infty(]0;T[,H^1_0(\Omega))$, que (u'_m) est bornée dans $L^2(]0;T[,L^2(\Omega)) \cap L^\infty(]0;T[,L^2(\Omega))$ et que (u''_m) est bornée dans $L^2(]0;T[,H^{-1}(\Omega))$. Il existe donc une fonction $u \in L^2(]0;T[,H^1_0(\Omega))$ telle que $u' \in L^2(]0;T[,L^2(\Omega))$ et $u'' \in L^2(]0;T[,H^{-1}(\Omega))$ et une sous-suite de (u_m) (que nous noterons encore u_m pour simplifier), telle que

$$ \left\{ \begin{array}{ll} u_m \rightharpoonup u & \text{faiblement dans } L^2(]0;T[,H^1_0(\Omega)), \\ u'_m \rightharpoonup u' & \text{faiblement dans } L^2(]0;T[,L^2(\Omega)), \\ u''_m \rightharpoonup u'' & \text{faiblement dans } L^2(]0;T[,H^{-1}(\Omega)). \end{array} \right. $$

On a aussi convergence faible-$*$ dans les espaces L^∞ correspondants pour u_m et u'_m mais nous n'utiliseront pas cette information. Par contre on a bien sûr $u \in L^\infty(]0;T[,H^1_0(\Omega))$ et $u' \in L^\infty(]0;T[,L^2(\Omega))$. On a aussi d'après le Théorème 6.32

$$ u \in C^0([0;T],L^2(\Omega)), \qquad u' \in C^0([0;T],H^{-1}(\Omega)). $$

Soit maintenant m' fixé et $v \in L^2(]0;T[,V_{m'})$. On a d'après $(O1)_m$ pour $m \geq m'$ et après intégration par rapport à t

$$ \int_0^T {}_{H^{-1}(\Omega)}\langle u''_m(t), v(t)\rangle_{H^1_0(\Omega)} dt + \int_0^T \int_\Omega \nabla u_m(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2(\Omega)} dt $$

qui s'écrit aussi

$$ {}_{L^2(]0;T[,H^{-1}(\Omega))}\langle u''_m, v\rangle_{L^2(]0;T[,H^1_0(\Omega))} + \langle u_m, v\rangle_{L^2(]0;T[,H^1_0(\Omega))} = \langle f, v\rangle_{L^2(]0;T[,L^2(\Omega))}. $$

On peut alors passer à la limite faible pour obtenir

$$ {}_{L^2(]0;T[,H^{-1}(\Omega))}\langle u'', v\rangle_{L^2(]0;T[,H_0^1(\Omega))} + \langle u, v\rangle_{L^2(]0;T[,H_0^1(\Omega))} = \langle f, v\rangle_{L^2(]0;T[,L^2(\Omega))}, \quad (7.22) $$

ceci pour tout $v \in L^2(]0;T[, V_{m'})$ avec m' quelconque. Comme ces fonctions sont denses dans $L^2(]0;T[, H_0^1(\Omega))$, on déduit bien que u vérifie (i).

Il reste à vérifier que u vérifie $(O2)$ et $(O3)$. Considérons une fonction régulière quelconque $v \in C^\infty([0;T], V_{m'})$, telle que $v(T) = v'(T) \equiv 0$. En intégrant (7.22) par parties, on obtient

$$ \int_0^T \langle v''(t), u(t)\rangle_{L^2(\Omega)} dt + \int_0^T \int_\Omega \nabla u(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2(\Omega)} dt \\ - \langle u(0), v'(0)\rangle + \langle u'(0), v(0)\rangle. \quad (7.23) $$

Or $(O1)_m$ donne de la même façon

$$ \int_0^T \langle v''(t), u_m(t)\rangle_{L^2(\Omega)} dt + \int_0^T \int_\Omega \nabla u_m(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2(\Omega)} dt \\ - \langle u_m(0), v'(0)\rangle + \langle u'_m(0), v(0)\rangle. \quad (7.24) $$

Passant à la limite faible et utilisant que $u_m(0) \to g$ dans $H_0^1(\Omega)$ et que $u'_m(0) \to h$ dans $L^2(\Omega)$ par construction, nous trouvons finalement :

$$ \int_0^T \langle v''(t), u(t)\rangle_{L^2(\Omega)} dt + \int_0^T \int_\Omega \nabla u(t) \cdot \nabla v(t) dt = \int_0^T \langle f(t), v(t)\rangle_{L^2(\Omega)} dt \\ - \langle g, v'(0)\rangle + \langle h, v(0)\rangle. \quad (7.25) $$

Ainsi $u(0) = g$ et $u'(0) = h$ car $v(0)$, $v'(0)$ et m' sont arbitraires. Ceci termine la démonstration de l'existence d'une solution faible.

Étape 4 : unicité de la solution faible.

Nous devons montrer que si $f = g = h = 0$ et u est une solution faible, alors nécessairement $u \equiv 0$. La preuve serait une facile adaptation de celle du Lemme 7.9 si nous savions que $u' \in L^2(]0;T[, H_0^1(\Omega))$ comme c'est le cas pour u'_m. Comme nous n'avons pas cette information, elle est un peu plus difficile.

Fixons un $t_0 \in]0;T[$ et introduisons la primitive temporelle v de $-u$ qui s'annule en t_0 :

$$ v(t) = \int_t^{t_0} u(s)\, ds. $$

Clairement $v' = -u \in L^2(]0;T[, H_0^1(\Omega)) \cap C^0([0;T], L^2(\Omega))$ et

$$ \nabla v = \int_t^{t_0} \nabla u(s)\, ds, \qquad \nabla v \in L^2(]0;T[, L^2(\Omega)) \cap C^0([0;T], H^{-1}(\Omega)). $$

Notons que

$$v(t_0) = 0 \quad \text{et} \quad \nabla v(t_0) = 0.$$

On applique maintenant $(O1)$ et on intègre sur $]0; t_0[$:

$$\int_0^{t_0} {}_{H^{-1}(\Omega)}\langle u''(s), v(s)\rangle_{H_0^1(\Omega)} ds + \int_0^{t_0} \int_\Omega \nabla u(s) \cdot \nabla v(s)\, ds = 0.$$

Une intégration par parties avec $u'(0) = 0$ et $v(t_0) = 0$ fournit

$$-\int_0^{t_0} {}_{H^{-1}(\Omega)}\langle u'(s), v'(s)\rangle_{H_0^1(\Omega)} ds + \int_0^{t_0} \int_\Omega \nabla u(s) \cdot \nabla v(s)\, ds = 0$$

donc puisque $v' = -u$

$$\int_0^{t_0} {}_{H^{-1}(\Omega)}\langle u'(s), u(s)\rangle_{H_0^1(\Omega)} ds - \int_0^{t_0} \langle v'(s), v(s)\rangle_{H_0^1(\Omega)}\, ds = 0.$$

Or

$${}_{H^{-1}(\Omega)}\langle u'(s), u(s)\rangle_{H_0^1(\Omega)} = \frac{1}{2}\frac{d}{ds}\langle u(s), u(s)\rangle_{L^2(\Omega)}$$

$$\langle v'(s), v(s)\rangle_{H_0^1(\Omega)} = \frac{1}{2}\frac{d}{ds}\langle v(s), v(s)\rangle_{H_0^1(\Omega)}$$

donc on trouve finalement puique $u(0) = 0$ et $v(t_0) = 0$

$$\|u(t_0)\|^2_{L^2(\Omega)} + \|v(0)\|^2_{H_0^1(\Omega)} = 0.$$

Comme t_0 était quelconque, on a bien

$$u \equiv 0$$

et la solution faible obtenue est unique.

