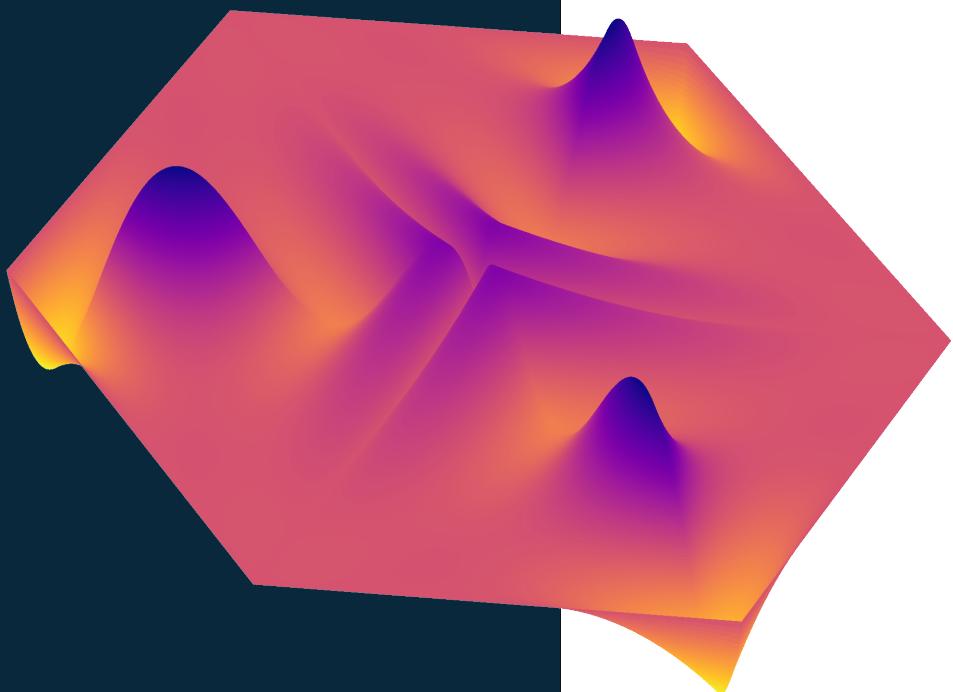


# Méthodes numériques pour les équations aux dérivées partielles

Christian Gout  
& Averil Aussedat  
**GM4**



Le programme de *Méthodes Numériques pour les Équations aux Dérivées Partielles* est riche et passionnant. L'intitulé de l'EC devrait être complété par le terme *linéaire*, car il s'agit principalement de généraliser les systèmes linéaires  $Ax = b$  en dimension infinie. Bien que ce programme s'énonce simplement, sa mise en pratique a occupé plusieurs générations de mathématiciens du XX<sup>ème</sup> siècle, qui ont bâti des bases solides sur lesquelles repose encore aujourd'hui la théorie des EDP. Ce sont ces bases que ce cours se propose d'introduire.

Le contenu est trop dense pour permettre d'être survolé. On suggère à l'étudiant·e de compléter le travail en CM et TD par la lecture régulière et indépendante du polycopié, voire des ressources bibliographiques, en s'efforçant de ne pas accumuler de lacunes. Les travaux dirigés vont rapidement s'accompagner d'implémentations : il est conseillé de prendre l'habitude de coder systématiquement "pour voir".

Ce cours s'appuie largement sur les quelques références ci-dessous. En particulier, [Bré10] est incontournable pour la théorie des équations différentielles linéaires. Certains exercices de TD sont inspirés de [Car], et les travaux pratiques héritent du travail de Sonia Fliss. N'hésitez pas à nous communiquer toute coquille ou imprécision, Christian se fera une joie de corriger le .tex.

## Bibliographie

- [All05] Grégoire Allaire. *Analyse Numérique et Optimisation: Une Introduction à La Modélisation Mathématique et à La Simulation Numérique*. Mathématiques Appliquées. École polytechnique ; Diffusion, Ellipses, Palaiseau : [Paris], 2005.
- [Bré10] Haïm Brézis. *Analyse fonctionnelle: théorie et applications*. Mathématiques appliquées pour le master. Dunod, Paris, nouvelle présentation, [nachdr.] édition, 2010.
- [Car] Pierre Cardaliaguet. Analyse fonctionnelle approfondie M1.
- [Mou22] Ayman Moussa. Approximation de fonctions et espaces de Sobolev. Notes de cours M2 Mathématiques de la Modélisation, Sorbonne Université, 2022.
- [Nic00] Serge Nicaise. *Analyse Numérique Et Equations Aux Dérivees Partielles*. Dunod, 2000.
- [RT88] Pierre-Arnaud Raviart and Jean-Marie Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques appliquées pour la maîtrise. Masson, Paris, 2. tirage édition, 1988.
- [Ste70] Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

# Table des matières

<b>Bienvenue en GM4</b>	<b>5</b>
0.1 Linéarité . . . . .	5
0.2 Séparabilité . . . . .	6
0.3 Compacité . . . . .	7
<b>I MNEDP1</b>	<b>8</b>
<b>1 Espaces de Sobolev</b>	<b>9</b>
1.1 Rappels d'analyse fonctionnelle . . . . .	9
1.2 Dérivée faible et espace $H^1$ . . . . .	12
1.3 Traitement du bord . . . . .	14
1.3.1 Régularité d'un ensemble . . . . .	14
1.3.2 Trace . . . . .	16
1.3.3 Outils subséquents . . . . .	19
1.4 Théorèmes de densité et d'injection . . . . .	20
<b>2 Problèmes linéaires bien posés</b>	<b>24</b>
2.1 Le théorème de Riesz . . . . .	24
2.2 Théorème de Lax-Milgram et Stampacchia . . . . .	24
2.3 Application à un problème elliptique . . . . .	26
<b>3 Méthode des éléments finis</b>	<b>29</b>
3.1 Principe de l'approximation de Galerkin . . . . .	29
3.2 Éléments d'analyse . . . . .	30
3.3 Éléments finis de Lagrange . . . . .	31
3.3.1 Éléments affine-équivalents . . . . .	33
3.3.2 Le cas de la dimension 1 . . . . .	35
3.3.3 Le cas de la dimension $d$ . . . . .	35
3.4 Construction de l'espace d'approximation . . . . .	41
3.5 Étude d'erreur . . . . .	43
3.5.1 Erreur d'interpolation des éléments finis de Lagrange . . . . .	43
3.5.2 Estimation d'erreur de la méthode des éléments finis d'ordre 1 . . . . .	46
3.6 Appendice: étude de l'erreur d'interpolation, cas général . . . . .	47
<b>4 Aspects pratiques</b>	<b>50</b>
4.1 Implémentation . . . . .	50
4.1.1 Assemblage des matrices . . . . .	50
4.1.2 Calcul des matrices élémentaires . . . . .	51
4.1.3 Pseudo-élimination . . . . .	52
<b>II MNEDP2</b>	<b>53</b>
<b>5 Théorie spectrale</b>	<b>55</b>
5.1 Adjoint et spectre . . . . .	55
5.2 Opérateurs compacts . . . . .	58

<b>6 Diagonalisation</b>	<b>63</b>
6.1 Décomposition spectrale des opérateurs compacts autoadjoints . . . . .	63
6.2 Problèmes spectraux . . . . .	64
6.3 Caractérisation des valeurs propres . . . . .	66
6.3.1 Quotients de Rayleigh . . . . .	66
6.3.2 [Non traité en cours] Approximation via des problèmes en dimension finie . . . . .	67
<b>7 Problèmes d'évolution</b>	<b>70</b>
7.1 Forme variationnelle . . . . .	70
7.2 Théorème de Hille-Yosida . . . . .	73
7.2.1 Définitions . . . . .	73
7.2.2 Un premier résultat . . . . .	74
7.2.3 Cas général . . . . .	76
7.3 Discrétisation en temps . . . . .	76
7.3.1 Cas parabolique . . . . .	77

# Bienvenue en GM4

Cette partie ne sera pas traitée en cours, et servira à remettre en mémoire de tout·e élève de bonne volonté les principaux outils nécessaires pour la suite. On suggère au lecteur d'aborder la suite confortablement installé, avec du temps et la boisson chaude de son choix, et en s'efforçant de tenir à distance l'angoisse de la reprise.

## 0.1 Linéarité

Commençons par quelques définitions fondamentales, mais utiles. Soit  $E$  un espace vectoriel.

**Définition 1 – Produit scalaire (réel)** Application  $\langle \cdot, \cdot \rangle : E^2 \rightarrow \mathbb{R}$  bilinéaire, symétrique et définie positive.

Tout produit scalaire induit une norme  $|\cdot|$  définie par  $|a| = \sqrt{\langle a, a \rangle}$ . L'étudiant·e venant de STPI aura déjà manipulé les produits scalaires entre fonctions, par exemple  $\langle f, g \rangle = \int_{x=0}^1 f(x)g(x)dx$ , pour calculer des coefficients de Fourier. Iel aura remarqué que la série de Fourier d'une fonction  $f$  converge *sous certaines hypothèses sur f*. L'ensemble des fonctions qui satisfont ces propriétés porte le nom d'espace de Dirichlet, et est muni d'un produit scalaire : c'est un espace pré-Hilbertien. Par contre, cet ensemble a la désagréable propriété de ne pas être fermé : on peut en prendre une suite de Cauchy qui convergera vers une fonction n'appartenant pas à l'espace de Dirichlet, mais à l'espace de Hilbert le contenant.

**Définition 2 – Espace de Hilbert** Espace vectoriel muni d'un produit scalaire complet pour la norme induit par ce produit scalaire.

Par abus de langage, on désigne parfois (toujours)  $H$  comme un espace de Hilbert quand le produit scalaire sous-jacent est évident. On se permettra d'écrire  $\langle \cdot, \cdot \rangle$ ,  $|\cdot|$  quand l'espace  $H$  est bien identifié.

### Exemples

- $(\mathbb{R}, \cdot)$  est un espace de Hilbert, où  $\cdot$  est la multiplication scalaire.
- $(L^2([0, 1]; \mathbb{R}^d), \langle \cdot, \cdot \rangle_{L^2})$  est un espace de Hilbert.
- Tout sous-ensemble vectoriel fermé d'un espace de Hilbert est de Hilbert pour le produit scalaire hérité.

### Contre-exemples

- $\mathbb{R}^2$  muni de l'application  $\langle (x_1, x_2), (y_1, y_2) \rangle \mapsto x_1 x_2$  n'est pas un Hilbert.
- $H^1([0, 1]; \mathbb{R})$ , muni du produit scalaire de  $L^2$ , n'est pas un espace de Hilbert, car pas complet.
- Tout espace de Hilbert est un espace de Banach, donc tout espace qui n'est pas de Banach n'est pas de Hilbert.
- Dans  $\mathbb{R}^d$  avec  $d > 1$ , les normes  $\left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$  pour  $p > 2$  ne sont pas induites par des produits scalaires.

Les espaces de Hilbert sont la généralisation en dimension infinie des espaces vectoriels munis d'un produit scalaire. Ils permettent de conserver la notion d'orthogonalité entre deux éléments de l'espace lui-même. De manière encore plus importante, on sait caractériser les formes linéaires d'un espace de Hilbert  $H$  (voir le théorème de Riesz à la fin de cette section).

Si l'on retire l'hypothèse que la norme est induite par un produit scalaire, on obtient un *espace de Banach*, qui est un espace vectoriel normé complet. Cette classe contient les espaces  $L^p$  pour  $p \neq 2$ , les espaces de mesures, les fonctions Lipschitz munies d'une certaine norme (laquelle ?)... Il est toujours possible, dans un espace de Banach, de considérer des applications linéaires, mais il n'est plus possible de représenter toute forme linéaire via un produit scalaire *contre un élément de l'espace* : le dual (ensemble des formes linéaires) perd la symétrie avec l'espace primal.

Si l'on retire maintenant l'hypothèse que la distance entre deux éléments est induite par une norme, ainsi que la structure vectorielle, on obtient un *espace métrique*, c'est-à-dire un espace muni d'une distance et complet. Il n'est plus question de dual, car si l'on ne peut pas additionner et multiplier deux éléments de l'espace, comment définir la linéarité ? On peut par exemple penser aux espaces d'ensembles, entre lesquels on peut mesurer une distance, ou les nuages de points si courants en statistiques.

Si l'on retire enfin l'hypothèse que la topologie (la notion de proximité) est définie via une distance, on obtient un *espace topologique*. Une topologie  $\tau$  sur un espace  $E$  est une famille d'ensembles de  $E$  que l'on définit comme étant les ouverts de  $E$ . Cette famille doit satisfaire

certaines propriétés minimales, qui généralisent celles des ouverts de la droite réelle : on demande à ce que l'ensemble vide et  $E$  appartiennent à  $\tau$ , et à ce que la famille soit stable par intersection finie (si  $(A_i)_{i \in [1, N]} \subset \tau$ , alors  $\cap_i A_i \in \tau$ ) et union quelconque (si  $(A_\lambda)_{\lambda \in \Lambda} \subset \tau$  avec  $\Lambda$  un ensemble d'indices qui peut ne pas être dénombrable, alors  $\cup_\lambda A_\lambda \in \tau$ ). Une topologie permet de donner une définition de convergence d'une suite, et rend ainsi compte de la proximité des points de l'espace. On retiendra qu'un même espace  $E$  "change de forme" si l'on en change la topologie : par exemple, considérons  $\mathbb{R}^2$  muni d'une part, de sa topologie métrique (les ouverts sont les boules ouvertes classiques pour la distance euclidienne), et de l'autre, de la topologie  $\tau$  donnée par la famille des ensembles de la forme  $\mathcal{O} \times \mathbb{R}$ , où  $\mathcal{O} \subset \mathbb{R}$  est un ouvert de  $\mathbb{R}$  au sens classique. On vérifie bien que  $\emptyset$  et  $E$  appartiennent à  $\tau$ , et les propriétés de stabilité. Par contre, la suite  $(x_n)_{n \in \mathbb{N}}$  donnée par  $x_n = (0, n)$  ne converge pas dans  $\mathbb{R}^2$  muni de sa topologie naturelle, mais converge dans  $\mathbb{R}^2$  muni de la topologie  $\tau$ . (Vérifiez-le !)

On retiendra que l'adjectif "topologique" qualifie les notions ayant trait à la continuité. Dans les espaces de Hilbert, la structure linéaire permet de définir un dual "général", qui serait l'ensemble des formes linéaires. Ce dual est qualifié *d'algébrique* pour souligner qu'il n'est défini que par les opérations algébriques d'addition et de multiplication. C'est un ensemble tellement large qu'il est difficile d'en manipuler les éléments, et on se restreindra à un sous-ensemble de formes linéaires *continues*.

**Définition 3 – Dual topologique** Soit  $E, F$  des espaces de Hilbert. On note  $\mathcal{L}(E; F)$  l'ensemble des applications de  $E$  dans  $F$  linéaires continues. On munit  $\mathcal{L}(E; F)$  de la norme

$$\|L\|_{\mathcal{L}(E; F)} := \sup_{\substack{x \in E \\ \|x\|_E = 1}} |Lx|_F.$$

On notera  $\mathcal{L}(E) = \mathcal{L}(E; E)$ .

Il n'est *a priori* pas évident que  $\|L\|_{\mathcal{L}(E; F)} < \infty$  pour tout  $L \in \mathcal{L}(E; F)$  : ce résultat est en effet difficile, et on l'admettra (le lecteur motivé pourra le déduire de [Bré10, Corollaire II.2]). Pour conclure sur la linéarité, on ne résiste pas à énoncer le plus beau théorème de l'analyse.

**Théorème 1 – de Riesz** Soit  $(H, \langle \cdot, \cdot \rangle)$  un espace de Hilbert et  $L \in \mathcal{L}(H; \mathbb{R})$ . Alors il existe un unique élément  $l \in H$  tel que

$$\langle l, x \rangle = Lx \quad \forall x \in H.$$

De plus, l'application  $L \mapsto l$  est linéaire, bijective, isométrique (i.e.  $|l| = \|L\|_{\mathcal{L}(H; \mathbb{R})}$ ) et bicontinue.

## 0.2 Séparabilité

Parmi les notions utiles héritées de l'intuition de  $\mathbb{R}$  se trouve la séparabilité. Cette propriété généralise le lien entre les rationnels  $\mathbb{Q}$  et les réels  $\mathbb{R}$  : les premiers forment un ensemble dénombrable, en bijection avec  $\mathbb{N}$ , donc *pas trop grand*, et tout de même dense dans  $\mathbb{R}$ .

**Définition 4 – Espace séparable** Soit  $(E, d)$  un espace métrique : il est séparable s'il existe une partie dénombrable dense.

Attention à ne pas confondre avec un espace séparé, dans lequel on impose que deux points distincts puissent être "séparés" par deux ouverts distincts.

### Exemples

- $\mathbb{R}$  muni de la valeur absolue : on peut prendre les rationnels, qui sont dénombrables.

### Contre-exemples

- Classiquement, les applications bornées de  $[0, 1]$  dans  $\mathbb{R}$ , muni de la norme sup, ne forment pas un espace séparable (en particulier,  $L^\infty$  n'est pas séparable, mais le quotient par la relation d'équivalence d'égalité presque partout n'est pas utile à l'argument). Plus précisément, on note  $|f|_\infty = \sup_{x \in [0, 1]} |f(x)|$ , et  $B := \{f : [0, 1] \rightarrow \mathbb{R} \mid |f|_\infty < \infty\}$ . Le lecteur s'amusera à vérifier que  $B$  est un espace vectoriel, et  $|\cdot|_\infty$  une norme sur  $B$ . Supposons par l'absurde qu'il existe une partie dénombrable dense  $(f_n)_{n \in \mathbb{N}}$  : grâce à l'absence totale de lien entre les images de  $f$  (pas de continuité ni d'intégrabilité), on peut construire une application  $f \in B$  qui soit à distance 1 de chacun des  $f_n$ . En effet, soit  $(x_n)_{n \in \mathbb{N}} \subset [0, 1]$  une suite telle que  $x_n \neq x_m$  for  $n \neq m$ . On pose

$$f(x) = \begin{cases} 0 & x \notin \{x_n \mid n \in \mathbb{N}\}, \\ 0 & x = x_n \text{ et } f_n(x_n) \notin [-1, 1], \\ 2 & x = x_n \text{ et } f_n(x_n) \in [-1, 1]. \end{cases}$$

On vérifie que  $f$  est bien bornée, et que  $|f - f_n|_\infty \geq |f(x_n) - f_n(x_n)| \geq 1$ . Donc la suite  $(f_n)_n$  n'est finalement pas dense.

La séparabilité intervient à chaque fois que l'on considère une base Hilbertienne, c'est-à-dire une famille *dénombrable* orthonormée et génératrice. Si l'espace n'était pas séparable, on ne pourrait pas en extraire une base dénombrable. Les physiciens ont l'habitude de résoudre ce problème en considérant des "bases" formelles, par exemple la famille des fonctions  $(e_x)_{x \in H}$  où  $e_x(y)$  vaut 0 si  $x \neq y$ , et 1 si  $x = y$ , qui permet de décomposer toute fonction  $f$  en  $f = \sum_{x \in H} f(x)e_x$  : les mathématiciens s'insurgent, protestent qu'il faut utiliser le dual, la théorie

des distributions de Laurent Schwartz, s'accordent qu'ils ont bien raison, et finissent enfin péniblement par démontrer les mêmes résultats avec un demi-siècle de retard. Toujours est-il que dans ce cours, on ne considèrera pratiquement que des espaces séparables.

### 0.3 Compacité

La dernière notion de ce chapitre préliminaire est, de loin, la plus importante. On propose au lecteur le jeu suivant : dessinons un carré sur une feuille. Dans ce carré, plaçons un point au hasard. Puis un autre point au hasard, qui peut coïncider avec le précédent ou non. Un troisième. Un quatrième. Et ainsi de suite : arrivé au 207<sup>ème</sup> point, le lecteur peut s'arrêter, et se contenter d'imaginer la suite infinie des futurs points. Qu'iel regarde alors le dessin ainsi formé. Quelque soit la manière dont les points ont été disposés, il existe au moins une région de l'espace où ils s'agglutinent. Plus précisément, pour tout  $\varepsilon > 0$  aussi petit que désiré, il existe un carré de côté  $\varepsilon$  contenant un nombre infini de nos points.

Cette propriété s'appelle la compacité.

La définition générale d'un ensemble compact est la suivante. Elle n'est pas (encore) en lien avec les suites de points, mais on verra que dans tout espace métrique complet, les deux points de vue coïncident.

**Définition 5 – Compact** Soit  $E$  un espace topologique. Un ensemble  $K \subset E$  est dit compact si pour tout recouvrement de  $K$  par une famille d'ouverts, il existe un sous-recouvrement fini.

En d'autres termes, si  $K$  est compact et que l'on dispose d'une famille d'ouverts  $(\mathcal{O}_\alpha)_{\alpha \in A} \subset E$  telle que  $K \subset \bigcup_\alpha \mathcal{O}_\alpha$ , alors il existe une sélection finie d'indices  $\alpha_1, \dots, \alpha_N \in A$  tels que  $K \subset \bigcup_i \mathcal{O}_{\alpha_i}$ . Une telle famille d'ouverts peut, par exemple, être la famille des boules ouvertes  $\mathcal{B}(x, \varepsilon)$ , indexées par  $x \in K$ , partageant le même  $\varepsilon > 0$ . En appliquant la définition de compacité, on montre donc qu'il existe une famille finie  $(x_1, \dots, x_N) \subset K$  telle que tout point de  $K$  est à distance  $\varepsilon$  d'un  $x_i$ .

**Remarque 1.** *Tout compact est séparable. En effet, soit  $K$  un compact d'un espace métrique  $(E, d)$ , et soit  $(\varepsilon_n)_{n \in \mathbb{N}} \subset \mathbb{R}_+^*$  une suite décroissante et convergeant vers 0. Pour chaque  $n \in \mathbb{N}$ , la famille  $(\mathcal{B}(x, \varepsilon_n))_{x \in K}$  est un recouvrement ouvert de  $K$ . Par compacité, on peut en extraire un sous-recouvrement fini, composé des boules  $\mathcal{B}(x_{n,j}, \varepsilon_n)$  pour  $j \in \llbracket 0, J_n \rrbracket$ . L'ensemble des  $x_{n,j}$  forme une partie dénombrable (par construction) et dense (car pour tout  $x \in K$  et tout  $\varepsilon > 0$ , il suffit de prendre  $n$  suffisamment grand pour que  $\varepsilon_n < \varepsilon$  pour obtenir que  $x$  se situe dans l'une des boules  $\mathcal{B}(x_{n,j}, \varepsilon_n)$ ).*

Le lien entre la définition "via les ouverts" et la définition "via les suites" (séquentielle) des compacts n'est pas trivial, et fait l'objet d'un théorème nommé. Pour l'énoncer, on considère également la définition intermédiaire suivante : un ensemble  $K$  est totalement borné si pour tout  $\varepsilon > 0$ , il existe un recouvrement fini de  $K$  par des ouverts de diamètre inférieur à  $\varepsilon$ .

**Théorème 2 – Bolzano-Weierstraß** Soit  $(E, d)$  un espace métrique complet, et  $K \subset E$  non vide. Alors les propriétés suivantes sont équivalentes :

- (a)  $K$  est compact,
- (b)  $K$  est totalement borné et fermé,
- (c) de toute suite de  $K$ , on peut extraire une sous-suite convergente dans  $K$ .

Bien que non immédiat, ce théorème est accessible au lecteur, qui peut en faire lui-même la démonstration. Pour l'encourager dans cette démarche, on lui propose un plan, ainsi que la correction en Annexe.

1. Montrer que (a) implique que  $K$  est totalement borné.
2. Par contraposée, montrer que compact implique fermé. On a ainsi (a)  $\implies$  (b).
3. Montrer par un procédé de construction que (b) implique (c).
4. Montrer que (c) implique (b).
5. En se rappelant que  $E$  est métrique, montrer que si  $K$  satisfait (b), alors  $K$  est séparable : il existe une suite  $(x_n)_{n \in \mathbb{N}}$  telle que pour tout  $y \in K$  et  $r > 0$ , il existe  $n \in \mathbb{N}$  tel que  $x_n \in \mathcal{B}(y, r)$ .
6. On dit d'un ensemble d'ouverts  $(\mathcal{O}_\alpha)_\alpha \subset E$  qu'il est une base si pour tout ouvert  $G \subset K$  et tout point  $x \in G$ , il existe  $\alpha$  tel que  $x \in \mathcal{O}_\alpha \subset G$ . Montrer qu'un espace séparable admet une base dénombrable.
7. Montrer alors que si  $K$  est totalement borné dans un espace métrique, tout recouvrement ouvert admet un sous-recouvrement dénombrable. (Les espaces possédant cette propriété sont dits de Lindelöf.)
8. En déduire que dans un espace métrique, (c) implique (a).

Arrivé au terme de ce chapitre, le lecteur (ou la lectrice, qui nous aura pardonné les multiples appels au genre "neutre"... ) est armé·e pour MNEDP. Soyez intransigeant·e·s avec vous-mêmes et vos professeur·e·s pour ne pas subir cette année sans en récolter ni compréhension durable, ni plaisir de l'étude. Les mathématiciens qui ont exhumé ce qui va suivre se sont, eux, beaucoup amusés, et seraient bien tristes de voir leur précieux et fragile butin, arraché avec peine à l'inconnu, se convertir en dogme indépassable !

## **Partie I**

### **MNEDP1**

# Chapitre 1

## Espaces de Sobolev

Les équations différentielles ont une fâcheuse tendance à admettre des solutions d'une régularité moindre que celle nécessaire à leur définition. Par exemple, si  $f$  n'est pas suffisamment lisse, l'équation de Poisson  $-u'' = f$  peut tout à fait admettre une solution  $u(\cdot)$  dont la dérivée seconde ne soit pas définie. À première vue, cet énoncé est absurde : comment dire que  $u$  est solution s'il n'est pas possible de le vérifier ? La réponse tient en la définition de ce que l'on appelle *solution*. Une fonction  $u(\cdot)$  qui n'est pas deux fois différentiable pourra tout de même être *solution faible*, en un sens à définir, de l'équation de Poisson. Pour introduire ce sens faible, on commence par préciser où se trouve  $u(\cdot)$ .

### 1.1 Rappels d'analyse fonctionnelle

Notons  $\Omega \subset \mathbb{R}^d$  un ensemble.

**Définition 6 – Espaces  $L^p$  pour  $p \in [1, \infty[$**  Soit  $p \in [1, \infty[$ . Notons  $\text{Pre}L^p(\Omega; \mathbb{R}^n)$  la classe des fonctions Lebesgue-mesurables et  $p$ -intégrables, au sens où

$$\|f\|_p^p := \int_{x \in \Omega} |f(x)|^p dx < \infty.$$

Deux fonctions  $f, g \in \text{Pre}L^p(\Omega; \mathbb{R}^n)$  sont dites équivalentes presque partout, ou égales presque partout, si  $\|f - g\|_p = 0$ , ce que l'on note  $f \sim g$ . On définit  $L^p(\Omega; \mathbb{R}^n)$  comme le quotient de  $\text{Pre}L^p(\Omega; \mathbb{R}^n)$  par la relation d'équivalence  $\sim$  : les éléments de  $L^p(\Omega; \mathbb{R}^n)$  sont les *classes d'équivalence* pour la relation  $\sim$ . On munit  $L^p$  de la norme  $\|\cdot\|_p$ .

Cette définition est technique au premier abord, mais permet d'éviter les situations pathologiques. Une fonction de  $L^p$  se doit d'être mesurable pour que l'on puisse l'intégrer en utilisant la théorie de l'intégrale de Lebesgue. Le critère de finitude de l'application  $\|\cdot\|_p$  forme le cœur de la définition. Le quotient par les classes d'équivalence est nécessaire pour que  $\|\cdot\|_p$  soit une norme : en effet, si  $f: \mathbb{R} \rightarrow \mathbb{R}$  sont définies par  $f \equiv 0$  et  $g(x) = 1$  si  $x = 0$ , et 0 sinon, on a bien  $f, g \in \text{Pre}L^p(\Omega; \mathbb{R}^n)$ , mais  $\|f - g\|_p = 0$ . En passant au quotient, ces deux fonctions deviennent deux représentants de la même classe d'équivalence, et sont toutes les deux notées 0 dans  $L^p$ .

**Remarque 2** (Autre mesure). *L'intérêt du passage au quotient est plus flagrant si l'on considère les espaces  $L_\mu^p$ , où  $\mu$  est une mesure qui remplace la mesure de Lebesgue. Par exemple, si  $\mu = \delta_0$  est la masse de Dirac en 0, l'espace  $\text{Pre}L_\mu^p(\mathbb{R}; \mathbb{R})$  devient l'espace des applications Borel-mesurables, qui satisfont toutes  $\int_{x \in \Omega} |f(x)|^p d\mu(x) = |f(0)|^p < \infty$ . Deux applications  $f, g \in \text{Pre}L_\mu^p(\mathbb{R}; \mathbb{R})$  satisferont  $f \sim_\mu g$  si  $f(0) = g(0)$ , ce que l'on désigne par  $f = g$   $\mu$ -presque partout. Ainsi, le passage au quotient permet de manipuler les classes d'équivalence pour  $\sim_\mu$ , et chaque membre de l'espace  $L_\mu^p(\mathbb{R}; \mathbb{R})$  correspond à une valeur possible en  $x = 0$ .*

**Définition 7 – Espace  $L^\infty$**  Soit  $f: \Omega \rightarrow \mathbb{R}^n$  une fonction Lebesgue-mesurable. En particulier, pour tout  $B \geq 0$ , l'ensemble  $\{|f| > B\}$  est Lebesgue-mesurable. On dit que  $B$  est une borne essentielle de  $f$  si  $|f| \leq B$  presque partout, i.e.  $\int_{x \in \Omega} \mathbb{1}_{\{|f(x)| > B\}} dx = 0$ . On définit alors  $\text{Pre}L^\infty(\Omega; \mathbb{R}^n)$  comme la classe des fonctions Lebesgue-mesurables qui

admettent un supremum essentiel, i.e. telles que

$$\|f\|_\infty := \inf \left\{ B \geq 0 \mid \int_{x \in \Omega} \mathbb{I}_{|f(x)| > B} dx = 0 \right\} < \infty.$$

Notons  $f \sim g$  si  $\|f - g\|_\infty = 0$ . On définit  $L^\infty(\Omega; \mathbb{R}^n)$  comme le quotient de  $\text{Pre}L^\infty(\Omega; \mathbb{R}^n)$  par la relation d'équivalence  $\sim$ , muni de la norme  $\|\cdot\|_\infty$ .

L'espace  $L^2(\Omega; \mathbb{R}^n)$  est un espace de Hilbert (espace vectoriel, muni d'un produit scalaire, complet pour la norme induite par ce produit scalaire) quand muni du produit scalaire

$$\langle f, g \rangle_{L^2} := \int_{x \in \Omega} \langle f(x), g(x) \rangle dx.$$

Les autres espaces  $L^p$  ne jouissent pas de cet avantage, bien qu'ils puissent être mis en dualité (ainsi que vu en cours d'Analyse fonctionnelle). Pour cette raison, on se concentrera dans ce cours sur l'espace  $L^2 = L^2(\Omega; \mathbb{R})$ . Notre but est d'étudier des équations différentielles, et pour ceci, il va nous falloir une notion de différentiabilité. Pour les équations linéaires, la notion la plus générale est celle de dérivée au sens des distributions. Pour les applications qui nous intéressent, c'est une notion trop faible, mais qui peut au moins servir de cadre pour introduire les bons espaces.

On définit le support d'une fonction  $\varphi : \Omega \rightarrow \mathbb{R}^m$  comme

$$\text{supp } \varphi := \overline{\{x \in \Omega \mid |\varphi(x)| > 0\}}.$$

**Définition 8 – Fonctions test** On note

$$\mathcal{D}(\Omega) := \{\varphi \in \mathcal{C}^\infty(\Omega, \mathbb{R}) \mid \text{supp } \varphi \text{ est compact}\}.$$

La condition de support compact permet deux choses : premièrement, comme  $\varphi$  est  $\mathcal{C}^\infty$ , sa valeur et la valeur de toutes ses dérivées sera bornée. Deuxièmement, comme  $\Omega$  est ouvert, le produit de  $\varphi$  et d'une fonction  $f$  quelconque s'annulera dans un voisinage ouvert de  $\partial\Omega$ , ce qui éliminera tout terme de bord dans les formules d'intégration par partie. On va considérer une forme de convergence particulière pour les fonctions tests :

**Définition 9 – Convergence dans  $\mathcal{D}(\Omega)$**  On dira que  $(\varphi^n)_n \subset \mathcal{D}(\Omega)$  converge vers  $\varphi \in \mathcal{D}(\Omega)$ , et on notera  $\varphi^n \xrightarrow{\mathcal{D}} \varphi$  si

- le support des fonctions  $\varphi^n$  reste dans un compact  $K \subset \mathbb{R}^d$  indépendant de  $n$ ,
- $(\varphi^n)_n$  converge vers  $\varphi$  uniformément sur  $K$ , et chaque suite des dérivées partielles (à n'importe quel ordre, pour n'importe quelle combinaison de variables) de  $\varphi^n$  converge uniformément vers la dérivée correspondante de  $\varphi$  uniformément sur  $K$ .

Ces fonctions très régulières permettent de définir les distributions comme des objets très irréguliers.

**Définition 10 – Distribution** L'ensemble des distributions est le dual topologique de  $\mathcal{D}(\Omega)$ , c'est-à-dire l'ensemble des formes linéaires continues sur  $\mathcal{D}(\Omega)$ . On note cet espace  $\mathcal{D}'(\Omega)$ .

Ici, la continuité d'une forme linéaire  $F \in \mathcal{D}'(\Omega)$  est entendue au sens suivant : pour chaque suite  $(\varphi^n)_n \subset \mathcal{D}(\Omega)$  qui converge vers  $\varphi \in \mathcal{D}(\Omega)$  au sens de la Définition 9, la suite  $(F(\varphi^n))_n$  converge vers  $F(\varphi)$  dans  $\mathbb{R}$ .

### Exemples

- À partir de toute fonction  $f \in L^2$ , on peut former la distribution  $F \in \mathcal{D}'(\Omega)$  par la définition

$$F : \varphi \in \mathcal{D}(\Omega) \rightarrow \mathbb{R}, \quad F(\varphi) := \int_{x \in \Omega} f(x)\varphi(x)dx.$$

Chaque  $\varphi \in \mathcal{D}(\Omega)$  est dans  $L^q$  pour tout  $q \in ]1, \infty[$ . Ainsi, en utilisant Cauchy-Schwarz avec  $1/p + 1/q = 1$ ,

$$\int_{x \in \Omega} |f(x)\varphi(x)| dx \leq \left( \int_{x \in \Omega} |f(x)|^p dx \right)^{1/p} \left( \int_{x \in \text{supp } \varphi} |\varphi(x)|^q dx \right)^{1/q} = \|f\|_p \|\varphi\|_q < \infty$$

et  $F$  est à valeurs réelles partout. C'est bien une fonction linéaire, et comme la convergence dans  $\mathcal{D}(\Omega)$  implique la convergence dans  $L^q$  (aidez-vous du support compact),  $F$  est bien continue. Cette "représentation" de fonctions dans l'espace des distributions est formalisée par l'application  $f \mapsto F$ , qui injecte  $L^2$  dans  $\mathcal{D}'(\Omega)$ .

- La plus connue des distributions est la masse de Dirac  $\delta_0$ , qui, à une fonction test  $\varphi$ , associe sa valeur  $\varphi(0)$ . On laisse le lecteur vérifier que  $\delta_0$  définit bien une distribution. Cet exemple se généralise à la distribution  $\varphi \mapsto \partial_x^3 \varphi(0)$  l'évaluation de la dérivée 3e, aux applications intégrales...

### Contre-exemples

- L'application  $\varphi \mapsto \varphi^2(0)$  n'est pas une distribution, car n'est pas linéaire.
- Soit  $\Omega = \mathbb{R}$ , et notons  $\partial_x^n \varphi(z)$  l'évaluation en  $z$  de la  $n^{\text{ième}}$  dérivée partielle de  $\varphi$ . L'application

$$F : \varphi \mapsto \sum_{m=0}^{\infty} \frac{\partial_x^m \varphi(1)}{m!} (0-1)^m$$

n'est pas une distribution : c'est bien une fonction linéaire, qui admet une valeur finie en chaque  $\varphi \in \mathcal{D}(\Omega)$ . Cependant, soit

$$g : \mathbb{R} \rightarrow \mathbb{R}, \quad g(x) = \begin{cases} e^{-\frac{1}{(1-x^2)^2}} & x \in ]-1, 1[, \\ 0 & \text{sinon.} \end{cases}$$

La fonction  $g$  est bien  $C^\infty$ , et son support est  $[-1, 1]$ . On pose  $g_n(x) = g(x-1/n)$ , et on vérifie que  $g_n \xrightarrow{\mathcal{D}} g$ . Pour tout  $n \in \mathbb{N}_*$ , la fonction  $g^n$  est analytique sur  $]-\frac{1}{2}, 1 + \frac{1}{2n}[$ , donc la fonction  $F(g_n)$  coïncide avec l'évaluation en 0 du développement de Taylor de  $g_n$  autour du point 1. On a donc  $F(g_n) = g_n(0)$  pour tout  $n$ , et  $F(g_n) \xrightarrow{n} g(0) > 0$ . Cependant,  $F(g) = 0$ , donc l'application  $F$  n'est pas continue.

On définit ainsi la dérivée au sens des distributions dans le but de préserver la validité de l'intégration par partie.

**Définition 11 – Dérivée au sens des distributions** Soit  $f \in \mathcal{D}'(\Omega)$ . La dérivée de  $f$  est la distribution  $g \in \mathcal{D}'(\Omega)$  définie par

$$\forall \varphi \in \mathcal{D}(\Omega), \quad g(\varphi) := -f(\partial_x \varphi).$$

On définit de manière analogue les dérivées partielles et d'ordre supérieur.

### Exemples

- Toute fonction  $f \in \mathcal{C}^1(\Omega; \mathbb{R})$  définit une distribution  $F \in \mathcal{D}'(\Omega)$  par  $F(\varphi) = \int_{x \in \Omega} f(x)\varphi(x)dx$ . On vérifie alors que la dérivée au sens des distributions coïncide avec la dérivée classique.
- La dérivée au sens des distributions de la fonction de Heaviside  $\mathbb{1}_{\mathbb{R}^+}$  est la masse de Dirac  $\delta_0$ .

Les distributions sont infiniment dérивables, ce qui constitue le moteur principal de leur introduction. Pour terminer cette section, il convient de clarifier un élément de langage important : on dit qu'une distribution  $F \in \mathcal{D}'(\Omega)$  est dans  $L^2$  s'il existe une fonction  $f \in L^2$  telle que  $F(\varphi) = \langle f, \varphi \rangle_{L^2}$ . La masse de Dirac  $\delta_0$  est un contre-exemple à garder en tête : c'est une distribution, mais s'il existait  $f \in L^2$  telle que  $\langle f, \varphi \rangle_{L^2} = \varphi(0)$  pour tout  $\varphi \in \mathcal{D}(\Omega)$ , alors  $\langle f, \varphi \rangle_{L^2} = 0$  pour tout  $\varphi \in \mathcal{D}(\Omega)$  telle que  $\varphi(0) = 0$ . On montre alors que nécessairement,  $f \equiv 0$ , ce qui est absurde. En résumé, la théorie de la dérivation au sens des distributions est un cadre très (trop) large pour les équations linéaires, et l'on va utiliser la définition plus exigeante de *dérivée faible*.

## 1.2 Dérivée faible et espace $H^1$

Toute fonction de  $L^2(\Omega; \mathbb{R})$  définit une distribution, donc admet des dérivées partielles  $\partial_{x_i} f \in \mathcal{D}'(\Omega)$  au sens de la Définition 11.

**Définition 12 – Dérivée faible** Soit  $\Omega \subset \mathbb{R}^d$  un ouvert. Une fonction  $u \in L^2(\Omega; \mathbb{R}^d)$  est dite *dérivable au sens faible* s'il existe  $d$  fonctions mesurables localement essentiellement bornées  $v_i$ ,  $i \in \llbracket 1, d \rrbracket$ , telles que

$$\int_{x \in \Omega} u(x) \partial_{x_i} \varphi(x) dx = - \int_{x \in \Omega} v_i(x) \varphi(x) dx \quad \forall \varphi \in \mathcal{D}(\Omega). \quad (1.1)$$

### Exemples

- L'égalité (1.1) est une généralisation de la formule d'intégration par partie, avec des termes de bord nuls grâce au choix des fonctions  $\varphi$  à support compact. Donc en particulier, toutes les fonctions  $u \in \mathcal{C}^1$  sont dérивables au sens faible, et les dérivées faibles sont les (classes d'équivalence dans  $L^2(\Omega; \mathbb{R})$  des) dérivées partielles classiques.
- Pour toutes fonctions  $u, v_i \in L^2(\Omega; \mathbb{R})$ , les applications

$$f, g : \mathcal{D}(\Omega) \rightarrow \mathbb{R}, \quad f(\varphi) := \int_{x \in \Omega} u(x) \varphi(x) dx, \quad g_i(\varphi) := \int_{x \in \Omega} v_i(x) \varphi(x) dx$$

définissent des distributions. L'égalité (1.1) s'écrit alors  $g_i(\varphi) = -f(\partial_{x_i} \varphi)$ , et en particulier,  $g_i$  est la dérivée (partielle) au sens des distributions de  $f$ . Par contre, la dérivée faible contient plus d'informations :  $f$  et  $g_i$  s'écrivent comme des produits scalaires avec des fonctions de  $L^2$ .

### Contre-exemples

- Les distributions ne sont pas toutes dérивables au sens faible : la fonction de Heaviside  $\mathbb{1}_{\mathbb{R}^+}$  appartient à  $L^2$ , et définit une distribution dont la dérivée (au sens des distributions) est la masse de Dirac, mais l'application  $\varphi \mapsto \langle \delta_0, \varphi \rangle = \varphi(0)$  ne s'écrit pas comme le produit scalaire de  $\varphi$  avec une fonction mesurable.

On s'intéressera en particulier à une classe de fonctions qui admettent une dérivée faible un peu plus régulière que "localement essentiellement bornée".

**Définition 13 – Espace  $H^1(\Omega; \mathbb{R}^n)$**  On note  $W^{1,p}$  l'espace de Sobolev

$$W^{1,p}(\Omega; \mathbb{R}^n) := \{f \in L^p(\Omega; \mathbb{R}^n) \mid \forall i \in \llbracket 1, d \rrbracket, \quad \partial_{x_i} f \in L^p(\Omega; \mathbb{R}^n)\},$$

c'est-à-dire l'espace des éléments de  $L^p$  dont les dérivées faibles appartiennent à  $L^p$ . On munit  $W^{1,p}$  de la norme

$$\|f\|_{W^{1,p}}^p := \|f\|_{L^p} + \sum_{i=1}^d \|\partial_{x_i} f\|_{L^p}.$$

On note  $H^1(\Omega; \mathbb{R}^n) := W^{1,2}(\Omega; \mathbb{R}^n)$ , qui est muni du produit scalaire

$$\langle f, g \rangle_{H^1} := \langle f, g \rangle_{L^2} + \sum_{i=1}^d \langle \partial_{x_i} f, \partial_{x_i} g \rangle_{L^2}.$$

L'espace  $H^1$  peut se comprendre d'une autre manière : on rappelle que la dérivée au sens des distributions d'une fonction  $f \in L^2$  existe, et est une forme linéaire. Si cette forme linéaire est *continue* par rapport à la norme  $L^2$ , alors par le théorème de Riesz, elle peut être représentée via un produit scalaire contre un élément de  $L^2$ , que l'on appellera le gradient de  $f$ . Ainsi, l'espace  $H^1$  revient à imposer une condition de continuité de la première dérivée par rapport à la norme  $L^2$ .

### Exemples

- Toute fonction  $\mathcal{C}^\infty$  à support compact appartient à  $H^1$ .
- La fonction  $f : x \mapsto |x|$  appartient à  $H^1([-1, 1]; \mathbb{R})$ . En effet, c'est une fonction de  $L^2([-1, 1]; \mathbb{R})$ . De plus, elle admet une dérivée faible localement bornée, donnée par  $f' : x \mapsto \text{signe}(x)$  (vérifiez-le) ! Comme  $\int_{x \in [-1, 1]} |f'(x)|^2 dx = 2 < \infty$ , (la classe

d'équivalence de) cette fonction appartient bien à  $L^2$ , et  $f \in H^1([-1,1];\mathbb{R})$ . Ici, la dérivée de  $f$  n'est pas définie en 0 !

#### Contre-exemples

- Une fonction discontinue ne peut pas appartenir à  $H^1$  : en effet, sa dérivée au sens des distributions contiendra une masse de Dirac au point de discontinuité, et ne peut pas être représentée via un produit scalaire contre un élément de  $L^2$ .

Plus généralement, on s'intéresse aux espaces de Hilbert d'ordre plus élevé et à leur semi-norme. Introduisons une notation compacte pour les dérivées partielles : pour chaque  $\alpha \in \llbracket 1, k \rrbracket^d$  un multi-indice, on note

$$\partial^\alpha u(x) = \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_k}^{\alpha_k} u(x).$$

Par exemple, si  $d = 2$  et  $k = 3$ ,  $\partial^{(1,2,1)} u$  est la dérivée partielle d'ordre 4  $\partial_{x_1} \partial_{x_2}^2 \partial_{x_1} u$ . Par convention, si  $\alpha = (0, \dots, 0)$ , on prend  $\partial^\alpha u(x) = u(x)$ .

**Définition 14 – Espace  $H^m$**  Soit  $m \in \mathbb{N}_*$ . L'espace  $H^m$  est l'ensemble des fonctions mesurables telles que pour tout multi-indice  $\alpha \in \llbracket 1, m \rrbracket^d$  avec  $\sum_{i=1}^d \alpha_i \leq m$ , la dérivée partielle  $\partial^\alpha u$  au sens des distributions soit représentable par un élément de  $L^2(\Omega)$ . On munit  $H^m$  de la norme

$$\|u\|_{H^m(\Omega;\mathbb{R})}^2 := \sum_{\alpha \in \llbracket 1, m \rrbracket^d} \|\partial^\alpha u\|_{L^2(\Omega;\mathbb{R})}^2,$$

et de la semi-norme

$$|u|_{\text{semi}, H^m}^2 := \sum_{\alpha \in \llbracket 1, m \rrbracket^d, \sum_{i=1}^d \alpha_i = m} \|\partial^\alpha u\|_{L^2(\Omega;\mathbb{R})}^2.$$

Sur le même modèle, on définit l'espace de Sobolev  $W^{m,p}$  comme l'ensemble des fonctions mesurables dont chaque dérivée partielle jusqu'à l'ordre  $m$  est représentable comme une fonction de  $L^p$ . Les espaces de fonctions régulières, et leurs parents de Sobolev, forment une famille très unie, liée par pléthore de théorèmes d'inclusion, d'injection, d'interpolation... Pour premier exemple, on a  $H^{m+1} \subset H^m$ . (Si ce n'est pas évident, prendre le temps de relire les définitions pour bien comprendre pourquoi.) Pour second exemple, on a le (très utile) résultat suivant.