Étape 5 : preuve de (7.17)**.**

L'inégalité (7.17) est obtenue par passage à la limite dans (7.18). ♢

Exercice 7.11. (Formules explicites pour la décomposition sur les modes propres du Laplacien). *Considérons une solution faible u de l'équation des ondes sur Ω. On décompose u sous la forme*

$$u(t) = \sum_{k\geq 1} \alpha_k(t) w_k$$

et on pose $\beta_k(t) = \langle f(t), w_k\rangle$.

1. *Trouver l'équation différentielle ordinaire vérifiée par α_k et écrire la forme de la solution pour tout k.*

2. *En déduire que u doit s'écrire sous la forme*

$$u(t) = V'(t)g + V(t)h + \int_0^t V(t-s)f(s)\,ds \tag{7.26}$$

où

$$V(t) = \sum_{k\geq 1} \frac{\sin(\sqrt{\lambda_k}t)}{\sqrt{\lambda_k}} |w_k\rangle\langle w_k|.$$

3. *Montrer que*

$$\|V(t)\|_{L^2(\Omega)\to L^2(\Omega)} \leq Ct \qquad \|V'(t)\|_{L^2(\Omega)\to L^2(\Omega)} \leq C,$$

$$\|V(t)\|_{L^2(\Omega)\to H_0^1(\Omega)} \leq C, \qquad \|V'(t)\|_{H_0^1(\Omega)\to H_0^1(\Omega)} \leq C.$$

Vérifier que $V''(t) = \Delta V(t) = V(t)\Delta$ et en déduire que

$$\|V''(t)\|_{H_0^1(\Omega)\to L^2(\Omega)} \leq C.$$

4. *Montrer que la formule* (7.26) *fournit une fonction telle que $u \in L^\infty(]0;T[, H_0^1(\Omega))$, $u' \in L^\infty(]0;T[, L^2(\Omega))$ et $u'' \in L^2(]0;T[, L^2(\Omega))$, qui est l'unique solution faible de l'équation des ondes sur Ω.*

Exercice 7.12. (Un théorème général)

1. *En s'inspirant de l'une des deux démonstrations du Théorème 7.6, démontrer le résultat général suivant :*

   Théorème 7.13. *Soit H et V deux espaces de Hilbert tel que $V \hookrightarrow H$ avec injection compacte et V est dense dans H. Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique continue et coercive dans V. Soit un temps final $T > 0$, une condition initiale $(g,h) \in V \times H$ et un terme source $f \in L^2(]0;T[, H)$. Il existe une unique solution faible $u \in L^2(]0;T[, V)$ telle que $u' \in L^2(]0;T[, H)$ et $u'' \in L^2(]0;T[, V')$ au problème*

$$\begin{cases} \frac{d^2}{dt^2}\langle u(t), v\rangle_H + a(u(t), v) = \langle f(t), v\rangle_H \quad \forall v \in V,\ t \in]0;T[\\ u(0) = g,\ u'(0) = h. \end{cases}$$

   De plus il existe une constante C telle que

$$\|u\|_{L^\infty([0;T],V)} + \|u'\|_{L^\infty([0;T],H)} + \|u''\|_{L^2(]0;T[,V')} \leq C(\|f\|_{L^2(]0;T[,H)} + \|g\|_V + \|h\|_H).$$

2. (Équation des ondes en milieu inhomogène). *Soit $\Omega \subset \mathbb{R}^n$ un ouvert borné régulier et A une fonction définie sur Ω à valeurs dans les matrices symétriques réelles définies positives de taille n, telle que*

$$\alpha I_n \leq A(x) \leq \beta I_n$$

p.p. $x \in \Omega$, *où* $\alpha, \beta > 0$ *et* I_n *est l'identité de* $\mathbb{R}^n$. *En déduire l'existence d'une unique solution faible au problème*

$$\begin{cases} \frac{\partial^2}{\partial t^2}u(t,x) - \operatorname{div}(A(x)\nabla u(t,x)) = f, & (t,x) \in]0;T[\times\Omega, \\ u(t,x) = 0, & (t,x) \in]0;T[\times\partial\Omega \\ u(0,x) = g(x),\ \frac{\partial}{\partial t}u(0,x) = h(x), & \end{cases}$$

où $g \in H^1(\Omega)$, $h \in L^2(\Omega)$ *et* $f \in L^2(]0;T[, L^2(\Omega))$.

7.2.2.2 Propriétés qualitatives des solutions faibles

Comme l'étude du cas 1D nous l'a montré, il n'y a pas de régularisation ou de propagation à vitesse infinie avec l'équation des ondes.

Théorème 7.14 (Réversibilité en temps)**.** *Soit* $\Omega \subset \mathbb{R}^n$ *un ouvert borné régulier. On suppose que* $f \in L^2(]0;T[, L^2(\Omega))$, $g \in H_0^1(\Omega)$ *et* $h \in L^2(\Omega)$. *Alors l'équation des ondes rétrograde en temps*

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}u(t,x) - \Delta u(t,x) = f(t,x), \quad (t,x) \in]0;T[\times\Omega, \\ u(t,x) = 0 \text{ si } x \in \partial\Omega, \\ u(T,x) = g(x), \\ \frac{\partial}{\partial t}u(T,x) = h(x), \end{cases} \tag{7.27}$$

admet une unique solution faible $\tilde{u} \in L^\infty(]0;T[, H_0^1(\Omega))$ *avec* $\tilde{u}' \in L^\infty(]0;T[, L^2(\Omega))$. *De plus si* u *est la solution de l'équation des ondes usuelle telle que* $u(T) = g$ *et* $u'(T) = h$, *alors on a* $u = \tilde{u}$.