**Lemme 1 – Densité des fonctions lisses à support compact** L'espace  $\mathcal{D}(\mathbb{R}^d)$  des fonctions de classe  $C^\infty$  à support compact est dense dans  $H^1(\mathbb{R}^d; \mathbb{R})$  au sens de la norme  $H^1$ . Par contre, si  $\Omega \subset \mathbb{R}^d$  est un ouvert borné, alors il se peut que  $\mathcal{D}(\Omega)$  ne soit pas dense dans  $H^1(\bar{\Omega}; \mathbb{R})$ .

#### Démonstration (Idée)

Premièrement, toute fonction de  $\mathcal{D}(\mathbb{R}^d)$  appartient à  $H^1$ . Pour montrer l'injection, il suffit de prouver que toute fonction  $u \in H^1(\mathbb{R}^d; \mathbb{R})$  est limite (au sens de la norme  $H^1$ ) d'une suite de fonctions de  $\mathcal{D}$ . Pour ceci, on procède en plusieurs étapes classiques.

- Troncature : on se ramène à un support compact en multipliant par une fonction  $\varphi^n \in C_c^\infty(\mathbb{R}^d; \mathbb{R})$ , avec pour propriété que  $\varphi^n(x) \rightarrow_n 1$  pour tout  $x$  et  $\|\nabla \varphi^n\|_{L^2} \rightarrow_n 0$ . Dans ce cas, la suite  $v^n := \varphi^n \times u$  appartient à  $H^1$ , et  $\|u - v_n\|_{H^1} \rightarrow_n 0$ .
- Régularisation : on approche chaque  $v^n$  par une fonction  $C^\infty$  à support compact. Pour ceci, on peut considérer une suite  $w^{n,m}(x)$  construite en intégrant une convolution de  $\nabla v^n$  par un noyau  $K^m$  lui-même  $C^\infty$  et à support compact. En faisant tendre le noyau  $K^m$  vers la masse de Dirac quand  $m \rightarrow \infty$ , les convolutions  $\nabla v^n * K^m$  tendront presque partout vers  $\nabla v^n$  : donc  $w^{n,m}$  tendra vers  $v^n$  dans  $H^1$  quand  $m \rightarrow \infty$ .
- Extraction : on choisit une suite diagonale de  $(w^{n,m})_{n,m}$ , qui fournira l'approximation désirée.

Chaque étape est relativement intuitive, mais cache des détails techniques pointus. Pour le cas borné, il suffit de considérer  $\Omega = ]0, 1[$  et la fonction  $u : x \mapsto x$ , qui appartient bien à  $H^1$ . Supposons qu'il existe une suite  $(\varphi_n)_{n \in \mathbb{N}} \subset \mathcal{D}(]0, 1[)$  qui tende vers  $u$  en norme  $H^1$ . En particulier, on a

$$u(x) - \varphi_n(x) = u(0) - \varphi_n(0) + \int_{y=0}^x (u'(y) - \varphi'_n(y)) dy \leq 0 + \left( \int_{y=0}^x 1 dy \right)^{1/2} \left( \int_{y=0}^x (u'(y) - \varphi'_n(y))^2 dy \right)^{1/2} \leq \|u - \varphi_n\|_{H^1} \xrightarrow{n \rightarrow \infty} 0,$$

et en inversant le rôle de  $u$  et  $\varphi_n$ , on obtient que  $(\varphi_n)_n$  converge uniformément vers  $u$ . Ainsi, pour tout  $\varepsilon > 0$ , il existe  $n_\varepsilon$  suffisamment grand tel que  $\varphi_n(x) \geq 1/2$  pour tout  $x \in [1 - \varepsilon, 1]$  et  $n \geq n_\varepsilon$ . Dès lors, par l'inégalité de Cauchy-Schwarz en sens

inverse, puis en utilisant le fait que  $\varphi_n(1) = 0$  car  $\varphi_n$  est à support compact dans  $]0, 1[$ ,

$$\|\varphi'_n - u'\|_{L^2} = \sqrt{\int_{x=0}^1 |\varphi'_n(x) - u'|^2 dx} \geq \int_{x=0}^1 |\varphi'_n(x) - 1| dx \geq \int_{x=1-\varepsilon}^1 |\varphi'_n(x) - 1| dx \geq \int_{x=1-\varepsilon}^1 \varphi'_n(x) - 1 dx = \varphi_n(1) - \varphi_n(1 - \varepsilon) - \varepsilon \geq \frac{1}{2} - \varepsilon.$$

Donc  $\varphi_n$  ne tend pas vers  $u$ . (On conseille au lecteur de se munir d'un dessin.)  $\square$

Le problème rencontré dans le cas d'un domaine borné vient du fait que les fonctions test s'annulent au bord, mais que les fonctions de  $H^1$  ne sont pas nécessairement nulles. On reverra plus en détail l'adhérence de  $\mathcal{D}(\Omega)$  dans  $H^1$  à la section 1.3.2.

## 1.3 Traitement du bord

Les équations aux dérivées partielles caractérisent un ensemble de solutions à l'aide de deux types de conditions : *l'équation* en elle-même, qui impose une relation entre la valeur et les dérivées de la fonction à l'intérieur du domaine, et *les conditions au bord*, qui fixent la valeur de la solution. Le sens des conditions au bord est une question au moins aussi délicate que le sens de l'équation elle-même, et qu'il ne faut pas prendre à la légère. Il est fréquent que les valeurs "admissibles" au bord doivent satisfaire des conditions de compatibilité assez fines.

Soit  $\Omega$  un domaine borné  $C^1$  par morceaux. Introduisons des notations générales : soient  $\mathcal{H} : H^1(\Omega; \mathbb{R}) \rightarrow L^2(\Omega; \mathbb{R})$  et  $\Psi : H^1(\Omega; \mathbb{R}) \rightarrow L^2(\partial\Omega; \mathbb{R})$  deux opérateurs linéaires. On s'intéresse au *problème aux limites*

$$\begin{cases} \mathcal{H}(u)(x) = f(x) \\ \Psi(u)(x) = g(x) \end{cases} \quad \begin{aligned} &x \in \Omega, \\ &x \in \partial\Omega. \end{aligned} \quad (1.2a)$$

$$(1.2b)$$

Notre but est de "résoudre" (1.2), c'est-à-dire trouver des classes de fonctions  $f, g$  telles qu'il existe une unique solution  $u$  dans un sens à préciser.

Dans ce cours, on s'intéressera à deux types d'opérateur  $\Psi$  :

- une condition de type Dirichlet, de la forme  $\Psi(u)(x) = u(x)$ ,
- une condition de type Neumann, de la forme  $\Psi(u)(x) = \partial_\eta u(x)$  la dérivée normale à la frontière.

On se retrouve alors face à deux sous-problèmes. Premièrement, il faut donner un sens à  $\Psi(u)(x)$  pour une fonction de  $H^1$  : on va voir que la définition de trace (pour le problème de Dirichlet) ou de trace normale (pour le problème de Neumann) dépend de la régularité de la frontière. Dans cette première phase, on considère le problème homogène : selon la définition de [RT88, §2.3], un problème aux limites est homogène si l'ensemble des fonctions qui vérifient la condition au bord forment un espace vectoriel. En pratique, cela revient à choisir  $g \equiv 0$ .

Deuxièmement, il faut déterminer quelles fonctions  $g$  sont suffisamment régulières pour donner un sens à (1.2b). Ce problème se traite usuellement grâce à un relèvement : on détermine d'abord une classe de fonctions  $g$  telles qu'il existe une fonction  $u_0 \in H^1(\Omega; \mathbb{R})$  vérifiant  $\Psi(u_0)(x) = g(x)$  au sens choisi précédemment. On exploite ensuite la linéarité de  $\Psi$  et  $\mathcal{H}$  pour introduire  $v = u - u_0$ , qui vérifie

$$\begin{cases} \mathcal{H}(v)(x) = f(x) - \mathcal{H}(u_0)(x) \\ \Psi(v)(x) = 0 \end{cases} \quad \begin{aligned} &x \in \Omega, \\ &x \in \partial\Omega. \end{aligned} \quad (1.3a)$$

$$(1.3b)$$

Ce nouveau problème aux limites est homogène, ce qui nous ramène au cas précédent. Pour être applicable, cette méthode exige de s'assurer que  $\mathcal{H}(u_0)$  a au moins la régularité de  $f$ , ce qui n'est pas trivial. Dans la suite de cette section, on se concentre sur le premier problème : donner une définition de trace.

### 1.3.1 Régularité d'un ensemble

Soit  $\Omega \subset \mathbb{R}^d$  un ouvert borné. On notera  $\Omega^c$  son complémentaire,  $\overline{\Omega} \subset \mathbb{R}^d$  son adhérence, et  $\partial\Omega = \overline{\Omega} \setminus \Omega$  sa frontière.

**Définition 15 – Ouvert borné régulier** Dénotons  $B = \mathbb{R}^{d-1} \times ]0, \infty[ = \{(x, x_d) \in \mathbb{R}^d \mid x_d > 0\}$  le demi-plan ouvert, et  $\partial B = \mathbb{R}^{d-1} \times \{0\}$  sa frontière. On dit que  $\Omega$  est de classe  $C^1$  s'il existe une famille finie d'ouverts bornés  $(\mathcal{O}_i)_{i \in \llbracket 1, I \rrbracket}$  tels que  $\overline{\Omega} \subset \bigcup_{i \in \llbracket 1, I \rrbracket} \mathcal{O}_i$ , et pour chaque  $i \in \llbracket 1, I \rrbracket$ , il existe une application  $\varphi_i : \mathcal{O}_i \rightarrow \mathbb{R}^d$  telle que

- $\varphi_i$  est injective, de classe  $C^1$  et  $\varphi_i^{-1} : \varphi(\mathcal{O}_i) \rightarrow \mathcal{O}_i$  est de classe  $C^1$ ,

- $\Omega \cap \mathcal{O}_i = \varphi_i^{-1}(\varphi(\mathcal{O}_i) \cap B)$ ,
- $\partial\Omega \cap \mathcal{O}_i = \varphi_i^{-1}(\varphi(\mathcal{O}_i) \cap \partial B)$ .

De manière analogue, on définit les ouverts de classe  $\mathcal{C}^m$  en imposant que  $\varphi_i$  et  $\varphi_i^{-1}$  soient de classe  $\mathcal{C}^m$ . On définit les ouverts réguliers à frontière  $\mathcal{C}^1$  par morceaux en permettant de choisir  $B$  parmi un demi-plan, un cône ouvert ou le complémentaire d'un cône fermé de  $\mathbb{R}^d$ . La famille des ouverts bornés  $\mathcal{C}^1$  par morceaux permet de travailler avec une large gamme de formes "simples" qui peuvent présenter des coins, tels que le carré, les lettres de l'alphabet ou un plan d'appartement. La définition permet d'éliminer les frontières trop irrégulières, comme les formes fractales, et les coins "dégénérés", comme l'union de deux disques tangents.

### Exemples

- En dimension  $d=1$ , les ouverts de classe  $\mathcal{C}^1$  sont les unions finies d'intervalles ouverts.
- En dimension  $d=2$ , la boule unité  $\mathcal{B}(0,1)$  est de classe  $\mathcal{C}^1$ . On peut prendre pour ouverts les  $(\mathcal{O}_i)_{i \in [0,4]}$ , où  $\mathcal{O}_0 = \mathcal{B}(0, \sqrt{3}/2)$  et

$$\mathcal{O}_1 = \mathcal{O}_0^c \cup \left[ -\frac{1}{2}, 2 \right]^2, \quad \mathcal{O}_2 = \mathcal{O}_0^c \cup \left[ -2, \frac{1}{2} \right] \times \left[ -\frac{1}{2}, 2 \right], \quad \mathcal{O}_3 = \mathcal{O}_0^c \cup \left[ -2, \frac{1}{2} \right]^2, \quad \mathcal{O}_4 = \mathcal{O}_0^c \cup \left[ -\frac{1}{2}, 2 \right] \times \left[ -2, \frac{1}{2} \right].$$

### Les applications

$$\varphi_0(x, y) = \begin{pmatrix} x \\ y+1 \end{pmatrix}, \quad \varphi_i(x, y) = \begin{pmatrix} \text{atan2}(y, x) \\ 1 - (x^2 + y^2) \end{pmatrix} \quad i \in \{1, 2, 3\}, \quad \varphi_i(x, y) = \begin{pmatrix} \text{atan2}(y, -x) \\ 1 - (x^2 + y^2) \end{pmatrix} \quad i \in \{2, 3\}$$

vérifient les conditions de la définition.

- En dimension  $d=2$ , le carré  $\{(x, y) \mid \max(|x|, |y|) < 1\}$  est de classe  $\mathcal{C}^1$  par morceaux. Soit  $B = ]0, \infty[^2$  le cône ouvert positif d'angle  $\pi/2$ . On peut prendre les ouverts  $(\mathcal{O}_i)_{i \in [1,4]}$  définis par

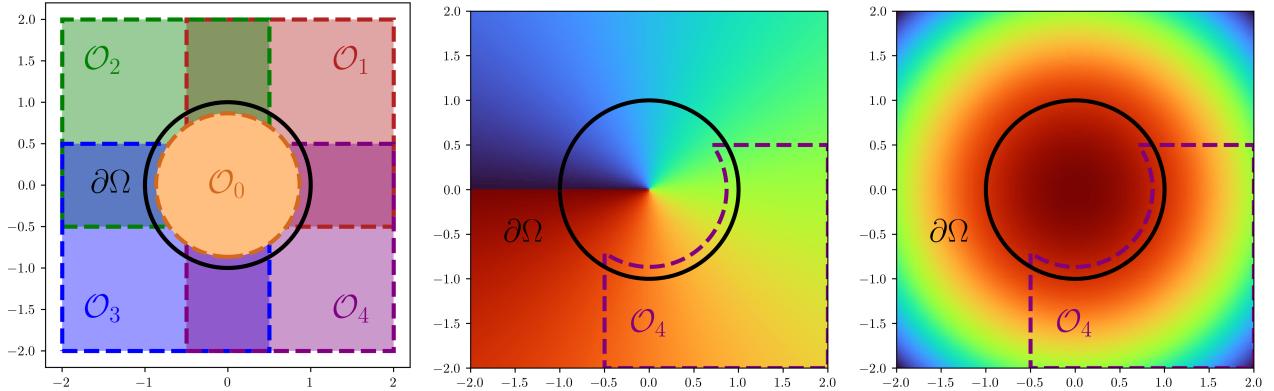
$$\mathcal{O}_1 = \left[ -\frac{1}{2}, 2 \right]^2, \quad \mathcal{O}_2 = \left[ -2, \frac{1}{2} \right] \times \left[ -\frac{1}{2}, 2 \right], \quad \mathcal{O}_3 = \left[ -2, \frac{1}{2} \right]^2, \quad \mathcal{O}_4 = \left[ -\frac{1}{2}, 2 \right] \times \left[ -2, \frac{1}{2} \right],$$

ainsi que

$$\varphi_1(x, y) = \begin{pmatrix} -x+1 \\ -y+1 \end{pmatrix}, \quad \varphi_2(x, y) = \begin{pmatrix} x+1 \\ -y+1 \end{pmatrix}, \quad \varphi_3(x, y) = \begin{pmatrix} x+1 \\ y+1 \end{pmatrix}, \quad \varphi_4(x, y) = \begin{pmatrix} -x+1 \\ y+1 \end{pmatrix}.$$

### Contre-exemples

- Toute frontière de dimension strictement inférieure à  $d-1$  ne sera pas régulière : en particulier, un ouvert percé d'un point n'est pas régulier selon la définition précédente.
- Les frontières à forme fractale, ou qui présentent des coins "dégénérés" sont évincées. Par exemple, l'ouvert  $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 \mid |x| \leq 1 \text{ et } x_2 < \sqrt{|x_1|}\}$  n'est pas régulier : le coin "entrant" ne peut pas être déformé en cône ou en demi-plan par une application suffisamment lisse.

Figure 1.1: Représentation des éléments de l'exemple  $\Omega = \mathcal{B}(0, 1)$ .

Dans la première figure, on a tracé  $\partial\Omega$  en trait plein noir, et les cinq ouverts  $\mathcal{O}_i$  en traits pointillés. Les ouverts se chevauchent et recouvrent  $\Omega$ , mais chacun ne contient (au plus) qu'une partie de la frontière. Les deux figures suivantes se concentrent sur  $\mathcal{O}_4$ , et représentent respectivement la première et la seconde coordonnée de  $\varphi_4$ . L'objectif de la fonction  $\varphi_4$  est de "redresser" la partie de la frontière  $\partial\Omega$  qui traverse  $\mathcal{O}_4$ , en l'envoyant sur  $\mathbb{R} \times \{0\}$ . On observe que c'est bien une fonction lisse, bijective et d'inverse lisse sur son domaine (pas sur tout  $\mathbb{R}^2$ , mais uniquement sur  $\mathcal{O}_4$ ). Ici, la première coordonnée correspond à (une détermination de) l'angle entre le vecteur  $(-1, 0)$  et  $(x, y)$ , qui varie de  $-\pi/6$  à  $2\pi/3$  sur  $\partial\Omega \cap \mathcal{O}_4$ . La seconde coordonnée s'annule sur  $\partial\Omega$ , donc l'image de  $\partial\Omega \cap \mathcal{O}_4$  par  $\varphi_4$  est bien contenue dans le plan  $\mathbb{R} \times \{0\}$ .

### 1.3.2 Trace

Considérons  $\Omega$  un ouvert borné de classe  $C^1$  par morceaux. Pour toute fonction  $\varphi \in \mathcal{D}(\mathbb{R}^d; \mathbb{R})$ , on peut définir sans problème la trace de  $\varphi$  sur  $\partial\Omega$  comme la fonction  $\partial\Omega \ni x \mapsto \varphi(x)$ . Pour les fonctions de  $H^1(\Omega; \mathbb{R})$ , deux problèmes se posent : techniquement, les membres de  $H^1$  sont des classes d'équivalence pour la relation d'égalité presque partout, donc parler de la valeur en un point est délicat. Deuxièmement, ladite valeur pourrait être égale à  $\pm\infty$ .

La théorie des traces est en général liée à celle des opérateurs de prolongement. Le raisonnement est le suivant : on sait que  $\mathcal{D}(\mathbb{R}^d)$  est dense dans  $H^1(\mathbb{R}^d; \mathbb{R})$ . D'autre part, chaque fonction  $\varphi \in \mathcal{D}(\mathbb{R}^d)$  admet une trace  $\gamma(\varphi)$  sur  $\partial\Omega$ , et l'opérateur de trace  $\gamma$  est linéaire continu (on verra de quelle continuité il convient de parler). Dès lors, par densité,  $\gamma$  s'étend en un opérateur linéaire continu sur  $H^1(\mathbb{R}^d)$ . Pour définir la trace des éléments de  $H^1(\Omega)$ , il suffit de les prolonger en des éléments de  $H^1(\mathbb{R}^d)$ , puis de définir la trace comme la restriction à  $H^1(\Omega; \mathbb{R})$  de l'extension de  $\gamma$ .

La régularité du bord intervient à deux endroits : la continuité que l'on arrive à récupérer sur  $\gamma$ , et la validité des théorèmes de prolongement. La régularité de  $\gamma$  va déterminer l'espace dans lequel les traces seront définies. Commençons par le lemme suivant, qui va servir de brique de base à la construction dans le cas d'un  $\Omega$  de classe  $C^1$ .

**Lemme 2 – Cas du demi-plan** Soit  $\varphi \in C^1(\mathbb{R}^d; \mathbb{R})$  à support compact, et  $B = \{(x, x_d) \mid x \in \mathbb{R}^{d-1}, x_d > 0\}$ . Notons  $\gamma_B : C^1(\mathbb{R}^d; \mathbb{R}) \rightarrow \mathcal{C}(\partial B; \mathbb{R})$  l'application trace. Alors

$$\|\gamma_B(\varphi)\|_{L^2(\partial B; \mathbb{R})}^2 \leq \|\varphi\|_{H^1(B; \mathbb{R})} \quad \forall \varphi \in C_c^1(\mathbb{R}^d; \mathbb{R}).$$

#### Démonstration

Soit  $y = (x, 0) \in \partial B$ . Comme  $\varphi$  est à support compact, il existe un  $c > 0$  suffisamment grand tel que  $\varphi(\bar{y}) = 0$  pour tout  $\bar{y} = (x, z)$  avec  $z \geq c$ . Par la régularité de  $z \mapsto \varphi(x, z)$ , on a

$$|\varphi(x, 0)|^2 - |\varphi(x, c)|^2 = |\varphi(x, 0)|^2 = \int_{z=0}^c \frac{d}{dz} [|\varphi(x, \cdot)|^2](z) dz = \int_{z=0}^c 2\varphi(x, z) \frac{d\varphi}{dz}(x, z) dz \leq \int_{z=0}^c |\varphi(x, z)|^2 + \left| \frac{d\varphi}{dz}(x, z) \right|^2 dz$$

où la dernière inégalité provient de  $2ab \leq a^2 + b^2$ . On peut faire tendre  $c \rightarrow \infty$  sans modifier la valeur du terme de droite. En intégrant sur  $x \in \mathbb{R}^{d-1}$ , on obtient

$$\|\gamma_B(\varphi)\|_{L^2(\partial B; \mathbb{R})}^2 = \int_{x \in \mathbb{R}^{d-1}} |\varphi(x, 0)|^2 dx \leq \int_{y \in B} |\varphi(y)|^2 dy + \int_{x \in \mathbb{R}^d, z \geq 0} \left| \frac{d\varphi}{dz}(x, z) \right|^2 dx dz \leq \|\varphi\|_{L^2(B; \mathbb{R})}^2 + \sum_{i=1}^d \|\partial_{x_i} \varphi\|_{L^2(B; \mathbb{R})}^2 = \|\varphi\|_{H^1(B; \mathbb{R})}^2.$$

D'où le résultat.  $\square$

**Lemme 3 – Cas d'un domaine borné  $\mathcal{C}^1$**  Soit  $\Omega$  borné de classe  $\mathcal{C}^1$ . Notons  $\gamma_{\partial\Omega} : \mathcal{D}(\mathbb{R}) \rightarrow \mathcal{C}(\partial\Omega; \mathbb{R})$  l'application trace. Il existe une constante  $C > 0$  indépendante de  $\varphi$  telle que

$$\|\gamma_{\partial\Omega}(\varphi)\|_{L^2(\partial\Omega; \mathbb{R})}^2 \leq C \|\varphi\|_{H^1(\Omega; \mathbb{R})} \quad \forall \varphi \in \mathcal{D}(\Omega).$$

### Démonstration

Notons  $B = \mathbb{R}^{d-1} \times ]0, \infty[$ . Par la régularité de  $\Omega$ , il existe un nombre fini d'ouverts bornés  $(\mathcal{O}_i)_{i \in [1, I]}$  qui recouvrent  $\Omega$ , et une famille d'applications bijectives  $(h_i)_{i \in [1, I]}$ , telles que  $h_i : \mathcal{O}_i \rightarrow \mathbb{R}^d$ ,  $h_i$  et  $h_i^{-1}$  soient de classe  $\mathcal{C}^1$ , et

$$\mathcal{O}_i \cap \Omega = h_i^{-1}(h_i(\mathcal{O}_i) \cap B), \quad \mathcal{O}_i \cap \partial\Omega = h_i^{-1}(h_i(\mathcal{O}_i) \cap \partial B).$$

Soit d'autre part une famille d'applications  $(\theta_i)_{i \in [1, I]}$  telles que  $\theta_i \in \mathcal{C}^\infty(\mathbb{R}^d; [0, 1])$ ,  $\theta_i$  est à support compact dans  $\mathcal{O}_i$ , et  $\sum_{i=1}^I \theta_i^2(x) = 1$  pour tout  $x$  dans l'union des  $\mathcal{O}_i$ . L'existence d'une telle partition de l'unité est classique, mais pas triviale (voir [Bré10, Lemme IX.3] pour des références). On pose

$$\varphi_i : h_i(\mathcal{O}_i) \rightarrow \mathbb{R}, \quad \varphi_i(x) = (\theta_i \varphi) \circ h_i^{-1}(x).$$

Par composition, chaque fonction  $\varphi_i$  est de classe  $\mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$ , et à support compact dans  $h_i(\mathcal{O}_i)$ . Ainsi, on peut prolonger naturellement  $\varphi_i$  en

$$\bar{\varphi}_i : \mathbb{R}^d \rightarrow \mathbb{R}, \quad \bar{\varphi}_i(x) := \begin{cases} \varphi_i(x) & x \in h_i(\mathcal{O}_i), \\ 0 & \text{sinon.} \end{cases}$$

Par le Lemme 2 appliqué à  $\bar{\varphi}_i$ , on a

$$\|\gamma_{\partial B} \bar{\varphi}_i\|_{L^2(\partial B; \mathbb{R})}^2 \leq \|\bar{\varphi}_i\|_{H^1(B; \mathbb{R})}^2.$$

Or d'une part,

$$\|\gamma_{\partial B} \bar{\varphi}_i\|_{L^2(\partial B; \mathbb{R})}^2 = \int_{x \in \partial B} |\bar{\varphi}_i(x)|^2 dx = \int_{x \in \partial B \cap h_i(\mathcal{O}_i)} |(\theta_i \varphi) \circ h_i^{-1}(x)|^2 dx = \int_{x \in \partial\Omega \cap \mathcal{O}_i} \theta_i^2(x) |\varphi(x)|^2 |\det \nabla h_i(x)| dx,$$

et de l'autre, en utilisant l'égalité  $\nabla \varphi_i(x) = \nabla h_i^{-1}(x) \nabla (\theta_i \varphi)(h_i^{-1}(x))$ ,

$$\|\bar{\varphi}_i\|_{H^1(B; \mathbb{R})}^2 = \int_{x \in B} |\nabla \bar{\varphi}_i(x)|^2 dx = \int_{x \in B \cap h_i(\mathcal{O}_i)} |\nabla h_i^{-1}(x) \nabla (\theta_i \varphi)(h_i^{-1}(x))|^2 dx \leq \int_{x \in \Omega \cap \mathcal{O}_i} \|\nabla h_i^{-1}(x)\|_2^2 |\nabla (\theta_i \varphi)(x)|^2 |\det \nabla h_i(x)| dx$$

Comme

$$|\nabla (\theta_i \varphi)(x)|^2 = (|\nabla \theta_i(x)| |\varphi(x)| + \theta_i(x) |\nabla \varphi(x)|)^2 \leq 2(|\nabla \theta_i(x)|^2 |\varphi(x)|^2 + \theta_i^2(x) |\nabla \varphi(x)|^2),$$

en sommant les inégalités obtenues sur  $i \in [1, I]$  et en se rappelant que  $\theta_i = 0$  sur  $\mathcal{O}_i^c$ , on obtient

$$\int_{x \in \partial\Omega} |\varphi(x)|^2 \sum_{i=1}^I \theta_i^2(x) |\det \nabla h_i(x)| dx \leq 2 \int_{x \in \Omega} \sum_{i=1}^I [\|\varphi(x)\|^2 \|\nabla h_i^{-1}(x)\|_2^2 |\nabla \theta_i(x)|^2 + |\nabla \varphi(x)|^2 \|\nabla h_i^{-1}(x)\|_2^2 \theta_i(x)^2] |\det \nabla h_i(x)| dx.$$

La fonction  $x \mapsto \sum_{i=1}^I \theta_i^2(x) |\det \nabla h_i(x)|$  est continue sur  $\partial\Omega$  et y est strictement positive, donc inférieurement bornée par un  $\alpha > 0$  qui ne dépend pas de  $\varphi$ . D'autre part, les quantités  $\|\nabla h_i^{-1}(x)\|_2^2$ ,  $|\nabla \theta_i(x)|^2$ ,  $\|\nabla h_i^{-1}(x)\|_2^2$ ,  $\theta_i(x)^2$  et  $|\det \nabla h_i(x)|$  sont toutes bornées sur  $\Omega$  indépendamment de  $\varphi$ . Ainsi, il existe un  $\beta \geq 0$  tel que

$$\alpha \|\varphi\|_{L^2(\partial\Omega; \mathbb{R})}^2 = \alpha \int_{x \in \partial\Omega} |\varphi(x)|^2 dx \leq 2\beta \left[ \int_{x \in \Omega} |\varphi(x)|^2 dx + \int_{x \in \Omega} |\nabla \varphi(x)|^2 dx \right] = 2\beta \|\varphi\|_{H^1(\Omega; \mathbb{R})}^2.$$

En posant  $C = 2\beta/\alpha$ , on obtient le résultat désiré.  $\square$

Comme  $\|\varphi\|_{H^1(\Omega; \mathbb{R})} \leq \|\varphi\|_{H^1(\mathbb{R}^d; \mathbb{R})}$ , le Lemme 3 nous permet d'étendre l'application  $\gamma_{\partial\Omega}$  en une application linéaire continue de  $H^1(\mathbb{R}^d; \mathbb{R})$  vers  $L^2(\partial\Omega; \mathbb{R})$ . Il nous reste à montrer que l'on peut étendre tout élément  $u \in H^1(\Omega; \mathbb{R})$  en un élément  $\bar{u} \in H^1(\mathbb{R}^d; \mathbb{R})$ . Pour les ouverts bornés de classe  $\mathcal{C}^1$ , l'argument le plus simple est une construction en deux temps très similaire au raisonnement des Lemmes 2 and 3. On se contente d'en donner l'argument clef.

**Lemme 4 – Prolongement par réflexion : cas du demi-plan** Notons  $B = \mathbb{R}^{d-1} \times ]0, \infty[$ . Soit  $u \in H^1(B; \mathbb{R})$ . Alors il

existe  $\bar{u} \in H^1(\mathbb{R}^d; \mathbb{R})$  telle que  $\bar{u} = u$  sur  $B$ , et

$$\|u\|_{H^1(\mathbb{R}^d; \mathbb{R})} \leq 2\|u\|_{H^1(B; \mathbb{R})}.$$

### Démonstration

Pour tout  $\mathbb{R}^d \ni y = (x, x_d)$ , posons

$$\bar{u}(y) = \begin{cases} u(x, x_d) & \text{si } x_d > 0, \\ u(x, -x_d) & \text{sinon.} \end{cases}$$

En particulier, on a  $\partial_{x_d} \bar{u}(x, x_d) = -\partial_{x_d} \bar{u}(x, -x_d)$ . On a alors

$$\begin{aligned} \int_{y \in \mathbb{R}^d} |\bar{u}(y)|^2 + \sum_{i=1}^d |\partial_{y_i} \bar{u}(y)|^2 dy &= \int_{x \in \mathbb{R}^{d-1}, x_d \in ]-\infty, 0] \cup [0, \infty[} \left[ |\bar{u}(x, x_d)|^2 + \sum_{i=1}^{d-1} |\partial_{x_i} \bar{u}(x, x_d)|^2 + |\partial_{x_d} \bar{u}(x, x_d)|^2 dx \right] dx_d \\ &= 2 \int_{x \in \mathbb{R}^{d-1}, x_d \in ]0, \infty[} \left[ |u(x, x_d)|^2 + \sum_{i=1}^{d-1} |\partial_{x_i} u(x, x_d)|^2 + |\partial_{x_d} u(x, x_d)|^2 \right] dx dx_d = 2\|u\|_{H^1(B; \mathbb{R})}. \end{aligned}$$

D'où le résultat.  $\square$

Ce résultat se généralise au cas d'un ouvert borné de classe  $\mathcal{C}^1$  par les mêmes arguments de cartes locales. Cet argument par réflexion est développé plus en détail dans [Bré10, Section IX.2]. Notons que cette construction ne permet pas de traiter les ouverts bornés de classe  $\mathcal{C}^1$  par morceaux, qui ne peuvent pas être redressés en plans. Il existe d'autres constructions, plus exigeantes, mais qui permettent d'atteindre des classes de régularités beaucoup plus larges. Le lecteur curieux pourra consulter [Ste70] pour une théorie adaptée aux ensembles de niveaux de fonctions Lipschitziennes, et [Mou22] pour plus de références.

**Théorème 3 – Opérateur de trace** Soit  $\Omega$  un ouvert borné de classe  $\mathcal{C}^1$  par morceaux. L'application  $\gamma: \mathcal{C}^1(\overline{\Omega}; \mathbb{R}) \rightarrow L^2(\partial\Omega; \mathbb{R})$  se prolonge par continuité en une application linéaire continue de  $H^1(\Omega; \mathbb{R})$  vers  $L^2(\partial\Omega; \mathbb{R})$ , encore notée  $\gamma$ .

La continuité de  $\gamma$  se traduit de la manière suivante : il existe une constante  $C$  telle que

$$\|\gamma_{\partial\Omega} u\|_{L^2} \leq C\|u\|_{H^1} \quad \forall u \in H^1(\Omega; \mathbb{R}). \quad (1.4)$$

Si  $\Omega$  est de classe  $\mathcal{C}^1$ , la démonstration résulte de la combinaison des Lemmes précédents. On en admettra la généralisation.

**Exemples** Soit  $\Omega = (\mathbb{R}_+)^d \cap \mathcal{B}(0, 1)$  en dimension  $d \geq 3$ , et  $u(x) = \frac{1}{|x|^\alpha}$  avec  $\alpha < \frac{d}{2} - 1$ . Notons  $C_d$  la mesure surfacique de  $(\mathbb{R}_+)^d \cap \partial\mathcal{B}(0, 1)$ . On a

$$\|u\|_{L^2(\Omega; \mathbb{R})}^2 = \int_{x \in \Omega} \frac{1}{|x|^{2\alpha}} dx = C_d \int_{r=0}^1 \frac{1}{r^{2\alpha}} r^{d-1} dr = C_d \left[ \frac{r^{d-2\alpha}}{d-2\alpha} \right]_0^1 = \frac{C_d}{(d-2\alpha)} < \infty,$$

et

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R})}^2 = \int_{x \in \Omega} \frac{\alpha^2}{|x|^{2\alpha+2}} dx = \alpha^2 C_d \int_{r=0}^1 \frac{1}{r^{2\alpha+2}} r^{d-1} dr = \frac{\alpha^2 C_d}{d-(2\alpha+2)} < \infty.$$

Donc  $u \in H^1(\Omega; \mathbb{R})$ . Comme  $u$  tend vers  $+\infty$  aux alentours de 0, et on ne peut pas assigner de valeur finie à  $u(0)$ . Par contre, la trace  $x \mapsto u(x)$  est bien dans  $L^2(\partial\Omega)$  : il suffit pour cela de s'assurer que

$$\int_{x \in (\mathbb{R}_+)^{d-1} \cap \mathcal{B}(0, 1)} \frac{1}{|x|^{2\alpha}} dx = C_{d-1} \int_{r=0}^1 \frac{1}{r^{2\alpha}} r^{d-1} dr = \frac{C_{d-1}}{d-2\alpha} < \infty.$$

**Définition 16 – Espace  $H_0^1$**  Soit  $\Omega$  ouvert. L'espace  $H_0^1(\Omega; \mathbb{R})$  est l'adhérence dans  $H^1(\Omega; \mathbb{R})$  de  $\mathcal{D}(\Omega)$ .

**Proposition 1 – Densité** Soit  $\Omega$  domaine ouvert borné régulier par morceaux. L'espace  $H_0^1(\Omega; \mathbb{R})$  est le noyau de l'application trace  $\gamma: H^1(\Omega; \mathbb{R}) \rightarrow L^2(\Omega; \mathbb{R})$ , c'est-à-dire l'ensemble des fonctions de  $H^1(\Omega; \mathbb{R})$  dont la trace est nulle dans  $L^2(\Omega; \mathbb{R})$ .

On se contente d'en donner la démonstration en dimension 1, et pour  $\Omega = ]0, 1[$ .

### Démonstration

D'une part, soit  $\bar{\varphi} \in H_0^1(\Omega; \mathbb{R})$ , et  $(\varphi_n)_{n \in \mathbb{N}} \subset \mathcal{D}(\Omega)$  une suite qui converge vers  $\bar{\varphi}$  pour la norme  $H^1(\Omega; \mathbb{R})$ . Par le Théorème 3,  $\bar{\varphi}$  admet une trace  $\gamma_{\partial\Omega}(\bar{\varphi})$ , et par l'inégalité (1.4), il existe une constante  $C$  indépendante de  $\varphi_n$  et  $\bar{\varphi}$  telle que

$$\|\gamma_{\partial\Omega}(\bar{\varphi}) - \gamma_{\partial\Omega}(\varphi_n)\|_{L^2(\partial\Omega; \mathbb{R})} \leq C\|\bar{\varphi} - \varphi_n\|_{H^1(\Omega; \mathbb{R})}.$$

Comme  $\gamma_{\partial\Omega}(\varphi_n) = 0$  pour tout  $n \in \mathbb{N}$ , on peut passer à la limite en  $n \rightarrow \infty$  pour obtenir que  $\|\gamma_{\partial\Omega}(\bar{\varphi})\|_{L^2(\partial\Omega; \mathbb{R})} = 0$ . Ainsi, la trace de  $\bar{\varphi}$  est nulle.

Réciproquement, soit  $\varphi \in H^1(\Omega; \mathbb{R})$  à trace nulle. On cherche à construire une suite  $(\varphi_n)_{n \in \mathbb{N}} \subset \mathcal{D}(\Omega)$  dont la limite soit  $\varphi$ . Comme  $\Omega$  est de dimension 1,  $\varphi$  est une fonction continue : en effet,

$$|\varphi(y) - \varphi(x)| \leq \left| \int_{z=x}^y \varphi'(z) dz \right| \leq \int_{z=x}^y |\varphi'(z)| dz \leq \sqrt{|y-x|} \|\varphi'\|_{L^2(\Omega; \mathbb{R})}.$$

La trace coïncide donc avec l'évaluation, et on a  $\varphi(0) = \varphi(1) = 0$ . Quitte à travailler sur chaque sous-intervalle sur lequel  $\varphi$  ne s'annule pas, on peut supposer que  $\varphi(x) \neq 0$  pour tout  $x \in ]0, 1[$ . Soit  $G: \mathbb{R} \rightarrow \mathbb{R}$  une fonction lisse telle que  $|G(x)| \leq |x|$ ,  $G_{[-1, 1]} \equiv 0$  et  $G_{]-\infty, -2] \cup [2, \infty[} \equiv id$ , et  $\varphi_n := \frac{1}{n} G(n\varphi)$ . Dans ce cas,

$$|\varphi_n(x) - \varphi(x)| = \begin{cases} 0 & \text{si } |\varphi(x)| \geq 1/n, \\ |G(n\varphi(x))/n - \varphi(x)| \leq 2|\varphi(x)| \leq 2/n & \text{sinon.} \end{cases}$$

Ainsi, par le théorème de convergence dominée de Lebesgue,  $\|\varphi_n - \varphi\|_{L^2} \rightarrow 0$ . De plus,

$$|\nabla \varphi_n(x) - \nabla \varphi(x)| = \left| \frac{1}{n} G'(n\varphi(x)) n\varphi'(x) - \varphi'(x) \right| \leq \begin{cases} 0 & \text{si } |\varphi(x)| \geq 1/n, \\ 2|\varphi'(x)| & \text{sinon.} \end{cases}$$

Ainsi, comme l'ensemble des  $x$  tels que  $|\varphi(x)| < 1/n$  est décroissant pour l'inclusion et tend vers  $\{0\} \cup \{1\}$  de mesure nulle, encore par la convergence dominée,

$$\|\nabla \varphi_n - \nabla \varphi\|_{L^2}^2 \leq \int_{x \in [0, 1], |\varphi(x)| < 1/n} 4|\varphi'(x)|^2 dx \xrightarrow{n \rightarrow \infty} 0.$$

On s'est donc ramené au cas d'une fonction à support compact dans l'ouvert  $]0, 1[$ , que l'on peut approcher par convolution sans détruire cette propriété. Comme la convolution produira une fonction lisse, la preuve est finie.  $\square$

### 1.3.3 Outils subséquents

**Proposition 2 – Formule de Green** Soient  $\Omega \subset \mathbb{R}^d$  ouvert borné régulier de classe  $\mathcal{C}^1$  et  $u, v \in H^1(\Omega; \mathbb{R})$ . Alors

$$\int_{x \in \Omega} u(x) \partial_{x_i} v(x) dx + \int_{x \in \Omega} v(x) \partial_{x_i} u(x) dx = \int_{x \in \partial\Omega} [\gamma_{\partial\Omega} u](x) [\gamma_{\partial\Omega} v](x) n_i(x) dx, \quad (1.5)$$

où  $n: \partial\Omega \rightarrow T\Omega$  est la normale unitaire extérieure en  $\partial\Omega$ .

Le schéma de la preuve est caractéristique des arguments employés dans les espaces de Hilbert et de Sobolev, et il est important de bien le comprendre pour tirer tous les avantages de la densité sans y voir une forme de magie universelle.

### Démonstration

On s'appuie sur

- la validité de cette égalité dans  $\mathcal{C}^1(\overline{\Omega}; \mathbb{R})$ ,
- la densité de  $\mathcal{C}^1(\overline{\Omega}; \mathbb{R})$  dans  $H^1(\Omega; \mathbb{R})$ ,
- la continuité des opérations impliquées dans (1.5) de  $H^1(\Omega; \mathbb{R})$  vers  $\mathbb{R}$ .

On suppose le premier point connu (voir [All05, Corollaire 3.2.3]). La densité de  $\mathcal{C}^1(\overline{\Omega}; \mathbb{R})$  dans  $H^1(\Omega; \mathbb{R})$  est déduite par le raisonnement suivant : grâce à la régularité de  $\Omega$ , toute fonction  $u$  de  $H^2(\Omega; \mathbb{R})$  se prolonge en une fonction  $\bar{u} \in H^1(\mathbb{R}^d; \mathbb{R})$ . Or  $\mathcal{D}(\mathbb{R}^d) \subset \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$  est dense dans  $H^1(\mathbb{R}^d; \mathbb{R})$ , donc on peut trouver une suite  $(\varphi^n)_{n \in \mathbb{N}} \subset \mathcal{C}^1(\mathbb{R}^d; \mathbb{R})$  qui tend vers  $\bar{u}$  en norme  $H^1(\mathbb{R}^d; \mathbb{R})$ . La suite des restrictions  $(\varphi_{|\Omega}^n)_{n \in \mathbb{N}}$  est alors une suite de  $\mathcal{C}^1(\overline{\Omega}; \mathbb{R})$  qui tend vers  $\bar{u}_{|\overline{\Omega}} = u$  en norme  $H^1(\Omega; \mathbb{R})$ .

Passons au troisième point. Pour tout  $i \in \llbracket 1, d \rrbracket$  et  $(u, v) \in (H^1(\Omega; \mathbb{R}))^2$ , l'application  $(u, v) \mapsto \int_{x \in \Omega} u(x) \partial_{x_i} v(x) dx$  est bilinéaire, donc

l'inégalité

$$\left| \int_{x \in \Omega} u(x) \partial_{x_i} v(x) dx \right| \leq \left( \int_{x \in \Omega} u^2(x) dx \right)^{1/2} \left( \int_{x \in \Omega} |\partial_{x_i} u(x)|^2 dx \right)^{1/2} \leq \|u\|_{H^1(\Omega; \mathbb{R})} \|v\|_{H^1(\Omega; \mathbb{R})}$$

suffit à montrer sa continuité. Ainsi, le membre de gauche de (1.5) ne contient que des termes continus. Notons que chaque coordonnée de la normale unitaire satisfait  $|n_i| \leq 1$ . Ainsi, en invoquant le Théorème 3,

$$\left| \int_{x \in \partial\Omega} [\gamma_{\partial\Omega} u](x) [\gamma_{\partial\Omega} v](x) n_i(x) dx \right| \leq \int_{x \in \partial\Omega} |\gamma_{\partial\Omega} u(x)| |\gamma_{\partial\Omega} v(x)| dx \leq \|\gamma_{\partial\Omega} u\|_{L^2(\partial\Omega; \mathbb{R})} \|\gamma_{\partial\Omega} v\|_{L^2(\partial\Omega; \mathbb{R})} \leq C_\gamma^2 \|u\|_{H^1(\Omega; \mathbb{R})} \|v\|_{H^1(\Omega; \mathbb{R})}$$

où  $C_\gamma$  est une constante de continuité de l'application trace  $\gamma_{\partial\Omega}$ .

On conclut ainsi que pour chaque  $(u, v) \in (H^2(\Omega; \mathbb{R}))^2$ , il existe une suite  $(u_n, v_n)_{n \in \mathbb{N}} \subset \mathcal{C}^1(\overline{\Omega}; \mathbb{R})$  qui converge vers  $(u, v)$  dans  $H^1(\Omega; \mathbb{R})$ , et que la continuité des opérations permet de passer à la limite dans l'inégalité (1.5) valide pour chaque  $n$ .  $\square$

**Proposition 3 – Inégalité de Poincaré** Soit  $\Omega \subset \mathbb{R}^d$  ouvert borné. Il existe une constante  $C_\Omega > 0$  telle que pour tout  $u \in H_0^1(\Omega; \mathbb{R})$ ,

$$\|u\|_{L^2(\Omega; \mathbb{R})} \leq C_\Omega \|\nabla u\|_{H^1(\Omega; \mathbb{R})}. \quad (1.6)$$

### Démonstration

Notons  $D_\Omega$  une constante suffisamment grande pour que  $\Omega \subset \overline{\mathcal{B}}(0, D_\Omega/2)$ . Soit  $u \in H_0^1(\Omega; \mathbb{R})$  et  $(u_n)_{n \in \mathbb{N}} \subset \mathcal{D}(\Omega)$  une suite régulière tendant vers  $u$  en norme  $H^1(\Omega; \mathbb{R})$ . Décomposons  $\Omega$  selon la première dimension en  $\Omega = \bigcup_{x \in \mathbb{R}^{d-1}} I_x \times \{x\}$ , où  $I_x \subset \mathbb{R}$  est soit vide, soit ouvert borné de diamètre inférieur à  $D_\Omega$ . Pour chaque  $n \in \mathbb{N}$  et  $x \in \mathbb{R}^{d-1}$  tel que  $I_x$  soit non vide, comme  $u_n$  est à support compact dans  $\Omega$  ouvert, la fonction restreinte  $I_x \ni r \mapsto u_n(r, x)$  se prolonge par 0 en une fonction de classe  $\mathcal{C}^\infty(\mathbb{R}; \mathbb{R})$  à support compact, donc

$$u_n(r, x) - 0 = \int_{s \geq r} \partial_{x_1} u_n(s, x) ds.$$

En élevant l'égalité précédente au carré et en utilisant Cauchy-Schwarz, puis la monotonie de l'intégrale,

$$|u_n(r, x)|^2 \leq \left( \int_{s \geq r} \partial_{x_1} u_n(s, x) ds \right)^2 \leq \int_{s \geq r} 1 ds \times \int_{s \geq r} |\partial_{x_1} u_n(s, x)|^2 ds \leq D_\Omega \times \int_{s \in \mathbb{R}} |\partial_{x_1} u_n(s, x)|^2 ds.$$

En intégrant maintenant sur  $x \in \mathbb{R}^{d-1}$ , on obtient

$$\int_{x \in \mathbb{R}^{d-1}} |u_n(r, x)|^2 dx \leq D_\Omega \times \int_{x \in \mathbb{R}^{d-1}} \int_{s \in \mathbb{R}} |\partial_{x_1} u_n(s, x)|^2 ds dx = D_\Omega \times \|\partial_{x_1} u_n\|_{L^2(\Omega; \mathbb{R})}^2 \leq D_\Omega \times \|\nabla u_n\|_{L^2(\Omega; \mathbb{R})}^2.$$

Il reste à intégrer en  $r \in [-D_\Omega/2, D_\Omega/2]$  pour obtenir

$$\|u_n\|_{L^2(\Omega; \mathbb{R})}^2 = \int_{r \in [-D_\Omega/2, D_\Omega/2]} \int_{x \in \mathbb{R}^{d-1}} |u_n(r, x)|^2 dx dr \leq D_\Omega^2 \times \|\nabla u_n\|_{L^2(\Omega; \mathbb{R})}^2.$$

Comme la convergence dans  $H^1(\Omega; \mathbb{R})$  implique la convergence dans  $L^2(\Omega; \mathbb{R})$ , on peut passer à la limite dans l'inégalité précédente pour obtenir (1.6) avec  $C_\Omega = D_\Omega$ .  $\square$

**Remarque 3** (Raffinement). La preuve de Proposition 3 suggère qu'il suffit que  $u$  s'annule sur une partie du bord qui permette la validité de la première étape, et que le même raisonnement permet de contrôler la valeur de  $\|u\|_{L^2}$  par la norme d'une seule de ses dérivées partielles.