Preuve : On fait le changement de variable $v(t) = \tilde{u}(T-t)$ pour se ramener à l'équation des ondes usuelle et obtenir l'existence et l'unicité d'une solution faible. L'équation ne change pas grâce à la dérivée d'ordre deux en temps. ♢

Si les données sont plus régulières, on peut montrer que la solution est elle même plus régulière. En résumé :

g	h	f	u
$H^{m+1}(\Omega)$	$H^m(\Omega)$	$\frac{d^k}{dt^k}f \in L^2(]0;T[, H^{m-k}(\Omega))$ $k = 0, ..., m$	$\frac{d^k}{dt^k}u \in L^\infty(]0;T[, H^{m+1-k}(\Omega))$ $k = 0, ..., m+1$

mais il faut des *conditions de compatibilité* pour que ceci soit vrai, voir par exemple [7]. Nous énonçons un résultat beaucoup plus simple, similaire au Théorème 6.60.

Théorème 7.15 (Régularité)**.** *On suppose que* Ω *est un ouvert borné de bord* C^∞ *et que* $g, h \in C_0^\infty(\Omega)$, $f \in C^\infty([0;T], C_0^\infty(\Omega))$. *Alors*

$$u \in C^\infty([0;T]\times\overline{\Omega}).$$

Preuve : À faire en exercice ! ♢

Comme annoncé précédemment, nous pouvons maintenant montrer que la solution faible est en fait plus régulière que prévue par rapport au temps.

Théorème 7.16 (Conservation de l'énergie)**.** *On se place sous les hypothèses du Théorème 7.6. Alors la solution u de l'équation vérifie*

$$u \in C^0([0;T], H_0^1(\Omega)), \qquad u' \in C^0([0;T], L^2(\Omega)),$$

et satisfait l'égalité

$$\int_\Omega \left(\left| \frac{\partial}{\partial t} u(t,x) \right|^2 + |\nabla u(t,x)|^2 \right) dx = \int_\Omega \left(h(x)^2 + |\nabla g(x)|^2 \right) dx + 2 \int_0^t \int_\Omega f(s,x) u'(s,x) dx\, ds \quad (7.28)$$

pour tout $t \in]0;T[$. En particulier si $f \equiv 0$, on a la conservation de l'énergie :

$$\int_\Omega \left(\left| \frac{\partial}{\partial t} u(t,x) \right|^2 + |\nabla u(t,x)|^2 \right) dx = \int_\Omega \left(h(x)^2 + |\nabla g(x)|^2 \right) dx \quad (7.29)$$

pour tout $t \in]0;T[$.

Preuve : Nous raisonnons une fois de plus par densité. Considérons des suites g_n, h_n et f_n régulières telles que

$$\lim_{n\to\infty} \|g_n - g\|_{H_0^1(\Omega)} = \lim_{n\to\infty} \|h_n - h\|_{L^2(\Omega)} = \lim_{n\to\infty} \|f_n - f\|_{L^2(]0;T[,L^2(\Omega))} = 0.$$

Comme la solution correspondante u_n est régulière (on a seulement besoin de $u' \in L^2(]0;T[, H_0^1(\Omega))$), il est facile de démontrer (7.28) en suivant la méthode de preuve utilisée pour u_m, cf (7.19).

En fait on a même en appliquant (7.17) à $u_n - u_m$

$$\|u_n' - u_m'\|_{L^\infty(]0;T[,L^2(\Omega))} + \|u_n - u_m\|_{L^\infty(]0;T[,H_0^1(\Omega))} \leq C \left(\|h_n - h_m\|_{L^2(\Omega)}^2 + \|g_n - g_m\|_{H_0^1(\Omega)}^2 + + \|f_n - f_m\|_{L^2(]0;T[,L^2(\Omega))} \right). \quad (7.30)$$

On en déduit que $u_n \to u$ et $u_n' \to u'$ respectivement dans $L^\infty(]0;T[, H_0^1(\Omega))$ et $L^\infty(]0;T[, L^2(\Omega))$. Or pour tout n, $u_n \in C^0([0;T], H_0^1(\Omega))$ et $u_n' \in C^0([0;T], L^2(\Omega))$. Donc $u \in C^0([0;T], H_0^1(\Omega))$ et $u' \in C^0([0;T], L^2(\Omega))$ comme limites uniformes de fonctions continues. On obtient alors l'égalité par passage à la limite. ♢

Si $f \equiv 0$ et g, h ont un support compact $K \subset \Omega$, on peut montrer que l'unique solution faible u coïncide avec la solution définie sur tout l'espace $\mathbb{R}^n$, et que la propagation a lieu à vitesse finie (cf Exercice 7.4) tant que cette solution ne touche pas le bord, donc sur un intervalle de temps $[0;\epsilon]$. Mais dès que la solution touche le bord, les deux solutions diffèrent à cause des conditions de Dirichlet sur $\partial\Omega$.

Chapitre 8

Méthode des éléments finis pour les équations d'évolution

On présente ici le principe de la méthode des éléments finis pour les équations d'évolution. Nous nous concentrons ici sur deux types d'équations à savoir l'équation de la chaleur et l'équation des ondes.

8.1 L'équation de la chaleur

8.1.1 Semi-discrétisation en espace

Nous commençons par seulement discrétiser en *espace* la formulation variationnelle de l'équation de la chaleur (6.12). Pour cela, nous considérons une suite d'espaces de dimension finie V_h, $h \to 0$, avec $V_h \subset H_0^1(\Omega)$. On suppose qu'il existe une application linéaire $r_h : H^k(\Omega) \to V_h$ telle que

$$\lim_{h\to 0} \|1 - r_h\|_{H^k(\Omega)\cap H_0^1(\Omega)\to H_0^1(\Omega)} = 0 \tag{8.1}$$

où k est assez grand.