Grâce à (1.6), on a pour toute fonction  $u \in H_0^1(\Omega; \mathbb{R})$  que

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R})}^2 \leq \|u\|_{L^2(\Omega; \mathbb{R})}^2 + \|\nabla u\|_{L^2(\Omega; \mathbb{R})}^2 = \|u\|_{H^1(\Omega; \mathbb{R})}^2 \leq (C_\Omega^2 + 1) \|\nabla u\|_{L^2(\Omega; \mathbb{R})}^2.$$

Ainsi, sur le sous-espace  $H_0^1(\Omega; \mathbb{R})$ , l'application  $u \mapsto \|\nabla u\|_{L^2(\Omega; \mathbb{R})}$  est une norme équivalente à celle induite par  $H^1(\Omega; \mathbb{R})$ . (Dans le reste de l'espace  $H^1(\Omega; \mathbb{R})$ , ce n'est qu'une semi-norme : pourquoi ?)

## 1.4 Théorèmes de densité et d'injection

Une des grandes vertus des espaces  $L^p$ ,  $H^k$ ,  $W^{m,p}$  et consorts est de pouvoir se ramener à des cas plus réguliers (typiquement, des fonctions continues, Höldériennes ou  $\mathcal{C}^k$ ) sous certaines conditions sur le degré d'intégrabilité et la dimension. On présente

ici deux résultats qui nous seront forts utiles par la suite.

**Proposition 4 – Théorème de Rellich** Soit  $\Omega$  un ouvert borné de classe  $C^1$  par morceaux. Alors l'injection de  $H^1(\Omega; \mathbb{R})$  dans  $L^2(\Omega; \mathbb{R})$  est compacte.

**Contre-exemples** Ce théorème est faux si  $\Omega = \mathbb{R}^d$  tout entier. En particulier, pour  $\Omega = \mathbb{R}$ , on peut considérer la suite  $f_n := f(x - n)$ , où  $f \in H^1$  est une fonction non nulle à support dans  $[-1/2, 1/2]$ . On a alors  $\|f_n\|_{H^1} = \|f\|_{H^1}$ , donc la suite  $(f_n)_n$  est bornée, mais  $\|f_n - f_m\|_{L^2} = \|f_n\|_{L^2} + \|f_m\|_{L^2} = 2\|f\|_{L^2}$  ne tend pas vers 0.

La preuve de ce théorème en dimension 1 fera l'objet d'un exercice de TD : on admettra le résultat en dimension supérieure, où l'injection dans les fonctions continues n'est plus valide. En effet, quand la dimension augmente, il est nécessaire d'augmenter également l'ordre de différentiabilité :

**Proposition 5 – Injection dans les fonctions continues** Soit  $\Omega \subset \mathbb{R}^d$  un ouvert de classe  $C^1$  par morceaux. Pour tout  $k > \frac{d}{2}$ , on a l'injection continue

$$H^k(\Omega; \mathbb{R}) \hookrightarrow \mathcal{C}(\Omega; \mathbb{R}).$$

**Exemples** En dimension  $d = 1$ , ce théorème dit que toute fonction de  $H^1$  est une fonction continue. On peut le vérifier en remarquant que sur chaque intervalle borné  $[a, b] \subset \Omega$ , on a par Cauchy-Schwarz

$$\int_{x=a}^b |u'(x)| dx \leq \sqrt{b-a} \sqrt{\int_{x=a}^b |u'(x)|^2 dx} \leq \sqrt{b-a} \|u\|_{H^1},$$

donc  $u|_{[a,b]}$  est absolument continue (ce point n'est pas à apprendre, mais il permet d'écrire la ligne suivante). Ainsi, pour tout  $x, y \in \Omega$ ,

$$|u(y) - u(x)| = \left| \int_{z=x}^y u'(z) dz \right| \leq \int_{z=x}^y |u'(z)| dz \leq \sqrt{|y-x|} \|u\|_{H^1},$$

et  $u$  est même Hölder-continue avec exposant 1/2.

**Contre-exemples** En dimension 2, il ne suffit pas d'être dans  $H^1$  pour être une fonction continue. Par exemple, pour  $\Omega = \overline{\mathcal{B}}(0, 1)$  et  $\alpha < 1$ , la fonction  $u(x) = \frac{1}{|x|^\alpha}$  vérifie

$$\|u\|_{L^2}^2 = \int_{x \in \overline{\mathcal{B}}(0,1)} \frac{1}{|x|^\alpha} dx = 2\pi \int_{r=0}^1 \frac{1}{r^\alpha} r dr = 2\pi \left[ \frac{r^{2-\alpha}}{2-\alpha} \right]_0^1 = \frac{2\pi}{2-\alpha},$$

et

$$\|\nabla u\|_{L^2}^2 = \int_{x \in \overline{\mathcal{B}}(0,1)} \frac{\alpha}{|x|^{\alpha+1}} dx = 2\pi \int_{r=0}^1 \frac{\alpha}{r^{\alpha+1}} dr = 2\pi \left[ \frac{r^{1-\alpha}}{1-\alpha} \right]_0^1 = \frac{2\pi}{1-\alpha}$$

Cependant,  $u$  n'est pas continue en 0.

Montrons ce théorème pour la dimension  $d = 2$  – ce qui impose  $k > 1$ . L'idée générale est toujours la même : établir une inégalité sur les fonctions de classe  $C^\infty$  à support compact qui ne dépende que des normes  $L^2$  des dérivées, ce qui permet ensuite de passer à la limite. Attention à ce point fondamental : dans l'inégalité obtenue ne doivent intervenir que des termes qui autorisent de passer à la limite en norme  $H^k$ . Par exemple, l'inégalité  $|u(y) - u(x)| \leq \int_{z=x}^y |u'(z)| dz$  est valide pour toute fonction de classe  $C^\infty$  à support compact dans  $\mathbb{R}$ , mais on ne peut pas passer à la limite en norme  $L^2$ , puisque  $u \mapsto \int_{z=x}^y |u'(z)| dz$  n'est pas continu (donc peut exploser) dans cette topologie.

On aura besoin du résultat fondamental suivant :

**Théorème 4 – Arzelà-Ascoli** Soit  $K$  un espace compact. Une famille  $F = \{f_\alpha\}_{\alpha \in A}$  de fonctions continues de  $K$  dans  $\mathbb{R}$  est relativement compacte pour la topologie de  $\mathcal{C}(K; \mathbb{R})$  si et seulement si

- $F$  est équibornée : pour tout  $x \in K$ , l'ensemble  $F_x := \{f_\alpha(x) \mid \alpha \in A\}$  est borné dans  $\mathbb{R}$ .
- $F$  est équicontinue : pour tout  $\varepsilon > 0$ , il existe  $\delta > 0$  tel que  $|x - y| \leq \delta$  implique  $|f_\alpha(x) - f_\alpha(y)| \leq \varepsilon$  uniformément

en  $\alpha \in A$ . Autrement dit,

$$\overline{\lim}_{\delta \searrow 0} \sup_{x, y \in K, |x-y| \leq \delta} \sup_{\alpha \in A} |f_\alpha(x) - f_\alpha(y)| = 0.$$

Le théorème d'Arzelà-Ascoli est (très, très) utile pour les EDO et leurs généralisations : en général, on s'attend à ce que la solution d'une équation différentielle ordinaire soit au moins continue et "un peu plus" en fonction du temps. Par exemple, il suffit que la dynamique  $f$  de l'EDO  $\dot{x}_t = f(x_t)$  soit bornée pour qu'une famille de solutions approchées, obtenues via un schéma type Euler, soit Lipschitz avec une même constante. Le Théorème 4 permet alors d'extraire une sous-suite convergente au sens (relativement fort !) de la topologie uniforme.

### Démonstration de la Proposition 5 en dimension 2

On va pouvoir utiliser le caractère *local* de la continuité pour se ramener à un support compact. Soit  $u \in H^2(\Omega; \mathbb{R})$  et  $x_0 \in \Omega$ . Soit  $r > 0$  tel que  $\mathcal{B}(x_0, r) \subset \Omega$ , et  $\chi : \mathbb{R}^2 \rightarrow [0, 1]$  une fonction de classe  $C^\infty$  à support compact dans  $\mathcal{B}(x_0, r)$  et telle que  $\chi|_{\mathcal{B}(x_0, r/2)} \equiv 1$ .

La fonction  $v : x \mapsto u(x)\chi(x)$  appartient à  $H^2(\mathbb{R}^2; \mathbb{R})$ , est à support compact, et coïncide avec  $u$  sur  $\mathcal{B}(x_0, r/2)$ .

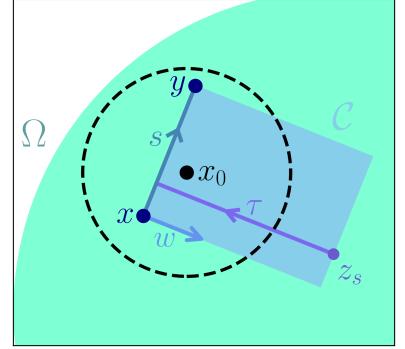
Pour montrer que  $v$  s'identifie (modulo sa classe d'équivalence) à une fonction continue, on considère une suite  $(\varphi_n)_{n \in \mathbb{N}} \subset C_c^\infty(\Omega; \mathbb{R})$  à support compact dans  $\mathcal{B}(x_0, r+1)$  qui converge vers  $v$  en norme  $H^2$ , et on établit une estimation uniforme qui montre que  $(\varphi_n)_n$  converge également vers une fonction continue. Soient  $x, y \in \mathbb{R}^2$ , et  $w \in \mathbb{R}^2$  un vecteur orthogonal à  $y-x$  de norme 1. Il existe  $T \leq 2(r+1)$  tel que pour tout  $s \in [0, 1]$ , le point  $z_s := ((1-s)x + sy) + Tw$  se situe hors de  $\mathcal{B}(x_0, r+1)$ . On a alors

$$\begin{aligned} \varphi_n(y) - \varphi_n(x) &= \int_{s=0}^1 \langle \nabla \varphi_n((1-s)x + sy), y-x \rangle ds = \int_{s=0}^1 \langle \nabla v(z_s) + \int_{\tau=0}^T \nabla^2 \varphi_n(z_s - \tau w)(-\tau w) d\tau, y-x \rangle ds \\ &= - \int_{s=0}^1 \int_{\tau=0}^T \langle \nabla^2 \varphi_n(z_s - \tau w)w, y-x \rangle d\tau ds, \end{aligned}$$

où la matrice  $\nabla^2 \varphi_n$  est la Hessienne de  $\varphi_n$ .

En utilisant  $|\langle Aa, b \rangle| \leq \|a\| \|A\| \|b\|$  pour toute matrice  $A$  de norme d'opérateur  $\|A\| = \sup_{\|a\|=1} \|Aa\|$ , et le fait que  $\|w\| = 1$ ,

$$\begin{aligned} |\varphi_n(y) - \varphi_n(x)| &\leq \int_{s=0}^1 \int_{\tau=0}^T \left| \langle \nabla^2 \varphi_n(z_s - \tau w)w, y-x \rangle \right| d\tau ds \\ &\leq \left( \int_{s=0}^1 \int_{\tau=0}^T \|\nabla^2 \varphi_n(z_s - \tau w)\|^2 d\tau ds \right)^{1/2} \left( T |y-x|^2 \right)^{1/2} \\ &= \sqrt{T} |y-x| \left( \frac{1}{|y-x|} \int_{z \in \mathcal{C}} \|\nabla^2 \varphi_n(z)\|^2 dz \right)^{1/2} \\ &\leq \sqrt{2(r+1)} |y-x| \|\varphi_n\|_{H^2}, \end{aligned}$$



où  $\mathcal{C} \subset \mathbb{R}^2$  désigne le rectangle paramétrisé par  $[0, 1]^2 \ni (s, \tau) \mapsto z_s - \tau w$ .

Ainsi, la suite  $(\varphi_n)_{n \in \mathbb{N}}$  est continue et partage le même module de continuité, donc équicontinue. En prenant  $x \notin \text{supp } \varphi_n$ , puis en passant au supremum sur  $y \in \mathbb{R}^2$ , on obtient que la suite  $(\varphi_n)_n$  est uniformément bornée. Les hypothèses du théorème d'Arzelà-Ascoli sont satisfaites, et une certaine sous-suite  $(\varphi_{n_m})_m$  converge uniformément vers une fonction continue. La convergence uniforme impliquant la convergence dans  $L^2$ , cette limite ne peut être qu'un représentant continu de la classe d'équivalence de  $v$ . Ainsi,  $v$  – et donc  $u$  – est continue dans  $\mathcal{B}(x_0, r/2)$ , et ceci étant valide pour tout  $x_0 \in \Omega$ , on en déduit la continuité sur l'ouvert.  $\square$

La généralisation aux dimensions supérieures se fait de la même manière : on se ramène d'abord à un support compact, puis on écrit  $\varphi_n(y) - \varphi_n(x)$  comme une suite d'intégrales qui partent d'un point hors du support. À chaque nouvel ordre de différentiation, on intègre sur une dimension supplémentaire, pour arriver enfin à une intégrale en dimension  $d$  à la  $d^{\text{ème}}$  dérivée.

Les deux propositions précédentes ne sont qu'un faible aperçu de la richesse des **injections de Sobolev**. Ces injections se déclinent en une foule de variations sur le thème suivant : "l'espace  $W^{m,p}(\Omega; \mathbb{R})$  des fonctions dont les dérivées jusqu'au  $m^{\text{ème}}$  ordre sont représentables par des éléments de  $L^p$  s'injecte continument/compactement dans l'espace  $L^q(\Omega; \mathbb{R})/C^q(\Omega; \mathbb{R})$ ", avec des conditions portant sur la régularité de  $\Omega$  (borné ou non, à frontière plus ou moins lisse), et sur le lien entre  $m$ ,  $p$ , la dimension  $d$  et  $q$ . Grâce aux efforts conjoints de Sobolev, Gagliardo, Nirenberg, Morrey, Rellich, Kondrachov, ... et leurs successeurs, c'est une véritable cartographie des injections qui a été obtenue, incluant les cas  $p, q \in [1, \infty]$  pas forcément entiers. On pourra en retenir les grandes lignes suivantes :

- Si  $\Omega$  est borné, les injections sont compactes ; sinon, elles ne sont que continues.
- L'espace d'arrivée ( $L^q$  ou  $C^q$ ) dépend du signe de la quantité  $\frac{m}{d} - \frac{1}{p}$ . Si  $\frac{m}{d} > \frac{1}{p}$ , on s'attend à une injection dans  $L^\infty$  ou les fonctions continues ; si  $\frac{m}{d} = \frac{1}{p}$ , on s'attend à une injection dans  $L^q$  pour tout  $q \in [p, \infty[$  ; et si  $\frac{m}{d} < \frac{1}{p}$ , l'injection dans  $L^q$  n'est plus valide que pour  $\frac{1}{q} \geq \frac{1}{p} - \frac{m}{d} > 0$ , donc  $q$  pas trop grand.

Le lecteur curieux pourra consulter avec profit le chapitre IX de [Bré10].

## Chapitre 2

# Problèmes linéaires bien posés

Ce court chapitre contient le cœur des théorèmes d'existence pour les équations elliptiques du second ordre : Lax-Milgram. La formulation nécessaire à l'application de ce théorème fait clairement apparaître le problème comme celui d'un *changement de produit scalaire*. Le lecteur a tout intérêt à garder clairement à l'esprit que la propriété fondamentale des opérateurs concernés est la *linéarité*, bien plus que l'ordre de différentiabilité. Cela explique le recours intensif aux outils des espaces de Hilbert et à leur interprétation géométrique.

### 2.1 Le théorème de Riesz

Commençons par la brique de base sur laquelle repose Lax-Milgram.

**Théorème 5 – Théorème de Riesz** Soit  $(H, \langle \cdot, \cdot \rangle)$  un espace de Hilbert. Pour chaque forme linéaire continue  $L: H \rightarrow \mathbb{R}$ , il existe un unique élément  $\ell \in H$  tel que

$$L(x) = \langle \ell, x \rangle \quad \forall x \in H.$$

L'application  $L \mapsto \ell$  est linéaire, bijective, et isométrique au sens où  $|\ell|_H = \|L\|_{\mathcal{L}(H)}$ .

On renvoie le lecteur au cours d'Analyse fonctionnelle pour la démonstration du Théorème 5 dans le cas général. Donnons tout de même une démonstration triviale dans le cas particulier où l'espace  $H$  est séparable, c'est-à-dire qu'il existe une famille dénombrable dense. Dans ce cas, par le procédé d'orthogonalisation de Gram-Schmidt, on peut construire une base Hilbertienne de  $H$ , que l'on notera  $(e_n)_{n \in \mathbb{N}}$ , telle que tout élément s'écrive  $u = \sum_{n \in \mathbb{N}} \langle u, e_n \rangle e_n$ . Soit alors une forme linéaire continue  $L: H \rightarrow \mathbb{R}$ . Pour tout  $v \in H$ , on a

$$L(v) = \sum_{n \in \mathbb{N}} \langle v, e_n \rangle L(e_n) = \langle v, \sum_{n \in \mathbb{N}} L(e_n) e_n \rangle.$$

Soit  $\ell_m := \sum_{n=0}^m L(e_n) e_n$ . En utilisant l'égalité de Parseval et l'orthonormalisation de la base  $(e_n)_n$ , on a

$$|\ell_m|_H^2 = \sum_{n=0}^m (\langle \ell_m, e_n \rangle)^2 = \sum_{n=0}^m (L(e_n))^2 = \langle \ell_m, \sum_{n \in \mathbb{N}} L(e_n) e_n \rangle = L(\ell_m) \leq \|L\|_{\mathcal{L}(H)} |\ell_m|_H.$$

Donc  $|\ell_m|_H \leq \|L\|_{\mathcal{L}(H)}$  indépendamment de  $m \geq 1$ . En passant à la limite en  $m \rightarrow \infty$ , on en déduit que l'élément  $\ell := \sum_{n \in \mathbb{N}} L(e_n) e_n$  appartient à  $H$ , et fournit le représentant souhaité. On laisse le lecteur vérifier les propriétés de l'application  $L \mapsto \ell$ .

### 2.2 Théorème de Lax-Milgram et Stampacchia

**Définition 17 – Coercivité** On dit d'une forme bilinéaire  $a(\cdot, \cdot): H^2 \rightarrow \mathbb{R}$  qu'elle est coercive s'il existe  $\alpha > 0$  tel que

$$a(u, u) \geq \alpha |u|_H^2.$$

**Remarque 4** (Conflit de lexique). En dimension finie, la terminologie est équivalente à la définition générale de coercivité d'une fonction, qui impose que  $\lim_{|x|_H \rightarrow \infty} a(x, x) = +\infty$ . En effet, on a trivialement  $\lim_{|u|_H \rightarrow \infty} a(u, u) \geq \alpha \lim_{|u|_H \rightarrow \infty} |u|_H^2 = +\infty$  avec la Définition 17. Réciproquement, supposons que  $\lim_{|u|_H \rightarrow \infty} a(u, u) = +\infty$  pour  $a(\cdot, \cdot)$  bilinéaire. En particulier, pour  $u = \lambda v$  avec  $|v|_H = 1$ , on a

$$\lim_{\lambda \rightarrow \infty} a(\lambda v, \lambda v) = \lambda^2 a(v, v) = +\infty$$

donc  $a(v, v) > 0$  pour tout  $v \in \partial\mathcal{B}(0, 1)$ . Comme en dimension finie,  $v \mapsto a(v, v)$  atteint sa valeur minimale  $\alpha > 0$  sur le compact  $\partial\mathcal{B}(0, 1)$ , on a bien équivalence. Cette coïncidence n'est plus valide en dimension infinie : si  $(e_n)_n$  est une base Hilbertienne infinie d'un espace  $H$ , la forme bilinéaire

$$a(u, v) := \sum_{n=1}^{\infty} \frac{1}{n^2} \langle u, e_n \rangle \langle v, e_n \rangle$$

n'est pas coercive (car  $a(e_m, e_m) = 1/m^2$ ).

Arrivé à ce niveau, le lecteur frétille sûrement d'impatience d'enfin connaître l'énoncé suivant.

**Théorème 6 – Lax-Milgram** Soient  $(H, \langle \cdot, \cdot \rangle_H)$  et  $(V, \langle \cdot, \cdot \rangle_V)$  deux espaces de Hilbert tels que  $V \subset H$  avec injection continue,  $a(\cdot, \cdot) : V^2 \rightarrow \mathbb{R}$  une forme bilinéaire, continue et coercive, et  $L : H \rightarrow \mathbb{R}$  une forme linéaire continue. Il existe un unique élément  $u \in V$  tel que

$$a(u, v) = L(v) \quad \forall v \in V, \tag{2.1}$$

et l'application  $L \mapsto u$  est continue.

### Démonstration

Pour chaque  $u \in V$ , l'application  $V \ni v \mapsto a(u, v)$  est une forme linéaire et continue. Le Théorème 5 assure l'existence d'une application linéaire, bijective et bicontinue telle que  $\varphi : V \rightarrow V$  telle que  $a(u, v) = \langle \varphi(u), v \rangle_V$ . Pour plus de clarté, notons  $\iota : V \rightarrow H$  l'injection de  $V$  dans  $H$ . Par hypothèse, il existe une constante  $[l]$  telle que  $|\iota(v)|_H \leq [l] |v|_V$ . Ainsi, l'application linéaire  $V \ni v \mapsto L(\iota(v))$  est bien continue, avec pour constante  $\|L\|_{\mathcal{L}(H)} [l]$ . Il existe donc un élément  $\ell \in V$  tel que  $L(\iota(v)) = \langle \ell, v \rangle_V$  pour tout  $v \in V$ . (Ici, on prendra bien garde à utiliser le bon produit scalaire !)

On s'est donc ramené au problème de trouver  $u \in V$  tel que  $\langle \varphi(u), v \rangle_V = \langle \ell, v \rangle_V$  pour tout  $v \in V$ , soit  $\varphi(u) = \ell$ . Pour prouver la surjectivité de  $\varphi$ , on emploie une astuce de point fixe qui fait usage de la coercivité de  $a(\cdot, \cdot)$ . Soit  $\rho > 0$  qui sera fixé plus tard : l'égalité précédente se réécrit  $u = u - \rho(\varphi(u) - \ell)$ . Soit  $\Psi(u) := u - \rho(\varphi(u) - \ell)$ . On a pour tout  $(v, w)$

$$\begin{aligned} |\Psi(v) - \Psi(w)|_V^2 &= \langle v - w - \rho\varphi(v - w), v - w - \rho\varphi(v - w) \rangle = |v - w|_V^2 - 2\rho \langle \varphi(v - w), v - w \rangle_V + \rho^2 |\varphi(v - w)|_V^2 \\ &= |v - w|_V^2 - 2\rho a(v - w, v - w) + \rho^2 |\varphi(v - w)|_V^2 \leq (1 - 2\rho\alpha + \rho^2) |v - w|_V^2. \end{aligned}$$

En choisissant  $\rho = \alpha$ , on obtient  $|\Psi(v) - \Psi(w)|_V^2 \leq (1 - \alpha^2) |v - w|_V^2$ , donc  $\Psi$  est une contraction stricte. Comme  $V$  est un espace de Hilbert, donc complet, le théorème de Picard s'applique et  $\Psi$  admet un unique point fixe  $u \in V$ .  $\square$

**Remarque 5** (Démonstration directe en dimension finie). En dimension finie, les formes bilinéaires (qui sont automatiquement continues) s'écrivent comme  $a(u, v) = \langle Au, v \rangle$  pour  $A \in \mathbb{M}_{d,d}$  une certaine matrice, et les formes linéaires comme  $L(v) = \langle \ell, v \rangle$ . En évaluant l'égalité (2.1) pour chaque  $v$  de la base canonique, on obtient que  $u$  doit satisfaire  $Au = \ell$ . Décomposons  $A = B + C$ , où  $B = (A + A^t)/2$  et  $C = (A - A^t)/2$ . Montrons que la condition de coercivité implique que  $A$  est inversible : en effet, pour tout vecteur  $v \in \mathbb{R}^d$ , on a

$$\langle (B - C)(B + C)v, v \rangle = \langle (B + C)v, (B - C)^t v \rangle = \langle Av, Av \rangle \geq \alpha |v|^2.$$

Ainsi, la matrice  $M := (B - C)(B + C)$  est symétrique et définie positive. Il existe donc une matrice  $P$  de changement de base telle que  $I = P^t MP = ((B + C)P)^t((B + C)P)$ . Donc  $(B + C)P$  est inversible, et par suite,  $A = B + C$  est inversible. L'élément  $u$  est directement donné par  $A^{-1}\ell$ .

Le cas particulier où  $a(\cdot, \cdot)$  est symétrique fournit une intuition géométrique encore plus précise. Dans ce cas, la forme bilinéaire  $a(\cdot, \cdot)$  est un produit scalaire sur  $V$ , et l'équation  $a(u, \cdot) - L(\cdot) = 0$  correspond à la condition d'optimalité du premier ordre  $\nabla \mathcal{E}(\cdot) = 0$ , où  $\mathcal{E}$  est une fonctionnelle faisant intervenir la norme induite par  $a(\cdot, \cdot)$ .

**Théorème 7 – Stampacchia** Soient  $(H, \langle \cdot, \cdot \rangle_H)$  et  $(V, \langle \cdot, \cdot \rangle_V)$  deux espaces de Hilbert tels que  $V \subset H$  avec injection continue,  $a(\cdot, \cdot) : V^2 \rightarrow \mathbb{R}$  une forme bilinéaire, symétrique, continue et coercive, et  $L : H \rightarrow \mathbb{R}$  une forme linéaire continue.

L'élément  $u^* \in V$  donné par le Théorème 6 est l'unique minimiseur de la fonction

$$\mathcal{E}: V \rightarrow \mathbb{R}, \quad \mathcal{E}(v) := \frac{1}{2} a(v, v) - L(v).$$

### Démonstration

Soient  $(u, v) \in V^2$  et  $\lambda \in [0, 1]$ . On a

$$a((1-\lambda)u + \lambda v, (1-\lambda)u + \lambda v) = a(u + \lambda(v-u), u + \lambda(v-u)) = a(u, u) + 2\lambda a(u, v-u) + \lambda^2 a(v-u, v-u),$$

donc  $v \mapsto a(v, v)$  est une fonction convexe. Comme c'est également une fonction continue et coercive par hypothèse, l'application  $\mathcal{E}$  admet un unique minimum  $v^*$  sur  $V$ , qui est caractérisé par la condition

$$\forall h \in V \setminus \{0\}, \quad \lim_{|h| \rightarrow 0} \frac{\mathcal{E}(v^* + h) - \mathcal{E}(v^*)}{|h|_V} = 0.$$

Dans notre cas, pour tout  $h \neq 0$ , la symétrie de  $a(\cdot, \cdot)$  permet d'écrire

$$\frac{\mathcal{E}(v^* + h) - \mathcal{E}(v^*)}{|h|_V} = \frac{\frac{1}{2}a(v^* + h, v^* + h) - L(v^* + h) - \frac{1}{2}a(v^*, v^*) + L(v^*)}{|h|_V} = \frac{|h|_V}{2} a\left(\frac{h}{|h|_V}, \frac{h}{|h|_V}\right) + a\left(v^*, \frac{h}{|h|_V}\right) - L\left(\frac{h}{|h|_V}\right).$$

Ainsi, en choisissant  $\frac{h}{|h|_V} = w \in V$  et en laissant  $|h|_V \rightarrow 0$ , on obtient que  $v^*$  satisfait  $a(v^*, w) = L(w)$  pour tout élément de la sphère unité de  $V$ . Par linéarité, l'égalité est satisfaite sur tout  $V$ . Comme  $u^* \in V$  est l'unique solution de ce problème, on retrouve  $v^* = u^*$ .  $\square$

La condition de symétrie dans le Théorème 7 n'est pas dispensable. En effet, en posant  $a^{\text{sym}}(u, v) := \frac{a(u, v) + a(v, u)}{2}$ , on vérifie facilement que  $a(v, v) = a^{\text{sym}}(v, v)$  pour tout  $v \in V$ . Ainsi, le problème de minimisation du Théorème 7 est "aveugle" à la partie antisymétrique de  $a(\cdot, \cdot)$ , là où la solution du problème (2.1) en dépendra (construire un exemple en dimension finie pour s'en convaincre !).

## 2.3 Application à un problème elliptique

Soit  $\Omega \subset \mathbb{R}^d$  un ouvert borné régulier.

**Définition 18 – Opérateur elliptique** Soient  $a_{ij} \in L^\infty(\Omega; \mathbb{R})$ ,  $(i, j) \in \llbracket 1, d \rrbracket^2$ , et  $a_0 \in L^\infty(\Omega; \mathbb{R}^+)$  telles que  $a_{ij} = a_{ji}$  et qui satisfont la condition de non-dégénérescence

$$\exists \alpha > 0, \quad \forall v \in \mathbb{R}^d, \quad \sum_{i=1}^d \sum_{j=1}^d v_i a_{ij}(x) v_j \geq \alpha |v|^2 \quad \forall x \in \Omega.$$

L'opérateur elliptique associé est défini par

$$A: \mathcal{D}'(\Omega) \rightarrow \mathcal{D}'(\Omega), \quad Au := - \sum_{i,j \in \llbracket 1, d \rrbracket^2} \partial_{x_i} (a_{ij} \partial_{x_j} u) + a_0 u.$$

On considère le problème suivant : pour  $f \in L^2(\Omega; \mathbb{R})$ , trouver  $u \in H^1(\Omega; \mathbb{R})$  solution de

$$\begin{cases} Au(x) = f(x) & x \in \Omega, \\ u(x) = 0 & x \in \partial\Omega. \end{cases} \quad (2.2a)$$

$$(2.2b)$$

**Remarque 6** (Interprétation). Les notations  $Au(x)$  et  $f(x)$  dans (2.2a) sont à prendre formellement : l'équation demande une égalité entre la distribution  $Au$  et (la distribution engendrée par)  $f \in L^2(\Omega; \mathbb{R})$ . L'égalité au sens des distributions est une condition très faible, et on va chercher à l'obtenir dans un sens plus fort qui permette des méthodes numériques. La méthode générale consiste à trouver un espace de fonctions tests  $V$  dans lequel on puisse obtenir un résultat d'existence de  $u$ .

Pour étudier (2.2), on se ramène d'abord à une forme variationnelle. Cette étape est formelle au sens où l'on ne connaît pas encore la régularité de  $u$ , et sert justement à établir dans quelle classe la solution "devrait" exister. Prenons formellement le produit scalaire de (2.2a) avec une fonction test  $v \in H_0^1(\Omega; \mathbb{R})$ . On obtient

$$\begin{aligned} \langle f, v \rangle_{L^2} &= \int_{x \in \Omega} f(x) v(x) dx \stackrel{(2.2a)}{=} \langle Au, v \rangle_{L^2} = - \sum_{i,j \in [1,d]^2} \int_{x \in \Omega} \partial_{x_i} (a_{ij} \partial_{x_j} u)(x) v(x) dx + \int_{x \in \Omega} a_0(x) u(x) v(x) dx \\ &= \sum_{i,j \in [1,d]^2} \int_{x \in \partial\Omega} [a_{ij} \partial_{x_j} u \cdot n_i](x) v(x) dx + \int_{x \in \Omega} a_{ij}(x) \partial_{x_j} u(x) \partial_{x_i} v(x) dx + \int_{x \in \Omega} a_0(x) u(x) v(x) dx \end{aligned}$$

Si la trace de  $a_{ij}(x) \partial_{x_j} u(x)$  sur  $\partial\Omega$  est dans  $L^2(\partial\Omega, \mathbb{R})$ , alors  $v \in H_0^1(\Omega; \mathbb{R})$  implique que le terme de bord s'annule. On aurait alors la *formulation variationnelle*

$$\langle f, v \rangle_{L^2} = \sum_{i,j \in [1,d]^2} \int_{x \in \Omega} a_{ij}(x) \partial_{x_j} u(x) \partial_{x_i} v(x) dx + \int_{x \in \Omega} a_0(x) u(x) v(x) dx =: a(u, v). \quad (2.3)$$

**Application de Lax-Milgram** L'application  $(u, v) \mapsto a(u, v)$  est une forme bilinéaire symétrique (grâce à la symétrie  $a_{ij} = a_{ji}$  supposée dans la Définition 18). Elle est bien définie si  $u$  et  $v$  appartiennent à  $H^1(\Omega; \mathbb{R})$ . Montrons que  $a$  est continue et définie positive dans cet espace : on a d'une part

$$\begin{aligned} |a(u, v)| &\leq \left| \sum_{i,j \in [1,d]^2} \int_{x \in \Omega} a_{ij}(x) \partial_{x_j} u(x) \partial_{x_i} v(x) dx \right| + \left| \int_{x \in \Omega} a_0(x) u(x) v(x) dx \right| \\ &\leq \sum_{i,j \in [1,d]^2} \|a_{ij}\|_{L^\infty} \|\partial_{x_j} u\|_{L^2} \|\partial_{x_i} v\|_{L^2} + \|a_0\|_{L^\infty} \|u\|_{L^2} \|v\|_{L^2} \\ &\leq \left( \sum_{j=1}^d \|\partial_{x_j} u\|_{L^2}^2 \right)^{1/2} \left( \sum_{i,j=1}^d \|a_{ij}\|_{L^\infty}^2 \right)^{1/2} \left( \sum_{i=1}^d \|\partial_{x_i} v\|_{L^2}^2 \right)^{1/2} + \|a_0\|_{L^\infty} \|u\|_{L^2} \|v\|_{L^2} \\ &\leq \|u\|_{H^1} \left( \sum_{i,j=1}^d \|a_{ij}\|_{L^\infty}^2 + \|a_0\|_{L^\infty}^2 \right) \|v\|_{H^1}. \end{aligned}$$

D'autre part, en utilisant l'inégalité de Poincaré (Proposition 3),

$$\begin{aligned} a(u, u) &= \sum_{i,j \in [1,d]^2} \int_{x \in \Omega} a_{ij}(x) \partial_{x_j} u(x) \partial_{x_i} u(x) dx + \int_{x \in \Omega} a_0(x) u(x) u(x) dx \\ &\geq \alpha \int_{x \in \Omega} \|\nabla u(x)\|^2 dx + 0 \geq \frac{\alpha}{2} \|\nabla u\|_{L^2}^2 + \frac{\alpha}{2C_\Omega^2} \|u\|_{L^2}^2 \geq \frac{\alpha}{2} \min\left(1, \frac{1}{C_\Omega^2}\right) \|u\|_{H^1}^2. \end{aligned}$$

Donc  $a$  est une forme bilinéaire continue définie positive. D'autre part, l'application  $v \mapsto \langle f, v \rangle_{L^2}$  est bien définie de  $H_0^1(\Omega; \mathbb{R})$  dans  $R$ , linéaire, et continue :

$$|\langle f, v \rangle_{L^2}| \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{L^2} \|v\|_{H^1}.$$

On peut donc appliquer le Théorème de Lax-Milgram (Théorème 6) au problème variationnel

$$a(u, v) = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega; \mathbb{R})$$

et obtenir l'existence et l'unicité d'une solution  $u \in H_0^1(\Omega; \mathbb{R})$ . N'est-ce pas formidable ?

**Retour au problème initial** Bien évidemment, le lecteur nous opposera qu'il n'est nul besoin de faire appel à cette machinerie pour résoudre l'équation  $u'' = f$  si  $f$  est une fonction continue. Cette remarque pertinente s'étend à la dimension supérieure : si le second membre  $f(\cdot)$  du problème (2.2a) est suffisamment régulier, la solution  $u$  sera elle-même régulière. Considérons le cas  $a_0 = 0$  et  $a_{ij} = \delta_{ij}$  pour simplifier, ce qui implique  $Au = -\Delta u$ . On a alors le résultat suivant : si  $f \in H^k(\Omega; \mathbb{R})$  pour  $k \geq 0$ , alors  $u \in H^{k+2}(\Omega; \mathbb{R})$ .

Pour établir cette régularité, on remarque que la solution faible  $u \in H_0^1(\Omega; \mathbb{R})$  satisfait

$$a(u, v) = \langle \nabla u, \nabla v \rangle = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega; \mathbb{R}).$$

Comme  $\mathcal{D}(\Omega) = \mathcal{C}_c^\infty(\Omega; \mathbb{R}) \subset H_0^1(\Omega; \mathbb{R})$ , cela implique que les dérivées seconde au sens des distributions de  $u$  sont représentables via le produit scalaire de  $f \in H^k \subset L^2$  et la fonction test  $v \in \mathcal{D}(\Omega)$ , donc que  $u \in H^2(\Omega; \mathbb{R})$ . De plus, pour  $2 \leq p \leq k+2$ , les dérivées  $p$ <sup>ièmes</sup> de  $u$  seront données par les dérivées  $(p-2)$ <sup>ièmes</sup> de  $f$ , qui appartiennent à  $H^{k-(p-2)}$ , donc  $u \in H^{k+2}$ . Si maintenant la dimension  $d$  est suffisamment faible pour que  $k+2 > \frac{d}{2}$ , la Proposition 5 implique que  $u \in \mathcal{C}(\Omega; \mathbb{R})$ . De même, si  $k+1 > \frac{d}{2}$ , alors chaque dérivée partielle de  $u$  est dans  $H^{k+1}(\Omega; \mathbb{R})$ , qui s'injecte dans les fonctions continues, donc  $u \in \mathcal{C}^1(\Omega; \mathbb{R})$ . On obtient ainsi la régularité minimale permettant d'assurer que la solution du problème (2.2) est classique, au sens où elle appartient à  $\mathcal{C}^2$  : il faut que  $f \in H^k$  pour  $k > \frac{d}{2}$ . On remarque ici que la régularité de Sobolev (espaces  $H^k$ ) de  $u$  ne dépend pas de la dimension, alors que la régularité au sens classique (espaces  $\mathcal{C}^k$ ) en dépend.

# Chapitre 3

## Méthode des éléments finis

### 3.1 Principe de l'approximation de Galerkin

Soient  $H, V$  deux espaces de Hilbert tels que  $V \subset H$  avec injection continue, et  $a(\cdot, \cdot) : V^2 \rightarrow \mathbb{R}$  une forme bilinéaire continue et coercive, et  $L : H \rightarrow \mathbb{R}$  une forme linéaire continue. Considérons le problème variationnel

$$\text{Trouver } u \in V \text{ tel que } a(u, v) = L(v) \text{ pour tout } v \in V. \quad (3.1)$$

Par le Théorème 6, le problème (3.1) est bien posé.

Le principe des éléments finis, et plus généralement des méthodes de Galerkin, est d'approcher l'espace  $V$  par un sous-espace de dimension finie  $V_h$ . Par comparaison, les *différences finies* peuvent être vues comme des méthodes de discréétisation des *opérateurs*, ce qui correspondrait ici à approcher  $a(\cdot, \cdot)$  et  $L(\cdot)$ . Une méthode de Galerkin *interne* consiste à choisir  $V_h \subset V$ , ce qui est le cadre de ce cours.

**Définition 19 – Famille d'espace d'approximation** On considère  $(V_h)_{h \in \mathbb{R}^+}$  une famille telle que  $V_h \subset V_{\bar{h}} \subset V_0 = V$  pour tout  $h \geq \bar{h} > 0$ , et pour tout  $h > 0$ , l'espace  $V_h$  est un espace vectoriel de dimension finie  $N_h$ .

Chaque espace  $V_h$  est muni du produit scalaire de  $V$ , pour lequel il devient un espace de Hilbert.

#### Exemples

- Soit  $V = \ell^2$  l'espace des suites de carré sommable, muni de sa base hilbertienne  $(e_i)_{i \in \mathbb{N}}$  donnée par  $e_i = (0, 0, \dots, 0, 1, 0, \dots)$  avec 1 sur la  $i^{\text{ème}}$  coordonnée. On peut considérer  $V_h = \text{Vect}\{e_i \mid i \in \llbracket 1, N_h \rrbracket\}$ , où  $h \mapsto N_h \in \mathbb{N}$  est une fonction décroissante qui tend vers  $+\infty$  quand  $h \rightarrow 0$ .
- Plus généralement, si  $V$  est un espace de Hilbert séparable, il admet une base hilbertienne et on peut appliquer la même construction que pour  $\ell^2$ .

**Proposition 6 – Solution approchée** Soit  $(V_h)_h$  une famille approximant  $V$  selon la Définition 19. Pour tout  $h > 0$ , il existe une unique solution  $u_h \in V_h$  du problème approché

$$\text{Trouver } u_h \in V_h \text{ tel que } a(u_h, v_h) = L(v_h) \text{ pour tout } v_h \in V_h. \quad (3.2)$$

De plus, l'erreur  $u - u_h$  est  $a(\cdot, \cdot)$ -orthogonale à l'espace  $V_h$ .

#### Démonstration

On a par définition  $V_h \subset V \subset H$ , donc  $V_h$  est un espace de Hilbert qui s'injecte continument dans  $H$ . Ainsi, toutes les hypothèses du Théorème 6 sont vérifiées, et il existe une unique solution. De plus, pour chaque  $v_h \in V_h$ , on a  $a(u, v_h) = L(v_h)$  et  $a(u_h, v_h) = L(v_h)$ . Par soustraction,

$$a(u - u_h, v_h) = 0$$

pour tout  $v_h \in V_h$ , d'où l'orthogonalité pour l'application  $a(\cdot, \cdot)$ . □

Soit  $h > 0$  fixé, et notons  $N = N_h$ . La méthode des éléments finis consiste à introduire une base  $(e_i)_{i \in [1, N]}$  de l'espace  $V_h$ , et de décomposer  $u_h = \sum_{i=1}^N U_i e_i$ . En injectant cette décomposition dans le problème (3.2), il vient

$$\sum_{i=1}^N U_i a(e_i, v_h) = L(v_h) \quad \forall v_h \in V_h.$$

En prenant maintenant successivement  $v_h = e_j$  pour chaque  $j$ , on obtient

$$\sum_{i=1}^N U_i a(e_i, e_j) = L(e_j) \quad \forall j \in [1, N].$$

En notant

$$U := \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix}, \quad \mathbb{A} := \begin{pmatrix} a(e_1, e_1) & a(e_2, e_1) & a(e_3, e_1) & \cdots & a(e_N, e_1) \\ a(e_1, e_2) & a(e_2, e_2) & a(e_3, e_2) & \cdots & a(e_N, e_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a(e_1, e_N) & a(e_2, e_N) & a(e_3, e_N) & \cdots & a(e_N, e_N) \end{pmatrix} \quad \text{et} \quad \mathbb{B} := \begin{pmatrix} L(e_1) \\ L(e_2) \\ \vdots \\ L(e_N) \end{pmatrix},$$

on obtient le système de dimension  $N \times N$

$$\mathbb{A}U = \mathbb{B}. \quad (3.3)$$

**Lemme 5 – Système bien posé** Le système (3.3) admet une unique solution  $U$ .

### Démonstration

Montrons que la matrice  $\mathbb{A}$  est définie positive : si tel est le cas, l'application linéaire  $V \rightarrow \mathbb{A}V$  sera injective, donc (en dimension finie) bijective. Soit  $U \in \mathbb{R}^N$  quelconque : par le raisonnement inverse de l'obtention du système en dimension finie, on pose  $u = \sum_{i=1}^N U_i e_i$  et on obtient

$$\langle \mathbb{A}U, U \rangle = \sum_{j=1}^N \sum_{i=1}^N \mathbb{A}_{ji} U_i U_j = \sum_{j=1}^N \sum_{i=1}^N a(e_i, e_j) U_i U_j = a(u, u) \geq \alpha \|u\|_V^2 \geq 0.$$

De plus,  $\langle \mathbb{A}U, U \rangle = 0$  si et seulement si  $\|u\|_V^2 = 0$ , ce qui implique que  $U = 0_{\mathbb{R}^N}$ . Ainsi (3.3) est bien posé.  $\square$

La méthode des éléments finis consiste à implémenter et résoudre le système (3.3). Plusieurs difficultés se posent :

- la taille du système (3.3) peut devenir gigantesque pour obtenir de bonnes précisions. Pour pallier à cette difficulté, on cherche à choisir convenablement la base  $(e_i)_{i \in [1, N]}$  pour rendre la matrice  $\mathbb{A}$  aussi creuse que possible, et profiter d'algorithmes optimisés pour ce cas.
- Le calcul des coefficients  $a(e_i, e_j)$  peut s'avérer coûteux. Pour diminuer le volume de calcul, on peut s'appuyer sur des méthodes de quadrature, qui exploitent elles-mêmes la régularité du second membre  $L$ .

## 3.2 Éléments d'analyse

Le résultat de convergence fondamental est donné par le lemme de Céa, du mathématicien Jean Céa (1932-2024).

**Lemme 6 – de Céa** Soient  $H, V$  deux espaces de Hilbert tels que  $V \subset H$  avec injection continue, et  $a(\cdot, \cdot) : V^2 \rightarrow \mathbb{R}$  une forme bilinéaire continue et coercive dans  $V$ ,  $L : H \rightarrow \mathbb{R}$  forme linéaire continue. Soient  $\alpha, M$  respectivement les constantes de coercivité et de continuité de  $a(\cdot, \cdot)$ . Dès lors,

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

### Démonstration

Soit  $v_h \in V_h$  arbitraire. On a

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h).$$

Comme  $v_h - u_h \in V_h$ , en utilisant la propriété de  $a(\cdot, \cdot)$ -orthogonalité de la Proposition 6, on obtient

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - v_h) + 0 \leq M \|u - u_h\|_V \|u - v_h\|_V.$$

En prenant l'infimum sur  $v_h \in V_h$ , on obtient le résultat désiré.  $\square$

L'intérêt du lemme de Céa est de se ramener à un problème d'approximation de  $V$  par  $V_h$ . On l'utilise de la manière suivante : supposons qu'il existe un opérateur d'approximation  $\Pi_h : V \rightarrow V_h$  tel que

$$\forall v \in V, \quad \lim_{h \searrow 0} \|v - \Pi_h v\|_V = 0.$$

Dès lors, la suite des solutions  $(u_h)_{h>0}$  converge vers  $u$  quand  $h \searrow 0$  : en effet,

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \|u - \Pi_h u\|_V \xrightarrow[h \searrow 0]{} 0.$$

**Exemples** Dans le cas où  $V = \ell^2$  et  $V_h$  est l'espace vectoriel engendré par les  $N_h$  premiers vecteurs de la base canonique, un tel opérateur peut être choisi comme  $\Pi_h u = (u_1, \dots, u_{N_h}, 0, 0, \dots)$ .