L'exemple typique est celui de l'approximation par des éléments finis. Soit Ω un ouvert borné connexe polyédrique de $\mathbb{R}^n$. Rappelons qu'une suite $(\mathcal{T}_h)$ est une suite de *maillages triangulaires réguliers* de Ω si pour chaque h, $\mathcal{T}_h = (K_i)$ où les K_i sont des tétraèdres formant un pavage de $\overline{\Omega}$. L'intersection de deux tétraèdres K_i et K_j est soit vide, soit un tétraèdre de dimension $m \leq n-1$ dont tous les sommets sont aussi des sommets de K_i et K_j. De plus on a

$$\max_i \operatorname{diam}(K_i) = h$$

$$\forall i, \quad \operatorname{diam}(K_i) \leq C \max_{B_r \subseteq K_i} r$$

c'est-à-dire que chaque K_i est de volume d'ordre h^n (il ne peut pas être aplati ou s'aplatir quand $h \to 0$). Introduisons alors l'espace d'approximation

$$V_{0h}^k := \left\{ u \in C^0(\overline{\Omega}) \mid u_{|\partial\Omega} = 0,\ u_{|K_i} \text{ est un polynôme de degré } k \right\}.$$

Pour chaque tel maillage, on peut définir une suite de points $(a_i)_{i=1}^p \in \Omega$ appelés noeuds ($p = \dim V_h$) et des fonctions $\varphi_i \in V_{0h}^k \subset H_0^1(\Omega)$ telles que $\varphi_i(a_j) = \delta_{ij}$. Pour toute fonction régulière v on pose alors

$$r_h v(x) = \sum_{i=1}^{p} v(a_i)\varphi_i(x).$$

Le résultat suivant est classique [1] :

Proposition 8.1. *Soit $(\mathcal{T}_h)$ une suite de maillages réguliers comme ci-dessus. On suppose que $k+1 > n/2$. Alors pour tout $v \in H^{k+1}(\Omega)$, l'interpolée $r_h v$ est bien définie et satisfait :*

$$\|v - r_h v\|_{H_0^1(\Omega)} \leq Ch^k \|v\|_{H^{k+1}(\Omega)\cap H_0^1(\Omega)}.$$

La propriété (8.1) est donc vraie pour une suite de maillages triangulaires réguliers, puisque la proposition ci-dessus signifie que

$$\|1 - r_h\|_{H^{k+1}(\Omega)\to H^1(\Omega)} \leq Ch^k$$

dès que $k+1 > n/2$.

Un autre exemple est celui d'une approximation de Galerkin (moins utilisée dans la pratique) où on pose simplement

$$V_h = \mathrm{Vect}(w_1, ..., w_m), \quad h = 1/m$$

où (w_k) est une base orthonormée de $L^2(\Omega)$ bien choisie (par exemple les fonctions propres du Laplacien comme nous l'avons fait dans les preuves précédentes). On peut alors poser simplement $r_h =$ le projecteur orthogonal sur V_h pour le produit scalaire de $H_0^1(\Omega)$.

Exercice 8.2. *Vérifier que la propriété* (8.1) *est bien vérifiée pour k assez grand, si on prend $V_h = \mathrm{Vect}(w_1, ..., w_m)$, $h = 1/m$ où les w_k sont les fonctions propres du Laplacien.*

La semi-discrétisation en espace consiste à résoudre le problème variationnel suivant

$$\begin{cases} \frac{d}{dt}\langle u_h(t), v_h\rangle_{L^2(\Omega)} + \int_\Omega \nabla u_h(t)\cdot\nabla v_h = \langle f(t), v_h\rangle_{L^2(\Omega)} & \forall v_h \in V_h \\ u_h(0) = g_h, \end{cases}$$

où $g_h \in V_h$ est telle que $g_h \to g$ dans $L^2(\Omega)$. Soit $(v_1, ..., v_m)$ une base orthonormée de V_h. Si on écrit

$$u_h(t) = \sum_{k=1}^{m} \alpha_k^h(t) v_k,$$

on trouve que les α_k^h doivent vérifier

$$\begin{cases} \sum_{k=1}^m \frac{d}{dt}\alpha_k^h(t)(S_h)_{k,\ell} + \sum_{k=1}^m \alpha_k^h(t)(K_h)_{k,\ell} = \langle f(t), v_k\rangle_{L^2(\Omega)} \quad k = 1, ..., m \\ \alpha_h^k(0) = \langle g_h, v_k\rangle, \end{cases}$$

où

$$(S_h)_{k,\ell} := \langle v_k, v_\ell\rangle_{L^2(\Omega)}, \qquad (K_h)_{k,\ell} := \int_\Omega \nabla v_k \cdot \nabla v_\ell.$$

Ce système d'équations différentielles ordinaires s'écrit sous la forme matricielle

$$\begin{cases} S_h \frac{d}{dt}\alpha^h(t) + K_h\alpha^h(t) = b^h(t) \\ \alpha(0) = a^h \end{cases} \tag{8.2}$$

en posant

$$\alpha^h = \begin{pmatrix} \alpha_1^h(t) \\ \vdots \\ \alpha_m^h(t) \end{pmatrix}, \qquad a^h = \begin{pmatrix} \langle g_h, v_1\rangle_{L^2(\Omega)} \\ \vdots \\ \langle g_h, v_m\rangle_{L^2(\Omega)} \end{pmatrix}, \qquad b^h = \begin{pmatrix} \langle f(t), v_1\rangle_{L^2(\Omega)} \\ \vdots \\ \langle f(t), v_m\rangle_{L^2(\Omega)} \end{pmatrix}.$$

L'existence et l'unicité ainsi qu'une formule explicite s'obtiennent par diagonalisation simultanée des matrices S_h et K_h. En pratique, on résout numériquement (8.2) par une discrétisation temporelle, comme pour tout système d'équations différentielles ordinaires.

On peut alors démontrer le résultat suivant :

Théorème 8.3. *Soient $f \in L^2(]0;T[, L^2(\Omega))$, $g \in L^2(\Omega)$ et $u \in L^2(]0;T[, H_0^1(\Omega)) \cap C^0([0;T], L^2(\Omega))$ l'unique solution faible de l'équation de la chaleur. Soit u_h l'unique solution variationnelle dans V_h. On suppose que la suite V_h satisfait* (8.1) *et que $\lim_{h\to 0} \|g_h - g\|_{L^2(\Omega)} = 0$. Alors on a*

$$\lim_{h\to 0} \|u_h - u\|_{L^2(]0;T[,H_0^1(\Omega))} = \lim_{h\to 0} \sup_{t\in[0;T]} \|u_h(t) - u(t)\|_{L^2(\Omega)} = 0.$$

Preuve : Soit $\epsilon > 0$ et $\tilde{f} \in C_0^\infty(]0;T[\times\Omega)$, $\tilde{g} \in C_0^\infty(\Omega)$ tels que

$$\left\|f - \tilde{f}\right\|_{L^2(]0;T[,L^2(\Omega))} \leq \epsilon, \quad \|g - \tilde{g}\|_{L^2(\Omega)} \leq \epsilon.$$

D'après la continuité par rapport aux données (cf le Théorème 6.53 pour le problème sur tout $H_0^1(\Omega)$ et une généralisation évidente sur V_h), on a

$$\|u - \tilde{u}\|_{L^\infty(]0;T[,L^2(\Omega))} + \|u_h - \tilde{u}_h\|_{L^\infty(]0;T[,L^2(\Omega))} \leq C\epsilon$$

où $\tilde{u}$ et $\tilde{u}_h$ sont les solutions faibles associées à $\tilde{g}$ et $\tilde{f}$. Il suffit donc de prouver le théorème pour $\tilde{u}$ et $\tilde{u}_h$. Pour simplifier les notations, nous supposerons $\tilde{f} = f \in C_0^\infty(]0;T[\times\Omega)$ et $\tilde{g} = g \in C_0^\infty(\Omega)$ de sorte que $u \in C^\infty([0;T] \times \overline{\Omega})$ d'après le Théorème 6.60.