On remarque ici toute l'importance des constantes  $\alpha$  et  $M$ , sur lesquelles repose l'analyse que nous venons de faire. En ce sens, la théorie de Galerkin telle que nous l'appliquons ici n'est rien d'autre qu'une généralisation en dimension infinie des systèmes linéaires définis positifs.

### 3.3 Éléments finis de Lagrange

Les éléments finis de Lagrange sont basés sur l'interpolation. Le principe général est le suivant : on construit un maillage de  $\Omega$  formé d'ensembles tous semblables, qui sont tous des déformations (translations, rotations, homothéties...) d'un unique ensemble de référence noté  $\hat{K}$ . On définit sur  $\hat{K}$  une famille de fonctions de base, qui seront déformées de la même manière que  $\hat{K}$  pour s'adapter à chacun des éléments du maillage. On obtient ainsi la base de l'espace  $V_h$ .

**Définition 20 – Élément fini de référence** Un élément fini de Lagrange est un triplet  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  tel que

- $\hat{K}$  est un fermé connexe de volume non nul,
- $\hat{\Sigma} = \{\hat{a}_i \mid i \in \llbracket 1, N \rrbracket\}$  est un ensemble de  $N$  points distincts de  $\hat{K}$ ,
- $\hat{\mathbb{P}} = \text{Vect}\{p_i\}_{i \in \llbracket 1, N \rrbracket}$  est un espace vectoriel de dimension  $N$  de fonctions de  $\hat{K}$  dans  $\mathbb{R}$ .

Moralement,  $\hat{K}$  est un modèle de triangle,  $\hat{\Sigma}$  est l'ensemble des points où l'on va interpoler la solution, et  $\hat{\mathbb{P}}$  est l'ensemble des fonctions de base "élémentaires" sur le triangle de référence.

**Exemples**

- Le triplet  $([0, 1], \{0, 1\}, \text{Vect}\{x \mapsto x, x \mapsto 1 - x\})$  est un élément fini.
- Le triplet  $([0, 1], \{1/3, 1/2, 2/3\}, \text{Vect}\{x \mapsto x^n \mid n \in \llbracket 0, 42 \rrbracket\})$  est un élément fini.

**Remarque 7** (Degrés de liberté). *De manière plus générale, les éléments de  $\hat{\Sigma}$  sont des degrés de liberté, c'est-à-dire des formes linéaires sur  $\mathbb{P}$ . Ici, la forme linéaire associée à un point  $\hat{a}_i \in \hat{\Sigma}$  est donnée par  $p \mapsto p(\hat{a}_i)$  l'interpolation. On vérifie bien que pour deux fonctions  $p, q$  et un réel  $\lambda$ , on a  $(\lambda p + q)(\hat{a}_i) = \lambda p(\hat{a}_i) + q(\hat{a}_i)$ . Mais les formes linéaires peuvent également être des intégrales sur un petit voisinage (fuzzy finite elements), des dérivées comme  $p \mapsto \partial_{x_1} p(\hat{a}_i)$  (éléments finis de Hermite), etc...*

**Définition 21 – Unisolvance** Un élément fini  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  est dit unisolvant si pour tout vecteur  $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$ , il existe une unique fonction  $p \in \hat{\mathbb{P}}$  telle que  $p(\hat{a}_i) = \alpha_i$  pour tout  $i \in \llbracket 1, N \rrbracket$ .

En d'autres termes,  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  est unisolvant si le système  $(p(\hat{a}_i))_i = \alpha$  admet une unique solution dans  $\hat{\mathbb{P}}$ . Un élément fini unisolvant est en quelque sorte "cohérent" : ses degrés de liberté permettent d'identifier une unique fonction de base. Par la suite, on travaillera toujours avec de tels éléments.

**Exemples** Considérons l'élément fini  $([0, 1], \{0, 1\}, \text{Vect}\{x \mapsto x, x \mapsto 1 - x\})$ . On a  $N = 2$  et  $\hat{a}_1 = 0$ ,  $\hat{a}_2 = 1$ . L'ensemble  $\hat{\mathbb{P}}$  est l'ensemble des fonctions affines de  $\hat{K}$  dans  $\mathbb{R}$ . Pour  $(\alpha_1, \alpha_2) \in \mathbb{R}^2$ , il existe bien une unique fonction affine  $p$  telle que  $p(0) = \alpha_1$  et  $p(1) = \alpha_2$ , donc l'élément fini est unisolvant. **Contre-exemples** Considérons l'élément fini  $([0, 1], \{1/3, 1/2, 2/3\}, \text{Vect}\{x \mapsto x^n \mid n \in \llbracket 0, 42 \rrbracket\})$ . Pour le vecteur  $\alpha = 0_{\mathbb{R}^3}$ , il existe une infinité de polynômes  $p$  d'ordre compris entre 4 et 42 tels que  $p(1/3) = p(1/2) = p(2/3) = 0$ . Donc l'élément fini n'est pas unisolvant.

Une fois donné un élément fini, on construit un maillage qui permet d'obtenir l'espace d'approximation  $V_h$ .

**Définition 22 – Maillage EF** Soit  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  un élément fini unisolvant. Un maillage élément fini de  $\Omega$  est défini par une famille  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in \llbracket 1, L \rrbracket}$  d'éléments finis, chacun associé à une transformation  $F_\ell : \hat{K} \rightarrow \Omega$ , tels que

- $K_\ell = F_\ell(\hat{K})$ ,
- $F_\ell$  est un difféomorphisme de classe  $\mathcal{C}^1$  (application  $\mathcal{C}^1$  bijective et dont l'inverse est  $\mathcal{C}^1$ ) de  $\hat{K}$  sur  $K_\ell$ ,
- $\Sigma_\ell = \{F_\ell(\hat{a}_i) \mid i \in \llbracket 1, N \rrbracket\}$ ,
- $\mathbb{P}_\ell = \{\hat{p} \circ F_\ell^{-1} : K_\ell \rightarrow \mathbb{R} \mid \hat{p} \in \hat{\mathbb{P}}\}$ ,

et tels que la famille  $K_\ell$  est un maillage de  $\Omega$ , c'est-à-dire

$$\overline{\Omega} = \bigcup_{\ell \in \llbracket 1, L \rrbracket} \overline{K}_\ell, \quad \text{et} \quad \forall \ell \neq \ell', \quad \overset{\circ}{K}_\ell \cap \overset{\circ}{K}_{\ell'} = \emptyset.$$

Dans le cas où les  $K_\ell$  sont des polyèdres, le maillage est dit *conforme* si toute face d'un polyèdre  $K_\ell$  est soit une face d'un autre polyèdre, soit une partie de la frontière de  $\Omega$ . Cette définition vise à interdire les situations où le sommet d'un polyèdre est situé sur l'intérieur d'une face d'un autre polyèdre.

**Remarque 8.** On suppose de manière implicite que le domaine  $\Omega$  peut être maillé de telle manière. Par exemple,  $\Omega$  sera polyédrique si l'on cherche des transformations  $F_\ell$  affines.

**Exemples** Supposons que l'on veuille mailler l'intervalle  $[a, b]$  avec des éléments finis équivalents à  $([0, 1], \{0, 1\}, \{x \mapsto x, x \mapsto 1 - x\})$ . On peut considérer la famille d'applications  $F_\ell : [0, 1] \rightarrow [a + \ell h, a + (\ell + 1)h]$ , où  $\ell \in \llbracket 0, L \rrbracket$  et  $h = (b - a)/(L + 1)$ . Ainsi, on aura  $K_\ell = [a + \ell h, a + (\ell + 1)h]$  un intervalle du maillage de pas  $h$ , et  $\Sigma_\ell = \{a + \ell h, a + (\ell + 1)h\} = \{a_{\ell,0}, a_{\ell,1}\}$ . Pour chaque  $\ell$ , l'inverse de  $F_\ell$  est donnée par  $F_\ell^{-1}(y) = (y - a_{\ell,0})/h$ , d'où  $\mathbb{P}_\ell = \left\{y \mapsto \frac{y - a_{\ell,0}}{h}, y \mapsto 1 - \frac{y - a_{\ell,0}}{h} = \frac{a_{\ell,1} - y}{h}\right\}$ . L'espace  $V_h$  engendré est l'espace des fonctions continues dont la restriction à chaque  $K_\ell$  est linéaire.

**Lemme 7 – Propriétés des transformations** Soit  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  un élément fini unisolvant et  $F : \hat{K} \rightarrow K$  un difféomorphisme. Alors l'élément transformé  $(K, \Sigma, \mathbb{P})$  défini par  $K = F(\hat{K})$ ,  $\Sigma = F(\hat{\Sigma})$  et  $\mathbb{P} = \hat{\mathbb{P}} \circ F^{-1}$  est encore un élément fini unisolvant.

### Démonstration

Premièrement,  $\hat{K} = F^{-1}(K) = \hat{K}$  est fermé et  $F^{-1}$  continue, donc  $K$  est fermé. De plus, l'image d'un connexe par une application continue est elle-même connexe. Enfin, en posant  $\hat{x} = F^{-1}(x)$ , on a

$$\text{Vol}(K) = \int_{x \in K} 1 dx = \int_{\hat{x} \in \hat{K}} 1 |\det \nabla F(\hat{x})| d\hat{x}.$$

Comme  $|\det \nabla F(\hat{x})| > 0$  en tout point, on en déduit  $\text{Vol}(K) > 0$ . Deuxièmement,  $\Sigma$  est une famille de  $N$  points de  $K$ , qui sont

distincts car les  $\hat{a}_i$  sont distincts et  $F$  est bijective. Troisièmement,  $\mathbb{P}$  est une famille d'application de  $K$  dans  $\mathbb{R}$ . C'est bien un espace vectoriel, car si  $p, q \in K$  et  $\lambda \in \mathbb{R}$ , alors

$$\lambda p + q = \lambda \hat{p} \circ F^{-1} + \hat{q} \circ F^{-1} = (\lambda \hat{p} + \hat{q}) \circ F^{-1} \in \mathbb{P}.$$

Montrons que  $\mathbb{P}$  est engendré par la base  $(p_i)_{i \in [1, N]} = (\hat{p}_i \circ F^{-1})_{i \in [1, N]}$ . Premièrement, toute fonction  $q \in \mathbb{P}$  s'écrit  $\hat{q} \circ F^{-1}$  pour un certain  $\hat{q} = \sum_{i=1}^N c_i \hat{p}_i \in \hat{\mathbb{P}}$ . Donc  $q = \sum_{i=1}^N c_i \hat{p}_i \circ F^{-1} = \sum_{i=1}^N c_i p_i$ , et  $\mathbb{P} \subset \text{Vect}\{p_i\}_i$ . D'autre part, s'il existe  $c \in \mathbb{R}^N$  tel que  $\sum_{i=1}^N c_i p_i = 0$ , alors  $\sum_{i=1}^N c_i \hat{p}_i \circ F^{-1} = 0$ . En composant avec  $F$ , on obtient  $\sum_{i=1}^N c_i \hat{p}_i = 0$ , donc  $c_i = 0$  pour tout  $i$  puisque  $(\hat{p}_i)_i$  est une base. En conclusion,  $(p_i)_i$  est une base de  $\mathbb{P}$ , qui est de dimension  $N$ .

Enfin, montrons l'unisolvance par analyse-synthèse. Soit  $\alpha \in \mathbb{R}^N$ , et supposons qu'il existe une fonction  $q \in \mathbb{P}$  telle que  $q(a_i) = \alpha_i$ . Soit  $\hat{q}$  l'unique fonction de  $\hat{\mathbb{P}}$  telle que  $q = \hat{q} \circ F^{-1}$  : on a donc  $\alpha_i = q(a_i) = \hat{q}(F^{-1}(a_i)) = \hat{q}(\hat{a}_i)$ . Par unisolvance de l'élément  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$ , il existe au plus un tel  $\hat{q}$  dans  $\hat{\mathbb{P}}$ , ce qui détermine  $q$ . Pour ce choix de  $\hat{q}$ , la fonction  $q = \hat{q} \circ F^{-1}$  satisfait bien  $q(a_i) = \alpha_i$ , d'où l'unisolvance de  $(K, \Sigma, \mathbb{P})$ .  $\square$

### 3.3.1 Éléments affine-équivalents

Considérons un élément fini de référence  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et un maillage  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in [1, L]}$  comme dans la Définition 22.

**Définition 23 – Éléments affine-équivalents** Deux éléments finis  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)$  sont dits affine-équivalents s'ils peuvent être mis en relation par une transformation  $F_\ell : \hat{K} \rightarrow K_\ell$  affine et bijective.

Les deux éléments  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)$  sont dits *équivalents*, car la bijection  $F_\ell$  permet de passer de l'un à l'autre. Si, de plus, cette bijection est affine, on dit que les deux éléments sont affine-équivalents. L'intérêt de ces transformations réside dans la facilitation des calculs d'intégrale. En effet, en posant  $x = F(\hat{x})$ ,

$$\int_{\hat{x} \in \hat{K}} \hat{\varphi}(\hat{x}) d\hat{x} = \int_{x \in K} \hat{\varphi}(F^{-1}(x)) |\nabla F^{-1}(x)| dx = \int_{x \in K} \varphi(x) |\nabla F^{-1}(x)| dx.$$

S'il se trouve que  $|\nabla F^{-1}(x)|$  est une constante sur  $K$ , comme par exemple pour  $F$  affine, on peut seulement faire des calculs sur l'élément de référence puis utiliser les transformations  $F_\ell$  pour en déduire les valeurs sur chaque élément du maillage. Ce choix est commode à la fois en pratique et dans la théorie, puisqu'il permet des estimations d'erreurs où interviennent des constantes géométriques de déformation du maillage.

**Définition 24 – Mesure de l'écrasement du maillage** Soit  $K \subset \mathbb{R}^d$  (respectivement  $\hat{K}$  ou  $K_\ell$ ). On note  $h \in \mathbb{R}^+$  (respectivement  $\hat{h}$  ou  $h_\ell$ ) le diamètre de  $K$ , c'est-à-dire

$$h := \inf \{d > 0 \mid \text{il existe une boule de diamètre } d \text{ contenant } K\}.$$

On note  $\rho$  (respectivement  $\hat{\rho}$ ,  $\rho_\ell$ ) le diamètre interne de  $K$ , c'est-à-dire

$$\rho := \sup \{d \geq 0 \mid \text{il existe une boule de diamètre } d \text{ contenue dans } K\}.$$

La quantité  $\sigma := \rho/h$  mesure l'écrasement de  $K$ .

La qualité d'un maillage peut être estimée en calculant la constante  $\sigma_\ell$  pour chacun des éléments  $K_\ell$ , voire en imposant  $\sigma_\ell \geq \sigma > 0$  pour un certain  $\sigma$  fixé *a priori*. Plus  $\sigma_\ell$  est faible, plus l'élément est étiré ou écrasé, et le maillage inhomogène.

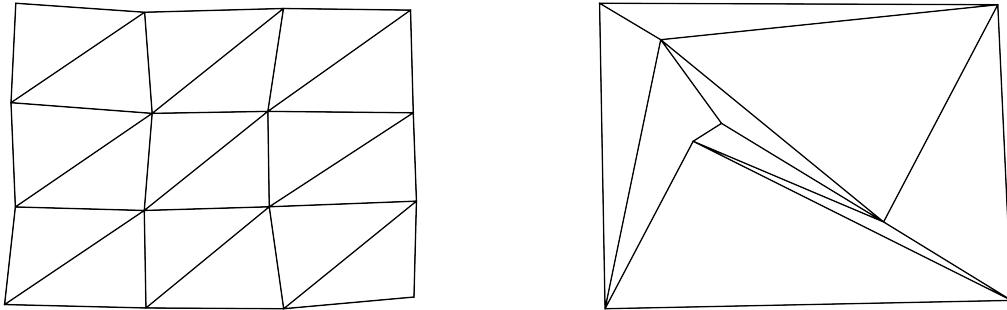
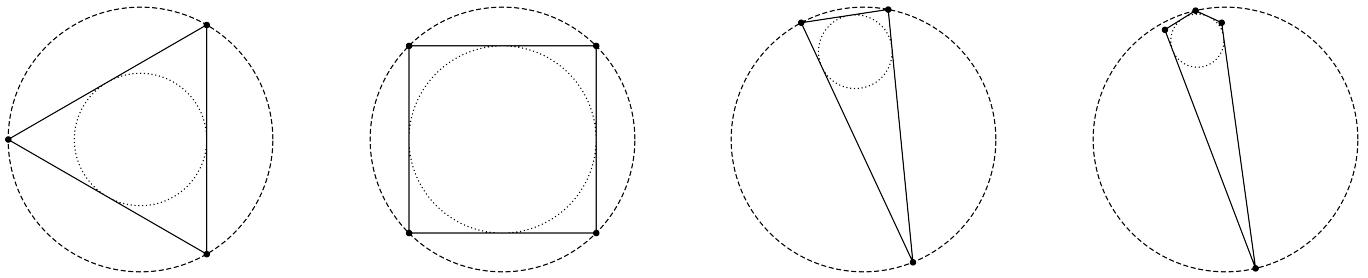


Figure 3.1: Maillage homogène (droite) et non homogène (gauche).

Figure 3.2: Éléments  $K_\ell$  non dégénérés (droite) et dégénérés (gauche).

La valeur  $\rho_\ell$  est le diamètre du cercle inscrit (en pointillés fins), et la valeur  $h_\ell$  est le diamètre du plus petit cercle contenant l'élément (en pointillés épais). Le rapport  $\sigma_\ell = \varepsilon_\ell / h_\ell$  fournit une mesure de la qualité du maillage : plus ce rapport est bas, et plus l'élément est dégénéré.

Dans le cas d'éléments affine-équivalents, les constantes géométriques  $\hat{h}$ ,  $h_\ell$ ,  $\hat{\rho}$  et  $\rho_\ell$  permettent d'estimer les déformations subies.

**Lemme 8 – Estimation des déformations** Soient  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)$  deux éléments affine-équivalents liés par la transformation  $F_\ell : \hat{K} \rightarrow K_\ell$ . Supposons que  $\hat{\rho} > 0$ , ce qui implique  $\rho_\ell > 0$ . Alors

$$\|\nabla F_\ell\| \leq \frac{h_\ell}{\hat{\rho}}, \quad \text{et} \quad \|\nabla F_\ell^{-1}\| \leq \frac{\hat{h}}{\rho_\ell}.$$

### Démonstration

Par définition de la norme d'opérateur, on a

$$\|\nabla F_\ell\| = \sup_{v \in \mathbb{R}^d, |v|=1} |\nabla F_\ell v| = \frac{1}{\hat{\rho}} \sup_{v \in \mathbb{R}^d, |v|=\hat{\rho}} |\nabla F_\ell v|.$$

Si  $|v| = \hat{\rho}$ , on peut trouver deux points  $\hat{x}, \hat{y}$  de  $\hat{K}$  tels que  $v = \hat{y} - \hat{x}$ . En effet, par définition du rayon interne, il existe une boule  $\mathcal{B}(\hat{c}, \hat{\rho}/2)$  contenue dans  $\hat{K}$  : il suffit de prendre  $\hat{y} = \hat{c} + v/2$  et  $\hat{x} = \hat{c} - v/2$ . Donc

$$\|\nabla F_\ell\| \leq \frac{1}{\hat{\rho}} \sup_{\hat{x}, \hat{y} \in \hat{K}, |\hat{y} - \hat{x}| = \hat{\rho}} |\nabla F_\ell \hat{y} - \nabla F_\ell \hat{x}|.$$

Or, comme  $F_\ell$  est affine, on a  $F_\ell(z) = \nabla F_\ell z + w$  pour un certain vecteur  $w$  fixé. Donc  $\nabla F_\ell \hat{y} - \nabla F_\ell \hat{x} = F_\ell(\hat{y}) - F_\ell(\hat{x})$ . Puisque  $F_\ell(\hat{y})$  et  $F_\ell(\hat{x})$  appartiennent à  $K_\ell$ , leur différence est bornée par  $h_\ell$ , et

$$\|F_\ell\| \leq \frac{1}{\hat{\rho}} \sup_{\hat{x}, \hat{y} \in \hat{K}, |\hat{y} - \hat{x}| = \hat{\rho}} |F_\ell(\hat{y}) - F_\ell(\hat{x})| \leq \frac{h_\ell}{\hat{\rho}}.$$

L'autre cas est symétrique. □

### 3.3.2 Le cas de la dimension 1

Soit  $\Omega = [a, b]$  un intervalle non trivial.

**Définition 25 – Base de Lagrange** Soit  $K = [0, 1]$  et  $\Sigma = \{a_1, \dots, a_N\} \subset K$  un ensemble fini de points distincts deux à deux. La *base de Lagrange* associée à  $\Sigma$  est l'ensemble des polynômes  $p_1, \dots, p_N$  de degré inférieur ou égal à  $N$  tels que  $p_i(a_j) = \delta_{ij}$  pour tout  $a_j \in \Sigma$ . Ces polynômes sont explicitement donnés par

$$p_i(x) = \frac{\prod_{j \neq i} (x - x_j)}{\prod_{j \neq i} (a_i - a_j)}.$$

Soit alors  $\mathbb{P} = \text{Vect}\{p_1, \dots, p_N\}$ . L'élément  $(K, \Sigma, \mathbb{P})$  est un élément fini de Lagrange.

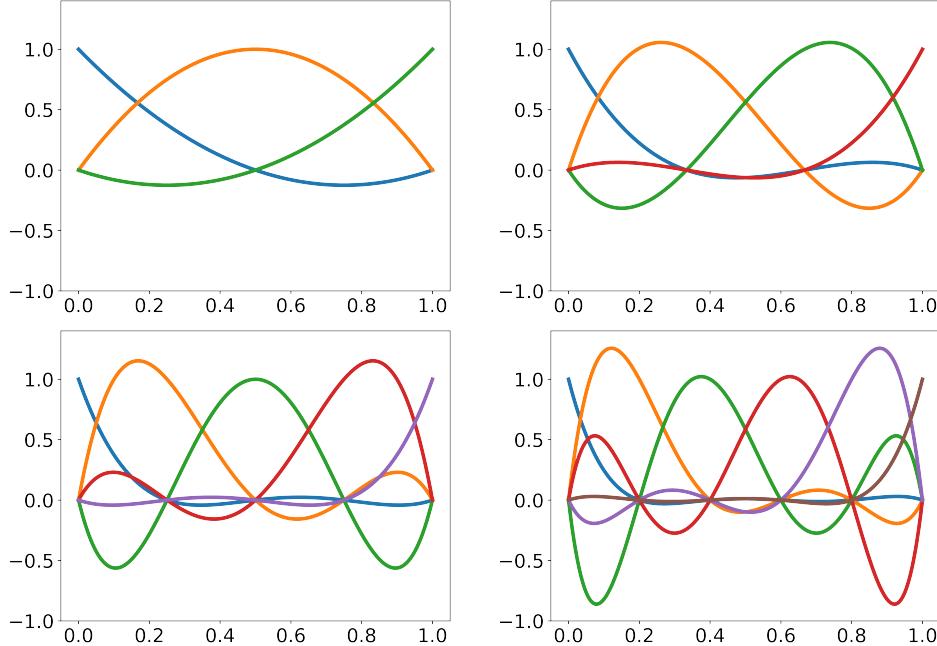


Figure 3.3: Bases de Lagrange pour une distribution homogène des points  $(a_i)_{i \in \llbracket 1, N \rrbracket}$ , avec  $N \in \{3, 4, 5, 6\}$ .

L'élément  $(K, \Sigma, \mathbb{P})$  est bien unisolvant : l'application  $v \mapsto \sum_{i=1}^L v_i p_i$  est une bijection entre  $\mathbb{R}^L$  et  $\mathbb{P}$ .

#### Exemples

- Soit  $K = [a_1, a_2]$  et  $\Sigma = \{a_1, a_2\}$ . La base de Lagrange est donnée par

$$p_1(x) = \frac{x - a_2}{a_1 - a_2}, \quad p_2(x) = \frac{x - a_1}{a_2 - a_1}.$$

Dans le cas particulier  $a_1 = 0$  et  $a_2 = 1$ , on retrouve les polynômes familiers  $p_1(x) = 1 - x$  et  $p_2(x) = x$ .

- Soit  $K = [a_1, a_2, a_3]$  et  $\Sigma = \{a_1, a_2, a_3\}$ . La base de Lagrange est donnée par

$$p_1(x) = \frac{(x - a_2)(x - a_3)}{(a_1 - a_2)(a_1 - a_3)}, \quad p_2(x) = \frac{(x - a_1)(x - a_3)}{(a_2 - a_1)(a_2 - a_3)}, \quad p_3(x) = \frac{(x - a_1)(x - a_2)}{(a_3 - a_1)(a_3 - a_2)}.$$

### 3.3.3 Le cas de la dimension $d$

En dimension supérieure à 1, les ensembles convexes ne sont plus seulement des intervalles. Afin de rester “simple”, les éléments finis sont souvent des transformations affines de triangles ou de carrés (en dimension 2), c'est-à-dire de simplexes ou de parallélétopopes (en dimension arbitraire).

### 3.3.3.1 Éléments finis simpliciaux

**Définition 26 – Simplexe** Notons  $d$  la dimension de l'espace ambiant et  $e_i = (0, 0, \dots, 0, 1, 0, \dots, 0)$  le  $i^{\text{ème}}$  vecteur de la base canonique. Un ensemble  $K \subset \mathbb{R}^d$  est un simplexe (non dégénéré) s'il existe une application affine inversible  $A: \mathbb{R}^d \rightarrow \mathbb{R}^d$  telle que  $K = A\hat{K}$ , où  $\hat{K}$  est l'enveloppe convexe des  $d+1$  points  $\{0, e_1, \dots, e_d\}$ .

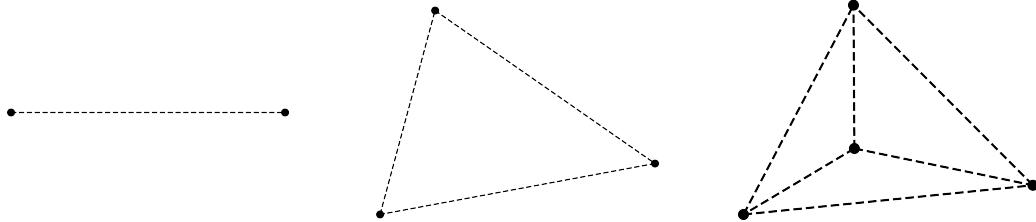


Figure 3.4: Simplexes en dimension  $d \in \{1, 2, 3\}$ .

De manière équivalente, un simplexe non dégénéré est l'enveloppe convexe de  $d+1$  points  $x^0, \dots, x^d$  non situés sur un même hyperplan, i.e. tels que la matrice

$$A = \begin{pmatrix} x_0^0 & \dots & x_0^d \\ \vdots & & \vdots \\ x_d^0 & \dots & x_d^d \\ 1 & \dots & 1 \end{pmatrix}$$

soit inversible.

#### Exemples

- En dimension 1, la condition de non-dégénérescence s'écrit  $x^0 \neq x^1$ . En dimension 2, elle est équivalente à ce que le vecteur  $x^2 - x^1$  ne soit pas colinéaire au vecteur  $x^3 - x^1$ .
- De manière générale,  $|\det A| = |\det \tilde{A}| = |\det B|$ , où les matrices  $\tilde{A}$  et  $B$  sont données par

$$\tilde{A} = \begin{pmatrix} x_0^0 & x_0^1 - x_0^0 & \dots & x_0^d - x_0^0 \\ \vdots & \vdots & & \vdots \\ x_d^0 & x_d^1 - x_d^0 & \dots & x_d^d - x_d^0 \\ 1 & 0 & \dots & 0 \end{pmatrix}, \quad B = \begin{pmatrix} x_0^1 - x_0^0 & \dots & x_0^d - x_0^0 \\ \vdots & & \vdots \\ x_d^1 - x_d^0 & \dots & x_d^d - x_d^0 \end{pmatrix}.$$

On retrouve ainsi l'intuition géométrique qu'un simplexe est non dégénéré si le volume des vecteurs formant ses arêtes est non nul.

Pour manipuler les fonctions de base, les coordonnées cartésiennes ne sont pas les plus adaptées. À la place, on considère les *coordonnées barycentriques*, qui sont les coefficients  $\lambda_0, \dots, \lambda_d$  de la combinaison linéaire  $x = \sum_{i=0}^d \lambda_i a^i$ . Pour fixer l'unicité de ce choix de  $d+1$  variables en dimension  $d$ , on impose  $\sum_{i=0}^d \lambda_i = 1$ . Ce raisonnement mène donc au système

$$A\lambda = \begin{pmatrix} x \\ 1 \end{pmatrix},$$

qui est inversible si le simplexe est non dégénéré. Il se trouve que ce système est inversible pour tout  $x \in \mathbb{R}^d$  : les points du simplexe sont caractérisés par des coordonnées barycentriques toutes positives.

**Définition 27 – Élément fini simplicial d'ordre  $k$**  Soit  $k \in \mathbb{N}_*$ . On définit l'ensemble  $\Sigma$  par

$$\Sigma = \left\{ a = \sum_{i=0}^d \lambda_i x^i \mid \text{Pour tout } i \in \llbracket 0, d \rrbracket, \lambda_i \text{ est de la forme } \frac{m}{k} \text{ pour un certain } m \in \llbracket 0, k \rrbracket, \text{ et } \sum_{i=0}^d \lambda_i = 1 \right\}.$$

Notons  $M \subset \llbracket 0, k \rrbracket^{d+1}$  l'ensemble des multi-indices  $m = (m_0, \dots, m_d)$  tels que  $\sum_{i=1}^d m_i = k$ . Chacun des points  $a \in \Sigma$  est uniquement déterminé par un indice  $m \in M$  tel que  $\lambda_i = m_i/k$  pour tout  $i \in \llbracket 0, d \rrbracket$ . La base lagrangienne est alors définie par les polynômes  $p_m$  pour  $m \in M$  par

$$p_m(x) = \frac{\prod_{i=0}^d \prod_{j < m_i} \left( \lambda_i(x) - \frac{j}{k} \right)}{\prod_{i=0}^d \prod_{j < m_i} \left( \frac{m_i}{k} - \frac{j}{k} \right)}.$$

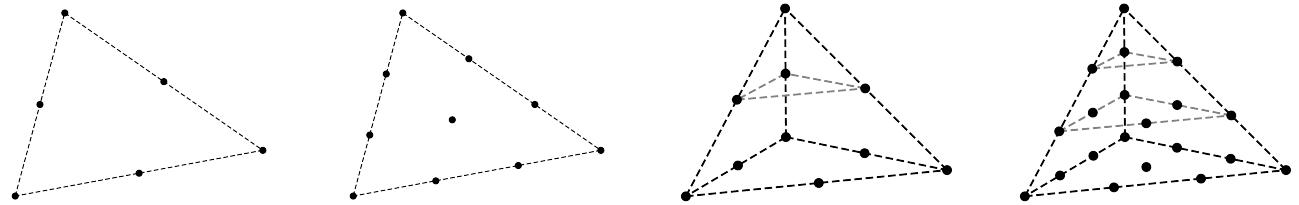


Figure 3.5: Simplexes en dimension  $d \in \{2,3\}$  pour un ordre  $k \in \{2,3\}$ .

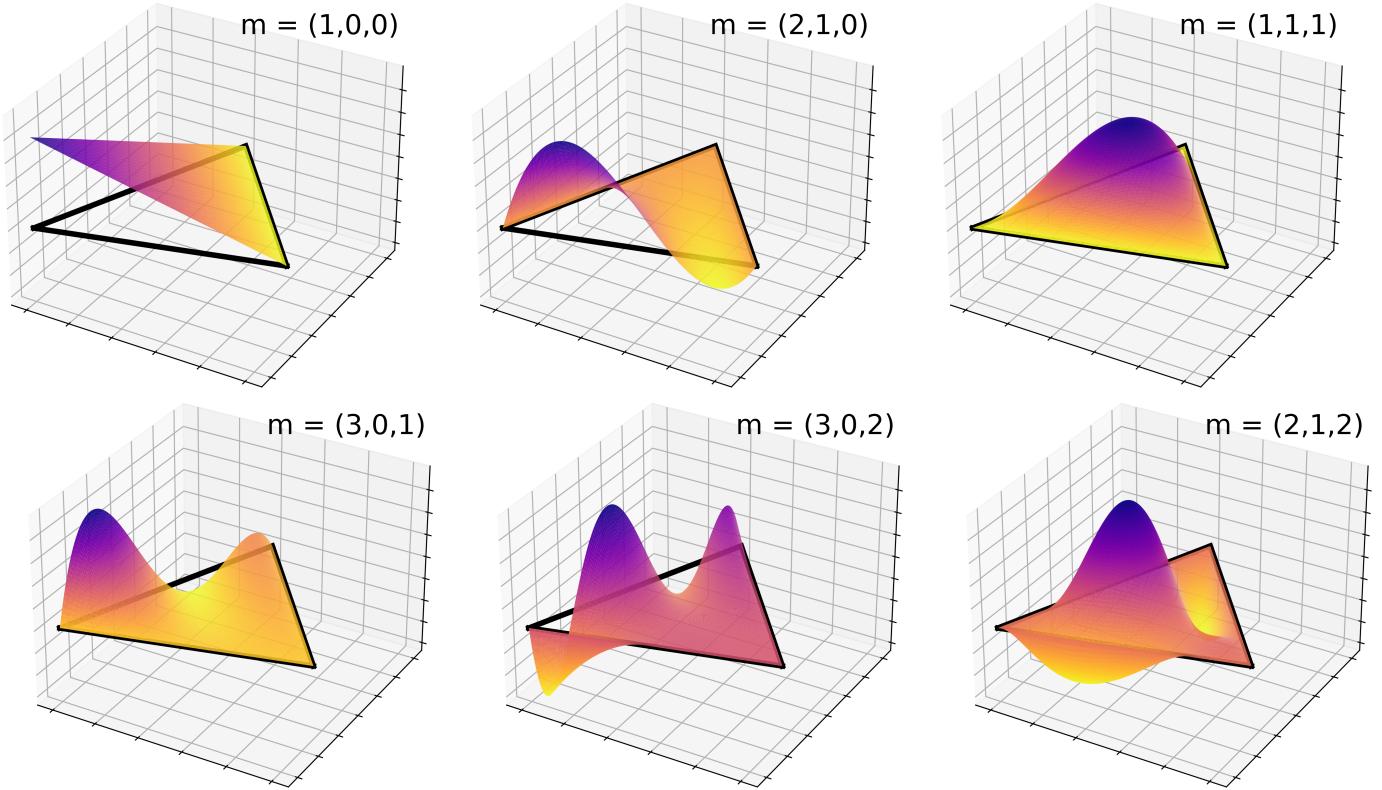


Figure 3.6: Quelques fonctions de base en dimension  $d = 2$ .

### Exemples

- En dimension  $d = 2$ , pour  $k = 1$  : les points  $a_m$  sont indexés par  $m^0 = (1,0,0)$ ,  $m^1 = (0,1,0)$  et  $m^2 = (0,0,1)$ . Cela correspond respectivement aux points  $x^0$ ,  $x^1$  et  $x^2$  qui sont les sommets du triangle. Le premier élément de la base lagrangienne s'écrit

$$p^0(x) = \frac{\prod_{i=0}^d \prod_{j < m_i} \left( \lambda_i(x) - \frac{j}{k} \right)}{\prod_{i=0}^d \prod_{j < m_i} \left( \frac{m_i}{k} - \frac{j}{k} \right)} = \lambda_0(x).$$

On vérifie de même que  $p^1 \equiv \lambda_1$  et  $p^2 \equiv \lambda_2$ .

- En dimension  $d = 2$ , pour  $k = 2$  : les points  $a_m$  sont indexés par  $(2, 0, 0)$ ,  $(1, 1, 0)$ ,  $(1, 0, 1)$ ,  $(0, 2, 0)$ ,  $(0, 1, 1)$  et  $(0, 0, 2)$ . Le premier élément de la base s'écrit

$$p^0(x) = \frac{\prod_{i=0}^d \prod_{j < m_i} \left( \lambda_i(x) - \frac{j}{k} \right)}{\prod_{i=0}^d \prod_{j < m_i} \left( \frac{m_i}{k} - \frac{j}{k} \right)} = \frac{\lambda_0(x) \left( \lambda_0(x) - \frac{1}{2} \right)}{1 \times \left( 1 - \frac{1}{2} \right)} = 2\lambda_0(x)(\lambda_0(x) - 1/2).$$

Pour  $m = (0, 1, 1)$  par exemple, on obtient

$$p(x) = 4\lambda_1(x)\lambda_2(x).$$

La base lagrangienne est constituée de polynômes de  $d$  variables d'ordre **total**  $k$ , c'est-à-dire de la forme

$$p(x) = \sum_{\substack{j_1, \dots, j_d \in \llbracket 0, k \rrbracket \\ \sum_{i=1}^d j_i \leq k}} \alpha_{j_1, \dots, j_d} x_1^{j_1} \cdots x_d^{j_d}.$$

**Lemme 9** La dimension de  $\Sigma$  est  $\frac{(d+k)!}{k!d!} = \binom{d+k}{d}$ .

### Démonstration

Procérons par dénombrement des cas. Considérons le vecteur  $(0, 0, \dots, 0)$  de taille  $d+1$  que l'on veut remplir avec  $k$  unités de masses. Si l'on en place  $i_0$  sur la première coordonnée, il en reste  $k - i_0$  à placer sur les  $d$  suivantes. Si l'on en place  $i_1$  parmi les  $k - i_0$  sur la seconde coordonnée, il en reste  $k - i_0 - i_1$  à placer sur les  $d-1$  suivantes, et ainsi de suite. En arrivant sur la dernière case, il en restera  $k - \sum_{j=0}^{d-1} i_j$  à placer, ce que l'on ne peut faire que d'une seule manière. On a donc

$$\text{Card}\Sigma = \sum_{i_0=0}^k \sum_{i_1=0}^{k-i_0} \cdots \sum_{i_{d-1}=0}^{k-\sum_{j=0}^{d-2} i_j} 1.$$

Pour  $\ell \geq 1$ , notons  $S_{d-\ell}^\alpha := \sum_{i_{d-\ell}=0}^\alpha \sum_{i_{d-\ell+1}=0}^{\alpha-i_{d-\ell}} \cdots \sum_{i_{d-1}=0}^{\alpha-\sum_{j=d-\ell}^{d-2} i_j} 1$ , et montrons par récurrence sur  $\ell$  que

$$S_{d-\ell}^\alpha = \frac{1}{\ell!} \frac{(\alpha+\ell)!}{\alpha!} = \binom{\alpha+\ell}{\ell}.$$

On a pour  $\ell = 1$  que  $S_{d-1}^\alpha = \alpha + 1$ , ce qui fournit l'initialisation. Supposons la propriété vraie pour  $\ell - 1 \geq 1$  donné, et considérons

$$S_{d-\ell}^\alpha = \sum_{i_{d-\ell}=0}^\alpha S_{d-\ell+1}^{\alpha-i_{d-\ell}} = \sum_{i_{d-\ell}=0}^\alpha \binom{\alpha-i_{d-\ell}+\ell-1}{\ell-1} = \sum_{p=0}^\alpha \binom{p+\ell-1}{\ell-1}.$$

Fixons  $\ell$  et procérons par récurrence sur  $\alpha$ . Pour  $\alpha = 0$ , on a  $\binom{\alpha+\ell}{\ell} = 1 = \binom{\alpha+\ell-1}{\ell-1}$ . Si maintenant  $\binom{\alpha+\ell}{\ell} = S_{d-\ell}^\alpha$  pour un certain  $\alpha \geq 0$ , alors

$$S_{d-\ell}^{\alpha+1} = \sum_{p=0}^{\alpha+1} \binom{p+\ell-1}{\ell-1} = S_{d-\ell}^\alpha + \binom{\alpha+1+\ell-1}{\ell-1} = \binom{\alpha+\ell}{\ell} + \binom{\alpha+\ell}{\ell-1} = \frac{(\alpha+\ell)!}{\ell!\alpha!} + \frac{(\alpha+\ell)!}{(\ell-1)!(\alpha+1)!} = \frac{(\alpha+1)(\alpha+\ell)! + \ell(\alpha+\ell)!}{\ell!(\alpha+1)!} = \binom{\alpha+1+\ell}{\ell}.$$

Donc l'égalité est vérifiée pour tout  $\alpha$ . En conséquence,  $S_{d-\ell}^\alpha = \binom{\alpha+\ell}{\ell}$  pour tout  $\ell$ , et en prenant  $\ell = d$ , on obtient

$$\text{Card}\Sigma = S_0^k = \binom{k+d}{d}.$$

D'où la propriété. □

**Lemme 10 – Unisolvance, cas simplicial** L'élément fini simplicial de degré  $k$  donné par la Définition 27 est unisolvant.

### Démonstration

Montrons que la base considérée vérifie bien la propriété de base lagrangienne, i.e.  $p_m(a_n) = \delta_{m,n}$  pour tout multi-indices  $m, n \in M$ . On a d'abord

$$p_m(a_m) = \frac{\prod_{i=0}^d \prod_{j < m_i} \left( \frac{m_i}{k} - \frac{j}{k} \right)}{\prod_{i=0}^d \prod_{j < m_i} \left( \frac{m_i}{k} - \frac{j}{k} \right)} = 1.$$

D'autre part, pour tout  $n \in M$  distinct de  $m$ , il existe  $i \in \llbracket 0, d \rrbracket$  et  $j < m_i$  tel que  $n_i = j$ . Sinon, par contradiction,  $n_i \geq m_i$  pour tout  $i$ . Comme  $n \neq m$ , il existe  $i_0 \in \llbracket 0, d \rrbracket$  tel que l'inégalité soit stricte, i.e/  $n_{i_0} \geq m_{i_0} + 1$ . En sommant sur les indices  $i$ , on obtient

$$k = \sum_{i=0}^d n_i \geq \sum_{i=0}^d m_i + \delta_{i,i_0} = k + 1,$$

ce qui est absurde. Donc pour un certain indice  $i \in \llbracket 0, d \rrbracket$ , le terme  $\prod_{j < m_i} (\lambda_i(a_n) - j/k) = \prod_{j < m_i} (n_i/k - j/k)$  s'annule, et  $p_m(a_n) = 0$ .

On en déduit que  $(K, \Sigma, \mathbb{P})$  est unisolvant : en effet, indexons par  $j \in \llbracket 1, N \rrbracket$  les multi-indices de  $M$ . Pour tout vecteur  $\alpha \in \mathbb{R}^N$ , il existe une unique fonction  $q \in \text{Vect}\{p_m \mid m \in M\}$  telle que  $q(a_{m_j}) = \alpha_j$  : il faut (et il suffit de) prendre  $q = \sum_{j=1}^N \alpha_j p_{m_j}$ .  $\square$

#### 3.3.3.2 Éléments finis parallélétopes

**Définition 28 – Parallélétope** Un ensemble  $K \subset \mathbb{R}^d$  est un parallélétope s'il existe une application affine inversible  $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$  telle que  $K = A\hat{K}$ , où  $\hat{K}$  est le cube unité possédant  $0, e_1, \dots, e_d$  pour sommets. On note  $x^0 = A(0)$  et  $x^i = A(e_i)$  pour  $i \in \llbracket 1, d \rrbracket$ .

La manipulation des parallélétopes est plus simple que celle des simplexes, mais il est plus compliqué de mailler un domaine avec de tels ensembles. On considère le système de coordonnée suivant : un point  $x \in K$  sera identifié par  $x = x^0 + \sum_{i=1}^d \mu_i(x^i - x^0)$ , où les coordonnées  $(\mu_i)_{i \in \llbracket 1, d \rrbracket}$  appartiennent à  $[0, 1]$ .

**Définition 29 – Élément fini parallélétope d'ordre  $k$**  Soit  $k \in \mathbb{N}_*$ . On définit l'ensemble  $\Sigma$  par

$$\Sigma := \left\{ a = x^0 + \sum_{i=1}^d \mu_i(x^i - x^0) \mid \text{Pour tout } i \in \llbracket 1, d \rrbracket, \mu_i \text{ est de la forme } \frac{m}{k} \text{ pour un certain } m \in \llbracket 0, k \rrbracket \right\}.$$

Notons  $M \subset \llbracket 0, k \rrbracket^d$ . Chacun des points  $a \in \Sigma$  est uniquement déterminé par un indice  $m \in M$ , tel que  $\mu_i = m_i/k$  pour tout  $i \in \llbracket 1, d \rrbracket$ . La base lagrangienne est alors définie par les polynômes  $p_m$  pour  $m \in M$  par

$$p_m(x) = \frac{\prod_{i=1}^d \prod_{\substack{j=1 \\ j \neq m_i}}^d \left( \mu_i(x) - \frac{j}{k} \right)}{\prod_{i=1}^d \prod_{\substack{j=1 \\ j \neq m_i}}^d \left( \frac{m_i}{k} - \frac{j}{k} \right)}.$$

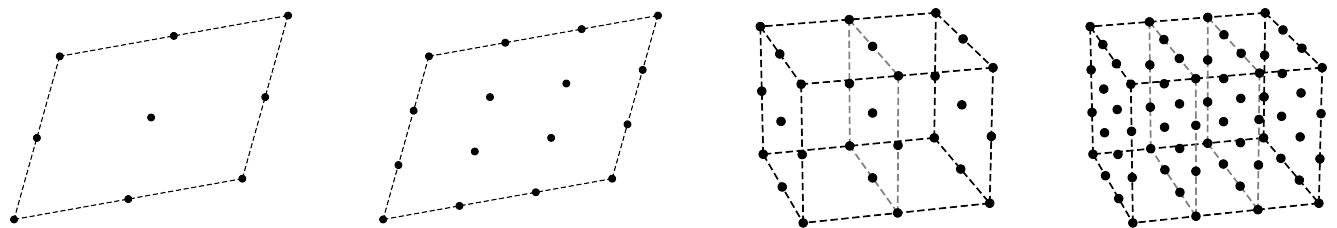


Figure 3.7: Parallélétopes en dimension  $d \in \{2, 3\}$  pour un ordre  $k \in \{2, 3\}$ .

La base lagrangienne est constituée de polynômes d'ordre inférieur ou égal à  $k$  **sur chacune de leurs variables**, c'est-à-dire de

la forme

$$p(x) = \sum_{j_1, \dots, j_d \in \llbracket 0, k \rrbracket} \alpha_{j_1, \dots, j_d} x_1^{j_1} \cdots x_d^{j_d}.$$

La dimension de cet espace, ainsi que le cardinal de  $\Sigma$ , coïncident et valent  $(k+1)^d$ .

**Lemme 11 – Unisolvance, cas paralléléotope** L'élément fini paralléléotope donné par la Définition 29 est unisolvant.

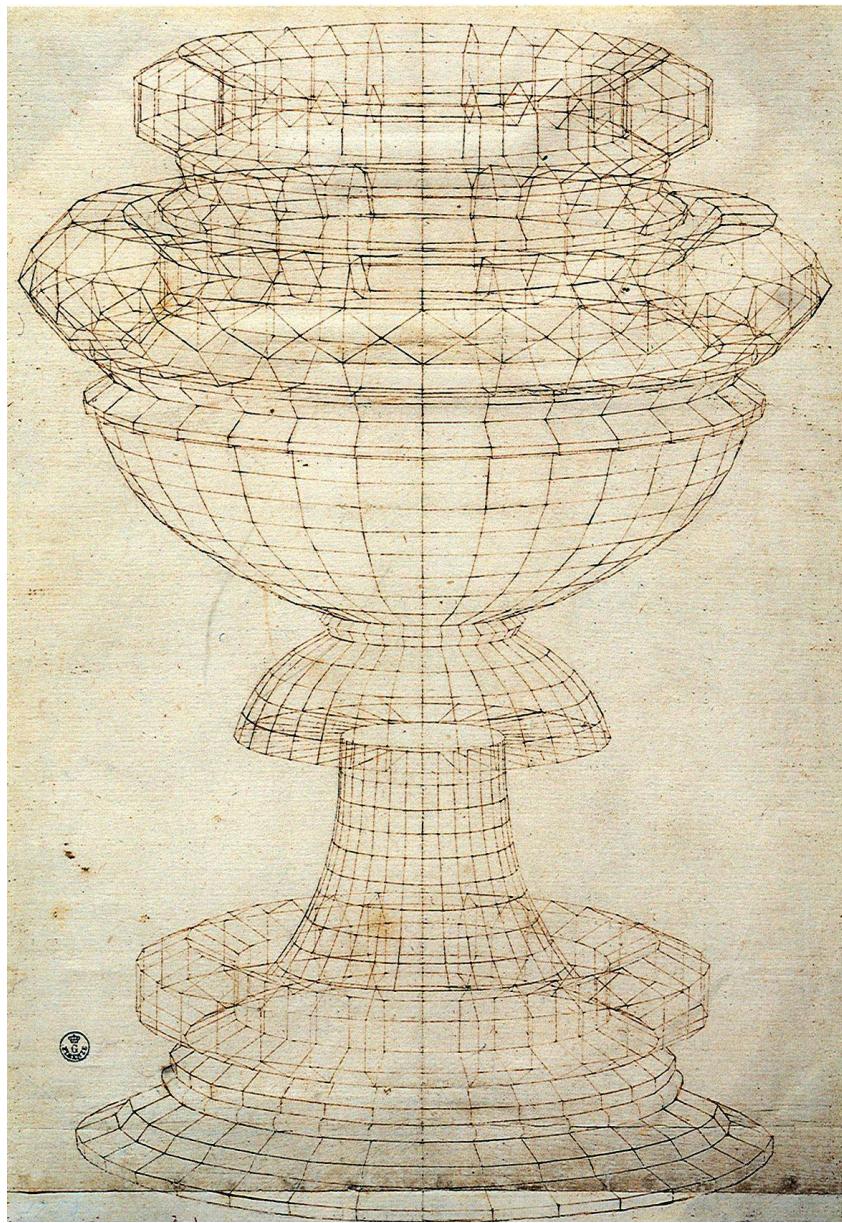
### Démonstration

Il suffit de montrer que la base des  $(p_m)_{m \in \mathbb{M}}$  vérifie la propriété de Lagrange. Hors, pour  $m \in M$  et  $a = x^0 + \sum_{i=1}^d \frac{m_i}{k} (x^i - x^0)$ , on a

$$p_m(a) = \frac{\prod_{i=1}^d \prod_{\substack{j=1 \\ j \neq m_i}}^d \left( \mu_i(x) - \frac{j}{k} \right)}{\prod_{i=1}^d \prod_{\substack{j=1 \\ j \neq m_i}}^d \left( \frac{m_i}{k} - \frac{j}{k} \right)} = 1.$$

D'autre part, si  $n \neq m$  est un élément de  $M$ , alors il existe  $i_0$  tel que  $m_{i_0} \neq n_{i_0}$  et le terme  $\mu_{i_0}(a_n) - \frac{n_{i_0}}{k} = 0$  annule le numérateur de  $p_m(a_n)$ .  $\square$

**Remarque 9** (Bien d'autres univers). *Les maillages sont aussi variés que les pièces à mailler. On peut imaginer mêler plusieurs types de mailles, avec plusieurs ordres d'approximation, ou des raffinements locaux... Bien entendu, cela demande plus de travail pour s'assurer que l'information circule de la bonne manière entre les mailles. Notons que l'idée visuelle d'un maillage était déjà présente au moyen âge, comme l'atteste cette étude de Paolo Uccello, 1450 !*



### 3.4 Construction de l'espace d'approximation

Une fois donné un maillage, on construit l'espace d'approximation  $V_h$ . Dans l'interpolation de Lagrange, un choix classique est le suivant.

**Définition 30 – Espace d'approximation associé à un maillage** Soit  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in [1, L]}$  un maillage du domaine  $\Omega$ , et  $h$  le diamètre maximal des éléments  $(K_\ell)_\ell$ . On pose

$$V_h := \{v \in \mathcal{C}(\Omega; \mathbb{R}) \mid \text{Pour tout } \ell \in [1, L], \text{ la restriction } v|_{K_\ell} \text{ appartient à } \mathbb{P}_\ell\}.$$

#### Exemples

- Soit  $\Omega = [a, b]$ , maillé par des éléments affine-équivalents à l'élément fini  $([0, 1], \{0, 1\}, \{x \mapsto x, x \mapsto 1 - x\})$ . Notons  $(K_i)_{i \in [0, L]}$  les intervalles du maillage, avec  $K_i = [a + ih, a + (i + 1)h]$ . L'espace  $V_h$  est l'espace des fonctions continues dont la restriction à chaque  $K_i$  est linéaire, donc déterminée par la valeur de  $u$  aux extrémités de  $K_i$ . Comme  $u$  est continue, elle s'identifie à

l'interpolation linéaire de ses valeurs aux points  $(a + i_h)_{i \in \llbracket 0, L+1 \rrbracket}$ . Donc la dimension de  $V_h$  est  $L+2$ .

L'espace  $V_h$  est de dimension finie, car la restriction de chaque  $v$  à un élément  $K_\ell$  est caractérisée par un vecteur de dimension  $N$ , donc l'ensemble  $V_h$  est de dimension au plus  $L \times N$ . Dans la pratique, la dimension peut être moindre, car les éléments  $K_\ell$  peuvent partager des degrés de liberté via la condition de continuité.

**Remarque 10** (Continuité). *Il est cohérent de prendre des fonctions de base continues si les solutions de l'équation que l'on cherche à approcher sont elles-mêmes continues. Cependant, ce n'est pas la seule option : les méthodes de Galerkin discontinues peuvent être utilisées, par exemple pour utiliser des projections sur chaque élément du maillage sans condition de raccord. Cela permet de réduire le coût de calcul, puisque chaque  $K_\ell$  est traité indépendamment des autres, mais rend l'analyse plus ardue.*

Dans le cadre de l'interpolation de Lagrange, on construit une base explicite de  $V_h$  de la manière suivante.

**Définition 31 – Base de  $V_h$  pour l'interpolation de Lagrange** Soit  $\Sigma = \bigcup_{\ell \in \llbracket 1, L \rrbracket} \Sigma_\ell$  l'union de tous les degrés de liberté du maillage. Renumérotons  $\Sigma = \{a_j \mid j \in \llbracket 1, J \rrbracket\}$  de manière globale : à chaque  $j \in \llbracket 1, J \rrbracket$  un indice global, il correspond au moins une paire  $(\ell, i) \in \llbracket 1, L \rrbracket \times \llbracket 1, N \rrbracket$  telle que  $a_j$  est le  $i^{\text{ème}}$  point de  $\Sigma_\ell$ . On note  $(\omega_j)_{j \in \llbracket 1, J \rrbracket}$  la famille des fonctions continues telles que

$$\omega_j(a_{j'}) = \delta_{j,j'} \text{ pour tous } j, j' \in \llbracket 1, J \rrbracket, \quad \forall \ell \in \llbracket 1, L \rrbracket, \quad \text{la restriction de } \omega_j \text{ à } K_\ell \text{ appartient à } \mathbb{P}_\ell.$$

On admet que l'élément fini  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$ , ainsi que la formation du maillage, permettent l'existence de telles fonctions. C'est en particulier le cas pour les maillages conformes et les exemples d'éléments finis développés plus haut.

### Exemples

- En dimension 1, pour  $N=2$ , les fonctions  $\omega_j$  sont les fameuses fonctions chapeau.
- En dimension supérieure, pour un simplexe muni d'un élément fini de Lagrange tel que donné par la Définition 27, ces fonctions sont bien continues.

### Contre-exemples

- En dimension 1, s'il n'y a pas de points  $\hat{a}_i$  sur les bords de l'intervalle  $[0, 1]$ , alors les fonctions de base lagangianes ne pourraient pas être continues.
- En dimension supérieure, si les points des arêtes ne coïncident pas, le même problème de continuité se pose.

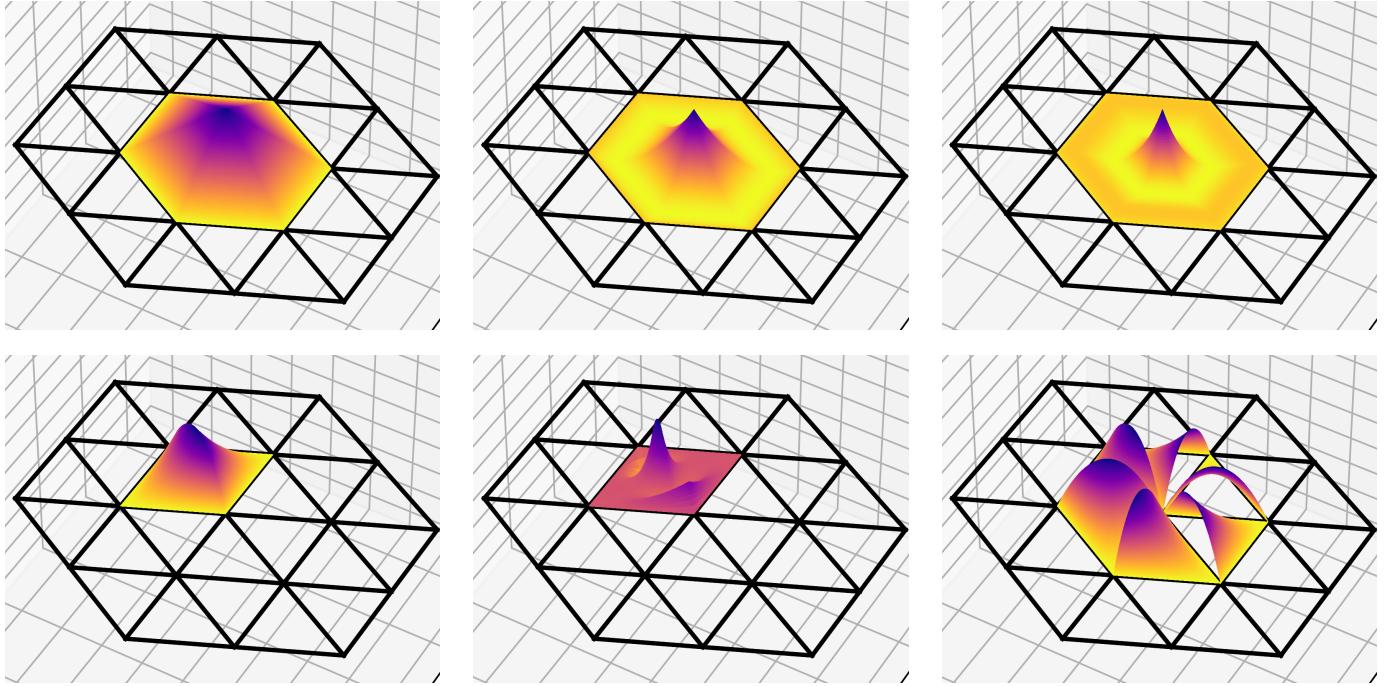


Figure 3.8: Exemples de fonctions de base sur un maillage triangulaire.

De gauche à droite et de haut en bas : recollement des fonctions de base associée au point central pour  $k \in \{1, 2, 3\}$ , puis recollement de seulement deux fonctions sur une arête commune pour  $k \in \{2, 5\}$ , et recollement avec discontinuité.

On appelle *support* d'une fonction de base  $\omega_j$  l'ensemble des éléments  $K_\ell$  sur lesquels  $\omega_j$  ne s'annule pas identiquement. Dans chacun des exemples que nous avons vus (dimension 1, éléments simpliciaux et éléments parallélétopes), le support de  $\omega_j$  est composé des triangles  $K_\ell$  auxquels appartient  $a_j$ . C'est donc un nombre relativement faible. En particulier, l'intégrale

$$\int_{x \in \Omega} \omega_j(x) dx$$

se décompose en une somme d'intégrale sur chaque  $K_\ell$  appartenant au support de  $\omega_j$ , somme qui contient relativement peu de termes. Cet avantage prendra tout son sens dans la partie suivante, où il permettra de construire des matrices creuses.

## 3.5 Étude d'erreur

Grâce au lemme de Céa (Lemme 6), on sait que l'erreur entre la solution exacte  $u$  et son approximation  $u_h$  est contrôlée par l'erreur entre  $u$  et son projeté  $\Pi_h u$  sur  $V_h$ . En particulier, si l'on trouve un élément  $\tilde{u}_h \in V_h$  tel que  $\|u - \tilde{u}_h\|_V$  tende vers 0 quand  $h$  tend vers 0, alors on aura montré la convergence. Dans le cadre des éléments finis de Lagrange,  $\tilde{u}_h$  peut par exemple être choisi comme l'interpolation de  $u$ .

### 3.5.1 Erreur d'interpolation des éléments finis de Lagrange

La démarche que nous suivrons est la suivante : on cherche à obtenir une estimation en norme  $L^2$  et en norme  $H^1$ , tout en supposant que nos éléments finis peuvent représenter exactement des polynômes de degré inférieur ou égal à 1. Pour simplifier les écritures, on note  $m \in \{0, 1\}$  l'ordre de l'erreur que l'on cherche (ordre 0 pour la norme  $L^2 = H^0$ , et ordre 1 pour la norme  $H^1$ ).

1. On commence par montrer que sur l'élément de référence  $\hat{K}$ , la norme  $H^m$  de l'erreur entre  $u \in H^2$  et son interpolée est contrôlée par la semi-norme dans  $H^2$  de  $u$ . En particulier, l'estimation ne dépend plus de l'interpolée.
2. On en déduit une estimation sur n'importe quel élément  $K_\ell$  en utilisant les relations du Lemme 8. Le point clé ici est que l'on utilisera des normes différentes à l'aller et au retour : norme  $H^m$  à l'aller, ce qui fera sortir un terme en  $(\hat{h}/\rho_\ell)^m$ , puis norme  $H^2$  au retour, ce qui fait sortir un  $(h_\ell/\hat{\rho})^2$ . Pour un maillage suffisamment régulier, cela permet de faire sortir un terme en  $h^{2-m}$ .

3. Par la relation de Chasles, l'erreur globale se décompose comme une somme des erreurs locales, donc l'estimation précédente permet d'obtenir une borne sur l'erreur globale.

On commence par étudier la régularité de l'opérateur d'interpolation.

**Lemme 12 – Continuité de l'interpolation** Soit  $d \in \{1, 2\}$  et  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  un élément fini simplicial de type  $\mathbb{P}_1$ , c'est-à-dire  $\hat{\Sigma} = \{\hat{a}_i\}_{i=0}^d$  et  $\hat{K} = \text{conv } \hat{\Sigma}$ . Alors l'opérateur d'interpolation est bien défini, linéaire et continu de  $H^2(\hat{K}; \mathbb{R})$  dans  $H^1(\hat{K}; \mathbb{R})$ .

### Démonstration

Grâce à la Proposition 5 et au fait que  $2 > d/2$ , l'espace  $H^2$  s'injecte continument dans  $C(\hat{K})$ , i.e. il existe une constante  $\bar{C}_{\hat{K}, d}$  telle que

$$|u|_{C(\hat{K}; \mathbb{R})} = \sup_{x \in \hat{K}} |u(x)| \leq \bar{C}_{\hat{K}, d} \|u\|_{H^2}.$$

Ceci implique que tout  $u \in H^2$  est en particulier une fonction continue, donc que les termes  $u(\hat{a}_i)$  sont bien définis. La linéarité est évidente par la formule explicite. Pour montrer la continuité, remarquons que

$$\|I_{\mathbb{P}} u\|_{H^1(\hat{K}; \mathbb{R})} = \left\| \sum_{i=0}^d u(\hat{a}_i) p_i \right\|_{H^1(\hat{K}; \mathbb{R})} \leq \sum_{i=0}^d |u(\hat{a}_i)| \|p_i\|_{H^1(\hat{K}; \mathbb{R})} \leq \left( \sum_{i=0}^d \|p_i\|_{H^1(\hat{K}; \mathbb{R})} \right) |u|_{C(\hat{K}; \mathbb{R})}.$$

Ainsi, en posant  $C_{\hat{K}, d} := \left( \sum_{i=0}^d \|p_i\|_{H^1(\hat{K}; \mathbb{R})} \right) \times \bar{C}_{\hat{K}, d}$ , on obtient bien

$$\|I_{\mathbb{P}} u\|_{H^1(\hat{K}; \mathbb{R})} \leq C_{\hat{K}, d} \|u\|_{H^2(\hat{K}; \mathbb{R})}.$$

Naturellement, comme  $\|I_{\mathbb{P}} u\|_{H^1(\hat{K}; \mathbb{R})} \geq \|I_{\mathbb{P}} u\|_{L^2(\hat{K}; \mathbb{R})}$ , l'opérateur  $I_{\mathbb{P}}$  est également continu de  $H^1$  dans  $L^2$ .  $\square$

On note maintenant  $|u|_{\text{semi}, H^m}$  la semi-norme  $H^m$ , qui vaut soit la norme  $L^2$  pour  $m=0$ , soit

$$|u|_{\text{semi}, H^m} = \sum_{\substack{\alpha \in \llbracket 1, m \rrbracket^d \\ \sum_{j=1}^d \alpha_j = m}} \|D^\alpha u\|_{L^2}$$

pour  $m > 1$ . En français, la semi-norme d'ordre  $m$  contient toutes les normes des dérivées d'ordre exactement  $m$ .

**Lemme 13 – Estimation sur l'élément de référence** Soit  $d \in \{1, 2\}$ ,  $m \in \{0, 1\}$  et  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  l'élément fini de Lagrange simplicial d'ordre 1. Alors il existe une constante  $C_{d, m}$  telle que pour toute fonction  $u \in H^2(\hat{K}; \mathbb{R})$ ,

$$|u - I_{\mathbb{P}} u|_{\text{semi}, H^m} \leq C_{d, m} |u|_{\text{semi}, H^2(\hat{K}; \mathbb{R})}.$$

### Démonstration

Notons que tout polynôme  $p$  d'ordre inférieur ou égal à 1 est exactement représenté par  $I_{\mathbb{P}}$ , i.e.  $I_{\mathbb{P}} p = p$ . Ainsi

$$u - I_{\mathbb{P}} u = u - p + I_{\mathbb{P}}(u - p) = (I - I_{\mathbb{P}})(u - p).$$

Par le Lemme 12, l'application  $I - I_{\mathbb{P}}$  est continue de  $H^2$  dans  $H^m$ , donc il existe une constante  $C_{d, m}$  telle que

$$|(I - I_{\mathbb{P}})(u - p)|_{\text{semi}, H^m} \leq \|(I - I_{\mathbb{P}})(u - p)\|_{H^m} \leq C_{d, m} \|u - p\|_{H^2(\hat{K}; \mathbb{R})}.$$

Ainsi, en prenant l'infimum sur tous les polynômes  $p$  d'ordre inférieur ou égal à 1,

$$|u - I_{\mathbb{P}} u|_{\text{semi}, H^m} \leq C_{d, m} \inf_{p \text{ poly d'ordre } \leq 1} \|u - p\|_{H^2(\hat{K}; \mathbb{R})}.$$

En utilisant l'inégalité de Poincaré-Friedrichs (vue en TD, voir également [Nic00, Théorème 3.48 et Corollaire 3.52]), il existe une (autre) constante telle que  $\inf_{p \text{ poly d'ordre } \leq 1} \|u - p\|_{H^2(\hat{K}; \mathbb{R})} \leq A |u|_{\text{semi}, H^2}$  pour tout  $u \in H^2$ , d'où la conclusion.  $\square$

Le Lemme 13 permet de passer d'une semi-norme  $H^m$  à une semi-norme  $H^2$ . L'intérêt est d'utiliser les propriétés suivantes des changements de variables entre chaque  $\hat{K}$  et  $K_\ell$ .

**Lemme 14 – Estimations des changements de variable** Soient  $\hat{K}$  et  $K_\ell$  deux domaines affine-équivalents liés par la transformation affine et bicontinue  $F_\ell : \hat{K} \rightarrow K_\ell$ . Notons  $\hat{u}(\hat{x}) = u(F_\ell(\hat{x}))$ . Alors pour  $k \in \{0, 1, 2\}$ , il existe une constante  $C_{k,d}$  telle que

$$|\hat{u}|_{\text{semi}, H^k(\hat{K}; \mathbb{R})} \leq C_{k,d} \|\nabla F_\ell\|^k |\det \nabla F_\ell^{-1}| |u|_{\text{semi}, H^k(K_\ell; \mathbb{R})}.$$

En intervertissant les rôles de  $K_\ell$  et  $\hat{K}$ , on déduit également que

$$|u|_{\text{semi}, H^k(K_\ell; \mathbb{R})} \leq C_{k,d} \|\nabla F_\ell^{-1}\|^k |\det \nabla F_\ell| |\hat{u}|_{\text{semi}, H^k(\hat{K}; \mathbb{R})}.$$

### Démonstration

Pour  $k=0$ , on a directement par changement de variable  $x = F_\ell(\hat{x})$

$$|\hat{u}|_{\text{semi}, H^0(\hat{K}; \mathbb{R})}^2 = \int_{\hat{x} \in \hat{K}} |u(F_\ell(\hat{x}))|^2 d\hat{x} = \int_{x \in K_\ell} |u(x)|^2 |\det \nabla F_\ell^{-1}| dx.$$

Comme  $F_\ell$  est linéaire,  $|\det \nabla F_\ell^{-1}|$  est une constante, et on obtient l'estimation souhaitée. Pour  $k=1$ , remarquons que

$$\partial_{\hat{x}_j} \hat{u}(\hat{x}) = \sum_{i=1}^d \partial_{x_i} u(F_\ell(\hat{x})) \partial_{\hat{x}_j} (F_\ell)_i(\hat{x}) = (\nabla F_\ell^t \nabla u(F_\ell(\hat{x})))_j$$

Ainsi

$$\|\nabla_{\hat{x}} \hat{u}(\hat{x})\| = \|\nabla F_\ell^t \nabla u(F_\ell(\hat{x}))\| \leq \|\nabla F_\ell\| \|\nabla u(F_\ell(\hat{x}))\|.$$

Via le même changement de variable, on obtient

$$\begin{aligned} |\hat{u}|_{\text{semi}, H^1(\hat{K}; \mathbb{R})}^2 &= \int_{\hat{x} \in \hat{K}} \|\nabla_{\hat{x}} \hat{u}(\hat{x})\|^2 d\hat{x} \leq \int_{\hat{x} \in \hat{K}} \|\nabla F_\ell\|^2 \|\nabla u(F_\ell(\hat{x}))\|^2 d\hat{x} \\ &= \|\nabla F_\ell\|^2 \int_{x \in K_\ell} \|\nabla u(x)\|^2 |\det \nabla F_\ell^{-1}| dx = \|\nabla F_\ell\|^2 |\det \nabla F_\ell^{-1}| |u|_{\text{semi}, H^1}^2. \end{aligned}$$

En poursuivant ce raisonnement avec les dérivées d'ordre 2 (et en remarquant que  $\partial_{\hat{x}_i} \partial_{\hat{x}_j} F_\ell(\hat{x}) = 0$  pour tout  $i, j$ , car  $F_\ell$  est affine), on obtient le résultat pour  $k=2$  (voir l'appendice 3.6 pour plus de détails).  $\square$

On en arrive donc aux estimations d'erreur.

**Théorème 8 – Estimation d'erreur pour les éléments  $\mathbb{P}_1$**  Soit  $d \in \{1, 2\}$  et  $m \in \{0, 1\}$ . Considérons un maillage  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in \llbracket 1, L \rrbracket}$  d'un domaine  $\Omega \subset \mathbb{R}^d$  par des éléments finis tous affine-équivalents à l'élément de Lagrange simplicial  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  d'ordre 1. Soit  $I_{\text{glob}} u$  l'interpolation globale de  $u$ , i.e. l'unique fonction continue telle que la restriction de  $I_{\text{glob}} u$  à chaque  $K_\ell$  coïncide avec l'interpolation de  $u|_{K_\ell}$ .

On suppose que le maillage est régulier, c'est-à-dire qu'il existe une constante  $\sigma$  telle que  $h_\ell \leq \sigma \rho_\ell$  pour tout  $\ell \in \llbracket 1, L \rrbracket$ , où  $h_\ell$  est le diamètre de  $K_\ell$  et  $\rho_\ell$  son diamètre interne (voir Définition 24). Alors il existe une constante  $C_{d,m}$  telle que pour tout  $u \in H^2(\hat{K}; \mathbb{R})$ ,

$$|u - I_{\text{glob}} u|_{\text{semi}, H^m(\Omega; \mathbb{R})} \leq C_{d,m,\sigma} \left( \sup_{\ell \in \llbracket 1, L \rrbracket} h_\ell \right)^{2-m} |u|_{\text{semi}, H^2(\Omega; \mathbb{R})}.$$

Ce théorème indique que les approximations par polynômes d'ordre 1 “approchent l'espace  $H^2$ ” avec un ordre 2 en norme  $L^2$ , et un ordre 1 en norme  $H^1$ .

### Démonstration

Tout d'abord, l'erreur globale sur  $\Omega$  se décompose en somme d'erreurs locales par la décomposition de Chasles :

$$|u - I_{\text{glob}} u|_{\text{semi}, H^m(\Omega; \mathbb{R})}^2 = \sum_{\ell \in \llbracket 1, L \rrbracket} |u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})}^2.$$

Soit  $\ell$  fixé. Par le Lemme 14 pour  $k=m \in \{0, 1\}$ , on a

$$|u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} \leq C_{m,d} \|\nabla F_\ell^{-1}\|^m |\det \nabla F_\ell|^{1/2} \widehat{|u - I_{K_\ell} u|}_{\text{semi}, H^m(\hat{K}; \mathbb{R})}.$$

Remarquons que

$$\widehat{u - I_{K_\ell} u} = (u - I_{K_\ell} u) \circ F_\ell = u \circ F_\ell - \left( \sum_{i=0}^d u(a_i) p_i \right) \circ F_\ell = \hat{u} - \sum_{i=0}^d u(F_\ell(F_\ell^{-1}(a_i))) p_i \circ F_\ell = \hat{u} - \sum_{i=0}^d \hat{u}(\hat{a}_i) \hat{p}_i = \hat{u} - I_{\mathbb{P}} \hat{u}.$$

Or, grâce au Lemme 13, on a

$$|\hat{u} - I_{\mathbb{P}} \hat{u}|_{\text{semi}, H^m(\hat{K}; \mathbb{R})} \leq \tilde{C}_{d,m} |\hat{u}|_{\text{semi}, H^2(\hat{K}; \mathbb{R})}.$$

En appliquant à nouveau le Lemme 14 pour  $k=2$ ,

$$|\hat{u}|_{\text{semi}, H^2(\hat{K}; \mathbb{R})} \leq C_{2,d} \|\nabla F_\ell\|^2 |\det \nabla F_\ell^{-1}|^{1/2} |u|_{\text{semi}, H^2(K_\ell; \mathbb{R})}.$$

En combinant les 3 dernières inégalités et en se rappelant que  $|\det A| |\det A^{-1}| = 1$ ,

$$\begin{aligned} |u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} &\leq C_{m,d} \|\nabla F_\ell^{-1}\|^m |\det \nabla F_\ell|^{1/2} \tilde{C}_{d,m} C_{2,d} \|\nabla F_\ell\|^2 |\det \nabla F_\ell^{-1}|^{1/2} |u|_{\text{semi}, H^2(K_\ell; \mathbb{R})} \\ &= \bar{C}_{m,d} \|\nabla F_\ell^{-1}\|^m \|\nabla F_\ell\|^2 |u|_{\text{semi}, H^2(K_\ell; \mathbb{R})}. \end{aligned}$$

On utilise maintenant l'hypothèse géométrique sur le maillage et le Lemme 8, qui implique que

$$\|\nabla F_\ell^{-1}\|^m \|\nabla F_\ell\|^2 \leq \frac{\hat{h}^m}{\rho_\ell^m} \frac{h_\ell^2}{\hat{\rho}^2} \leq \frac{\hat{h}^m}{\hat{\rho}^2} h_\ell^{2-m} \sigma^m.$$

En injectant cette inégalité dans la précédente, et en sommant leur carré sur  $\ell \in \llbracket 1, L \rrbracket$ , on conclut que

$$|u - I_{\text{glob}} u|_{\text{semi}, H^m(\Omega; \mathbb{R})}^2 \leq \sum_{\ell \in \llbracket 1, L \rrbracket} \bar{C}_{m,d}^2 \left( \frac{\hat{h}^m}{\hat{\rho}^2} h_\ell^{2-m} \sigma^m \right)^2 |u|_{\text{semi}, H^2(K_\ell; \mathbb{R})}^2 \leq \bar{C}_{m,d}^2 \left( \frac{\hat{h}^m}{\hat{\rho}^2} \sup_{\ell \in \llbracket 1, L \rrbracket} h_\ell^{2-m} \sigma^m \right)^2 |u|_{\text{semi}, H^2(\Omega; \mathbb{R})}^2.$$

En prenant pour dernière constante  $C_{d,m,\sigma} := \bar{C}_{m,d} \frac{\hat{h}^m}{\hat{\rho}^2} \sigma^m$ , on obtient l'estimation désirée.  $\square$

### 3.5.2 Estimation d'erreur de la méthode des éléments finis d'ordre 1

Pour conclure sur les estimations d'erreur, voyons ce que l'on peut en déduire sur l'approximation des solutions d'EDP. Considérons un problème sous forme variationnelle

$$\text{Trouver } u \in H^1(\Omega; \mathbb{R}) \text{ tel que } a(u, v) = L(v) \quad \text{pour tout } v \in H^1(\Omega; \mathbb{R}).$$

**Théorème 9 – Approximation d'erreur** Supposons que  $a(\cdot, \cdot) : H^1(\Omega; \mathbb{R}) \times H^1(\Omega; \mathbb{R}) \rightarrow \mathbb{R}$  soit une forme bilinéaire, continue et coercive, et que  $L : L^2(\Omega; \mathbb{R}) \rightarrow \mathbb{R}$  soit une forme linéaire continue. Soit  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in \llbracket 1, L \rrbracket}$  un maillage élément fini satisfaisant aux hypothèses du Théorème 8, pour lequel on note  $h = \sup_{\ell \in \llbracket 1, L \rrbracket} h_\ell$ . Soit  $u_h \in V_h$  l'unique solution du problème variationnel

$$\text{Trouver } u_h \in V_h \text{ tel que } a(u_h, v_h) = L(v_h) \quad \text{pour tout } v \in V_h.$$

Si  $u \in H^2(\Omega; \mathbb{R})$ , alors il existe une constante  $C = C_{d,m,\sigma,\alpha,M}$  telle que pour  $m \in \{0, 1\}$ ,

$$\|u - u_h\|_{H^1(\Omega; \mathbb{R})} \leq C h \|u\|_{\text{semi}, H^m(\Omega; \mathbb{R})}.$$

#### Démonstration

Par le Lemme de Céa (Lemme 6), on a

$$\|u - u_h\|_{H^1(\Omega; \mathbb{R})} \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega; \mathbb{R})}.$$

En prenant  $v_h = I_{\text{glob}} u \in V_h$ , on obtient via le Théorème 8 que

$$\|u - v_h\|_{H^1(\Omega; \mathbb{R})} \leq C_{d,m,\sigma} h |u|_{\text{semi}, H^2(\Omega; \mathbb{R})}.$$

On conclut en combinant les deux inégalités.  $\square$

### 3.6 Appendice: étude de l'erreur d'interpolation, cas général

Pour l'étude d'erreur locale, on aura besoin de manipuler deux types de semi-normes : la semi-norme classique

$$|u|_{\text{semi},H^m} = \left( \sum_{|\alpha|=m} |\partial^\alpha u|_{L^2}^2 \right)^{1/2},$$

et la semi-norme quotient

$$|u|_{\text{quot},H^m} = \inf_{p \in \mathbb{P}_{m-1}} \|u - p\|_{H^m}.$$

Comme tout polynôme d'ordre inférieur ou égal à  $m-1$  a toutes ses dérivées  $m^{\text{ièmes}}$  identiquement nulles, on a

$$|u|_{\text{quot},H^m}^2 = \inf_{p \in \mathbb{P}_{m-1}} \sum_{|\alpha| < m} |\partial^\alpha(u - p)|_{L^2}^2 + \sum_{|\alpha|=m} |\partial^\alpha u|_{L^2}^2 \geq |u|_{\text{semi},H^m}^2.$$

D'autre part, l'inégalité de Poincaré-Friedrichs donne l'existence d'une constante  $C = C_{n,m}$  telle que pour tout  $u \in H^m$ ,

$$|u|_{\text{quot},H^m} \leq C_{n,m} |u|_{\text{semi},H^m}.$$

Ces deux semi-normes sont donc équivalentes. On renvoie le lecteur curieux à [Nic00, Théorème 3.48 et Corollaire 3.52] pour plus de détails.

**Lemme 15 – Erreur locale sur l'élément de référence** Soit  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  un élément fini de Lagrange, avec  $\hat{\Sigma} = \{\hat{a}_i\}_{i \in [1,N]}$ . Supposons que  $\hat{\mathbb{P}}$  contienne  $\mathbb{P}_k$  l'ensemble des polynômes d'ordre inférieur ou égal à  $k$  pour un certain  $k > d/2 - 1$ , et que l'opérateur d'interpolation  $I_{\mathbb{P}}$  soit linéaire et continu de  $H^{k+1}(\hat{K}; \mathbb{R})$  dans  $H^m(\hat{K}; \mathbb{R})$  pour un certain  $m \leq k+1$ . Alors pour tout  $u \in H^{k+1}(\hat{K}; \mathbb{R})$  et  $I_{\mathbb{P}}u = \sum_{i=1}^N u(a_i)p_i$  son interpolée, on a

$$\|u - I_{\mathbb{P}}u\|_{H^m(\hat{K}; \mathbb{R})} \leq \|I - I_{\mathbb{P}}\|_{\mathcal{L}(H^{k+1}; H^m)} |u|_{\text{quot},H^{k+1}(\hat{K}; \mathbb{R})}, \quad (3.4)$$

où  $|u|_{\text{quot},H^{k+1}}$  est la norme dans l'espace quotient  $H^{k+1}/\mathbb{P}_k$ , donnée par

$$|u|_{\text{quot},H^{k+1}}^{k+1} = \inf_{p \in \mathbb{P}_k} \|u - p\|_{H^{k+1}(\hat{K}; \mathbb{R})}.$$

#### Démonstration

Sous l'hypothèse que  $k+1 > d/2$ , l'espace  $H^{k+1}(\hat{K}; \mathbb{R})$  s'injecte continument dans  $\mathcal{C}(\hat{K}; \mathbb{R})$  (voir Proposition 5), ce qui assure que l'opérateur d'interpolation soit bien défini. Comme tout polynôme  $p \in \mathbb{P}_k$  est exactement représenté dans l'espace  $\hat{\mathbb{P}}$ , on a  $I_{\mathbb{P}}p = p$ , et

$$u - I_h u = u - p - I_{\mathbb{P}}(u - p) = (I - I_{\mathbb{P}})(u - p).$$

En passant à la norme  $H^m$ , puis en utilisant la continuité de  $I - I_{\mathbb{P}}$  de  $H^{k+1}$  dans  $H^m$ , il vient

$$\|u - I_{\mathbb{P}}u\|_{H^m} \leq \|I - I_{\mathbb{P}}\|_{\mathcal{L}(H^{k+1}; H^m)} \|u - p\|_{H^{k+1}}.$$

En passant à l'inf sur  $p$ , on obtient le résultat désiré. □

La constante  $\|I - I_{\mathbb{P}}\|_{\mathcal{L}(H^{k+1}; H^m)}$  dépend de  $k, m$  et  $\hat{K}$ , mais pas de  $u$  ni du maillage élément fini. Considérons maintenant un maillage  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in [1,L]}$  dont chaque élément est affine-équivalent à  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  via une transformation notée  $F_\ell$ . Le lemme suivant permet de passer de l'élément de référence à l'élément  $K_\ell$ , et inversement.

**Lemme 16 – Changement de variable** Soient  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)$  deux éléments affine-équivalents pour une transformation  $F_\ell$ . Alors il existe une constante  $C = C_{m,d}$  telle que pour tout  $v \in H^m(K_\ell, \mathbb{R})$ , en notant  $\hat{v} = v \circ F_\ell$ , on a

$$|\hat{v}|_{\text{semi},H^m(\hat{K}; \mathbb{R})} \leq C_{m,d} \|\nabla F_\ell\|^m |\det \nabla F_\ell^{-1}|^{1/2} |v|_{\text{semi},H^m(K_\ell; \mathbb{R})}.$$

En intervertissant les rôles de  $\hat{K}$  et  $K_\ell$ , il vient également

$$|v|_{\text{semi},H^m(K_\ell; \mathbb{R})} \leq C_{m,d} \|\nabla F_\ell^{-1}\|^m |\det \nabla F_\ell|^{1/2} |\hat{v}|_{\text{semi},H^m(\hat{K}; \mathbb{R})}.$$

### Démonstration

Utilisons la formule de changement de variable pour passer de  $K_\ell$  à  $\hat{K}$ . Notons  $v := u - I_{K_\ell} u \in H^m(K_\ell; \mathbb{R})$ , et  $\hat{v} := v \circ F_\ell$ . Alors pour tout  $\hat{x} \in \hat{K}$

$$\partial_{\hat{x}_i} \hat{v}(\hat{x}) = \sum_{j \in \llbracket 1, d \rrbracket} \partial_{x_j} v(F_\ell(\hat{x})) \partial_{x_i} (F_\ell)_j(\hat{x}).$$

En notant  $B = \nabla F_\ell$ , on obtient  $\partial_{\hat{x}_i} \hat{v} = \sum_{j=1}^d \partial_{x_j} v(F_\ell) B_{ji}$ . Comme  $B$  est une matrice constante, on peut réitérer l'opération, et obtenir que pour toute suite  $(i_1, \dots, i_m) \in \llbracket 1, d \rrbracket^m$ ,

$$\partial_{x_{i_m}} \cdots \partial_{x_{i_1}} \hat{v}(\hat{x}) = \sum_{j_1, \dots, j_m \in \llbracket 1, d \rrbracket} \partial_{x_{j_1} \cdots x_{j_m}} v(F_\ell^{-1}(\hat{x})) B_{j_m i_m} B_{j_{m-1} i_{m-1}} \cdots B_{j_1 i_1}.$$

Chaque indice  $i$  est associé à une multiplication par la colonne  $B_{:,i}$  correspondante. En conséquence,

$$\begin{aligned} \left| \partial_{x_{i_m}} \cdots \partial_{x_{i_1}} \hat{v}(\hat{x}) \right| &\leq \left| \sum_{j_1, \dots, j_m \in \llbracket 1, d \rrbracket} \partial_{x_{j_1} \cdots x_{j_m}} v(F_\ell^{-1}(\hat{x})) B_{j_m i_m} B_{j_{m-1} i_{m-1}} \cdots B_{j_1 i_1} \right| \\ &\leq \|B\| \left( \sum_{j_m \in \llbracket 1, d \rrbracket} \left| \sum_{j_1, \dots, j_{m-1} \in \llbracket 1, d \rrbracket} \partial_{x_{j_1} \cdots x_{j_m}} v(F_\ell^{-1}(\hat{x})) B_{j_m i_m} B_{j_{m-1} i_{m-1}} \cdots B_{j_1 i_1} \right|^2 \right)^{1/2} \\ &\leq \cdots \\ &\leq \|B\|^m \left( \sum_{j_1, \dots, j_m \in \llbracket 1, d \rrbracket} \left| \partial_{x_{j_1} \cdots x_{j_m}} v(F_\ell^{-1}(\hat{x})) \right|^2 \right)^{1/2}. \end{aligned}$$

En prenant le carré et en intégrant pour  $\hat{x} \in \hat{K}$ ,

$$\begin{aligned} \int_{\hat{x} \in \hat{K}} \left| \partial_{x_{i_m}} \cdots \partial_{x_{i_1}} \hat{v}(\hat{x}) \right|^2 d\hat{x} &\leq \|B\|^{2m} \sum_{j_1, \dots, j_m \in \llbracket 1, d \rrbracket} \int_{\hat{x} \in \hat{K}} \left| \partial_{x_{j_1} \cdots x_{j_m}} v(F_\ell^{-1}(\hat{x})) \right|^2 d\hat{x} \\ &= \|B\|^{2m} \sum_{j_1, \dots, j_m \in \llbracket 1, d \rrbracket} \int_{x \in K_\ell} \left| \partial_{x_{j_1} \cdots x_{j_m}} v(x) \right|^2 \left| \det \nabla F_\ell^{-1} \right| dx. \end{aligned}$$

Comme  $\nabla F_\ell^{-1} = B^{-1}$ , en sommant sur les multi-indices  $i_1, \dots, i_m$ , on a obtenu

$$|\hat{v}|_{\text{semi}, H^m(\hat{K}; \mathbb{R})} \leq \|B\|^m \left| \det B^{-1} \right|^{1/2} C_{m,d} |v|_{H^m(K_\ell; \mathbb{R})}$$

où  $C_{m,d} = \binom{d+m-1}{d}$  est le nombre de multi-indices  $(i_1, \dots, i_m)$  donnant lieu à des dérivées différentes.  $\square$

La combinaison des deux résultats précédents mène à l'étude d'erreur locale. On note à nouveau  $\hat{h}$  (resp.  $h_\ell$ ) le diamètre de  $\hat{K}$  (resp.  $K_\ell$ ), et  $\hat{\rho}$  (resp.  $\rho_\ell$ ) son diamètre interne (voir Définition 24).

**Lemme 17 – Étude locale sur un élément du maillage** Considérons les hypothèses du Lemme 15. Il existe une constante  $C = C_{n,m,d,k,\hat{K}}$  telle que pour tout  $\ell \in \llbracket 1, L \rrbracket$  et  $u \in H^{k+1}(K_\ell; \mathbb{R})$ ,

$$|u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} \leq C_{n,m,d,k,\hat{K}} \|\nabla F_\ell^{-1}\|^m \|\nabla F_\ell\|^{k+1} |u|_{\text{semi}, H^{k+1}(K_\ell; \mathbb{R})}.$$

### Démonstration

Par le lemme de changement de variable, on a

$$|u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} \leq C_{m,d} \|\nabla F_\ell^{-1}\|^m |\det \nabla F_\ell|^{1/2} |\hat{u} - I_{\hat{K}} \hat{u}|_{\text{semi}, H^m(\hat{K}; \mathbb{R})}.$$

Comme

$$I_{\hat{K}} \hat{u} = I_{K_\ell} u \circ F_\ell = \sum_{i=1}^N u(a_i) p_i \circ F_\ell = \sum_{i=1}^N u(F_\ell(\hat{a}_i)) \hat{p}_i = I_{\hat{K}} \hat{u},$$

on peut appliquer le Lemme 15 pour obtenir

$$|\hat{u} - I_{\hat{K}} \hat{u}|_{\text{semi}, H^m(\hat{K}; \mathbb{R})} \leq \|I - I_{\hat{K}}\|_{\mathcal{L}(H^{k+1}; H^m)} |\hat{u}|_{\text{quot}, H^{k+1}(\hat{K}; \mathbb{R})}.$$

Par l'équivalence des semi-normes et une nouvelle application du lemme de changement de variable,

$$|\hat{u}|_{\text{quot}, H^{k+1}(\hat{K}; \mathbb{R})} \leq C_{n,m} |\hat{u}|_{\text{semi}, H^{k+1}(\hat{K}; \mathbb{R})} \leq C_{n,m} C_{m,d} \|\nabla F_\ell\|^{k+1} \left| \det \nabla F_\ell^{-1} \right|^{1/2} |u|_{\text{semi}, H^{k+1}(K_\ell; \mathbb{R})}.$$

Comme  $|\det A| |\det A^{-1}| = 1$ , en prenant  $C_{n,m,d,k,\hat{K}} := C_{n,m} C_{m,d}^2 \|I - I_{\hat{K}}\|_{\mathcal{L}(H^{k+1}; H^m)}$ , on obtient bien le résultat désiré.  $\square$

En conséquence, s'il existe  $\sigma_* > 0$  tel que  $\frac{h_\ell}{\rho_\ell} \leq \sigma_*$  pour tout  $\ell \in \llbracket 1, L \rrbracket$ , on peut utiliser les estimations du Lemme 8 pour obtenir

$$\begin{aligned} |u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} &\leq C_{n,m,d,k,\hat{K}} \|\nabla F_\ell^{-1}\|^m \|\nabla F_\ell\|^{k+1} |u|_{\text{semi}, H^{k+1}(K_\ell; \mathbb{R})} \leq C_{n,m,d,k,\hat{K}} \frac{\hat{h}_\ell^m}{\rho_\ell^m} \frac{h_\ell^{k+1}}{\hat{\rho}^{k+1}} |u|_{\text{semi}, H^{k+1}(K_\ell; \mathbb{R})} \\ &\leq \tilde{C}_{n,m,d,k,\hat{K}} h_\ell^{k+1-m} \sigma_*^m |u|_{\text{semi}, H^{k+1}(K_\ell; \mathbb{R})}, \end{aligned}$$

où

$$\tilde{C}_{n,m,d,k,\hat{K}} := C_{n,m,d,k,\hat{K}} \frac{\hat{h}^m}{\hat{\rho}^{k+1}}.$$

### Exemples

- Pour  $d = 1$ , les hypothèses du Lemme 15 imposent  $k > 0$ . Considérons  $k = 1$ , i.e. les éléments finis d'ordre 1. L'espace  $H^{k+1} = H^2$  s'injecte dans les fonctions  $\mathcal{C}^1$ , donc l'opérateur d'interpolation est bien défini et linéaire. À l'aide du théorème des accroissements finis, on montre que dans ce cas, l'interpolation est également continue de  $H^2$  dans  $H^m$  pour  $m \in \{0, 1\}$ . En dimension 1, on a toujours  $h_\ell = \rho_\ell$ , donc  $\sigma \equiv 1$ . Pour un maillage régulier de pas  $h$ , on obtient donc via Lemme 17 que sur chaque  $K_\ell$ ,

$$|u - I_{\mathbb{P}} u|_{\text{semi}, H^m} \leq \bar{C} h^{2-m}$$

pour une certaine constante  $\bar{C}$ .

On peut enfin généraliser à l'ensemble du domaine  $\Omega$ . L'étude globale s'appuie sur la relation de Chasles, qui permet d'évaluer l'erreur globale en sommant les erreurs locales sur chacune des composantes.

### Théorème 10 – Estimation d'erreur

Supposons que

- $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in \llbracket 1, L \rrbracket}$  soit un maillage élément fini généré par l'élément fini de Lagrange  $(\hat{K}, \hat{\Sigma}, \hat{\mathbb{P}})$  et les transformations affines  $(F_\ell)$ ,
- il existe  $\sigma_* > 0$  tel que  $\frac{h_\ell}{\rho_\ell} \leq \sigma_*$  pour tout  $\ell \in \llbracket 1, L \rrbracket$ ,
- il existe  $k > d/2 - 1$  tel que les polynômes d'ordre inférieur ou égal à  $k$  définis sur  $\mathbb{K}_\ell$  soient inclus dans chaque  $\mathbb{P}_\ell$ .