On a pour tout $v_h \in V_h$

$$\langle u'(t) - u_h'(t), v_h \rangle_{L^2(\Omega)} + \int_\Omega \nabla(u(t) - u_h(t)) \cdot \nabla v_h = 0.$$

Soit maintenant π_h le projecteur orthogonal sur V_h, pour le produit scalaire $\langle v, w \rangle_{H_0^1(\Omega)} = \int_\Omega \nabla v \cdot \nabla w$. On a donc pour tout $v \in H^k(\Omega) \cap H_0^1(\Omega) \subseteq V \subset H_0^1(\Omega)$, avec k assez grand,

$$\begin{aligned} \|(1-\pi_h)v\|_{L^2(\Omega)} \leq C \|(1-\pi_h)v\|_{H_0^1(\Omega)} &\leq C \|(1-r_h)v\|_{H_0^1(\Omega)} \\ &\leq C \|1-r_h\|_{H^k(\Omega)\to H_0^1(\Omega)} \|v\|_{H^k(\Omega)} \end{aligned} \quad (8.3)$$

d'après l'inégalité de Poincaré et l'hypothèse (8.1). Nous avons aussi utilisé que

$$\forall v_h \in V_h, \quad \|v - \pi_h v\|_{H_0^1(\Omega)} \leq \|v - v_h\|_{H_0^1(\Omega)}$$

d'après la caractérisation de la projection orthogonale, donc en particulier que

$$\|v - \pi_h v\|_{H_0^1(\Omega)} \leq \|v - r_h v\|_{H_0^1(\Omega)}.$$

On a pour tout $v_h \in V_h$

$$\langle \pi_h u'(t) - u_h'(t), v_h \rangle_{L^2(\Omega)} + \int_\Omega \nabla(\pi_h u(t) - u_h(t)) \cdot \nabla v_h = \langle (\pi_h - 1)u'(t), v_h \rangle_{L^2(\Omega)}$$

puisque

$$\forall v_h \in V_h, \quad \int_\Omega \nabla(\pi_h - 1)u(t) \cdot \nabla v_h = 0$$

par définition de π_h. Prenons maintenant $v_h = \pi_h u - u_h \in V_h$. On trouve

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\pi_h u(t) - u_h(t)\|_{L^2(\Omega)}^2 + \int_\Omega \nabla(\pi_h u(t) - u_h(t)) \cdot \nabla(\pi_h u(t) - u_h(t)) \\ \leq \frac{1}{2}\|\pi_h u(t) - u_h(t)\|_{L^2(\Omega)}^2 + \frac{1}{2}\|(\pi_h - 1)u'(t)\|_{L^2(\Omega)}^2. \end{aligned} \quad (8.4)$$

D'après l'inégalité de Gronwall, on déduit que

$$\begin{aligned} \|\pi_h u(t) - u_h(t)\|_{L^2(\Omega)}^2 &\leq e^t \left(\|\pi_h g - g_h\|_{L^2(\Omega)}^2 + \int_0^t \|(\pi_h - 1)u'(s)\|_{L^2(\Omega)}^2 \, ds \right) \\ &\leq C \Bigg(\|\pi_h g - g\|_{L^2(\Omega)}^2 + \|g - g_h\|_{L^2(\Omega)}^2 \\ &\qquad + \int_0^t \|(\pi_h - 1)u'(s)\|_{L^2(\Omega)}^2 \, ds \Bigg) \end{aligned}$$

Comme g est régulière, $g \in H^k(\Omega)$ et de même $u'(s) \in H^k(\Omega)$ d'après le Théorème 6.60 uniformément par rapport à s, donc

$$\|\pi_h u(t) - u_h(t)\|^2_{L^2(\Omega)} \leq C \|g - g_h\|^2_{L^2(\Omega)} + C \|1 - r_h\|^2_{H^k(\Omega)\to H^1(\Omega)} \left(\|g\|^2_{H^k(\Omega)} + \int_0^t \|u'(s)\|^2_{H^k(\Omega)}\, ds \right).$$

En utilisant (8.1), on trouve que

$$\lim_{h\to 0} \max_{t\in[0;T]} \|\pi_h u(t) - u_h(t)\|_{L^2(\Omega)} = 0$$

quand $h \to 0$, et donc que

$$\lim_{h\to 0} \max_{t\in[0;T]} \|u(t) - u_h(t)\|_{L^2(\Omega)} = 0$$

puisque

$$\lim_{h\to 0} \|\pi_h u(t) - u(t)\|_{L^2(\Omega)} = 0$$

d'après (8.3), u étant régulière. De même on trouve d'après (8.4)

$$\lim_{h\to 0} \max_{t\in[0;T]} \|u(t) - u_h(t)\|_{H^1_0(\Omega)} = 0$$

donc en particulier que

$$\lim_{h\to 0} \|u - u_h\|_{L^2(]0;T[,H^1_0(\Omega))} = 0.$$

Ceci termine la preuve du théorème. ♢

8.1.2 Discrétisation totale en espace-temps

Après avoir discrétisé le problème par rapport à la variable spatiale, il nous faut maintenant discrétiser en temps le système d'équations différentielles ordinaires (8.2). Pour simplifier, nous supprimons la référence à la variable h décrivant la taille du maillage.

Nous considérons une grille uniforme en temps de pas de temps $\Delta t = T/n$, et nous posons $t_n = n\Delta t$. Notons a^n l'approximation de $\alpha(t_n)$ calculée par un schéma. Pour calculer numériquement des solutions approchées de (8.2), la méthode la plus simple et la plus utilisée est celle du *θ-schéma* :

$$S\frac{a^{n+1} - a^n}{\Delta t} + K(\theta a^{n+1} + (1-\theta)a^n) = \theta b(t_{n+1}) + (1-\theta)b(t_n)$$

qui peut se réécrire sous la forme

$$(S + \theta\Delta t K)a^{n+1} = (S - (1-\theta)\Delta t K)a^n + \Delta t(\theta b(t_{n+1}) + (1-\theta)b(t_n)). \qquad (8.5)$$

Lorsque $\theta = 0$, on trouve le schéma d'*Euler explicite*, lorsque $\theta = 1$, il s'agit du schéma d'*Euler implicite*, alors que si $\theta = 1/2$, on parle de schéma de *Crank-Nicholson* ou du *point milieu*.