Soit  $u \in H^{k+1}(\Omega; \mathbb{R})$  et  $Iu$  l'unique élément de  $\mathcal{C}(\Omega; \mathbb{R})$  tel que la restriction à chaque  $K_\ell$  de  $Iu$  coïncide avec  $I_{K_\ell} u$ . Alors pour chaque  $m \leq k+1$  tel que les interpolations  $I_{K_\ell}$  soient bien définies, linéaires et continues de  $H^{k+1}(K_\ell; \mathbb{R})$  dans  $H^m(K_\ell; \mathbb{R})$ , il existe une constante  $\bar{C} = \bar{C}_{n,m,d,k,\hat{K},\sigma_*}$  telle que

$$|u - Iu|_{\text{semi}, H^m(\Omega; \mathbb{R})} \leq \bar{C} \max_{\ell \in \llbracket 1, L \rrbracket} h_\ell^{k+1-m} |u|_{\text{semi}, H^m(\Omega; \mathbb{R})}.$$

### Démonstration

Par Chasles,

$$|u - Iu|_{\text{semi}, H^m(\Omega; \mathbb{R})}^2 = \sum_{\ell \in \llbracket 1, L \rrbracket} |u - Iu|_{\text{semi}, H^m(K_\ell; \mathbb{R})}^2 = \sum_{\ell \in \llbracket 1, L \rrbracket} |u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})}^2.$$

En utilisant le Lemme 17,

$$|u - I_{K_\ell} u|_{\text{semi}, H^m(K_\ell; \mathbb{R})} \leq \bar{C} h_\ell^{k+1-m} |u|_{\text{semi}, H^m(K_\ell; \mathbb{R})},$$

où  $\bar{C} = C_{n,m,d,k,\hat{K}} \frac{\hat{h}^m}{\hat{\rho}^{k+1}} \sigma_*^m$ . En majorant  $h_\ell$  par  $\sum_\ell h_\ell$  et en injectant cette inégalité dans la décomposition de Chasles, on obtient le résultat désiré.  $\square$

# Chapitre 4

## Aspects pratiques

### 4.1 Implémentation

Rappelons que pour un espace  $V_h = \text{Vect} \{ \omega_j \mid j \in \llbracket 1, J \rrbracket \}$  donné, la formulation du problème variationnel discret

$$\text{Trouver } u_h \in V_h \text{ t.q. } a(u_h, v_h) = L(v_h) \quad \forall v_h \in V_h$$

s'écrit  $\mathbb{A}U = \mathbb{B}$ , où  $\mathbb{A}_{ij} = a(\omega_j, \omega_i)$  et  $\mathbb{B}_i = L(\omega_i)$ . Dans la pratique, si l'on considère des équations de la forme

$$u - \Delta u = f, \quad \text{de formulation variationnelle} \quad \langle u, v \rangle_{L^2} + \langle \nabla u, \nabla v \rangle_{L^2} = \langle f, v \rangle_{L^2} \quad \forall v \in V,$$

on nomme  $\mathbb{M} = (\langle \omega_j, \omega_i \rangle_{L^2})_{ij}$  la matrice de masse, et  $\mathbb{K} = (\langle \nabla \omega_j, \nabla \omega_i \rangle)_{ij}$  la matrice de rigidité.

#### 4.1.1 Assemblage des matrices

Notons  $a(\cdot, \cdot)$  une forme bilinéaire continue et positive (par exemple  $\langle \cdot, \cdot \rangle_{L^2}$  ou  $\langle \nabla \cdot, \nabla \cdot \rangle_{L^2}$ ). Avec un peu d'abus de notation, notons  $\omega|_K$  la fonction définie par  $\omega|_K(x) = \omega(x)$  si  $x \in K$ , et  $\omega|_K(x) = 0$  sinon. On a

$$a(\omega_j, \omega_i) = a\left(\sum_{K_\ell \in \text{supp } \omega_j} (\omega_j)|_{K_\ell}, \sum_{K_m \in \text{supp } \omega_i} (\omega_i)|_{K_m}\right) = \sum_{K_\ell \in \text{supp } \omega_j \cap \text{supp } \omega_i} a((\omega_j)|_{K_\ell}, (\omega_i)|_{K_\ell}).$$

De plus, dans le cadre des bases de Lagrange, la (vraie) restriction  $(\omega_j)|_{K_\ell}$  coïncide avec une fonction de base de l'élément  $K_\ell$ . Ainsi, si l'on calcule tous les termes  $a(p_j, p_i)$  pour les fonctions de base  $p_i, p_j \in \mathbb{P}_\ell$ , et ceci pour chaque triangle  $K_\ell$ , on pourra en déduire les termes de la matrice  $\mathbb{A}$ . De plus, chacun des termes  $a(p_j, p_i)$  ainsi calculés n'apparaît qu'une seule fois dans la matrice  $\mathbb{A}$ . Pour détailler plus avant, nous aurons besoin de préciser la numérotation.

**Définition 32 – Numérotation locale et globale** Soit  $(K_\ell, \Sigma_\ell, \mathbb{P}_\ell)_{\ell \in \llbracket 1, L \rrbracket}$  le maillage élément fini de  $\Omega$ , et  $(\omega_j)_{j \in \llbracket 1, J \rrbracket}$  l'ensemble des fonctions de base de  $V_h$ . On note

$$\text{LocalVersGlobal} : \llbracket 1, L \rrbracket \times \llbracket 1, N \rrbracket \rightarrow \llbracket 1, J \rrbracket, \quad \text{LocalVersGlobal}(\ell, i) = j$$

l'application qui, à un numéro d'élément  $\ell$  et à un numéro de fonction de base *dans cet élément*  $i$ , fait correspondre l'unique numéro de fonction de base  $j$  tel que  $(\omega_j)|_{K_\ell} = p_i$ .

On pourrait également définir l'application  $\text{GlobalVersLocal}$  qui, à un numéro  $j$  de fonction de base  $\omega_j$  donné, fait correspondre l'ensemble des paires  $(\ell, i)$  telles que  $K_\ell$  appartient au support de  $\omega_j$ , et la restriction de  $\omega_j$  à  $K_\ell$  est  $p_i$ . Cependant, le calcul de  $\text{GlobalVersLocal}$  est bien plus coûteux que celui de  $\text{LocalVersGlobal}$ . On procède donc en faisant une boucle sur tous les éléments  $K_\ell$ .

**Algorithm 1:** Algorithme d'assemblage générique

---

```

1  $\mathbb{A} \leftarrow 0_{J \times J}$ 
2 for  $\ell \in [1, L]$  do
3   Considérer la base  $(p_i)_{i \in [1, I]}$  de  $\mathbb{P}_\ell$  sur le triangle  $K_\ell$ .
4   for  $(i, i') \in [1, I]^2$  do
5     Déterminer les indices globaux  $j = \text{LocalVersGlobal}(\ell, i)$  et  $j' = \text{LocalVersGlobal}(\ell, i')$ .
6      $\mathbb{A}_{j, j'} \leftarrow \mathbb{A}_{j, j'} + a(p_{i'}, p_i).$ 

```

---

Pour chaque élément  $K_\ell$ , il est parfois plus simple de calculer tous les termes  $a(p_{i'}, p_i)$  dans une petite matrice, dite “matrice élémentaire”. L'algorithme est alors modifié de la façon suivante : le calcul de la matrice élémentaire  $\mathbb{A}_\ell = (a(p_{i'}, p_i))_{i, i' \in [1, I]}$  intervient juste après la ligne 3, et les valeurs  $a(p_{i'}, p_i)$  de la ligne 6 sont prises directement dans  $\mathbb{A}_\ell$ .

## 4.1.2 Calcul des matrices élémentaires

Concentrons-nous maintenant sur le calcul d'un terme  $a(p_{i'}, p_i)$ , où  $p_i, p_{i'} \in \mathbb{P}_\ell$  sont des fonctions de base sur l'élément  $K_\ell$ .

### 4.1.2.1 Changements de variable

On s'intéresse dans un premier temps à une application  $a(\cdot, \cdot)$  de la forme

$$a(u, v) := \int_{x \in \Omega} u(x)v(x)dx.$$

Par construction du maillage fini, il existe  $F_\ell : \hat{K} \rightarrow K_\ell$  un difféomorphisme de classe  $C^1$  tel que  $K_\ell = F_\ell(\hat{K})$ , et  $p_i = \hat{p}_i \circ F_\ell^{-1}$  pour une unique fonction  $\hat{p}_i \in \hat{\mathbb{P}}_\ell$ . On peut donc écrire

$$a(p_{i'}, p_i) = \int_{x \in K_\ell} p_{i'}(x)p_i(x)dx = \int_{\hat{x} \in \hat{K}} p_{i'}(F_\ell(\hat{x}))p_i(F_\ell(\hat{x}))|\det \nabla F_\ell(\hat{x})|d\hat{x} = \int_{\hat{x} \in \hat{K}} \hat{p}_{i'}(\hat{x})\hat{p}_i(\hat{x})|\det \nabla F_\ell(\hat{x})|d\hat{x}. \quad (4.1)$$

Pour peu que la transformation  $F_\ell$  soit affine, on aurait  $|\det \nabla F_\ell(\hat{x})|$  indépendant de  $\hat{x}$ , et le terme  $a(p_{i'}, p_i)$  deviendrait proportionnel au terme  $a(\hat{p}_{i'}, \hat{p}_i)$ . Même dans le cas où  $|\det \nabla F_\ell(\hat{x})|$  n'est pas constant, ce changement de variable permet de ne calculer les fonctions de base que sur l'élément de référence, ce qui devient précieux quand l'expression des  $\hat{p}_i$  est compliquée. De la même manière, considérons

$$a(u, v) := \int_{x \in \Omega} \langle \nabla u(x), \nabla v(x) \rangle dx.$$

Dans ce cas, le même changement de variable implique

$$a(p_{i'}, p_i) = \int_{x \in K_\ell} \langle \nabla p_{i'}(x), \nabla p_i(x) \rangle dx = \int_{\hat{x} \in \hat{K}} \langle (\nabla p_{i'})(F_\ell(\hat{x})), (\nabla p_i)(F_\ell(\hat{x})) \rangle |\det \nabla F_\ell(\hat{x})|d\hat{x}.$$

Comme on a

$$\nabla p_i = \nabla(\hat{p}_i \circ F_\ell^{-1}) = \nabla^t(F_\ell^{-1})\nabla \hat{p}_i(F_\ell^{-1}) = (\nabla F_\ell)^{-t}\nabla \hat{p}_i(F_\ell^{-1}),$$

on en déduit

$$a(p_{i'}, p_i) = \int_{\hat{x} \in \hat{K}} \langle (\nabla F_\ell)^{-t}(F_\ell(\hat{x}))\nabla \hat{p}_{i'}(\hat{x}), (\nabla F_\ell)^{-t}(F_\ell(\hat{x}))\nabla \hat{p}_i(\hat{x}) \rangle |\det \nabla F_\ell(\hat{x})|d\hat{x}. \quad (4.2)$$

Si la transformation  $F_\ell$  est affine et de la forme  $F_\ell(\hat{x}) = A_\ell \hat{x} + b_\ell$  pour  $A_\ell$  inversible, l'expression se simplifie en

$$a(p_{i'}, p_i) = \int_{\hat{x} \in \hat{K}} \langle A_\ell^{-t} \nabla \hat{p}_{i'}(\hat{x}), A_\ell^{-t} \nabla \hat{p}_i(\hat{x}) \rangle |\det A_\ell| d\hat{x}$$

### 4.1.2.2 Formules de quadrature

Pour compléter la description de l'implémentation, il reste à calculer numériquement les termes obtenus en (4.1) et (4.2). De même, le calcul du second membre  $\mathbb{B}$  peut faire intervenir des produits scalaires entre  $f$  et les fonctions de base. De manière générale, on est ramenés à approximer numériquement des termes de la forme

$$\int_{\hat{x} \in \hat{K}} \varphi(\hat{x}) d\hat{x},$$

où  $\varphi : \hat{K} \rightarrow \mathbb{R}$  est connue.

L'idée des méthodes de quadrature est de construire une approximation

$$\int_{\hat{x} \in \hat{K}} \varphi(\hat{x}) d\hat{x} \approx \sum_{k \in [1, K]} \varphi(\hat{x}_k) \mu_k, \quad (4.3)$$

où les points  $(\hat{x}_k)_{k \in [1, K]} \subset \hat{K}$  sont appelés points ou nœuds de quadrature, et les réels  $(\mu_k)_{k \in [1, K]}$  sont les poids de quadrature. La construction (4.3) est choisie de manière à être exacte sur une certaine classe de fonction  $\mathcal{F}_K : \hat{K} \rightarrow \mathbb{R}$ , qui grandit avec  $K$  pour devenir une base de l'espace dans lequel on souhaite choisir  $\varphi$ . Par exemple, on peut prendre  $\mathcal{F}_K$  un espace de fonctions trigonométrique en analyse de Fourier, ou des polynômes d'ordre de plus en plus élevé.

Pour plus de détails, vous êtes invité·e·s à consulter le cours d'Analyse Numérique de GM3 !

### 4.1.3 Pseudo-élimination

La méthode "générique" des éléments finis se décline en de multiples versions selon les équations considérées, chacune donnant lieu à des astuces pour diminuer le coût mémoire ou le temps de calcul. Voyons ici une de ces astuces dédiée à la prise en compte de la condition de Dirichlet. Supposons que l'on considère une équation de la forme

$$u - \Delta u = f \quad \text{dans } \Omega, \quad u = 0 \quad \text{sur } \partial\Omega. \quad (4.4)$$

Notons  $V_h$  un espace d'approximation pour le problème  $u - \Delta u = f$  muni de conditions de Neumann homogène  $\partial_\nu u = 0$  sur  $\partial\Omega$ . La formulation variationnelle du problème (4.4) est posée dans l'espace  $V_h^0$  des fonctions de  $V_h$  qui s'annulent au bord de  $\Omega$ . On pourrait tout à fait traiter ce problème de manière classique, en exhibant une base de  $V_h^0$  et en calculant les matrices associées. Cependant, on peut aussi exploiter l'algorithme d'assemblage des matrices de masse et de rigidité pour le problème avec conditions de Neumann, et imposer *a posteriori* les conditions de Dirichlet.

Pour ceci, décomposons le système  $\mathbb{A}U = \mathbb{B}$  suivant l'emplacement des points  $a_j$  associés aux fonctions de base  $w_j$ . On note  $U^{\text{int}} = U_{j \in J_{\text{int}}}$  la sélection des coordonnées de  $U$  telles que  $a_j \in \bar{\Omega}$  pour tout  $j \in J_{\text{int}}$ , et  $U^{\text{bord}} = U_{i \in J_{\text{bord}}}$  l'ensemble des coordonnées  $j$  que  $a_j \in \partial\Omega$ . La condition de Dirichlet impose que  $U^{\text{bord}} = 0$ , et la formulation variationnelle de (4.4) impose que  $\mathbb{A}^{\text{int}}U^{\text{int}} = \mathbb{B}^{\text{int}}$ . En combinant les deux, on obtient le système

$$\begin{pmatrix} \mathbb{I} & 0 \\ 0 & \mathbb{A}^{\text{int}} \end{pmatrix} \begin{pmatrix} U^{\text{bord}} \\ U^{\text{int}} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbb{B}^{\text{int}} \end{pmatrix}.$$

La technique de la pseudo-élimination consiste à passer de la matrice  $\mathbb{A}$  à la matrice  $\begin{pmatrix} \mathbb{I} & 0 \\ 0 & \mathbb{A}^{\text{int}} \end{pmatrix}$ , ainsi que du vecteur  $\mathbb{B}$  au vecteur  $\begin{pmatrix} 0 \\ \mathbb{B}^{\text{int}} \end{pmatrix}$ , en modifiant les lignes et les colonnes des indices de  $J_{\text{bord}}$ .

---

#### Algorithm 2: Pseudo-élimination

---

- 1 **for**  $j \in J_{\text{bord}}$  **do**
- 2     Mettre la ligne  $j$  de la matrice  $\mathbb{A}$  à 0,
- 3     Mettre la colonne  $j$  de la matrice  $\mathbb{A}$  à 0,
- 4     Mettre la valeur de  $\mathbb{A}_{j,j}$  à 1,
- 5     Mettre la valeur de  $\mathbb{B}_j$  à 0.

On peut ainsi assembler le même système quelles que soient les conditions au bord, puis faire l'étape de pseudo-élimination pour intégrer les conditions de Dirichlet.

**Remarque 11** (Conditionnement). *Le choix de la valeur 1 à la ligne 4 est arbitraire, et l'on pourrait y substituer n'importe quelle valeur non nulle. En particulier, on peut choisir une valeur qui ne modifie pas le conditionnement de la matrice  $\mathbb{A}$ .*

# **Partie II**

# **MNEDP2**

Les équations différentielles linéaires peuvent être traitées de manière assez générique avec la méthode des éléments finis. Par discrétisation, on se ramène à un problème d'algèbre linéaire, et l'on tire profit de la structure des matrices obtenues pour le résoudre efficacement. Ladite structure est directement liée aux propriétés des bases EF : en pratique, si les fonctions de base ont un petit support, alors la matrice à inverser sera creuse. On retiendra l'idée générale suivante : **les éléments finis résolvent des projections de dimension finie intelligentes de systèmes linéaires en dimension infinie**.

L'objet de la théorie spectrale est de généraliser à la dimension infinie les notions de valeurs et vecteurs propres, et d'en tirer profit pour construire des méthodes numériques. On retiendra l'idée générale suivante : **les méthodes spectrales cherchent des coefficients dans une base de vecteurs propres**. Ces méthodes utilisent beaucoup plus la structure du problème que les éléments finis, et sont bien plus efficaces en pratique : seulement, les cas dans lesquels il est possible de les appliquer sont assez restreints, souvent avec des géométries simples (là où les éléments finis sont très flexibles). Ainsi, les méthodes spectrales peuvent être utilisées pour obtenir une solution "de référence" à laquelle on pourra comparer d'autres schémas. Attention au vocabulaire : il existe une classe d'éléments finis dits "spectraux", mais qui ne s'appuient pas sur les méthodes spectrales (ils en ont simplement les ordres de convergence).

La seconde partie de ce cours s'attache en premier lieu à formaliser la diagonalisation en dimension infinie. C'est un passage nécessaire, car les opérateurs linéaires en dimension infinie sont beaucoup plus compliqués qu'en dimension finie, et le chemin vers la généralisation s'est révélé semé d'embûches. La dernière partie du cours se concentre sur la variable temporelle, avec la philosophie suivante : "bien que les problèmes d'évolution ne rentrent pas dans le cadre diagonalisable (ils ne sont pas elliptiques), on peut s'en sortir sans trop d'efforts".

# Chapitre 5

## Théorie spectrale

Le lecteur connaît déjà toute une zoologie de théorèmes portant sur la diagonalisation de matrices. La situation canonique est celle d'une matrice symétrique définie positive, dont on sait qu'elle est diagonalisable dans une base orthonormée de vecteurs propres. Pour étendre ce théorème, il est nécessaire de donner une définition de "symétrique" (qui deviendra *autoadjoint*), "définie positive" (qui deviendra *coercif* ou *elliptique non dégénéré*) et d'ajouter une condition, présente mais cachée en dimension infinie, pour retrouver la compacité perdue des images. Procédons dans l'ordre.

### 5.1 Adjoint et spectre

Soit  $H$  un espace de Hilbert muni du produit scalaire  $\langle \cdot, \cdot \rangle$ .

**Adjoint et autoadjoint** Dans  $\mathbb{R}^n$ , on a la relation  $\langle Ax, y \rangle = y^t Ax = (A^t y)^t x = \langle x, A^t y \rangle$ . En français, le produit scalaire entre une transformation linéaire de  $x$ , i.e. l'élément  $Ax$ , et un autre vecteur quelconque  $y$ , est égal au produit scalaire entre l'élément de base et la transformation  $A^t y$ .

**Définition 33 – Adjoint** Soit  $T \in \mathcal{L}(H) = \mathcal{L}(H; H)$ . Son opérateur adjoint  $T^*$  est l'application de  $H$  dans  $H$  définie par

$$\langle T^* x, y \rangle_H = \langle x, Ty \rangle_H.$$

On retiendra que l'adjoint généralise la *transposée*.

**Proposition 7 – Existence et propriété de  $T^*$**  L'adjoint  $T^*$  est bien défini, linéaire et continu.

#### Démonstration

Soit  $x \in H$  fixé. Comme  $T$  est continu, l'application  $y \mapsto \langle x, Ty \rangle_H$  est linéaire et continue. Par le Théorème 5, il existe donc un unique élément  $\ell = \ell_x \in H$  tel que

$$\langle \ell_x, y \rangle_H = \langle x, Ty \rangle_H$$

pour tout  $y \in H$ . On pose  $T^* x = \ell_x$ . On a alors pour tout  $\alpha \in \mathbb{R}$  et  $(x_1, x_2) \in H^2$

$$\langle T^*(\alpha x_1 + x_2), y \rangle_H = \langle \alpha x_1 + x_2, Ty \rangle_H = \alpha \langle x_1, Ty \rangle_H + \langle x_2, Ty \rangle_H = \langle \alpha T^* x_1 + T^* x_2, y \rangle_H,$$

et par unicité,  $T^*(\alpha x_1 + x_2) = \alpha T^* x_1 + T^* x_2$ . De plus, pour les applications linéaires, la continuité est équivalente à la finitude de  $\|\cdot\|_{\mathcal{L}(H)}$ . Ainsi, puisque

$$\|T^*\|_{\mathcal{L}(H)} = \sup_{\substack{x \in H \\ \|x\|_H=1}} \sqrt{\langle T^* x, T^* x \rangle} = \sup_{\substack{x \in H \\ \|x\|_H=1}} \sqrt{\langle x, TT^* x \rangle} \leq \sup_{\substack{x \in H \\ \|x\|_H=1}} \sqrt{|x| |TT^* x|} \leq \sqrt{\|T\|_{\mathcal{L}(H)}} \sup_{\substack{x \in H \\ \|x\|_H=1}} \sqrt{|T^* x|} = \sqrt{\|T\|_{\mathcal{L}(H)} \|T^*\|_{\mathcal{L}(H)}},$$

on déduit que  $\|T^*\|_{\mathcal{L}(H)} \leq \|T\|_{\mathcal{L}(H)}$ . □

Dans notre cadre, la proposition suivante paraît simple, mais elle peut devenir fausse quand les applications considérées ne sont définies que sur un domaine dense dans  $H$ .

**Proposition 8 – Biadjoint** Soit  $H$  un espace de Hilbert et  $T \in \mathcal{L}(H)$ . Alors  $(T^*)^* = T$ .

### Démonstration

Pour tout  $(x, y) \in H^2$ , on a

$$\langle (T^*)^* x, y \rangle = \langle x, T^* y \rangle = \langle Tx, y \rangle.$$

Ainsi  $\langle (T^*)^* x - Tx, y \rangle = 0$  pour tout  $y$ , et en prenant  $y = (T^*)^* x - Tx$ , on en déduit l'égalité.  $\square$

Soit  $T \in \mathcal{L}(H)$ . On utilisera les notations

- $\ker T := \{x \in H \mid Tx = 0\}$  le noyau de  $T$ ,
- $Im T := \{Tx \mid x \in H\}$  l'image de  $T$ ,
- l'orthogonal d'un ensemble  $S \subset H$  sera noté  $S^\perp := \{x \in H \mid \langle x, y \rangle = 0 \quad \forall y \in S\}$ .

On rappelle que  $S^\perp$  est un espace vectoriel fermé, et que  $(S^\perp)^\perp = \overline{\text{Vect } S}$ .

Remarquons que dans  $\mathbb{R}^d$ , si  $x$  appartient au noyau de  $A$ , alors  $\langle Ax, y \rangle = 0$  pour tout  $y$ . Donc  $\langle x, A^t y \rangle = 0$  pour tout  $y$ , et  $x$  appartient à l'orthogonal de l'image de  $A^t$ . Ce petit calcul se généralise de la manière suivante :

**Lemme 18 – Relations d'orthogonalité** Soit  $A \in \mathcal{L}(H)$ . Alors

$$\ker A = (Im A^*)^\perp, \quad \ker A^* = (Im A)^\perp, \quad (\ker A)^\perp = \overline{Im A^*}, \quad (\ker A^*)^\perp = \overline{Im A}.$$

### Démonstration

Soit  $x \in \ker A$ . Alors pour tout  $y \in H$ , on a

$$0 = \langle Ax, y \rangle = \langle x, A^* y \rangle.$$

Donc  $x$  est orthogonal à tout élément de l'image de  $A^*$ , et  $\ker A \subset (Im A^*)^\perp$ . Réciproquement, si  $x \in (Im A^*)^\perp$ , alors pour tout  $y \in H$ ,  $0 = \langle x, A^* y \rangle = \langle Ax, y \rangle$ , et en prenant  $y = Ax$ , on déduit que  $Ax = 0$ . D'où l'égalité. Les autres égalités se déduisent respectivement en prenant  $A = A^*$  et en passant au complémentaire orthogonal.  $\square$

**Définition 34 – Opérateur autoadjoint** Soit  $T \in \mathcal{L}(H)$  un opérateur linéaire continu (en particulier,  $\text{dom}(T) = H$ ). On dit que  $T$  est autoadjoint si  $T^* = T$ , au sens où  $\langle Tu, v \rangle = \langle u, Tv \rangle$  (en identifiant  $H$  et son dual).

### Exemples

- L'identité est toujours autoadjointe.
- Par linéarité du produit scalaire, les opérateurs autoadjoints forment un espace vectoriel.
- Dans l'espace  $\mathbb{R}^d$  muni du produit scalaire canonique, les opérateurs autoadjoints sont représentés par  $x \mapsto Ax$ , où  $A \in \mathbb{M}_{d,d}$  est symétrique.
- Dans  $L^2([0, 1], \mathbb{R})$ , toute application  $u \mapsto \kappa u$  avec  $\kappa \in L^\infty([0, 1], \mathbb{R}^d)$  est autoadjointe.

### Contre-exemples

- Les opérateurs de rotation sont antiadjoints, i.e.  $\langle Tu, v \rangle = -\langle u, Tv \rangle$ .
- Dans le cas de fonctions à valeurs complexes, la bonne notion est celle d'opérateur hermitien. Ceci se déduit de la définition du produit scalaire  $\langle x, y \rangle_{\mathbb{C}} := x\bar{y}$ , pour lequel on vérifie que  $\langle Ax, y \rangle_{\mathbb{C}} = \langle x, A^* y \rangle_{\mathbb{C}}$  pour  $A^* = \overline{A^t}$ .

**Spectre** En dimension finie, toute matrice symétrique définie positive est caractérisée par ses valeurs propres et ses vecteurs propres. On rappelle que  $\lambda \in \mathbb{R}$  est valeur propre de  $A$  s'il existe  $v \in \mathbb{R}^d \setminus \{0\}$  tel que  $Av = \lambda v$ . Via le théorème du rang, cette condition est équivalente aux trois suivantes :

- l'application  $v \mapsto (A - \lambda I)v$  n'est pas bijective ;
- l'application  $v \mapsto (A - \lambda I)v$  n'est pas injective ;
- l'application  $v \mapsto (A - \lambda I)v$  n'est pas surjective.

En dimension infinie, cette équivalence **n'est plus vérifiée**. Il peut exister  $\lambda \in \mathbb{R}$  tel que  $A - \lambda I : H \rightarrow H$  ait une image strictement contenue dans  $H$ , mais un noyau réduit à 0, et inversement. Les exemples canoniques sont les *shifts*, ou décalages : dans  $H = \ell^2$  les suites bornées, l'application "décalage à gauche" qui à  $x = (x_1, x_2, \dots)$  associe  $(x_2, x_3, \dots)$  est surjective mais pas injective, et l'application "décalage à droite"  $x \mapsto (0, x_1, x_2, \dots)$  est injective mais pas surjective.

On distingue donc le spectre, qui contient toutes les valeurs  $\lambda \in \mathbb{R}$  faisant perdre la bijectivité, des valeurs propres, qui restent liées à des vecteurs propres non nuls.

**Définition 35 – Spectre** Soit  $T \in \mathcal{L}(H)$ .

- l'ensemble résolvant de  $T$  est l'ensemble des valeurs  $\lambda \in \mathbb{R}$  telles que  $T - \lambda I$  est bijectif de  $H$  dans  $H$ .
- le spectre de  $T$ , noté  $\sigma(T)$ , est le complémentaire dans  $\mathbb{R}$  de l'ensemble résolvant.
- les valeurs propres de  $T$  sont les valeurs  $\lambda$  du spectre telles que le noyau  $\ker(T - \lambda I)$  est non réduit à  $\{0\}$ .
- l'espace propre associé à une valeur propre est  $\ker(T - \lambda I)$ .

### Exemples

- En reprenant les opérateurs de shift dans  $\ell^2$ , on observe que 0 est dans le spectre des deux shifts, mais est valeur propre pour le shift à gauche (pour le vecteur propre  $(1, 0, 0, \dots)$ ) et pas pour le shift à droite (puisque  $(0, x_1, x_2, \dots) = 0$  implique  $x=0$ ).

Dans la suite, on se restreindra à des classes d'opérateurs pour lesquels le spectre est relativement facile à manipuler. En particulier, dès que  $T$  est linéaire, continu et autoadjoint, on peut identifier deux valeurs qui font partie du spectre et l'encadrent.

**Lemme 19 – Encadrement du spectre** Soit  $T \in \mathcal{L}(H)$  un opérateur autoadjoint. On pose

$$m := \inf_{\substack{u \in H \\ |u|=1}} \langle Tu, u \rangle, \quad M := \sup_{\substack{u \in H \\ |u|=1}} \langle Tu, u \rangle.$$

Alors  $\{m, M\} \subset \sigma(T) \subset [m, M]$ , i.e.  $m$  et  $M$  font partie du spectre de  $T$ , et toute valeur du spectre est encadrée par  $m$  et  $M$ .

### Démonstration

Soit  $\lambda > M$  : par définition, on a  $\langle Tu, u \rangle \leq M|u|^2$ , donc

$$\langle \lambda u - Tu, u \rangle \geq (\lambda - M) \langle u, u \rangle = (\lambda - M)|u|^2.$$

Ainsi,  $\lambda I - T$  est une forme bilinéaire continue, symétrique et (par le calcul ci-dessus) coercive. En appliquant le lemme de Lax-Milgram (Théorème 6), on déduit que  $\lambda I - T$  est bijective. (Écrivez les détails, et posez-moi des questions si ce n'est pas clair !) Donc  $\lambda$  n'est pas une valeur du spectre.

La propriété  $M \in \sigma(T)$  s'obtient par contradiction. Remarquons que pour chaque  $u \in H$ , on a  $\langle Tu, u \rangle \leq M \langle u, u \rangle$ , donc la forme bilinéaire symétrique  $a(u, v) := \langle Mu - Tu, v \rangle$  est positive. En lui appliquant Cauchy-Schwarz,

$$|\langle Mu - Tu, v \rangle| \leq (\langle Mu - Tu, u \rangle)^{1/2} (\langle Mu - Tu, v \rangle)^{1/2}.$$

En prenant le sup sur  $v \in H$  tel que  $|v|=1$ , puisque (par continuité)  $\|M - T\| < \infty$ , pour tout  $|u|=1$  on a

$$|Mu - Tu| \leq \|M - T\|^{1/2} (\langle Mu - Tu, u \rangle)^{1/2} = \|M - T\|^{1/2} (M - \langle Tu, u \rangle)^{1/2}.$$

Si maintenant  $M \notin \sigma(T)$ , alors  $M - T$  est bijectif, et pour tout élément  $u \in H$ , on a  $u = (M - T)^{-1}(Mu - Tu)$ . Soit  $(u_n)_{n \geq 0} \subset H$  une suite maximisante pour la définition de  $M$ , i.e. telle que  $\langle Tu_n, u_n \rangle \rightarrow_n M$  et  $|u_n|=1$ . Par ce qui précède,  $|Mu_n - Tu_n| \leq \|M - T\|^{1/2} (M - \langle Tu_n, u_n \rangle)^{1/2} \rightarrow_n 0$ . Donc

$$1 = |u_n| = \left| (M - T)^{-1}(Mu_n - Tu_n) \right| \leq \|M - T\|^{-1} \|Mu_n - Tu_n\| \xrightarrow{n \rightarrow \infty} 0,$$

ce qui est absurde, et  $M \in \sigma(T)$ . Ici, on a utilisé le théorème de l'application ouverte pour obtenir que [linéaire + continue + bijective] implique [inverse continue].

On obtient les mêmes propriétés pour  $m$  en considérant  $-T$  au lieu de  $T$ .  $\square$

## 5.2 Opérateurs compacts

**Rappels : compacité** Le recours à la compacité est souvent implicite en dimension finie. Le premier réflexe en dimension infinie est de récupérer de la compacité *quelque part* ; en l'occurrence, dans l'image des opérateurs considérés. Pour bien commencer, assurons-nous d'avoir les définitions à l'esprit.

**Définition 36 – Ensemble compact** Soit  $(E, d)$  un espace métrique. Un ensemble  $K \subset E$  est dit compact si pour tout recouvrement ouvert  $(O_i)_{i \in I}$  de  $K$ , c'est-à-dire toute famille d'ouverts telle que  $K \subset \bigcup_{i \in I} O_i$ , on peut extraire un sous-recouvrement fini, c'est-à-dire trouver  $J \subset I$  un ensemble fini tel que  $K \subset \bigcup_{j \in J} O_j$ .

### Exemples

- $[0, 1]$  muni de la distance induite par la valeur absolue.
- Dans un espace vectoriel de dimension finie, tout fermé borné est compact.
- Les inégalités de Sobolev permettent d'établir que si  $\Omega \subset \mathbb{R}^d$  est borné et régulier, l'injection canonique  $\iota : H^1(\Omega; \mathbb{R}) \rightarrow L^2(\Omega; \mathbb{R})$  est un opérateur compact (on parle d'injection compacte).

### Contre-exemples

- Dans un espace vectoriel de dimension infinie, la boule unité n'est pas compacte. Par exemple, si  $E = l^\infty$  l'ensemble des suites bornées, muni de la distance  $d(x, y) = \sup_j |x(j) - y(j)|$ , on peut considérer la suite (de suites)  $x_n \in l^\infty$  où  $x_n(j) = 1$  si  $n = j$ , et  $x_n(j) = 0$  sinon. Dès lors,  $d(x_n, 0) = \sup_j |x_n(j) - 0| = 1$ , donc la suite  $(x_n)_n$  est bien bornée dans  $l^\infty$ . Cependant,  $d(x_n, x_m) = 2$ , donc on ne peut pas extraire de sous-suite convergente. On remarque qu'ici, l'absence de compacité vient de la possibilité de "fuir à l'infini", uniquement présente dans les espaces de dimension infinie.

On dit d'un ensemble  $K$  qu'il est relativement compact si  $\overline{K}$  est compact, et qu'il est totalement borné si pour tout  $\varepsilon > 0$ , il existe un recouvrement fini de  $K$  par des ouverts de diamètre inférieur ou égal à  $\varepsilon$ .

**Remarque 12** (Stabilité). Entre autres bonnes propriétés, les compacts sont stables par les applications continues, c'est-à-dire que l'image d'un compact  $K \subset E$  par une fonction  $f : E \rightarrow F$  continue sera un compact de  $F$ . Ce n'est pas le cas des ensembles fermés (par exemple, l'image de  $\mathbb{R}$  par l'application arctan est l'ouvert  $] -1, 1 [$ ). En effet, soit  $(V_n)_n \subset F$  un recouvrement ouvert de  $f(K)$ . Alors, comme l'image réciproque d'un ouvert par une application continue est un ouvert, la famille  $(f^{-1}(V_n))_n \subset E$  est un recouvrement ouvert de  $K$ , et par compacité, on peut en extraire un sous-recouvrement  $(f^{-1}(V_{n_i}))_{i \in [1, N]}$  fini. On vérifie alors que la famille finie  $(V_{n_i})_{i \in [1, N]}$  est un recouvrement de  $f(K)$ .

**Théorème 11 – Bolzano-Weierstrass** Soit  $(E, d)$  un espace métrique complet, et  $K \subset E$ . Alors les propriétés suivantes sont équivalentes :

- $K$  est compact,
- $K$  est totalement borné et fermé,
- de toute suite de  $K$ , on peut extraire une sous-suite convergente dans  $K$ .

### Exemples

- $[0, 1]$  muni de la distance induite par la valeur absolue.
- Dans un espace vectoriel de dimension finie, tout fermé borné est compact.

### Contre-exemples

- Dans un espace vectoriel de dimension infinie, la boule unité n'est pas compacte. Par exemple, si  $E = l^\infty$  l'ensemble des suites bornées, muni de la distance  $d(x, y) = \sup_j |x(j) - y(j)|$ , on peut considérer la suite (de suites)  $x_n \in l^\infty$  où  $x_n(j) = 1$  si  $n = j$ , et  $x_n(j) = 0$  sinon. Dès lors,  $d(x_n, 0) = \sup_j |x_n(j) - 0| = 1$ , donc la suite  $(x_n)_n$  est bien bornée dans  $l^\infty$ . Cependant,

$d(x_n, x_m) = 2$ , donc on ne peut pas extraire de sous-suite convergente. On remarque qu'ici, l'absence de compacité vient de la possibilité de "fuir à l'infini", uniquement présente dans les espaces de dimension infinie.

Pour conclure ces rappels, on remarque que les *ouverts* sont liés à la topologie : ainsi, changer la topologie permet de retrouver de la compacité perdue. Dans ce cours, on travaillera toujours avec la topologie induite par la norme des espaces de Hilbert.

**Opérateurs compacts** Cette section commence par une mauvaise nouvelle. Si l'espace de Hilbert  $H$  est de dimension infinie, le résultat suivant (à ne pas confondre avec le Théorème 5, dit *de représentation de Riesz*) nous dit que les fermés bornés peuvent ne pas être compacts.

**Lemme 20 – de Riesz (voir preuve en TD)** Soit  $H$  un espace de Hilbert. Alors  $H$  est de dimension finie si et seulement sa boule unité est compacte.

Il ne sera donc plus suffisant de travailler avec des applications continues et coercives pour obtenir l'existence d'un minimum. Pour se ramener à ce cadre confortable, on se concentre sur une classe d'applications linéaires continues, désignées à partir de maintenant sous le nom *opérateurs*, qui redonnent la compacité perdue.

**Définition 37 – Opérateur compact** Soit  $H$  un espace de Hilbert. Un opérateur  $T \in \mathcal{L}(H)$  est dit compact si l'image de la boule unité de  $H$  par  $T$  est relativement compacte.

On notera  $\mathcal{K}(H)$  l'ensemble des opérateurs compacts de  $H$  dans  $H$ .

#### Exemples

- Si  $H = \mathbb{R}^d$ , toute application linéaire  $T$  est un opérateur compact (voir en TD).
- Quelque soit l'espace  $H$ , l'application nulle  $T: e \mapsto 0_H$  est compacte.
- L'adjoint d'un opérateur compact est lui-même compact (Théorème de Schauder, voir la preuve en TD).

#### Contre-exemples

- Dans un espace  $H$  de dimension infinie, l'application identité (bien que fort linéaire et continue) n'est pas compacte, puisque  $T\mathcal{B}(0, 1) = \overline{\mathcal{B}(0, 1)}$  n'est pas compacte (voir les exemples de la Définition 36).

**Remarque 13** (Continuité). *Remarquons que tout opérateur compact est continu. En effet, soit*

$$A := \left\{ |T(x)| \mid x \in \overline{\mathcal{B}(0, 1)} \right\} \subset \left\{ |\tau| \mid \tau \in \overline{T(\mathcal{B}(0, 1))} \right\}.$$

Comme  $\overline{T(\mathcal{B}(0, 1))}$  est un ensemble compact et  $|\cdot|$  une application continue, l'ensemble  $A$  est précompact dans  $\mathbb{R}^+$ , donc admet un majorant noté  $[T]$ . Pour  $x \in H$  arbitraire, on a alors  $|T(x)| = |x| \left| T \left( \frac{x}{|x|} \right) \right| \leq |x| [T]$ , et par linéarité,  $T$  est continu.

Chez les opérateurs compacts, les propriétés du spectre sont plus proches de celle de la dimension finie. Par exemple, en dimension  $d < \infty$ , tout opérateur admet au plus  $d$  valeurs distinctes. C'est faux en dimension infinie : pour prendre un exemple abusif mais éclairant, le Laplacien  $\Delta$  défini sur  $H^3([-1, 1]; \mathbb{R}) \subset C^2([-1, 1]; \mathbb{R})$  admet un spectre *continu* : en effet, pour  $\varphi(x) := \cos(\mu x)$ , on a

$$(\Delta\varphi)(x) = -\mu^2 \cos(\mu x) = -\mu^2 \varphi(x),$$

donc  $-\mu^2$  est valeur propre.

Chez les opérateurs compacts, les valeurs propres peuvent être en nombre infini, mais sont au plus *dénombrables*. De plus, elles sont en nombre fini dans tout ensemble de la forme  $]-\infty, \varepsilon] \cup [\varepsilon, \infty[$  pour  $\varepsilon > 0$ .

**Lemme 21 – Les valeurs propres non nulles sont isolées ([Bré10, Lemme VI.2 p. 95])** Soit  $T \in \mathcal{K}(H)$  et une suite  $(\lambda_n)_{n \in \mathbb{N}} \subset \mathbb{R}$  de valeurs propres de  $T$  qui sont toutes non nulles, distinctes deux à deux et qui convergent vers un certain  $\bar{\lambda} \in \mathbb{R}$ . Alors  $\bar{\lambda} = 0$ .

L'argument de la preuve sera utilisé plusieurs fois par la suite en manipulant des opérateurs compacts : par l'absurde, on va construire une suite de  $T(\overline{\mathcal{B}(0, 1)})$ , et montrer qu'elle n'admet aucune sous-suite convergente.

**Démonstration**

Comme chaque  $\lambda_n \neq 0$  est valeur propre de  $T$ , on peut trouver  $e_n$  de norme 1 tel que  $Te_n = \lambda_n e_n$ . Montrons que chaque  $e_{n+1}$  est linéairement indépendant des  $(e_m)_{m \in [0, n]}$ . C'est trivialement le cas pour  $e_0$ . Supposons alors que ce soit le cas au rang  $n$ , et par l'absurde, qu'il existe des coefficients  $(a_m)_{m \in [0, n]}$  tels que  $e_{n+1} = \sum_{m=0}^n a_m e_m$ . Dès lors, par linéarité,

$$\sum_{m=0}^n a_m \lambda_{n+1} e_m = \lambda_{n+1} e_{n+1} = Te_{n+1} = \sum_{m=0}^n a_m Te_m = \sum_{m=0}^n a_m \lambda_m e_m.$$

Comme les  $(e_m)_{m \in [0, n]}$  sont indépendants, on a  $a_m (\lambda_m - \lambda_{n+1}) = 0$  pour tout  $m$ , donc  $a_m = 0$  et  $e_{n+1} = 0$ , ce qui est absurde. Supposons maintenant que  $|\lambda_n| \rightarrow c > 0$  pour tout  $n$ . Par le procédé de Gram-Schmidt, on peut donc construire une suite orthogonale  $(f_n)_{n \in \mathbb{N}}$  telle que  $f_n = \sum_{m=0}^n \alpha_m^n e_m$  et  $|f_n| = 1$ . Dès lors, la suite des  $(Tf_n)_n$  appartient à l'image de la boule unité de  $H$  par  $T$ ,

$$Tf_n - \lambda_n f_n = \sum_{m=0}^{n-1} (\lambda_m - \lambda_n) \alpha_m^n e_m,$$

donc pour  $m < n$ ,

$$\left| \frac{Tf_n}{\lambda_n} - \frac{Tf_m}{\lambda_m} \right| = \left| f_n + \frac{Tf_n - \lambda_n f_n}{\lambda_n} - \frac{Tf_m}{\lambda_m} \right| = |f_n - g_{m,n}|,$$

où  $g_{m,n} = \frac{Tf_m}{\lambda_m} - \frac{Tf_n - \lambda_n f_n}{\lambda_n}$  est combinaison linéaire des  $e_i$  pour  $i < n$ . En conséquence,  $\langle f_n, g_{m,n} \rangle = 0$ , et

$$\left| \frac{Tf_n}{\lambda_n} - \frac{Tf_m}{\lambda_m} \right| = \sqrt{|f_n|^2 + |g_{m,n}|^2} \geq 1.$$

Comme enfin

$$\left| \frac{Tf_n}{\lambda_n} - \frac{Tf_m}{\lambda_m} \right| \leq \left| \frac{Tf_n}{c} - \frac{Tf_m}{c} \right| + \left| \frac{Tf_n}{\lambda_n} - \frac{Tf_n}{c} \right| + \left| \frac{Tf_m}{\lambda_n} - \frac{Tf_m}{c} \right| = \frac{1}{c} |Tf_n - Tf_m| + \|T\| \left( \left| \frac{1}{\lambda_n} - \frac{1}{c} \right| + \left| \frac{1}{\lambda_m} - \frac{1}{c} \right| \right),$$

pour  $m$  suffisamment grand, on a  $1 \leq \frac{1}{c} |Tf_n - Tf_m| + 1/2$ , donc la suite  $(Tf_n)_n$  ne contient pas de sous-suite convergente, et  $c = 0$ .

□

**L'alternative de Fredholm** En dimension finie, tous les opérateurs linéaires sont continus et compacts, donc la résolution de l'équation

$$\text{pour } b \text{ donné, trouver } x \text{ tel que } x + Tx = b \quad (5.1)$$

où  $T$  est une matrice ne dépend que de l'inversibilité de  $I - T$ . Plusieurs cas peuvent se produire : soit  $I - T$  est inversible, et (5.1) admet une solution pour chaque  $b$ . Si  $I - T$  n'est pas de rang plein, il est quand même possible de résoudre (5.1) dans le cas où  $b$  appartient à l'image de  $I - T$  : par exemple, en dimension 2, prendre  $T = \begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix}$  donne  $I - T = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ . Si  $b = \alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

pour un certain  $\alpha \in \mathbb{R}$ , alors tout vecteur de la forme  $x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ , où  $\beta \in \mathbb{R}$ , est solution de l'équation. De manière générale, si  $I - T$  n'est pas inversible, (5.1) admet une solution si et seulement si  $b$  appartient à l'image de  $I - T$ , et dans ce cas, l'ensemble des solutions est de la forme  $x_0 + \ker(I - T)$ .

En dimension infinie, une partie de ces propriétés se transfèrent aux opérateurs compacts.

**Théorème 12 – Alternative de Fredholm ([Bré10, Théorème VI.6])** Soit  $H$  Hilbert, et  $T \in \mathcal{K}(H)$ . Notons  $I : H \mapsto H$  l'application identité. Alors

- T12.1) le noyau de  $I - T$  est de dimension finie.
- T12.2) l'image de  $I - T$  est fermée.
- T12.3) les suites  $(\ker(I - T)^n)_n$  et  $(\text{Im}(I - T)^n)_n$  sont stationnaires à partir d'un certain rang, i.e. il existe  $n \geq 1$  tel que  $\ker(I - T)^i = \ker(I - T)^n$  pour tout  $i \geq n$ .
- T12.4)  $\ker(I - T) = \{0\}$  si et seulement si  $\text{Im}(I - T) = H$ .

**Remarque 14** (Cas de la dimension finie). Si  $H = \mathbb{R}^d$ , le théorème du rang nous dit que  $\dim(\text{Im}(T)) + \dim(\ker(T)) = d$ , ce qui est plus fort que T12.4.