Dans la formule précédente, on suppose pour simplifier que f est assez régulière, de sorte que $b(t_n)$ ait un sens (sinon b n'est *a priori* seulement définie presque partout). Si f n'est pas régulière, il faut remplacer $b(t_n)$ par une moyenne de $b(t)$ sur un intervalle autour de t_n.

Notons que comme la matrice S n'est en général pas diagonale, il est souvent nécessaire de résoudre un système linéaire à chaque étape même pour le schéma explicite. Bien sûr, on peut utiliser des schémas ne rentrant pas dans la classe (8.5).

Une propriété importante des schémas est leur ordre. Le lecteur vérifiera en exercice que le θ-schéma est toujours d'ordre un, sauf quand $\theta = 1/2$ où il est d'ordre 2.

La solution exacte du système d'équations différentielles (8.2) est une fonction $\alpha \in C^0([0;T],\mathbb{R}^m)$ donc bornée sur $[0;T]$ (ceci est vrai même quand $b(t)$ n'est que dans $L^2(]0;T[)$). Une propriété importante d'un schéma est sa stabilité. Nous dirons qu'un schéma est stable si $\|a\|$ possède la même propriété et reste uniformément bornée si on augmente le nombre de points de discrétisation.

Définition 8.4 (Stabilité)**.** *Un schéma est dit* stable *si on a*

$$\sup_n \langle Sa^n, a^n\rangle \leq C$$

pour une constante C ne dépendant pas de Δt (mais qui peut dépendre des données $\alpha(0)$ et $b(t)$ et de T).

On a choisi la norme associée à la matrice de masse S (que l'on suppose inversible) car c'est celle qui correspond à la norme de $L^2(\Omega)$. Mais bien sûr toutes les normes sont équivalentes en dimension finie.

Lemme 8.5 (Stabilité des θ-schémas pour l'équation de la chaleur)**.** *Si $1/2 \leq \theta \leq 1$, le θ-schéma est inconditionnellement stable. Si $0 \leq \theta < 1/2$ il est stable sous la condition*

$$\max \lambda_i \Delta t \leq \frac{2}{1-2\theta} \tag{8.6}$$

où les λ_i sont les valeurs propres de $S^{-1/2}KS^{-1/2}$.

Preuve : On peut choisir de travailler une base orthonormée pour le produit scalaire associé à S et qui est orthogonale pour celui associé à K. Dans cette base, le θ-schéma s'écrit

$$(I + \theta\Delta t D)\tilde{a}^{n+1} = (I - (1-\theta)\Delta t D)\tilde{a}^n + \Delta t\tilde{b}^n$$

où $\tilde{b}^n$ contient les coordonnées de $(\theta b(t_{n+1}) + (1-\theta)b(t_n))$ dans la nouvelle base et

$$D = \mathrm{diag}(\lambda_i)$$

contient les valeurs propres de K pour le produit scalaire de S, c'est-à-dire celles de $S^{-1/2}KS^{-1/2}$. Donc on obtient

$$\tilde{a}^{n+1} = \tilde{D}a^n + (I + \theta\Delta t D)^{-1}\Delta t\tilde{b}^n \tag{8.7}$$

$$\tilde{D} = \text{diag}\left(\frac{1-(1-\theta)\Delta t\lambda_i}{1+\theta\Delta t\lambda_i}\right) = \text{diag}\left(1 - \frac{\Delta t\lambda_i}{1+\theta\Delta t\lambda_i}\right).$$

La condition de stabilité pour un système sous la forme (8.7) est alors que $\left\|\tilde{D}\right\| \leq 1$, c'est-à-dire

$$-1 \leq 1 - \frac{\Delta t\lambda_i}{1+\theta\Delta t\lambda_i} \leq 1$$

pour tout i. Comme tous les $\lambda_i \geq 0$, ceci se réduit à

$$\frac{\Delta t\lambda_i}{1+\theta\Delta t\lambda_i} \leq 2$$

C'est-à-dire

$$\forall i = 1...m, \qquad (1-2\theta)\Delta t\lambda_i \leq 2.$$

♢

Il reste à démontrer la convergence du schéma vers la solution de l'équation de la chaleur.

Théorème 8.6 (Convergence)**.** *Soit Ω un ouvert régulier, $T > 0$ un temps final. On suppose que $g \in L^2(\Omega)$ et $f \in L^2(]0;T[, L^2(\Omega))$ sont suffisamment régulières et on note u l'unique solution faible de l'équation de la chaleur. On suppose aussi que V_h satisfait* (8.1) *et que $\lim_{h\to 0}\|g_h - g\|_{L^2(\Omega)} = 0$. On note u_h^n la solution de l'équation de la chaleur totalement discrétisée par un θ-schéma avec $\theta \in [0;1]$, en supposant la condition de stabilité du Lemme 8.5 vérifiée. Alors on a*

$$\lim_{\substack{\Delta t\to 0,\\ h\to 0}} \max_{0\leq t_n = n\Delta t\leq T} \|u_h^n - u(t_n)\|_{L^2(\Omega)} = 0.$$

Pour la preuve et des estimées explicites, voir [13].

8.2 L'équation des ondes

8.2.1 Semi-discrétisation en espace

Comme pour l'équation de la chaleur, on peut discrétiser l'équation des ondes seulement en espace. On introduit comme à la section 8.1.1 un espace V_h approchant $H_0^1(\Omega)$ et on suppose que (8.1) est satisfaite.

On considère l'approximation variationnelle suivante : trouver u_h vérifiant

$$\begin{cases} \frac{d^2}{dt^2}\langle u_h, v_h\rangle + \int_\Omega \nabla u_h \cdot \nabla v_h = \langle f(t), v_h\rangle \quad \forall v_h \in V_h,\ t \in]0;T[\\ u_h(0) = g_h,\ \frac{\partial}{\partial t}u_h(0) = g_h'. \end{cases}$$

où

$$\lim_{h\to 0} \|g_h - g\|_{H_0^1(\Omega)} = 0, \quad \lim_{h\to 0} \|g_h' - g'\|_{L^2(\Omega)} = 0.$$

Soit $(v_1, ..., v_m)$ une base orthonormée de V_h. Si on écrit

$$u_h(t) = \sum_{k=1}^{m} \alpha_k^h(t) v_k,$$

on trouve que les α_k^h doivent vérifier

$$\begin{cases} \sum_{k=1}^m \frac{d^2}{dt^2}\alpha_k^h(t)(S_h)_{k,\ell} + \sum_{k=1}^m \alpha_k^h(t)(K_h)_{k,\ell} = \langle f(t), v_k\rangle_{L^2(\Omega)} \quad k = 1, ..., m \\ \alpha_h^k(0) = \langle g_h, v_k\rangle, \ \frac{d}{dt}\alpha_h^k(0) = \langle g_h', v_k\rangle, \end{cases}$$

où

$$(S_h)_{k,\ell} := \langle v_k, v_\ell\rangle_{L^2(\Omega)}, \qquad (K_h)_{k,\ell} := \int_\Omega \nabla v_k \cdot \nabla v_\ell.$$

Ce système d'équations différentielles ordinaires s'écrit sous la forme matricielle

$$\begin{cases} S_h \frac{d^2}{dt^2}\alpha^h(t) + K_h \alpha^h(t) = b^h(t) \\ \alpha(0) = a^h, \ \alpha'(0) = A^h \end{cases} \tag{8.8}$$

en posant

$$\alpha^h = \begin{pmatrix} \alpha_1^h(t) \\ \vdots \\ \alpha_m^h(t) \end{pmatrix}, \ a^h = \begin{pmatrix} \langle g_h, v_1\rangle_{L^2(\Omega)} \\ \vdots \\ \langle g_h, v_m\rangle_{L^2(\Omega)} \end{pmatrix},$$

$$A^h = \begin{pmatrix} \langle g_h', v_1\rangle_{L^2(\Omega)} \\ \vdots \\ \langle g_h, v_m\rangle_{L^2(\Omega)} \end{pmatrix}, \ b^h = \begin{pmatrix} \langle f(t), v_1\rangle_{L^2(\Omega)} \\ \vdots \\ \langle f(t), v_m\rangle_{L^2(\Omega)} \end{pmatrix}.$$

L'existence et l'unicité ainsi qu'une formule explicite s'obtiennent par diagonalisation simultanée des matrices S_h et K_h. En pratique, on résout numériquement (8.8) par une discrétisation temporelle, comme pour tout système d'équations différentielles ordinaires.

Remarque 8.7. *Quand $f \equiv 0$ (donc $b^h \equiv 0$), l'équation semi-discrétisée en espace* (8.8) *décrit un système Hamiltonien. En effet, on peut la réécrire*

$$\begin{cases} q'(t) = \frac{\partial H}{\partial p}(q(t), p(t)) \\ p'(t) = -\frac{\partial H}{\partial q}(q(t), p(t)) \end{cases}$$

avec

$$H(q,p) = \frac{1}{2}\big\langle (S_h)^{-1/2} K_h (S_h)^{-1/2} q, q\big\rangle_{\mathbb{R}^n} + \frac{1}{2}\|p\|_{\mathbb{R}^n}^2$$

où $q(t) = (S_h)^{1/2}\alpha^h(t)$ et $p(t) = q'(t)$.

On peut maintenant démontrer un résultat similaire au Théorème 8.3.

Théorème 8.8. *Soient $f \in L^2(]0;T[,L^2(\Omega))$, $g \in H_0^1(\Omega)$, $g' \in L^2(\Omega)$ et $u \in L^2(]0;T[,H_0^1(\Omega)) \cap C^0([0;T],L^2(\Omega))$ l'unique solution faible de l'équation des ondes sur Ω. Soit u_h l'unique solution variationnelle dans V_h définie ci-dessus. On suppose que la suite V_h satisfait* (8.1) *et que*

$$\lim_{h\to 0} \|g_h - g\|_{H_0^1(\Omega)} = 0, \qquad \lim_{h\to 0} \|g_h' - g'\|_{L^2(\Omega)} = 0.$$

Alors on a

$$\lim_{h\to 0} \sup_{t\in[0;T]} \|u_h - u\|_{H_0^1(\Omega)} = \lim_{h\to 0} \sup_{t\in[0;T]} \|u_h' - u'\|_{L^2(\Omega)} = 0.$$

Preuve : La démonstration suit celle du Théorème 8.3 : on commence par se ramener au cas où f, g et g' sont régulières grâce à la continuité par rapport aux données.

On introduit ensuite, comme précédemment, le projecteur orthogonal π_h sur V_h pour le produit scalaire de $H_0^1(\Omega)$. On a alors

$${}_{H^{-1}(\Omega)}\langle u''(t) - u_h''(t), v_h\rangle_{H_0^1(\Omega)} + \langle u - u_h, v_h\rangle_{H_0^1(\Omega)} = 0$$

pour presque tout $t \in]0;T[$ et tout $v \in V_h$. On prend alors $v_h = \pi_h u' - u_h' \in L^2(]0;T[,H_0^1(\Omega))$ car u est très régulière et $\pi_h : H^k(\Omega) \to H_0^1(\Omega)$ pour k assez grand d'après (8.1). On obtient

$${}_{H^{-1}(\Omega)}\langle u''(t) - u_h''(t), \pi_h u' - u_h'\rangle_{H_0^1(\Omega)} + \langle \pi_h u - u_h, \pi_h u' - u_h'\rangle_{H_0^1(\Omega)} = 0$$

c'est-à-dire

$$\frac{1}{2}\frac{d}{dt}\left(\|\pi_h u'(t) - u_h'(t)\|_{L^2(\Omega)}^2 + \|\pi_h u(t) - u_h(t)\|_{H_0^1(\Omega)}^2\right) = \langle (\pi_h - 1)u''(t), \pi_h u' - u_h'\rangle_{L^2(\Omega)} \quad (8.9)$$

car u est très régulière donc $(1-\pi_h)u''(t) \in L^2(]0;T[,H_0^1(\Omega))$ d'après (8.1).

Le lemme de Gronwall nous donne

$$\begin{aligned}
&\|\pi_h u'(t) - u_h'(t)\|_{L^2(\Omega)}^2 + \|\pi_h u(t) - u_h(t)\|_{H_0^1(\Omega)}^2 \\
&\quad \le e^t \left(\|\pi_h g' - g_h'\|_{L^2(\Omega)}^2 + \|\pi_h g - g_h\|_{H_0^1(\Omega)}^2 + \int_0^t \|(\pi_h - 1)u''(t)\|_{L^2(\Omega)}^2\right) \\
&\quad \le C\left(\|g' - g_h'\|_{L^2(\Omega)}^2 + \|g - g_h\|_{H_0^1(\Omega)}^2\right) \\
&\qquad + C\,\|1 - \pi_h\|_{H^k(\Omega)\to H_0^1(\Omega)}^2 \left(\|g'\|_{H^k(\Omega)}^2 + \|g\|_{H^k(\Omega)}^2 + \|u''\|_{L^2(]0;T[,H^k(\Omega))}^2\right)
\end{aligned}$$

qui permet de conclure. ♢

8.2.2 Discrétisation totale en espace-temps

Il faut discrétiser le système d'équations différentielles ordinaires (8.8). Pour simplifier, nous supprimons la référence à la variable h décrivant la taille du maillage.