### Démonstration

**T12.1)** Le noyau  $\ker(I - T) = \{x \in H \mid x - Tx = 0\}$  est un sous-espace vectoriel de  $H$ . Montrons que la boule unité  $\overline{\mathcal{B}}_{\ker(I-T)}(0, 1)$  est compacte : soit  $x \in \ker(I - T)$  tel que  $|x| \leq 1$ . Comme  $x = Tx$ , on a  $x \in T\overline{\mathcal{B}}(0, 1)$  l'image de la boule unité de  $H$  par  $T$ . Ainsi, la boule unité du noyau satisfait  $\overline{\mathcal{B}}_{\ker(I-T)}(0, 1) \subset T(\overline{\mathcal{B}}(0, 1))$ , donc est (précompacte et fermée donc) compacte. Par le Lemme 20, la dimension de  $\ker(I - T)$  est finie.

**T12.2)** On cherche à montrer que toute suite de Cauchy de  $Im(I - T)$  converge vers un élément de  $Im(I - T)$ . Soit  $f_n = u_n - Tu_n$  une suite de Cauchy : comme  $H$  est complet,  $f_n$  converge vers un certain  $f \in H$ . On veut montrer qu'il existe  $u \in H$  tel que  $f = u - Tu$ . Posons  $d_n = d(u_n, \ker(I - T))$ . Comme  $\ker(I - T)$  est de dimension finie, il existe  $v_n \in \ker(I - T)$  tel que  $d_n = |u_n - v_n|$ . On a

$$f_n = (u_n - v_n) - T(u_n - v_n). \quad (5.2)$$

Montrons que  $(u_n - v_n)_n$  est une suite bornée.

Par l'absurde, on suppose qu'il existe une sous-suite qui diverge, i.e.  $0 < |u_{n_k} - v_{n_k}| \rightarrow \infty$ . Posons  $w_n = \frac{u_n - v_n}{|u_n - v_n|}$ . En divisant (5.2) par  $|u_{n_k} - v_{n_k}|$ , on obtient

$$w_{n_k} - Tw_{n_k} = \frac{u_{n_k} - v_{n_k}}{|u_n - v_n|} - T \frac{u_{n_k} - v_{n_k}}{|u_{n_k} - v_{n_k}|} = \frac{f_{n_k}}{|u_{n_k} - v_{n_k}|} \xrightarrow{k \rightarrow \infty} 0$$

puisque  $(f_{n_k})_k$  est une suite convergente, donc bornée. Comme  $T$  est un opérateur compact et que la suite  $(w_{n_k})$  reste bornée, la suite  $(Tw_{n_k})_k$  admet une sous-suite  $(Tw_{n_{k_j}})_j$  qui converge vers un certain  $z \in Im(T)$ . Donc  $w_{n_{k_j}} \rightarrow_j z$ , et en passant à la limite,  $z - Tz = 0$ , donc  $z \in \ker(Id - T)$ . D'autre part,

$$d(w_{n_{k_j}}, \ker(I - T)) = \inf_{\alpha \in \ker(I - T)} |w_{n_{k_j}} - \alpha| = \frac{\inf_{\alpha \in \ker(I - T)} |u_{n_{k_j}} - v_{n_{k_j}} - |u_{n_{k_j}} - v_{n_{k_j}}|\alpha|}{|u_{n_{k_j}} - v_{n_{k_j}}|} = \frac{\inf_{\beta \in \ker(I - T)} |u_{n_{k_j}} - \beta|}{|u_{n_{k_j}} - v_{n_{k_j}}|} = \frac{|u_{n_{k_j}} - v_{n_{k_j}}|}{|u_{n_{k_j}} - v_{n_{k_j}}|} = 1,$$

où l'on a utilisé le fait que  $v_{n_{k_j}} \in \ker(I - T)$  pour écrire que  $\alpha \mapsto v_{n_{k_j}} + |u_{n_{k_j}} - v_{n_{k_j}}|\alpha|$  est une bijection de  $\ker(I - T)$  dans lui-même. En passant à la limite en  $j \rightarrow \infty$ , on obtient la contradiction  $d(z, \ker(I - T)) = 1$ . Donc  $|u_n - v_n|$  est une suite bornée dans  $\mathbb{R}$ .

Comme  $T$  est compact, on en déduit qu'il existe une sous-suite telle que  $T(u_{n_k} - v_{n_k}) \rightarrow_k l$  pour un certain  $l \in Im(T)$ . En passant à la limite en  $k \rightarrow \infty$  dans (5.2), on obtient  $\lim_{k \rightarrow \infty} u_{n_k} - v_{n_k} = f + l$ . Donc  $f = f + l - T(f + l) = (I - T)(f + l)$ , et  $f$  appartient bien à l'image de  $Id - T$ , qui est donc un ensemble fermé.

**T12.3)** Notons  $K_n = \ker(I - T)^n$ . Remarquons déjà que si  $x \in K_n$ , alors  $(I - T)^{n+1}(x) = 0 - T(0) = 0$ , donc  $K_n \subset K_{n+1}$ . Pour montrer la stationnarité, on suppose par l'absurde que  $K_n \neq K_{n+1}$  pour tout  $n \in \mathbb{N}_*$ , et on construit une suite bornée de l'image de  $T$  dont aucune sous-suite ne converge, contredisant la compacité de  $T$ . Ainsi, supposons que  $D_n := K_{n+1} \setminus K_n \neq \emptyset$ . Pour chaque  $n \geq 1$ , on choisit un point  $v_n \in D_n$  tel que  $|v_n| = 1$  et  $\text{dist}(v_n, K_n) \geq 1/2$ .

On peut construire un tel élément en piochant  $u_n \in D_n$  au hasard et en remarquant que  $K_n$  fermé implique  $\text{dist}(u_n, K_n) > 0$ . On choisit alors  $p_n \in K_n$  tel que  $\text{dist}(u_n, K_n) \leq |u_n - p_n| \leq 2\text{dist}(u_n, K_n)$ , ce qui est toujours possible par la définition d'infimum, et en posant  $v_n = \frac{u_n - p_n}{|u_n - p_n|}$ . On a alors que  $|v_n| = 1$  et

$$\text{dist}(v_n, K_n) = \frac{\text{dist}(u_n - p_n, K_n)}{|u_n - p_n|} = \frac{\text{dist}(u_n, K_n)}{|u_n - p_n|} \geq 1/2.$$

On a utilisé ici le fait que  $\text{dist}(\cdot, K_n)$  est positivement homogène, et que l'adhérence de  $K_n$  est un espace vectoriel. (Pas  $K_n$  car  $0 \neq K_n$ , par exemple...)

Comme  $T$  est Lipschitzien et  $(v_n)_n$  bornée, la suite  $(T(v_n))_n$  est également bornée. Soient  $n > m \geq 0$ . On a  $|T(v_n) - T(v_m)| = |v_n - ((I - T)(v_n) + T(v_m))|$ . On utilise le fait que  $T((I - T)(v)) = T(v - T(v)) = T(v) - T^2(v) = (I - T)(T(v))$  pour conclure que  $T$  et  $I - T$  commutent. Remarquons alors que

$$(I - T)^n((I - T)(v_n) + T(v_m)) = (I - T)^{n+1}(v_n) + T((I - T)^n(v_m)) = 0 + T(0) = 0,$$

donc  $(I - T)(v_n) + T(v_m) \in \ker(I - T)^n = K_n$ , et par construction,  $|T(v_n) - T(v_m)| \geq \text{dist}(v_n, K_n) \geq 1/2$ . Donc la suite  $(T(v_n))_{n \geq 1}$  n'admet aucune sous-suite convergente, ce qui contredit la compacité de  $T$ .

La preuve pour  $Im(I - T)^n$  est similaire, et utilise T12.2) pour pouvoir construire  $(v_n)_n$ .

**T12.4)** Supposons que  $\ker(I - T) = \{0\}$ , i.e. que  $I - T$  est injective, et montrons que  $I - T$  est surjective. Supposons qu'il existe  $z \in H \setminus Im(I - T)$ . Par le point T12.3), on sait que  $(Im(I - T)^n)_{n \in \mathbb{N}_*}$  est stationnaire à partir d'un certain rang, i.e. il existe  $n_0$  tel que  $Im(I - T)^{n_0+1} = Im(I - T)^{n_0}$ . Donc pour ce  $n_0$  il existe un  $x \in H$  tel que  $(I - T)^{n_0}(z) = (I - T)^{n_0+1}(x)$ . Comme  $I - T$  est injectif,

$$(I - T)((I - T)^{n_0-1}(z)) = (I - T)((I - T)^{n_0}(x)) \implies (I - T)^{n_0-1}(z) = (I - T)^{n_0}(x).$$

En itérant  $n_0$  fois cet argument, on obtient  $z = (I - T)(x)$ , ce qui est absurde. Pour l'autre implication, on suppose que  $\text{Im}(I - T) = H$ . Par les relations d'orthogonalité du Lemme 18,

$$\ker(I - T)^* = \text{Im}(I - T)^\perp = H^\perp = \{0\}.$$

Comme  $T^*$  est compact (voir TD), on peut appliquer la preuve précédente à  $I - T^*$ , et obtenir que  $\text{Im}(I - T^*) = H$ . En réutilisant le Lemme 18, on conclut que  $\ker(I - T) = \text{Im}(I - T^*)^\perp = H^\perp = \{0\}$ .  $\square$

Comme conséquence de Théorème 12, on obtient que sur les opérateurs compacts, les valeurs propres et le spectre coïncident “presque”, et ne peuvent différer qu'en 0.

**Lemme 22 – Spectre d'un opérateur compact ([Bré10, Théorème VI.8 p. 95])** Soit  $T : H \rightarrow H$  un opérateur compact. Alors toute valeur non nulle du spectre  $\sigma(T)$  est valeur propre non nulle de  $T$ , i.e.  $\sigma(T) \setminus \{0\} = VP(T) \setminus \{0\}$ , et  $\sigma(T) \setminus \{0\}$  est soit vide, soit fini, soit forme une suite qui tend vers 0.

### Démonstration

Soit  $\lambda \in \sigma(T) \setminus \{0\}$ . Pour montrer que  $\lambda \in VP(T)$ , on doit s'assurer que  $\dim \ker(T - \lambda I) > 0$ . Par l'absurde, si  $\ker(T - \lambda I) = \{0\}$ , alors T12.4 implique que  $\text{Im}(T - \lambda I) = H$ . Donc  $T - \lambda I$  est à la fois injectif et surjectif, donc bijectif, ce qui contredit  $\lambda \in \sigma(T)$ .

Montrons que le spectre de  $T$  est contenu dans  $[-\|T\|, \|T\|]$ , donc borné. Soit  $f \in H$  arbitraire. Pour tout  $\lambda \in \mathbb{R}$  avec  $|\lambda| > \|T\|$ , considérons l'équation

$$Tx - \lambda x = f \quad \Leftrightarrow \quad x = \frac{1}{\lambda} (Tx - f).$$

L'application  $x \mapsto \frac{1}{\lambda} (Tx - f)$  est continue et strictement contractante, donc admet un unique point fixe. Ceci montre que  $T - \lambda I$  est bijectif, donc  $\lambda \notin \sigma(T)$ . Par contraposée,  $\sigma(T) \subset [-\|T\|, \|T\|]$ .

Considérons les ensembles  $S_k := \sigma(T) \cap \{\lambda \in \mathbb{R} \mid |\lambda| \geq 1/k\}$ . Soit  $S_k$  est vide, soit il est fini : en effet, s'il existait une suite de valeurs toutes distinctes  $(\lambda_n)_n \subset S_k$ , alors

- les  $\lambda_n$  sont tous des valeurs propres de  $T$  par ce qui précède,
- comme  $\sigma(T)$  est borné, on peut extraire une sous-suite  $(\lambda_{n_k})_k$  qui converge vers un certain  $\bar{\lambda}$ , qui est tel que  $|\bar{\lambda}| \geq 1/k$ .

Par le Lemme 21, on a également  $\bar{\lambda} = 0$ , ce qui est absurde. On peut donc construire un ensemble (éventuellement vide) en concaténant les  $S_k$  dans l'ordre croissant de  $k$ , ce qui fournit au plus une suite tendant vers 0.  $\square$

# Chapitre 6

## Diagonalisation

Cette section introduit les problèmes spectraux, dont l'étude numérique fait l'objet de la suite du cours.

### 6.1 Décomposition spectrale des opérateurs compacts autoadjoints

La théorie spectrale abstraite qui précède aboutit au théorème suivant, qui permet de traiter beaucoup d'équations linéaires.

**Théorème 13 – Décomposition spectrale, Théorème de Riesz-Fredholm ([Bré10, Théorème VI.11 p. 97])**

Supposons que  $H$  est séparable. Soit  $T$  un opérateur autoadjoint compact. Alors  $H$  admet une base hilbertienne formée de vecteurs propres de  $T$  telles que les valeurs propres associées forment une suite qui tend vers 0.

On sait déjà par le Lemme 22 que les valeurs propres de  $T$  sont au plus dénombrables. On va montrer que si  $T$  est autoadjoint, alors les sous-espaces propres associés à des valeurs propres distinctes sont orthogonaux : il restera à montrer que l'espace engendré par les sous-espaces propres est dense dans  $H$ , puis à utiliser la séparabilité pour obtenir une base hilbertienne de chacun de ces sous-espaces.

**Remarque 15** (Suite qui tend vers 0). *Attention au vocabulaire : la suite des valeurs propres peut être égale à 0 à partir d'un certain rang. Par exemple, l'opérateur  $T : H \rightarrow H$  donné par  $T(x) = 0$  est linéaire, continu, compact et autoadjoint, et en effet, on peut trouver une base hilbertienne  $(e_m)_m$  de  $H$  (par l'hypothèse de séparabilité) telle que  $T(e_m) = 0 \times e_m$  pour tout  $m$ , et la suite des valeurs propres est identiquement nulle.*

#### Démonstration

Soit  $\lambda_0 = 0$  et  $(\lambda_n)_{n \geq 1}$  la suite (éventuellement vide ou finie) des valeurs propres non nulles de  $T$  donnée par Lemme 22, qui sont distinctes par le Lemme 21. Pour chaque  $n \in \mathbb{N}$ , on pose  $E_n = \ker(T - \lambda_n I)$ . Par la définition des valeurs propres et T12.1), on a  $0 < \dim(E_n) < \infty$  pour tout  $n \geq 1$ .

Supposons que  $u \in E_n$  et  $v \in E_m$  avec  $m \neq n$ . Alors

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

et  $\langle u, v \rangle = 0$ . Ainsi, les sous-espaces propres sont orthogonaux deux à deux.

Soit maintenant  $F = \text{Vect}\{E_n\}_{n \geq 0}$ . En particulier,  $\ker T = E_0 \subset F$ , et pour tout  $y \in F$ , on a  $Ty \in F$ . On cherche à savoir si  $F$  est dense dans  $H$ . On s'intéresse à l'orthogonal  $F^\perp := \{x \in H \mid \langle x, y \rangle = 0 \ \forall y \in F\}$ . Cet ensemble est stable par  $T$ , car si  $x \in F^\perp$  et  $y \in F$ , alors

$$\langle Tx, y \rangle = \langle x, Ty \rangle = 0.$$

On peut donc définir l'opérateur restreint  $T|_{F^\perp} : F^\perp \rightarrow F^\perp$ , qui est autoadjoint compact. De plus, s'il existe une valeur  $\lambda \in \sigma(T|_{F^\perp}) \setminus \{0\}$ , alors (par le Lemme 22)  $\lambda$  est valeur propre de  $T|_{F^\perp}$  avec un sous-espace propre non vide. Mais ce sous-espace propre est également un sous-espace propre de  $T$ , ce qui contredit la construction de  $F$ . Donc  $\sigma(T|_{F^\perp}) = \{0\}$ . Par le Lemme 19, on a  $m = M = 0$ , où

$$m = \inf_{\substack{x \in H \\ |x|=1}} \langle Tx, x \rangle, \quad M = \sup_{\substack{x \in H \\ |x|=1}} \langle Tx, x \rangle.$$

Ainsi  $\langle Tx, x \rangle = 0$  pour tout  $x \in F^\perp$ . Comme par le caractère autoadjoint,

$$2 \langle Tx, y \rangle = \langle T(x+y), x+y \rangle - \langle Tx, x \rangle - \langle Ty, y \rangle = 0,$$

on en déduit que  $T|_{F^\perp} \equiv 0$ , donc  $Tx = 0$  pour tout  $x \in F^\perp$ . De manière équivalente,  $F^\perp \subset \ker T$ . Mais comme  $\ker T \subset F$ , on a  $F^\perp \subset F$ , et  $F^\perp = \{0\}$ . Ainsi,  $F$  est dense dans  $H$ .

Enfin, grâce à la séparabilité de l'espace, il existe une base orthonormée de chaque sous-espace propre, qui est même finie pour tout  $n \geq 1$ . On peut donc concaténer ces bases dans l'ordre décroissant des valeurs propres, ce qui donne une base hilbertienne de l'espace constituée de vecteurs propres de  $T$ . Par le Lemme 22, la suite des valeurs propres non nulles est soit vide, soit finie (et dans ce cas, on complète par les éléments propres associés à la valeur 0), soit tend vers 0.  $\square$

## 6.2 Problèmes spectraux

Faisons maintenant le lien entre la théorie spectrale et l'étude de problèmes variationnels.

**Corollaire 1 – Problèmes aux valeurs propres** Soient  $H, V$  deux espaces de Hilbert tels que  $H$  soit séparable et  $V \hookrightarrow H$  avec injection compacte. Soit  $a: V \times V \rightarrow \mathbb{R}$  une forme bilinéaire symétrique continue et coercive dans  $V$ . Alors le problème aux valeurs propres

$$\text{Trouver } (\omega, \mu) \in V \times \mathbb{R} \text{ tels que } a(\omega, v) = \mu \langle \omega, v \rangle_H \quad (6.1)$$

admet une famille au plus dénombrable  $(\omega_n, \mu_n)_n$  de solutions telle que  $(\mu_n)_n$  est une suite croissante, qui tend vers  $+\infty$  si la famille est de cardinal infini, et  $(\omega_n)_{n \in \mathbb{N}}$  est une base hilbertienne de  $H$ .

Ici, il est implicite que si  $H$  est de dimension finie, la famille  $(\omega_n, \mu_n)_n$  compte un nombre fini de termes.

### Démonstration

Pour plus de précision, on note  $\iota: V \rightarrow H$  l'injection canonique, qui est linéaire et compacte par hypothèse. Grâce au Théorème de Lax-Milgram (Théorème 6), on sait que pour tout  $f \in H$ , le problème

$$\text{Trouver } u \in V \text{ tel que } a(u, v) = \langle f, \iota(v) \rangle_H \text{ pour tout } v \in V \quad (6.2)$$

admet une unique solution  $u \in V$ .

### Opérateur inverse

Notons

$$\Phi: H \rightarrow V, \quad \Phi(f) = u$$

l'opérateur qui à  $f$ , associe la solution  $u$  du problème (6.2). En particulier,  $\langle f, \iota(v) \rangle_H = a(\Phi(f), v)$  pour tout  $v \in V$ . Montrons que  $\iota \circ \Phi$  satisfait les hypothèses du Théorème 13. En premier lieu,  $\Phi$  est linéaire, car pour tout  $v \in V$ ,

$$a(\alpha \Phi(f) + \Phi(g), v) = \alpha a(\Phi(f), v) + a(\Phi(g), v) = \langle \alpha f + g, v \rangle_H = a(\Phi(\alpha f + g), v)$$

et en prenant  $v = \alpha \Phi(f) + \Phi(g) - \Phi(\alpha f + g)$ , on obtient  $a(v, v) = 0$ , donc  $v = 0$ . Comme  $\iota$  est linéaire, la composition  $\iota \circ \Phi$  l'est également. De plus,  $\Phi$  est continue de  $H$  dans  $V$  : en effet, en notant  $\alpha > 0$  la constante de coercivité de  $a(\cdot, \cdot)$ ,

$$|\Phi(f)|_V^2 = |u|_V^2 = \frac{1}{\alpha} a(u, u) = \frac{1}{\alpha} \langle f, \iota(u) \rangle_H \leq \frac{\|f\|_H}{\alpha} |u|_H \leq \frac{\|f\|_H}{\alpha} |u|_V,$$

donc  $|\Phi(f)|_V \leq \|f\|_H / \alpha$ . Comme l'injection  $\iota: V \rightarrow H$  est compacte, la composition  $\iota \circ \Phi: H \rightarrow H$  est un opérateur linéaire, continu, compact. Enfin, pour tout  $g \in H$ ,

$$\langle \iota \circ \Phi(f), g \rangle_H = a(\Phi(f), \Phi(g)) = \langle f, \iota \circ \Phi(g) \rangle_H,$$

donc  $\iota \circ \Phi$  est autoadjoint dans  $H$ .

**Application de Riesz-Fredholm** Par le Théorème 13, il existe une suite  $(\omega_n, \lambda_n)_n \subset H \times \mathbb{R}$  telle que  $\lim_{n \rightarrow \infty} \lambda_n = 0$ ,  $(\omega_n)_n$  est une base hilbertienne de  $H$ , et

$$\iota \circ \Phi(\omega_n) = \lambda_n \omega_n.$$

Comme  $\omega_n$  appartient à l'image de  $\iota$ , qui est bijective sur son image, on a  $\Phi(\omega_n) = \lambda_n \omega_n \in V$  pour tout  $n \in \mathbb{N}$ . (On considère maintenant que  $\omega_n \in V$ .) Comme  $V \subset H$ , la famille  $(\omega_n)_{n \in \mathbb{N}}$  est également une base de  $V$ . De plus, pour tout  $v \in V$ ,

$$\langle \iota(\omega_n), \iota(v) \rangle_H = a(\Phi(\omega_n), v) = \lambda_n a(\omega_n, v). \quad (6.3)$$

Pour conclure, il reste à diviser l'égalité précédente par  $\lambda_n$ . Pour ce faire, montrons que  $\lambda_n > 0$  : en effet, en choisissant  $v = \omega_n$  dans (6.3),

$$\|\iota(\omega_n)\|_H^2 = 1 = \lambda_n a(\omega_n, \omega_n) \iff \lambda_n = \frac{1}{a(\omega_n, \omega_n)} > 0.$$

On pose alors  $\mu_n := 1/\lambda_n$ , et l'on obtient le résultat désiré.  $\square$

**Remarque 16** (Base de  $V$ ). *Remarquons que*

$$a(\omega_n, \omega_n) = \mu_n \langle \iota(\omega_n), \iota(\omega_n) \rangle_H = \mu_n$$

et que pour tout  $n \neq m$ ,

$$a(\omega_n, \omega_m) = \mu_n \langle \iota(\omega_n), \iota(\omega_m) \rangle_H = 0.$$

Donc la famille  $\left(\frac{1}{\sqrt{\mu_n}}\omega_n\right)_{n \in \mathbb{N}}$  est une base de  $V$  qui est orthonormée pour le produit scalaire  $a(\cdot, \cdot)$ .

L'intérêt des problèmes aux valeurs propres est le suivant. Considérons maintenant le problème variationnel

Trouver  $u \in V$  tel que  $a(u, v) = \langle f, v \rangle_H$  pour tout  $v \in V$ .

Soit  $(\omega_n, \mu_n)_{n \in \mathbb{N}} \subset V \times \mathbb{R}$  la suite donnée par le Corollaire 1. Grâce au fait que  $\left(\frac{1}{\sqrt{\mu_n}}\omega_n\right)_n$  est une base  $a(\cdot, \cdot)$ -orthonormale de  $V$ , l'unique solution  $u$  du problème variationnel s'écrit

$$u = \sum_{n \in \mathbb{N}} a\left(u, \frac{\omega_n}{\sqrt{\mu_n}}\right) \frac{\omega_n}{\sqrt{\mu_n}} = \sum_{n \in \mathbb{N}} \frac{1}{\mu_n} a(u, \omega_n) \omega_n = \sum_{n \in \mathbb{N}} \frac{1}{\mu_n} \langle f, \omega_n \rangle_V \omega_n. \quad (6.4)$$

Ainsi, pour connaître la solution  $u$ , il suffit de calculer les produits scalaires  $\langle f, \omega_n \rangle_H$ , ce qui se fait numériquement par des formules de quadrature. C'est le principe des méthodes spectrales, qui cherchent à résoudre le problème aux valeurs propres numériquement, pour ensuite résoudre les problèmes variationnels très rapidement pour des seconds membres  $f \in H$  différents. Contrairement aux éléments finis, une méthode spectrale fonctionne en deux temps : une étape de calcul des éléments propres  $(\omega_m, \lambda_m)_{m \in \mathbb{N}}$ , généralement coûteuse, puis des calculs de quadrature pour appliquer la formule (6.4). Heuristiquement, les méthodes spectrales sont plus rapides et plus précises que les éléments finis. Malheureusement, l'implémentation numérique du problème aux valeurs propres n'est vraiment efficace que sur des géométries simples, ce qui restreint beaucoup les possibilités.

### Exemples

- Considérons  $H = V = \mathbb{R}^d$  muni du produit scalaire canonique. Soit  $A$  une matrice symétrique définie positive. L'application  $a: V \times V \rightarrow \mathbb{R}$  définie par  $a(u, v) = \langle Au, v \rangle$  satisfait bien les conditions du Corollaire 1, donc il existe une famille  $(\omega_n, \mu_n)_n$  telle que  $(\omega_n)_n$  soit une base de  $\mathbb{R}^d$ , ce qui impose que  $n \in [1, d]$ , et  $\mu_n$  est une suite croissante telle que

$$A\omega_n = \lambda_n \omega_n.$$

Notons  $O := (\omega_1, \dots, \omega_n)$  la matrice dont les colonnes sont les vecteurs  $\omega_n$ , et  $\Lambda$  la matrice dont la diagonale est la suite  $(\lambda_1, \dots, \lambda_n)$ , et les termes extradiagonaux sont nuls. Comme  $\langle \omega_n, \omega_m \rangle = 1$  si  $n = m$  et 0 sinon, on a  $O^t O = O O^t = Id$ , donc la matrice  $O$  est orthogonale. L'égalité précédente s'écrit

$$AO = O\Lambda \iff O^t AO = \Lambda.$$

On a donc diagonalisé  $A$ . Comme le Corollaire 1 est valide pour des espaces de Hilbert, on parle ici aussi de théorème de diagonalisation.

- Prenons un exemple canonique de problème elliptique. Soit  $\Omega \subset \mathbb{R}^d$  un ouvert borné régulier. On considère données des fonctions  $a_{ij} \in L^\infty(\Omega; \mathbb{R})$ ,  $(i, j) \in [1, d]^2$ , et  $a_0 \in L^\infty(\Omega; \mathbb{R}^+)$  telles que  $a_{ij} = a_{ji}$  et qui satisfont la condition de non-dégénérescence

$$\exists \alpha > 0, \quad \forall v \in \mathbb{R}^d, \quad \sum_{i=1}^d \sum_{j=1}^d v_i a_{ij}(x) v_j \geq \alpha |v|^2 \quad \forall x \in \Omega.$$

Posons

$$A: \mathcal{D}'(\Omega) \rightarrow \mathcal{D}'(\Omega), \quad Au := - \sum_{i, j \in [1, d]^2} \partial_{x_i} (a_{ij} \partial_{x_j} u) + a_0 u, \quad (6.5)$$

et considérons le problème suivant : pour  $f \in L^2(\Omega; \mathbb{R})$ , trouver  $u \in H^1(\Omega; \mathbb{R})$  solution de

$$\begin{cases} Au(x) = f(x) & x \in \Omega, \\ u(x) = 0 & x \in \partial\Omega. \end{cases} \quad (6.6a)$$

$$(6.6b)$$

On choisit maintenant  $H := L^2(\Omega; \mathbb{R})$ ,  $V := H_0^1(\Omega; \mathbb{R})$  et on considère la formulation variationnelle de (6.6), qui s'écrit

$$a(u, v) = \langle f, v \rangle_H$$

où la forme bilinéaire symétrique définie positive  $a: V \times V \rightarrow \mathbb{R}$  est donnée par

$$a(u, v) := \sum_{i,j \in \llbracket 1, d \rrbracket^2} \int_{x \in \Omega} a_{ij}(x) \partial_{x_j} u(x) \partial_{x_i} v(x) dx + \int_{x \in \Omega} a_0(x) u(x) v(x) dx.$$

Comme l'injection  $V \hookrightarrow H$  est compacte (par la Proposition 4), le Corollaire 1 s'applique.

## 6.3 Caractérisation des valeurs propres

Dans toute cette partie, on note  $(H, \langle \cdot, \cdot \rangle_H)$  un espace de Hilbert,  $(V, \langle \cdot, \cdot \rangle_V)$  un sous-espace de Hilbert tel que l'injection canonique de  $V$  dans  $H$  soit compacte, et  $a: V \times V \rightarrow \mathbb{R}$  une forme bilinéaire continue symétrique et coercive dans  $V$ .

### 6.3.1 Quotients de Rayleigh

**Définition 38 – Quotient de Rayleigh** Le quotient de Rayleigh de  $a(\cdot, \cdot)$  est l'application  $R: V \setminus \{0_V\} \rightarrow \mathbb{R}^+$  définie par

$$R(v) = \frac{a(v, v)}{\|v\|_H^2} \quad \forall v \in V \setminus \{0_V\}.$$

**Lemme 23 – Première caractérisation ([RT88, (6.2-18) p. 139])** Par le Corollaire 1, il existe une famille  $(\omega_n, \mu_n)_{n \in \mathbb{N}} \subset H \times \mathbb{R}$  telle que la suite  $(\mu_n)_n$  soit positive et croissante, et  $a(\omega_n, v) = \mu_n \langle \omega_n, v \rangle_H$  pour tout  $v \in V$ . Définissons par récurrence la suite de sous-espaces

$$E_0 := \emptyset, \quad E_n = \text{Vect}(\{\omega_1, \dots, \omega_n\}).$$

Alors

$$\mu_n = \max\{R(v) \mid v \in E_n\} = \min\{R(v) \mid v \in E_{n-1}^{\perp a}, v \neq 0\}, \quad \text{où} \quad E_{n-1}^{\perp a} := \{v \in V \mid a(v, w) = 0 \quad \forall w \in E_{n-1}\}.$$

#### Démonstration

Par Remarque 16, la suite  $(\frac{1}{\sqrt{\mu_n}} \omega_n)_n$  est une base  $a(\cdot, \cdot)$ -orthonormale de  $V$ . Décomposons tout  $v \in V$  en  $v = \sum_i \alpha_i \omega_i$ . Si  $v \in E_n$ , alors  $\alpha_i = 0$  pour tout  $i > n$ , et

$$R(v) = \frac{a(v, v)}{\|v\|_H^2} = \frac{\sum_{i \leq n} \alpha_i^2 a(\omega_i, \omega_i)}{\|v\|_H^2} = \frac{\sum_{i \leq n} \alpha_i^2 \mu_i \times 1}{\sum_{i \leq n} \alpha_i^2} \leq \mu_n \frac{\sum_{i \leq n} \alpha_i^2}{\sum_{i \leq n} \alpha_i^2} = \mu_n.$$

Comme  $\omega_n \in E_n$  et que  $R(\omega_n) = \lambda_n$ , le sup est atteint.

D'autre part, si  $v \in E_{n-1}^{\perp a}$ , alors pour tout  $j \in \llbracket 1, n-1 \rrbracket$

$$a(v, \omega_j) = \sum_{i \geq n} \alpha_i a(\omega_i, \omega_j) = 0,$$

donc  $v = \sum_{i \geq n} \alpha_i \omega_i$  pour tout  $v \in E_{n-1}^{\perp a}$ . Pour un tel  $v$ ,

$$R(v) = \frac{a(v, v)}{\|v\|_H^2} = \frac{\sum_{i \geq n} \alpha_i^2 a(\omega_i, \omega_i)}{\|v\|_H^2} = \frac{\sum_{i \geq n} \alpha_i^2 \mu_i \times 1}{\sum_{i \geq n} \alpha_i^2} \geq \mu_n \frac{\sum_{i \geq n} \alpha_i^2}{\sum_{i \geq n} \alpha_i^2} = \mu_n.$$

On en déduit que  $\sup_{v \in E_n} R(v) \leq \lambda_n \leq \inf_{v \in E_{n-1}^{\perp a}} R(v)$ . Comme  $\omega_n \in E_{n-1}^{\perp a}$  et que  $R(\omega_n) = \lambda_n$ , l'inf est atteint.  $\square$

**Théorème 14 – Théorème de Courant-Fischer ou principe du min-max ([RT88, Théorème 6.2-2])** Notons  $\mathcal{V}_n$  l'ensemble des sous-espaces vectoriels de  $V$  de dimension  $n$ . Alors

$$\mu_n = \min_{V_n \in \mathcal{V}_n} \max_{v \in V_n \setminus \{0_V\}} R(v) = \max_{V_{n-1} \in \mathcal{V}_{n-1}} \min_{v \in V_{n-1}^{\perp a} \setminus \{0_V\}} R(v).$$

### Démonstration

Si  $V_n$  est un sous-espace de dimension  $n > 0$ , alors  $V_n \cap E_{n-1}^{\perp a} \neq \{0\}$ . En effet, on aurait sinon  $V_n \subset E_{n-1}$ , ce qui est impossible car  $\dim E_{n-1} = n-1$ . Soit  $v^* \neq 0$  tel que  $v^* \in V_n$  et  $v^* \in E_{n-1}^{\perp a}$  : par le Lemme 23,  $R(v^*) \geq \mu_n$ . Donc pour tout  $V_n \in \mathcal{V}_n$ , on a  $\mu_n \leq \max_{v \in V_n \setminus \{0\}} R(v)$ , et comme  $V_n$  est arbitraire,

$$\mu_n \leq \inf_{V_n \in \mathcal{V}_n} \max_{v \in V_n \setminus \{0\}} R(v).$$

En prenant  $V_n = E_n$  et  $v = \omega_n$ , on obtient l'égalité.

D'autre part, soit  $V_{n-1} \in \mathcal{V}_{n-1}$ . Par le même raisonnement de dimension,  $V_{n-1}^{\perp a} \cap E_n \neq \{0\}$ , donc il existe  $v^* \in (V_{n-1}^{\perp a} \cap E_n) \setminus \{0\}$ . Par le Lemme 23,  $\mu_n \geq R(v^*)$ , et

$$\mu_n \geq \min_{v \in V_{n-1}^{\perp a}} R(v) \quad \forall V_{n-1} \in \mathcal{V}_{n-1} \quad \Rightarrow \quad \mu_n \geq \sup_{V_{n-1} \in \mathcal{V}_{n-1}} \min_{v \in V_{n-1}^{\perp a}} R(v).$$

En prenant  $V_{n-1} = E_{n-1}$  et  $v = \omega_n$ , on obtient l'égalité, et le sup est un max.  $\square$

### 6.3.2 [Non traité en cours] Approximation via des problèmes en dimension finie

On s'intéresse à une approximation des valeurs propres du problème (6.1) par une version discrète du principe du min-max. Soit  $(V_h)_{h \in \mathbb{R}^+}$  une suite de sous-espaces de  $V$ , tels que  $V_h \subset V$  soit de dimension finie. On note encore  $a(\cdot, \cdot)$  la restriction de  $a(\cdot, \cdot)$  à  $V_h$ . Par le Théorème 14, on sait que les valeurs propres discrètes associées au problème

$$\text{Trouver } (\omega_h, \mu_h) \in \mathbb{R} \times V_h \text{ tels que } a(\omega_h, v_h) = \mu_h \langle \omega_h, v_h \rangle_H \quad \forall v_h \in V_h \quad (\text{PVPh})$$

forment une suite croissante  $(\mu_{h,i})_{i \in \llbracket 1, \dim V_h \rrbracket}$ , et sont caractérisées par

$$\mu_{h,m} = \min_{E_{m,h} \in \mathcal{V}_{m,h}} \max_{v_h \in E_{m,h} \setminus \{0\}} R(v_h),$$

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

**Remarque 17** (Ordre naturel). *On a l'inclusion  $\mathcal{V}_{m,h} \subset \mathcal{V}_m$ , donc (le minimum sur un sous-ensemble réduit est plus grand que sur tout l'ensemble) les valeurs propres discrètes satisfont naturellement  $\mu_{m,h} \geq \mu_m$ . Pour obtenir un résultat de convergence, il suffit donc de trouver une inégalité de la forme  $\mu_{m,h} \leq \mu_m + \epsilon(h)$ .*

L'hypothèse suivante est la plus naturelle que l'on puisse formuler pour espérer une convergence des valeurs propres.

**Densité des sous-espaces de dimension finie** La suite d'espaces  $(V_h)_h$  converge vers  $V$  au sens suivant :

(A1)

$$\forall u \in V, \quad \lim_{h \searrow 0} \inf_{v_h \in V_h} \|u - v_h\|_V^2 = 0.$$

**Théorème 15 – Convergence de l'approximation des valeurs propres ([RT88, Théorème 6.4-2])** Supposons que les hypothèses du Corollaire 1 soient satisfaites, ainsi que (A1). Alors il existe une constante  $c(m)$  telle que pour  $h$  suffisamment petit,

$$0 \leq \mu_{m,h} - \mu_m \leq 2 \frac{\mu_m M}{\alpha} c(m) \sup_{\substack{v \in V_m \\ \|v\|_H=1}} \inf_{v_h \in V_h} \|v - v_h\|_V^2 \xrightarrow[h \searrow 0]{} 0.$$

Pour démontrer ce théorème, on s'accorde deux lemmes. Pour chaque  $V_h$ , notons  $\Pi_h : V \rightarrow V_h$  l'opérateur de projection orthogonale dans l'espace  $V$  muni du produit scalaire  $a(\cdot, \cdot)$ , i.e. l'unique application telle que

$$\forall (u, v_h) \in V \times V_h, \quad a(v_h, u - \Pi_h u) = 0.$$

**Lemme 24** Soit  $V_m \subset V$  un sous-espace de  $V$  de dimension  $m$ . Posons  $\sigma_{m,h} = \min_{v \in V_m \setminus \{0\}} |\Pi_h v|_H^2 / |v|_H^2$ . Si  $\sigma_{m,h} > 0$ , alors  $\mu_{m,h} \leq \mu_m / \sigma_{m,h}$ .

### Démonstration

La restriction de  $\Pi_h$  à  $V_m$  est une application linéaire d'un espace vectoriel de dimension finie dans un autre espace vectoriel de dimension finie : on a donc équivalence entre  $\dim \Pi_h V_m = \dim V_m = m$  et le caractère bijectif de  $(\Pi_h)|_{V_m}$ . Supposons que  $\dim \Pi_h V_m < m$  : alors  $\dim(\ker_{V_m}(\Pi_h)) \neq 0$ , et il existe un vecteur  $v^* \in V_m$  non nul tel que  $\Pi_h v^* = 0$ . D'où  $\sigma_{m,h} = 0$ . Par contraposée, si  $\sigma_{m,h} > 0$ , le sous-espace  $\Pi_h V_m \subset V_h$  est de dimension  $m$ , donc dans  $\mathcal{V}_{m,h}$ . Dès lors,

$$\mu_{m,h} = \min_{E_{m,h} \in \mathcal{V}_{m,h}} \max_{v_h \in E_{m,h} \setminus \{0\}} R(v_h) \leq \max_{v_h \in \Pi_h V_m \setminus \{0\}} R(v_h) = \max_{v_h \in \Pi_h V_m \setminus \{0\}} \frac{a(v_h, v_h)}{|v_h|_H^2}.$$

Notons  $v_h = \Pi_h v$  pour  $v \in V_m$ . Par le caractère contractant de la projection, on a  $a(v_h, v_h) \leq a(v, v)$ , ce qui mène à

$$\mu_{m,h} = \max_{v \in V_m \setminus \{0\}} \frac{a(v_h, v_h)}{|v_h|_H^2} \leq \max_{v \in V_m \setminus \{0\}} \frac{a(v, v)}{|\Pi_h v|_H^2} \leq \left( \max_{v \in V_m \setminus \{0\}} \frac{a(v, v)}{|v|_H^2} \right) \times \left( \max_{v \in V_m \setminus \{0\}} \frac{|v|_H^2}{|\Pi_h v|_H^2} \right) = \frac{\mu_m}{\sigma_{m,h}}.$$

□

**Lemme 25** Pour tout entier  $m \geq 1$ , il existe une constante  $c(m) > 0$  telle que pour tout sous-espace  $V_h \subset V$  de dimension  $i \geq m$ , on ait

$$\sigma_{m,h} \geq 1 - c(m) \times \max_{\substack{w \in V_m \\ |w|_H=1}} |w - \Pi_h w|_V^2.$$

### Démonstration

Soit  $v \in V_m$ . On a

$$0 \leq |v|_H^2 - |\Pi_h v|_H^2 = \langle v - \Pi_h v, v + \Pi_h v \rangle_H = -\langle v - \Pi_h v, v - \Pi_h v \rangle_H + 2 \langle v - \Pi_h v, v \rangle_H.$$

Comme  $\langle v - \Pi_h v, v - \Pi_h v \rangle_H \geq 0$ , on déduit  $\langle v - \Pi_h v, v \rangle_H \geq 0$ . Estimons ce terme grâce à la continuité de  $a(\cdot, \cdot)$ . Notons encore  $(\omega_i)_i$  la base des vecteurs propres associés à  $a(\cdot, \cdot)$ . En décomposant  $V_m \ni v = \sum_{i=1}^m \alpha_i \omega_i$ , et en utilisant la propriété  $a(v - \Pi_h v, \Pi_h \omega_i) = 0$ , on obtient

$$\langle v - \Pi_h v, v \rangle_H = \sum_{i=1}^m \alpha_i \langle v - \Pi_h v, \omega_i \rangle_H = \sum_{i=1}^m \frac{\alpha_i}{\mu_i} a(v - \Pi_h v, \omega_i) = \sum_{i=1}^m \frac{\alpha_i}{\mu_i} a(v - \Pi_h v, \omega_i - \Pi_h \omega_i).$$

Notons  $M$  une constante telle que  $a(u, v) \leq M |u|_V |v|_V$  pour tout  $(u, v) \in V^2$ . Par l'inégalité de Cauchy-Schwarz,

$$\sum_{i=1}^m |\alpha_i| \leq \left( \sum_{i=1}^m 1 \right)^{1/2} \times \left( \sum_{i=1}^m \alpha_i^2 \right)^{1/2} = \sqrt{m} |v|_H.$$

De plus,  $\mu_i \geq \mu_1 > 0$ , et  $|\omega_i|_H = 1$ . Ainsi

$$|\langle v - \Pi_h v, v \rangle_H| \leq \sum_{i=1}^m \frac{|\alpha_i|}{\mu_i} M |v|_H \frac{|v - \Pi_h v|_V}{|v|_H} \frac{|\omega_i - \Pi_h \omega_i|_V}{|\omega_i|_H} \leq M \frac{\sqrt{m}}{\mu_1} |v|_H^2 \times \max_{w \in V_m \setminus \{0\}} \frac{|w - \Pi_h w|_V^2}{|w|_H^2}.$$

Comme  $\beta |w - \Pi_h w|_V = |\beta w - \Pi_h(\beta w)|_V$  pour tout  $\beta > 0$ , il suffit de prendre le max sur les vecteurs  $v \in V_m$  tels que  $|v|_H = 1$ . Posons alors  $c(m) = \frac{2M\sqrt{m}}{\mu_1}$ . On a obtenu

$$|\Pi_h v|_H^2 \geq |v|_H^2 - c(m) |v|_H^2 \max_{\substack{w \in V_m \\ |w|_H=1}} |w - \Pi_h w|_V^2.$$

En divisant par  $|v|_H^2$  et en passant au minimum sur  $v \in V_m \setminus \{0\}$ , on obtient l'estimation désirée. □

Revenons maintenant à la preuve du Théorème 15.

### Démonstration du Théorème 15

Par minimalité de la projection,  $a(u - \Pi_h u, u - \Pi_h u) \leq a(u - v_h, u - v_h)$  pour tout  $v_h \in V_h$ . D'où, par les hypothèses sur  $a(\cdot, \cdot)$ ,

$$\alpha |u - \Pi_h u|_V^2 \leq a(u - \Pi_h u, u - \Pi_h u) \leq a(u - v_h, u - v_h) \leq M |u - v_h|_V^2$$

pour tout  $v_h \in V_h$ , et en passant à l'infimum, on en déduit que  $|u - \Pi_h u|_V \leq \sqrt{\frac{M}{\alpha}} \inf_{v \in V_h} |u - v|_V^2$ . Montrons que cette approximation point par point (pour chaque  $u$ ) devient uniforme sur  $V_m$  : on pose  $\varepsilon(h) = \sum_{i \in \llbracket 1, m \rrbracket} |\omega_i - \Pi_h \omega_i|_V^2 \xrightarrow[h \searrow 0]{} 0$ , par somme sur un nombre fini d'éléments. Pour tout  $v = \sum_{i=1}^m \alpha_i \omega_i$ , on a

$$|v - \Pi_h v|_V^2 = \left| \sum_{i=1}^m \alpha_i (\omega_i - \Pi_h \omega_i) \right|_V^2 \leq \left( \sum_{i=1}^m \alpha_i^2 \right) \left( \sum_{i \in \llbracket 1, m \rrbracket} |\omega_i - \Pi_h \omega_i|_V^2 \right) = |v|_H^2 \varepsilon(h).$$

D'où

$$\lim_{h \searrow 0} \sup_{v \in V_m \setminus \{0\}} \frac{|v - \Pi_h v|}{|v|_H^2} = 0.$$

Grâce aux lemmes précédents,

$$\mu_{m,h} \leq \frac{\mu_m}{\sigma_{m,h}} \leq \frac{\mu_m}{1 - c(m) \max_{\substack{w \in V_m \\ |w|_H=1}} |w - \Pi_h w|_V^2} =: \frac{\mu_m}{1 - a_h}.$$

En développant  $\frac{1}{1-a_h}$  en la série  $1 + \sum_{n \geq 1} a_h^n = 1 + a_h \sum_{n \geq 0} a_h^n$ , et en prenant  $h$  suffisamment petit pour que  $\sum_{n \geq 0} a_h^n \leq 2$ , on obtient  $\mu_{m,h} \leq \mu_m (1 + 2a_h)$ , c'est-à-dire le résultat désiré.  $\square$

# Chapitre 7

## Problèmes d'évolution

On considère maintenant la classe de problèmes suivants : les problèmes paraboliques

$$\left\{ \begin{array}{l} \partial_t u(t, x) - \Delta u(t, x) = f(t, x) \\ u(0, x) = u_0(x) \\ u(t, x) = 0 \end{array} \right. \quad \begin{array}{l} (t, x) \in ]0, T[ \times \Omega, \\ x \in \Omega, \\ (t) \in ]0, T[ \times \partial\Omega. \end{array} \quad \begin{array}{l} (7.1a) \\ (7.1b) \\ (7.1c) \end{array}$$

et les problèmes hyperboliques

$$\left\{ \begin{array}{l} \partial_t^2 u(t, x) - \Delta u(t, x) = f(t, x) \\ u(0, x) = u_0(x), \quad \partial_t u(0, x) = u_1(x) \\ u(t, x) = 0 \end{array} \right. \quad \begin{array}{l} (t, x) \in ]0, T[ \times \Omega, \\ x \in \Omega, \\ (t) \in ]0, T[ \times \partial\Omega. \end{array} \quad \begin{array}{l} (7.2a) \\ (7.2b) \\ (7.2c) \end{array}$$

### 7.1 Forme variationnelle

Les problèmes (7.1) et (7.2) ne peuvent pas se mettre sous la forme d'un problème elliptique, et on ne peut pas directement appliquer la théorie précédente. Par contre, en séparant les rôles du temps et de l'espace, on peut déduire des résultats d'existence et des formules explicites. Pour cela, on procède de manière analogue au cas elliptique : on introduit une formulation variationnelle, que l'on résout dans un certain espace fonctionnel avec les outils de l'analyse spectrale. On s'intéresse au problème parabolique à titre d'exemple.