Nous considérons une grille uniforme en temps de pas de temps $\Delta t = T/n$, et posons comme précédemment $t_n = n\Delta t$. Nous notons aussi a^n l'approximation de $\alpha(t_n)$ calculée par le schéma considéré.

Pour $0 \leq \theta \leq 1/2$, on peut considérer comme avant le *θ-schéma*

$$S\frac{a^{n+1}-2a^n+a^{n+1}}{\Delta t^2}+K(\theta a^{n+1}+(1-2\theta)a^n+\theta a^{n-1}) = \theta b(t_{n+1})+(1-2\theta)b(t_n)+\theta b(t_{n-1})$$

avec les conditions initiales

$$a^0 = \alpha(0) \quad \text{et} \quad a^1 = \alpha'(0).$$

Un schéma plus fréquemment utilisé est celui de *Newmark* :

$$\begin{aligned} S\frac{a^{n+1}-2a^n+a^{n+1}}{\Delta t^2} + K(\theta a^{n+1} + (1/2+\delta-2\theta)a^n + (1/2-\delta+\theta)a^{n-1}) \\ = \theta b(t_{n+1}) + (1/2+\delta-2\theta)b(t_n) + (1/2-\delta+\theta)b(t_{n-1}) \end{aligned}$$

qui généralise les θ-schémas obtenus en prenant $\delta = 1/2$.

Lemme 8.9 (Stabilité du schéma de Newmark pour l'équation des ondes)**.** *Si $\delta < 1/2$, le schéma de Newmark est toujours instable. Si $\delta \geq 1/2$, le schéma est stable si*

$$\delta \leq 2\theta \leq 1$$

ou si

$$0 \leq 2\theta < \delta \quad et \quad \max_i \lambda_i(\Delta t)^2 < \frac{2}{\delta - 2\theta}$$

où les λ_i sont les valeurs propres de $S^{-1/2}KS^{-1/2}$.

Preuve : Comme précédemment, on se place dans une base orthonormée pour S et orthogonale pour K, et on note $\tilde{b}^n$ le vecteur contenant les coordonnées de $\theta b(t_{n+1}) + (1/2+\delta-2\theta)b(t_n) + (1/2-\delta+\theta)b(t_{n-1})$ dans cette base. On introduit aussi les matrices

$$A_i = \begin{pmatrix} \frac{2-\lambda_i(\Delta t)^2(1/2+\delta-2\theta)}{1+\theta\lambda_i(\Delta t)^2} & \frac{1+\lambda_i(\Delta t)^2(1/2-\delta+\theta)}{1+\theta\lambda_i(\Delta t)^2} \\ 1 & 0 \end{pmatrix}.$$

On voit alors facilement que le schéma de Newmark s'écrit

$$\begin{pmatrix} \tilde{a}_i^{n+1} \\ \tilde{a}_i^n \end{pmatrix} = A_i \begin{pmatrix} \tilde{a}_i^n \\ \tilde{a}_i^{n-1} \end{pmatrix} + \frac{(\Delta t)^2}{1+\theta\lambda_i(\Delta t)^2} \begin{pmatrix} \tilde{b}_i^n \\ 0 \end{pmatrix}$$

où $\tilde{a}^n$ contient les coordonnées de a^n dans la base considérée. Le schéma est donc stable quand $\mathrm{Sp}(A_i) \subset [-1;1]$ pour tout i. Vérifier en exercice que l'on tombe bien sur les conditions données dans l'énoncé. $\diamondsuit$

Comme pour l'équation de la chaleur, on peut étudier la convergence des schémas que nous venons de présenter. Nous renvoyons à [13] pour plus de détails.

Théorème 8.10 (Convergence)**.** *Soit Ω un ouvert régulier, $T > 0$ un temps final. On suppose que $g \in H_0^1(\Omega)$, $g' \in L^2(\Omega)$ et $f \in L^2(]0;T[, L^2(\Omega))$ sont suffisamment régulières et on note u l'unique solution faible de l'équation des ondes. On suppose aussi que V_h satisfait* (8.1) *et que $\lim_{h\to 0} \|g_h - g\|_{H_0^1(\Omega)} = \lim_{h\to 0} \|g_h' - g'\|_{L^2(\Omega)} = 0$. On note u_h^n la solution de l'équation des ondes totalement discrétisée par un schéma de Newmark, en supposant les conditions de stabilité du Lemme 8.9 vérifiées. Alors on a*

$$\lim_{\substack{\Delta t\to 0,\\ h\to 0}} \max_{0\le t_n = n\Delta t \le T} \|u_h^n - u(t_n)\|_{L^2(\Omega)} = 0.$$

Bibliographie

[1] G. Allaire, Analyse numérique et optimisation (Editions de l'Ecole Polytechnique, 2005).

[2] A. Quarteroni et A. Valli, Numerical Approximation of Partial Differential Equations, (Springer 1994).

[3] H. Brézis, Analyse fonctionnelle (Dunod, Paris, 1999).

[4] M. Chipot, Microstructures and Calculus of Variations (Monografia, Univ. di Roma "La Sapienza", 2001).

[5] R. Dautray et J.-L. Lions, Analyse mathématique et calcul numérique pour les sciences et les techniques, vol. 8 : Évolution : semi-groupe, variationnel (Masson, Paris, 1988).

[6] E.B. Davies, Spectral Theory and Differential Operators (Cambridge University Press, 1995).

[7] L.C. Evans, Partial differential equations (Graduate Studies in Mathematics, 19, American Mathematical Society, Providence, RI, 1998).

[8] P.D. Hislop et I.M. Sigal, Introduction to Spectral Theory with Application to Schrödinger Operators (Springer-Verlag, Applied Mathematical Science, 113, 1996).

[9] T. Kato, Perturbation Theory for Linear Operators (Springer-Verlag, 1976).

[10] C. Le Bris, Systèmes multi-échelles : modélisation et simulation (SMAI, Mathématiques et Applications 47, 2005).

[11] F. Legoll, Equations aux dérivées partielles et Eléments finis (Cours de première année de l'ENPC, 2017).

[12] E.H. Lieb et M. Loss, Analysis, 2nd ed. (Graduate Studies in Mathematics, vol. 14, American Mathematical Society, 2001).

[13] P.-A. Raviart et J.-M. Thomas, Introduction à l'analyse numérique des équations aux dérivées partielles (Masson, Paris, 1983).

[14] G. Stoltz, Analyse numérique et Calcul scientifique (Cours de première année de l'ENPC, 2015).

[15] M. Reed et B. Simon, Methods of Modern Mathematical Physics. I. Functional Analysis (Academic Press, 1980).