Prenons le produit scalaire de (7.1a) dans  $L^2(\Omega; \mathbb{R})$  avec une fonction test  $v$  qui ne dépend que de l'espace, pour l'instant considérée aussi régulière que désiré. On obtient par intégration par partie formelle

$$\begin{aligned} \langle \partial_t u(t, \cdot), v \rangle_{L^2} - \langle \Delta u(t, \cdot), v \rangle_{L^2} &= \langle f(t, \cdot), v \rangle_{L^2} \\ \partial_t \langle u(t, \cdot), v \rangle_{L^2} + \langle \nabla u(t, \cdot), \nabla v \rangle_{L^2} &= \langle f(t, \cdot), v \rangle_{L^2}. \end{aligned}$$

Dans la suite, on va séparer les variables de temps et d'espace en considérant que  $u$  est une fonction du temps à valeurs dans les fonctions de l'espace, c'est-à-dire que pour tout  $t$ ,  $u(t)$  est une fonction de  $x$ . Le choix d'un espace fonctionnel pour  $u$  et  $f$  ressemblera donc à  $\mathcal{C}([0, T]; H_0^1(\Omega))$ , ou  $L^2(0, T; W^{4,\infty}(\Omega))$ ...

De manière générale, on s'intéressera à la formulation variationnelle suivante.

**Définition 39 – Problème parabolique** Soient  $H, V$  deux espaces de Hilbert tels que  $V \subset H$  avec injection compacte. Soit  $a(\cdot, \cdot) : V \rightarrow \mathbb{R}$  une forme bilinéaire symétrique continue et coercive dans  $V$ . Soient  $f \in L^2(0, T; H)$  et  $u_0 \in H$ . On s'intéresse au problème sous forme variationnelle

$$\text{Trouver } u \in L^2(0, T; V) \cap \mathcal{C}([0, T]; H) \text{ telle que } \partial_t \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H, \quad u(0) = u_0. \quad (7.3)$$

Ici, la dérivée en temps est prise au sens des distributions : pour tout  $\varphi \in \mathcal{C}_c^\infty([0, T[; \mathbb{R})$ ,

$$0 = \int_{t=0}^T [-\varphi'(t) \langle u(t), v \rangle_H + \varphi(t) (a(u(t), v) - \langle f(t), v \rangle_H)] dt.$$

**Théorème 16 – Existence et unicité du cas parabolique** On considère les hypothèses de la Définition 39, et on note  $\alpha > 0$  la constante de coercivité de  $a(\cdot, \cdot)$  dans  $V$ . Soit  $(\mu_n, \omega_n)_{n \in \mathbb{N}}$  une famille d'éléments propres de  $a(\cdot, \cdot)$  donnée par le Corollaire 1. Le problème (7.3) admet une unique solution, donnée par

$$\bar{u}(t, x) = \sum_{n \in \mathbb{N}} \left[ \exp(-\mu_n t) \langle u_0, \omega_n \rangle_H + \int_{s=0}^t \exp(\mu_n(s-t)) \langle f(s), \omega_n \rangle_H ds \right] \omega_n, \quad (7.4)$$

et qui satisfait l'estimation d'énergie

$$\|u\|_{L^2(0, T; V)}^2 \leq \frac{1}{\alpha} \|u_0\|_H^2 + \frac{2}{\alpha \mu_1} \|f\|_{L^2(0, T; H)}^2.$$

Idée de la preuve : raisonnement par analyse-synthèse. Soit  $U$  l'ensemble des solutions de (7.1), et  $\bar{u}$  la fonction donnée par (7.4).

- On suppose d'abord que  $U$  est non vide, et on montre que tout élément  $u \in U$  est égal à  $\bar{u}$  : pour ceci, on décompose  $u(t) = \sum_{n \in \mathbb{N}} \xi_n(t) \omega_n$  dans la base des vecteurs propres, puis on montre que  $\xi_n$  est solution de l'EDO  $\dot{\xi}_n + \mu_n \xi_n = \langle f(t), \omega_n \rangle_H$ , et on résout explicitement cette EDO pour trouver la forme de  $\bar{u}$ . À ce stade,  $U \subset \{\bar{u}\}$ .
- On montre ensuite que  $\bar{u}$  est solution, donc que  $U$  est non vide. Pour ceci, la construction de  $\bar{u}$  assure que l'équation (7.1) est satisfaite au sens des distributions. Il reste à montrer que  $\bar{u} \in L^2(0, T; V) \cap C([0, T]; H)$ . Pour montrer que  $\bar{u} \in L^2(0, T; V)$ , on obtient une majoration du type  $\alpha \|\bar{u}\|_{L^2(0, T; V)}^2 \leq \|u_0\|_H^2 + \frac{2}{\mu_1} \|f\|_{L^2(0, T; H)}^2$ , où  $\alpha > 0$  est la constante de coercivité de  $a(\cdot, \cdot)$ , et  $\mu_1 > 0$  la plus petite valeur propre. Pour montrer la continuité, on emploie le théorème de convergence dominée : on majore  $\|u(t) - u(s)\|_H^2$  par une série dont chaque terme tend vers 0 quand  $s \rightarrow t$ , puis on montre que la série est bornée uniformément, et on en déduit la convergence.

### Démonstration

On procède de la manière suivante : soit  $U$  l'ensemble des solutions. On montre premièrement que si  $u \in U$ , alors  $u = \bar{u}$  où  $\bar{u}$  est donné par (7.4). Cela montre que  $U \subset \{\bar{u}\}$ , mais à ce stade,  $U$  pourrait être vide. Pour conclure, on montre que  $\bar{u} \in U$ , ce qui prouve que  $U = \{\bar{u}\}$ .

**Étape 1 : Décomposition dans une base** Soit  $u \in L^2(0, T; V) \cap C([0, T]; H)$  satisfaisant (7.3). Par le Corollaire 1,  $H$  admet une base hilbertienne  $(\omega_n)_n$  formée de vecteurs propres de  $a(\cdot, \cdot)$ . Notons  $\mu_n$  la valeur propre associée à  $\omega_n$ . Comme  $u(t) \in H$  pour tout  $t$ , on a

$$u(t) = \sum_{n \in \mathbb{N}} \langle u(t), \omega_n \rangle_H \omega_n =: \sum_{n \in \mathbb{N}} \xi_n(t) \omega_n.$$

Comme  $t \mapsto u(t)$  est continue dans  $H$ , on a en particulier  $t \mapsto \langle u(t), \omega_n \rangle = \xi_n(t)$  continue, donc  $\xi_n \in C([0, T]; \mathbb{R})$  pour tout  $n \in \mathbb{N}$ . En particulier, la condition initiale  $u(0) = u_0$  nous indique que

$$\langle u(0), \omega_m \rangle_H = \xi_m(0) = \langle u_0, \omega_m \rangle_H.$$

Ici, la continuité de  $u(\cdot)$  permet de manipuler chaque  $u(t)$  : par contre, pour  $f \in L^2(0, T; H)$ , on ne peut pas parler de la valeur de  $f(t)$  pour tout  $t$ , seulement du produit scalaire de  $f$  avec une fonction test de  $L^2(0, T; H)$ .

Injectons la décomposition de  $u$  dans l'équation (7.4). Pour tout  $\varphi \in C_c^\infty([0, T]; \mathbb{R})$ , on a

$$0 = \int_{t=0}^T [-\varphi'(t) \langle u(t), v \rangle_H + \varphi(t) (a(u(t), v) - \langle f(t), v \rangle_H)] dt = \int_{t=0}^T \left[ -\varphi'(t) \sum_{n \in \mathbb{N}} \xi_n(t) \langle \omega_n, v \rangle_H + \varphi(t) \left( \sum_{n \in \mathbb{N}} \xi_n(t) a(\omega_n, v) - \langle f(t), v \rangle_H \right) \right] dt.$$

En particulier, pour  $v = \omega_m$ , on a

$$\langle \omega_n, \omega_m \rangle_H = \delta_{n,m} := \begin{cases} 1 & \text{si } n = m, \\ 0 & \text{sinon} \end{cases}, \quad a(\omega_n, \omega_m) = \mu_n \langle \omega_n, \omega_m \rangle_H = \mu_n \delta_{n,m}.$$

Ainsi,

$$0 = \int_{t=0}^T [-\varphi'(t) \xi_m(t) + \varphi(t) (\mu_m \xi_m(t) - \langle f(t), \omega_m \rangle_H)] dt. \quad (7.5)$$

Ici apparaît le terme  $\int_{t=0}^T \varphi(t) \langle f(t), \omega_m \rangle_H dt = \langle f, \varphi \otimes \omega_m \rangle_{L^2(0, T; H)}$ . La fonction test  $(t, x) \mapsto \varphi(t) \omega_m(x)$  appartient bien à  $L^2(0, T; H)$ , donc le calcul est licite.

**Étape 2 : Résolution en temps** L'équation (7.5) est exactement la formulation au sens des distributions de l'EDO

$$\frac{d}{dt} \xi_m(t) + \mu_m \xi_m(t) = \langle f(t), \omega_m \rangle_H. \quad (7.6)$$

Vérifions que cette EDO est bien posée : (7.6) s'écrit  $\frac{d}{dt} \xi(t) = F_m(t, \xi(t))$ , où  $F_m(t, r) = \langle f(t), \omega_m \rangle_H - \mu_m r$ . La fonction  $F_m$  est Lipschitz-continue en la variable  $r$  et mesurable en temps (car  $t \mapsto \langle f(t), \omega_m \rangle_H \in L^2(0, T; \mathbb{R})$ ) : on peut appliquer le théorème d'existence et d'unicité de Cauchy-Lipschitz. D'autre part, la méthode de la variation de la constante permet de déterminer que l'unique solution s'écrit

$$\xi_m(t) = \exp(-\mu_m t) \xi_m(0) + \int_{s=0}^t \exp(\mu_m(s-t)) \langle f(s), \omega_m \rangle_H ds.$$

Encore une fois, l'écriture est licite, car la fonction  $(s, x) \mapsto \exp(\mu_m(s-t)) \omega_m(x)$  appartient bien à  $L^2(0, T; H)$ . On sait que la condition initiale  $\xi_m(0)$  est donnée par  $\langle u_0, \omega_m \rangle_H$ , donc la fonction  $\xi_m$  est bien uniquement déterminée. On a montré que si  $u$  est solution de (7.4), alors

$$u(t) = \sum_{m \in \mathbb{N}} \left[ \exp(-\mu_m t) \langle u_0, \omega_m \rangle_H + \int_{s=0}^t \exp(\mu_m(s-t)) \langle f(s), \omega_m \rangle_H ds \right] \omega_m = \bar{u}(t).$$

**Étape 3 : Régularité de la solution** Montrons que la fonction  $u$  de (7.4) est bien solution de l'équation, ce qui comporte plusieurs conditions : l'égalité au sens des distributions, qui est assurée par construction, mais également la *bonne régularité de  $u$* .

Ici par exemple, on ne peut pas assurer que  $u$  soit dans  $C^\infty([0, T]; H)$  : montrons que l'on a bien  $u \in L^2(0, T; V) \cap C([0, T]; H)$ . Pour montrer que  $u \in L^2(0, T; V)$ , on utilise la coercivité du produit scalaire  $a(\cdot, \cdot)$  pour simplifier les calculs. Soit  $\alpha > 0$  une constante telle que  $a(v, v) \geq \alpha |v|_V^2$ . On a

$$\|u\|_{L^2(0, T; V)}^2 = \int_{t=0}^T \|u(t)\|_V^2 dt \leq \frac{1}{\alpha} \int_{t=0}^T a(u(t), u(t)) dt = \frac{1}{\alpha} \int_{t=0}^T \sum_{n \in \mathbb{N}} \sum_{m \in \mathbb{N}} \xi_n(t) \xi_m(t) a(\omega_n, \omega_m) dt.$$

En utilisant à nouveau  $a(\omega_n, \omega_m) = \mu_n \delta_{n,m}$ , il vient

$$\|u\|_{L^2(0, T; V)}^2 \leq \frac{1}{\alpha} \int_{t=0}^T \sum_{n \in \mathbb{N}} \mu_n \xi_n^2(t) dt = \frac{1}{\alpha} \sum_{n \in \mathbb{N}} \mu_n \|\xi_n\|_{L^2(0, T; \mathbb{R})}^2.$$

Calculons donc

$$\begin{aligned} \|\xi_n\|_{L^2(0, T; \mathbb{R})}^2 &= \int_{t=0}^T \left| \xi_n(0) e^{-\mu_n t} + \int_{s=0}^t \langle f(s), \omega_n \rangle e^{\mu_n(s-t)} ds \right|^2 dt \\ &\leq 2 \int_{t=0}^T |\xi_n(0)|^2 e^{-2\mu_n t} dt + 2 \int_{t=0}^T \left| \int_{s=0}^t \langle f(s), \omega_n \rangle e^{\mu_n(s-t)} ds \right|^2 dt. \end{aligned}$$

On emploie l'astuce de calcul suivante : on écrit  $e^{\mu_n(s-t)} = e^{\mu_n(s-t)/2} e^{\mu_n(s-t)/2}$ , et on applique Cauchy-Schwarz pour avoir

$$\begin{aligned} \int_{s=0}^t \langle f(s), \omega_n \rangle e^{\mu_n(s-t)/2} e^{\mu_n(s-t)/2} ds &\leq \left( \int_{s=0}^t |\langle f(s), \omega_n \rangle|^2 e^{\mu_n(s-t)} ds \right)^{1/2} \left( \int_{s=0}^t e^{\mu_n(s-t)} ds \right)^{1/2} \\ &\leq \left( \int_{s=0}^t |\langle f(s), \omega_n \rangle|^2 e^{\mu_n(s-t)} ds \right)^{1/2} \frac{1}{\sqrt{\mu_n}}. \end{aligned}$$

Puis, grâce au théorème de Fubini,

$$\begin{aligned} \|\xi_n\|_{L^2(0, T; \mathbb{R})}^2 &\leq 2 \int_{t=0}^T |\xi_n(0)|^2 e^{-2\mu_n t} dt + 2 \int_{t=0}^T \left( \int_{s=0}^t |\langle f(s), \omega_n \rangle|^2 e^{\mu_n(s-t)} ds \right) \frac{1}{\mu_n} dt \\ &\leq \frac{|\xi_n(0)|^2}{\mu_n} + 2 \int_{s=0}^T |\langle f(s), \omega_n \rangle|^2 \int_{t=s}^T e^{\mu_n(s-t)} dt \frac{1}{\mu_n} ds \leq \frac{|\xi_n(0)|^2}{\mu_n} + \frac{2}{\mu_n^2} \int_{s=0}^T |\langle f(s), \omega_n \rangle|^2 ds. \end{aligned}$$

Ainsi, en se rappelant que la suite  $(\mu_n)_n$  est croissante, donc que  $\frac{1}{\mu_n} \leq \frac{1}{\mu_1}$ , on obtient

$$\sum_{n \in \mathbb{N}} \mu_n \|\xi_n\|_{L^2(0, T; \mathbb{R})}^2 \leq \sum_{n \in \mathbb{N}} |\xi_n(0)|^2 + \frac{2}{\mu_1} \int_{s=0}^T |\langle f(s), \omega_n \rangle|^2 ds = \|u_0\|_H^2 + \frac{2}{\mu_1} \int_{s=0}^T \|f(s)\|_H^2 ds = \|u_0\|_H^2 + \frac{2}{\mu_1} \|f\|_{L^2(0, T; H)}^2 < \infty,$$

et la fonction  $u$  appartient bien à  $L^2(0, T; V)$ .

Pour montrer la continuité, utilise le théorème de la convergence dominée. Pour  $0 \leq s \leq t \leq T$ , en utilisant que  $|e^{-a} - e^{-b}| \leq 1$  pour  $a, b \geq 0$ , on a

$$\begin{aligned} |u(t) - u(s)|_H^2 &= \sum_{n \in \mathbb{N}} \left| (e^{-\mu_n t} - e^{-\mu_n s}) \langle u_0, \omega_n \rangle + \int_{r=0}^t e^{\mu_n(r-t)} \langle f(r), \omega_n \rangle_H dr - \int_{r=0}^s e^{\mu_n(r-s)} \langle f(r), \omega_n \rangle_H dr \right|^2 \\ &= \sum_{n \in \mathbb{N}} \left| (e^{-\mu_n t} - e^{-\mu_n s}) \langle u_0, \omega_n \rangle + \int_{r=s}^t e^{\mu_n(r-t)} \langle f(r), \omega_n \rangle_H dr + \int_{r=0}^s (e^{\mu_n(r-t)} - e^{\mu_n(r-s)}) \langle f(r), \omega_n \rangle_H dr \right|^2 \\ &\leq \sum_{n \in \mathbb{N}} 3 |\langle u_0, \omega_n \rangle|^2 + 3(t-s) \int_{r=0}^t |\langle f(r), \omega_n \rangle_H|^2 dr + 3s \int_{r=s}^t |\langle f(r), \omega_n \rangle_H|^2 dr \leq 3 \|u_0\|_H^2 + 3t \|f\|_{L^2(0,T;H)}^2. \end{aligned} \quad (7.7)$$

Donc la somme est dominée par une série intégrable. Comme chaque terme de la somme en (7.7) tend vers 0 quand  $t \rightarrow s$ , la convergence dominée implique la continuité de  $u(\cdot)$  dans  $H$ , ce qui conclut la preuve.  $\square$

Le problème hyperbolique abstrait généralisant (7.2) se traite de la même manière, avec des EDO à coefficients d'ordre 2.

## 7.2 Théorème de Hille-Yosida

La section précédente se base sur l'utilisation de la méthode spectrale. Dans cette section, on présente une autre direction pour résoudre les problèmes d'évolution, qui se base sur la dissipativité des opérateurs.

### 7.2.1 Définitions

**Définition 40 – Opérateur maximal monotone** Soit  $(H, \langle \cdot, \cdot \rangle)$  un espace de Hilbert. Soit  $A : D(A) \subset H \rightarrow H$  une application linéaire définie de son domaine  $D(A) \subset H$  dans  $H$ . L'opérateur  $A$  est dit monotone si

$$\langle Au, u \rangle \geq 0 \quad \forall u \in D(A),$$

et maximal si  $I + A$  est surjectif, i.e. si pour tout  $f \in H$ , il existe  $u \in D(A)$  tel que  $u + Au = f$ .

Ici, on ne demande pas à ce que  $A$  soit continu, donc il est possible que  $D(A) \neq H$ . On considérera toujours que  $D(A)$  est un sous-espace vectoriel de  $H$ , mais potentiellement non fermé.

#### Exemples

- L'application nulle  $A \equiv 0$  et l'identité sont deux opérateurs monotones maximaux dont le domaine est  $H$ .
- Supposons que  $\langle Au, v \rangle = a(u, v)$  pour tout  $u \in D(A)$  et  $v \in H$ , où  $a(\cdot, \cdot)$  est une forme positive. Alors  $\langle Au, u \rangle = a(u, u) \geq 0$ , et l'opérateur  $A$  est monotone. Il est maximal si pour tout  $f \in H$ , il existe au moins une solution  $u \in H$  à l'équation  $(I + A)u = f$ , qui s'écrit également  $\langle u, v \rangle + a(u, v) = \langle f, v \rangle$  pour tout  $v \in H$ .

#### Contre-exemples

- Soit  $A : D(A) \rightarrow \mathbb{R}^2$  l'opérateur défini par  $Ax = 0$  et  $D(A) = \{x = (x_0, x_1) \in \mathbb{R}^2 \mid x_0 = 0\}$ .  $A$  est trivialement monotone. Par contre, pour  $f = (1, 0) \in \mathbb{R}^2$ , l'équation  $x + Ax = f$  n'admet pas de solution dans  $D(A)$ , donc  $A$  n'est pas maximal. On retient de cet exemple que **le domaine de  $A$  fait partie intégrante de la définition, et ne peut pas être négligé**.

Les opérateurs monotones maximaux sont liés à l'étude du problème de Cauchy

$$\begin{cases} \frac{d}{dt} u(t) + Au(t) = 0 & t \in ]0, \infty[, \\ u(0) = u_0 \in H. \end{cases} \quad (7.8a)$$

$$(7.8b)$$

De manière intuitive, on peut garder en tête que l'opérateur  $A$  prend pour argument des éléments réguliers de l'espace pour les transformer en éléments moins réguliers. L'opérateur inverse est alors régularisant.

**Lemme 26 – Effet régularisant** Soit  $A : D(A) \subset H \rightarrow H$  maximal monotone. Alors  $D(A)$  est dense dans  $H$ , et pour tout  $\lambda > 0$ , l'application  $I + \lambda A$  est bijective de  $D(A)$  sur  $H$  avec  $\|(I + \lambda A)^{-1}\|_{\mathcal{L}(H)} \leq 1$ .

**Démonstration**

**Densité** Pour montrer que  $D(A)$  est dense, il suffit de montrer que si  $L: H \rightarrow H$  est une forme linéaire continue qui s'annule sur  $D(A)$ , alors  $L \equiv 0$ . Par Riesz, chacune de ces formes est représentable par  $Lv = \langle \ell, v \rangle$  pour un certain  $\ell \in H$ . Supposons donc que  $\langle \ell, u \rangle = 0$  pour tout  $u \in D(A)$  : comme  $A$  est maximal, il existe une solution  $v \in D(A)$  à l'équation  $v + Av = \ell$ . Moralement,  $\langle Av, v \rangle \geq 0$  indique que  $Av$  pointe dans la même direction que  $v$ . Pourtant,  $\ell$  est orthogonal aux éléments de  $D(A)$ , dont fait partie  $v$ . En prenant le produit scalaire avec  $v$ , on obtient

$$\langle v, v \rangle + \langle Av, v \rangle = \langle \ell, v \rangle = 0, \quad \text{d'où} \quad |v|^2 + 0 \leq 0,$$

donc  $v = 0$  et  $\ell = v + Av = 0$ . Ainsi  $D(A)$  est dense.

**Effet régularisant** Supposons que pour  $\lambda > 0$ ,  $I + \lambda A$  soit surjectif. En particulier, pour tout  $f \in H$ , il existe une solution à  $u + \lambda Au = f$ . Alors

$$\langle u, u \rangle + \langle \lambda Au, u \rangle = \langle f, u \rangle \implies |u|^2 \leq |f| |u|,$$

et  $|(I - \lambda A)^{-1} f| = |u| \leq |f|$ . Comme ceci est valide pour tout  $f \in H$ , on obtient la borne uniforme sur  $\|(I - \lambda A)^{-1}\|_{\mathcal{L}(H)}$ .

**Bijjectivité** Montrons d'abord que  $I + \lambda A$  est injectif, i.e.  $u + \lambda Au = 0$  implique  $u = 0$ . En prenant le produit scalaire avec  $u$ , on a

$$0 = |u|^2 + \langle \lambda Au, u \rangle \geq |u|^2,$$

donc  $u = 0$ .

Montrons par induction que  $I + \lambda A$  est surjectif pour tout  $\lambda > 0$  : comme  $A$  est maximal, on sait que c'est le cas pour  $\lambda = 1$ . Supposons que ce soit le cas pour  $\lambda_0 > 0$ . Pour  $f \in H$  quelconque, l'équation  $u + \lambda_0 Au = f$  se réécrit

$$u + \frac{\lambda}{\lambda_0} \lambda_0 Au = f \Leftrightarrow u + \frac{\lambda}{\lambda_0} (u - u + \lambda_0 Au) = f \Leftrightarrow u + \lambda_0 Au = \frac{\lambda_0}{\lambda} (f - u) + u = \frac{\lambda_0}{\lambda} f + u \left(1 - \frac{\lambda_0}{\lambda}\right),$$

donc  $(I + \lambda_0 A)u = \frac{\lambda_0}{\lambda} f + \left(1 - \frac{\lambda_0}{\lambda}\right)u$ . Résoudre cette équation revient à chercher  $u$  tel que

$$u = (I - \lambda_0 A)^{-1} \left( \frac{\lambda_0}{\lambda} f + \left(1 - \frac{\lambda_0}{\lambda}\right)u \right) =: K(u), \quad \text{avec} \quad \|K(u) - K(v)\| = \left|1 - \frac{\lambda_0}{\lambda}\right| \|(I - \lambda_0 A)^{-1} (u - v)\| \leq \left|1 - \frac{\lambda_0}{\lambda}\right| \|u - v\|.$$

Si  $\left|1 - \frac{\lambda_0}{\lambda}\right| < 1$ , on obtient une équation de point fixe pour un opérateur strictement contractant, donc qui admet une unique solution  $u$ , qui appartient au domaine de  $A$  car de la forme  $(I - \lambda_0 A)^{-1} g$ . Or  $\left|1 - \frac{\lambda_0}{\lambda}\right| < 1$  précisément quand  $\lambda > \frac{\lambda_0}{2}$ . Ainsi, par récurrence,  $I + \lambda A$  est bijectif pour  $\lambda \in \{1\}$ , puis  $(1/2, \infty)$ , puis  $(1/4, \infty)$ , etc... et donc pour tout  $\lambda > 0$ .  $\square$

Le Lemme 26 permet de définir  $(I - \lambda A)^{-1}$  pour tout  $\lambda > 0$ . Pour mieux manipuler cet opérateur, on lui donne un nom.

**Définition 41 – Résolvante de  $A$**  Soit  $A: D(A) \subset H \rightarrow H$  un opérateur maximal monotone, et  $\lambda > 0$ . La résolvante de  $A$  est définie par

$$J_\lambda := (I + \lambda A)^{-1}: H \rightarrow D(A).$$

**Remarque 18** (La résolvante commute avec  $A$ ). Soit  $A: D(A) \subset H \rightarrow H$  un opérateur maximal monotone et  $J_\lambda$  sa résolvante. Alors  $\forall u \in H$ , il existe  $v \in D(A)$  tel que  $u = (I + \lambda A)v$ . Ainsi

$$J_\lambda A u = J_\lambda A(I + \lambda A)v = (I + \lambda A)^{-1}(I + \lambda A)Av = Av = AJ_\lambda u.$$

Par récurrence, on a donc

$$A(I + \lambda A)^{-n} u = (I + \lambda A)^{-n} Au$$

pour tout  $n \in \mathbb{N}$  et  $u \in H$ .

### 7.2.2 Un premier résultat

Soient  $u \in H$  et  $\lambda, \mu > 0$ . Posons  $z = J_\mu \left( \frac{\mu}{\lambda} u + \frac{\lambda - \mu}{\lambda} J_\lambda u \right)$ . Alors

$$(I + \mu A)z = \frac{\mu}{\lambda} u + \frac{\lambda - \mu}{\lambda} J_\lambda u, \quad \text{et} \quad (I + \lambda A)(I + \mu A)z = \frac{\mu}{\lambda} (I + \lambda A)u + \frac{\lambda - \mu}{\lambda} u = u + \mu Au.$$

Ainsi, comme  $I + \lambda A$  et  $I + \mu A$  commutent, et  $I + \mu A$  est inversible, on obtient  $(I + \lambda A)z = u$ , donc  $z = J_\lambda u$ . Ceci montre l'équation de la résolvante

$$J_\lambda = J_\mu \left( \frac{\mu}{\lambda} I + \frac{\lambda - \mu}{\lambda} J_\lambda \right).$$

**Lemme 27 – Estimation du schéma** Soient  $\lambda \leq \mu > 0$  et  $n > m > 0$  avec  $n, m \in \mathbb{N}$ . Notons  $\alpha = \frac{\mu}{\lambda}$  et  $\beta = \frac{\lambda - \mu}{\lambda}$ . Alors

$$\|J_\lambda^n u - u\| \leq n\lambda \|Au\| \quad \forall u \in D(A),$$

et

$$\|J_\mu^n u - J_\lambda^m u\| \leq \sum_{j=0}^{m-1} \alpha^j \beta^{n-j} \binom{n}{j} \|J_\lambda^{m-j} u - u\| + \sum_{j=m}^n \alpha^m \beta^{j-m} \binom{j-1}{m-1} \|J_\mu^{n-j} u - u\|.$$

### Démonstration

Soit  $u$  appartenant au domaine de  $A$ . On commence par  $n = 1$ : posons  $z := J_\lambda u = (I - \lambda A)u$ . On a

$$\|J_\lambda u - u\| = \|z - (I - \lambda A)z\| = \lambda \|Az\| \leq \lambda \|J_\lambda\| \|Au\| \leq \lambda \|Au\|.$$

Pour  $n > 1$  entier arbitraire, on en déduit

$$\|J_\lambda^n u - u\| = \left\| \sum_{j=0}^{n-1} J_\lambda^{j+1} u - J_\lambda^j u \right\| \leq \sum_{j=0}^{n-1} \|J_\lambda^j (J_\lambda u - u)\| \leq n\lambda \|Au\|.$$

Soient  $j, k$  des entiers tels que  $0 \leq j \leq n$  et  $0 \leq k \leq m$ . Notons  $a_{k,j} := \|J_\mu^j u - J_\lambda^k u\|$ . On a

$$a_{k,j} = \|J_\mu^j u - J_\mu \left( \alpha J_\lambda^{k-1} u + \beta J_\lambda^k u \right)\| \leq \|J_\mu^{j-1} u - (\alpha J_\lambda^{k-1} u + \beta J_\lambda^k u)\| \leq \alpha \|J_\mu^{j-1} u - J_\lambda^{k-1} u\| + \beta \|J_\mu^{j-1} u - J_\lambda^k u\| = \alpha a_{k-1,j-1} + \beta a_{k,j-1}.$$

Par récurrence,

$$a_{m,n} \leq \sum_{j=0}^{m-1} \alpha^j \beta^{n-j} \binom{n}{j} a_{m-j,0} + \sum_{j=m}^n \alpha^m \beta^{j-m} \binom{j-1}{m-1} a_{0,n-j},$$

avec la deuxième somme égale à 0 par convention si  $m > n$ . □

**Lemme 28 – Estimation des coefficients, admis** Pour  $\alpha, \beta > 0$  tels que  $\alpha + \beta = 1$ , et  $n \geq m > 0$  des entiers, on a

$$\begin{aligned} \sum_{j=0}^m \alpha^j \beta^{n-j} \binom{n}{j} (m-j) &\leq \sqrt{(n\alpha - m)^2 + n\alpha\beta}, \\ \sum_{j=m}^n \alpha^m \beta^{j-m} \binom{j-1}{m-1} (n-j) &\leq \sqrt{\frac{m\beta}{\alpha^2} + \left(\frac{m\beta}{\alpha} + m - n\right)^2}. \end{aligned}$$

**Théorème 17 – Convergence du schéma** Soit  $A : D(A) \subset H \rightarrow H$  maximal monotone, et  $u_0 \in D(A)$ . Pour tout  $n \in \mathbb{N}_*$ , considérons le schéma implicite

$$u^n(t) = J_{t/n}^n u_0.$$

Alors la famille de courbes  $(u^n)_n$  admet une limite pour  $n \rightarrow \infty$ .

### Démonstration

Pour tout  $t > 0$ , on a l'estimation uniforme  $\|u^n(t)\| = \|J_{t/n}^n u_0\| \leq \|J_{t/n}\|^n \|u_0\| \leq \|u_0\|$  par le Lemme 26. De plus, à  $t > 0$  fixé, on

prend  $\mu = t/n$  et  $\lambda = t/m$  dans le Lemme 27. Il vient  $\alpha = \mu/\lambda = m/n$  et  $\beta = (\lambda - \mu)/\lambda = (1 - m/n)$ , et

$$\begin{aligned} \|u^n(t) - u^m(t)\| &= \|J_{t/n}^n u_0 - J_{t/m}^m u_0\| \leq \sum_{j=0}^{m-1} \alpha^j \beta^{n-j} \binom{n}{j} \|J_{t/m}^{m-j} u_0 - u_0\| + \sum_{j=m}^n \alpha^m \beta^{j-m} \binom{j-1}{m-1} \|J_{t/n}^{n-j} u_0 - u_0\| \\ &\leq \sum_{j=0}^{m-1} \alpha^j \beta^{n-j} \binom{n}{j} (m-j) \frac{t}{m} \|Au_0\| + \sum_{j=m}^n \alpha^m \beta^{j-m} \binom{j-1}{m-1} \frac{t}{n} (n-j) \|Au_0\| \\ &\leq \left[ \sqrt{(na - m)^2 + na\beta} \frac{1}{m} + \sqrt{\frac{m\beta}{\alpha^2} + \left( \frac{m\beta}{\alpha} + m - n \right)^2} \frac{1}{n} \right] t \|Au_0\|. \end{aligned}$$

Ici,  $na - m = 0$  et  $\frac{m\beta}{\alpha} + m - n = 0$  par le choix de  $\alpha$  et  $\beta$ , et  $na\beta = m(1 - \frac{m}{n})$ , ainsi que  $m \frac{\beta}{\alpha^2} = \frac{1}{m}(1 - \frac{m}{n})$ . D'où

$$\|u^n(t) - u^m(t)\| \leq \sqrt{\frac{1}{m} - \frac{1}{n}} \left( 1 + \frac{1}{n} \right) t \|Au_0\| \leq t \sqrt{\frac{1}{m} - \frac{1}{n}} \|Au_0\|.$$

Donc  $\lim_{m \rightarrow \infty} \sup_{n > m} \|u^n(t) - u^m(t)\| \leq \lim_{m \rightarrow \infty} \frac{t \|Au_0\|}{\sqrt{m}} = 0$ , et la famille de courbes  $u^n$  admet une limite  $u$  point par point.  $\square$

### 7.2.3 Cas général

On se contente d'énoncer sans démonstration le résultat suivant.

**Théorème 18 – Hille-Yosida, cas général ([Bré10, Théorème VII.4 p. 105])** Soit  $A$  un opérateur maximal monotone dans un espace de Hilbert  $H$ . Alors pour tout  $u_0 \in D(A)$ , il existe une unique fonction

$$u \in C^1([0, \infty); H) \cap C([0, \infty); D(A))$$

solution au sens classique de (7.8).

De même, si  $A$  est autoadjoint, on peut montrer qu'une telle courbe existe pour tout  $u \in H$  (même en dehors du domaine de  $A$ ) et se situe dans  $D(A)$  pour tout  $t > 0$ . Ceci illustre la "régularisation immédiate" du problème de Cauchy (7.8) : par exemple, si  $A = -\Delta$ , la solution de l'équation de la chaleur est dans  $H^2$  pour tout temps positif, même si  $u_0$  n'est que dans  $L^2$ .

L'espace dans lequel se situe la courbe  $u(\cdot)$  est à prendre au sens suivant :  $u(\cdot)$  est continue et à valeurs dans  $D(A)$ , et pour tout temps  $t > 0$ , il existe un ouvert de  $]0; \infty[$  contenant  $t$  et tel que  $u(\cdot)$  soit de classe  $C^1$  sur cet ouvert. Par contre, la norme de  $\frac{d}{dt}u(\cdot)$  peut exploser quand  $t$  tend vers 0 (on peut penser à  $t \mapsto \sqrt{t}$ ).

Le Théorème 18 s'applique à diverses variations de (7.8). Par exemple, soit  $\lambda \in \mathbb{R}$ . Considérons le problème

$$\begin{cases} \frac{d}{dt}u(t) + \lambda u(t) + Au(t) = 0 & t \in ]0, \infty[, \\ u(0) = u_0 \in H. \end{cases} \quad (7.9a)$$

$$(7.9b)$$

On pose  $v(t) := e^{\lambda t} u(t)$ . Dès lors,

$$\frac{d}{dt}v(t) = \lambda e^{\lambda t} u(t) + e^{\lambda t} \frac{d}{dt}u(t) = \lambda e^{\lambda t} u(t) + e^{\lambda t} (-\lambda u(t) - Au(t)) = -Av(t).$$

Ainsi, on peut appliquer le Théorème 18 à l'équation satisfaite par  $v(\cdot)$ , et en déduire que  $u(t) := e^{-\lambda t} v(t)$  est l'unique solution du problème (7.9).

## 7.3 Discrétisation en temps

De même que l'analyse théorique tire parti de la spécificité de la variable temporelle, les schémas numériques peuvent exploiter le caractère parabolique ou hyperbolique de l'équation. Rappelons que de manière générale, un schéma se doit d'être *consistant*, c'est-à-dire d'approcher la bonne équation, et *stable*, ce qui se traduit souvent par des inégalités contrôlant la norme de la solution dans différents espaces.

La construction d'un schéma numérique implementable doit contenir deux discrétisations : une discrétisation temporelle et une discrétisation spatiale. Les étapes intermédiaires, appelées schémas semi-discrets, permettent d'analyser la convergence dans

un cadre général : par exemple, étudier un schéma semi-discret en temps permet d'obtenir des résultats indépendants de la discrétisation en espace qui viendra par la suite.

### 7.3.1 Cas parabolique

Considérons le cas d'une équation parabolique sous forme variationnelle

$$\frac{d}{dt} \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H \quad \forall v \in V.$$

On reprend la terminologie (et les idées) des schémas pour les EDOs. Soit  $N \in \mathbb{N}_*$  et  $\Delta t := T/N$ . On note  $t_n := n\Delta t$  pour  $n \in \llbracket 0, N \rrbracket$ . Notons

$$f^n := \frac{1}{\Delta t} \int_{s=t^n}^{t^{n+1}} f(s) ds \in H.$$

**Définition 42 –  $\theta$ -schéma pour le cas parabolique** Soit  $\theta \in [0, 1]$ . On pose  $u^0 = u_0 \in V$ , et pour tout  $n \in \llbracket 1, N \rrbracket$ , on définit  $u^{n+1} \in V$  comme l'unique solution de

$$\left\langle \frac{u^{n+1} - u^n}{\Delta t}, v \right\rangle_H + a(\theta u^{n+1} + (1-\theta)u^n, v) = \langle f^n, v \rangle_H \quad v \in V, \quad n \in \llbracket 0, N-1 \rrbracket. \quad (7.10)$$

La suite  $(u^n)_{n \in \llbracket 0, N \rrbracket}$  est bien définie par le théorème de Lax-Milgram. Dans le cas  $\theta = 0$ , le schéma est dit explicite : les produits scalaires  $\langle u^{n+1}, v \rangle_H$ , qui déterminent complètement  $u^{n+1}$ , sont calculables directement sans résoudre de système. Dans le cas  $\theta = 1$ , le schéma est dit implicite (même si théoriquement, (7.10) est implicite dès que  $\theta > 0$ ). Le cas  $\theta = 1/2$  est appelé schéma de Crank-Nicolson.

**Définition 43 – Stabilité du schéma** On dit que le schéma est stable (ici, en norme de  $V$ ) si pour tout  $u_0 \in V$  et  $f \in L^2(0, T; H)$ , il existe une constante  $C > 0$  indépendante de  $N$  et  $\Delta t$  telle que

$$\max_{n \in \llbracket 0, N \rrbracket} \|u^n\|_V^2 \leq C \left( \|u_0\|_V^2 + \|f\|_{L^2(0, T; H)}^2 \right).$$

Remarquons que  $\frac{1}{M} \sqrt{a(u^n, u^n)} \leq \|u^n\|_V \leq \frac{1}{\alpha} \sqrt{a(u^n, u^n)}$ , ce qui nous permet d'utiliser  $a(\cdot, \cdot)$  en lieu et place de la norme de  $V$  dans l'estimation de stabilité.

**Lemme 29 – Stabilité du cas parabolique** Le  $\theta$ -schéma est inconditionnellement stable pour  $\theta \in ]1/2, 1]$ . Si  $\theta \leq 1/2$ , alors le  $\theta$ -schéma est stable sous la condition de Courant-Friedrich-Lowy

$$\sup_{\substack{\mu_m \text{ valeur propre de } a(\cdot, \cdot)}} \left( \frac{1}{2} - \theta \right) \Delta t \mu_m < 1. \quad (7.11)$$

Si la dimension de  $V$  est infinie, la condition (7.11) n'est jamais valide, car la suite  $(\mu_m)_m$  tend vers l'infini. En revanche, si  $a(\cdot, \cdot)$  admet un nombre fini de valeurs propres, (7.11) peut être réalisée (bien que très restrictive).

#### Démonstration

Imitons l'analyse spectrale du cas continu. Soit  $(\mu_m, \omega_m)_{m \in \mathbb{N}}$  une famille d'éléments propres pour  $a(\cdot, \cdot)$  telle que  $(\omega_m)_m$  soit une base hilbertienne de  $H$ . On décompose

$$u^n := \sum_{m \in \mathbb{N}} \xi_{m,n} \omega_m, \quad \xi_{m,n} := \langle u^n, \omega_m \rangle_H, \quad f^n := \sum_{m \in \mathbb{N}} \langle f^n, \omega_m \rangle_H \omega_m.$$

**Résolution du système discret** Comme dans le cas continu, les coefficients initiaux  $\xi_{m,0}$  sont fixés par la condition initiale via  $\xi_{m,0} = \langle u_0, \omega_m \rangle_H$ . En injectant ces décompositions dans (7.10) et en considérant pour fonction test chaque  $\omega_m$ , on obtient le

système discret

$$\begin{aligned} \frac{\xi_{m,n+1} - \xi_{m,n}}{\Delta t} + \mu_m (\theta \xi_{m,n+1} + (1-\theta) \xi_{m,n}) &= \langle f^n, \omega_m \rangle_H \\ \xi_{m,n+1} &= \xi_{m,n} \frac{1 - \Delta t \mu_m (1-\theta)}{1 + \Delta t \mu_m \theta} + \frac{\Delta t}{1 + \Delta t \mu_m \theta} \langle f^n, \omega_m \rangle_H. \end{aligned}$$

Notons  $\tau_m := \frac{1 - \Delta t \mu_m (1-\theta)}{1 + \Delta t \mu_m \theta}$ . Les coefficients sont donnés par

$$\xi_{m,n} = \tau_m^n \xi_{m,0} + \frac{\Delta t}{1 + \Delta t \mu_m \theta} \sum_{j=0}^{n-1} \tau_m^{n-1-j} \langle f^j, \omega_m \rangle_H. \quad (7.12)$$

**Borne sur  $\tau_m$**  Vérifions donc dans quel intervalle se situe  $\tau_m$  : considérons la fonction

$$\tau(r) = \frac{1 - \Delta t r (1-\theta)}{1 + \Delta t r \theta} = 1 - \frac{\Delta t r}{1 + \Delta t r \theta}, \quad r > 0.$$

On a  $\tau'(r) = -\frac{\Delta t}{(1 + \Delta t r \theta)^2} < 0$ , donc  $\tau(r) < 1$  pour tout  $r > 0$ . Si  $a(\cdot, \cdot)$  admet un nombre infini de valeurs propres, alors elles forment une suite qui tend vers  $+\infty$  (voir Corollaire 1). Dans ce cas, on a  $|\tau(r)| \leq 1$  si et seulement si  $\theta \geq 1/2$ . Autrement, soit  $\mu_{\max}$  la valeur maximale des valeurs propres de  $a(\cdot, \cdot)$  : alors

$$\max_{\substack{\mu_m \text{ valeur propre}}} |\tau(\mu_m)| \leq |\tau(\mu_{\max})| < 1 \iff \frac{1 - \Delta t \mu_{\max} (1-\theta)}{1 + \Delta t \mu_{\max} \theta} > -1 \iff 1 > \Delta t \mu_{\max} \left( \frac{1}{2} - \theta \right).$$

**Estimation d'énergie** Supposons maintenant que  $|\tau_m| \leq |\tau_{\max}| < 1$  pour tout  $m$ . On écrit

$$\sum_{n \in [0, N]} \|u^n\|_V^2 \leq \frac{1}{\alpha} \sum_{n \in [0, N]} a(u^n, u^n) = \frac{1}{\alpha} \sum_{n \in [0, N]} \sum_{m \in \mathbb{N}} |\xi_{n,m}|^2 \mu_m.$$

Or, par Cauchy-Schwarz,

$$\begin{aligned} |\xi_{n,m}|^2 &\leq 2 |\tau_m|^{2n} |\xi_{m,0}|^2 + 2 \left( \frac{\Delta t}{1 + \Delta t \mu_m \theta} \sum_{j=0}^{n-1} \tau_m^{n-1-j} \langle f^j, \omega_m \rangle_H \right)^2 \\ &\leq 2 |\tau_m|^{2n} |\xi_{m,0}|^2 + 2 \left( \frac{\Delta t}{1 + \Delta t \mu_m \theta} \right)^2 \left( \sum_{j=0}^{n-1} |\tau_m|^{n-1-j} |\langle f^j, \omega_m \rangle_H|^2 \right) \left( \sum_{j'=0}^{n-1} |\tau_m|^{n-1-j'} \right). \end{aligned}$$

On a en particulier

$$\sum_{j=0}^{n-1} |\tau_m|^{n-1-j} = \sum_{j=0}^{n-1} |\tau_m|^j = \frac{1 - |\tau_m|^n}{1 - |\tau_m|} \leq \frac{1}{1 - |\tau_m|},$$

qui est positif et fini par hypothèse. En revenant à la somme en temps,

$$\begin{aligned} \sum_{n \in [0, N]} \mu_m |\xi_{n,m}|^2 &\leq 2 \sum_{n \in [0, N]} \mu_m |\tau_m|^{2n} |\xi_{m,0}|^2 + 2 \mu_m \left( \frac{\Delta t}{1 + \Delta t \mu_m \theta} \right)^2 \sum_{n \in [0, N]} \sum_{j=0}^{n-1} |\tau_m|^{n-1-j} |\langle f^j, \omega_m \rangle_H|^2 \frac{1}{1 - |\tau_m|} \\ &\leq \frac{2}{1 - |\tau_{\max}|^2} \mu_m |\xi_{m,0}|^2 + 2 \mu_m \left( \frac{\Delta t}{1 + \Delta t \mu_m \theta} \right)^2 \sum_{j=0}^{N-1} |\langle f^j, \omega_m \rangle_H|^2 \frac{1}{1 - |\tau_m|} \underbrace{\sum_{n=j+1}^N |\tau_m|^{n-(j+1)}}_{=\sum_{k=0}^{N-(j+1)} |\tau_m|^k \leq \frac{1}{1 - |\tau_m|}} \\ &\leq \frac{2}{1 - |\tau_{\max}|^2} \mu_m |\xi_{m,0}|^2 + 2 \frac{\Delta t \mu_m}{1 + \Delta t \mu_m \theta} \frac{\Delta t}{1 + \Delta t \mu_m \theta} \sum_{j=0}^{N-1} |\langle f^j, \omega_m \rangle_H|^2 \frac{1}{(1 - |\tau_{\max}|)^2}. \end{aligned}$$

Le terme  $\frac{\Delta t \mu_m}{1 + \Delta t \mu_m \theta}$  est toujours inférieur à  $1/\theta$ . De plus, dans le cas  $\theta < 1/2$ , la condition CFL implique  $1 + \theta \Delta t \mu_m \geq \frac{1}{2} \Delta t \mu_m$ , donc  $\frac{\Delta t \mu_m}{1 + \Delta t \mu_m \theta} \leq 2$  dans chacun des cas. En minorant grossièrement  $1 + \Delta t \mu_m \theta \geq 1$ , il vient

$$\begin{aligned} \sum_{n \in [0, N]} \|u^n\|_V^2 &\leq \frac{1}{\alpha} \frac{2}{1 - |\tau_{\max}|^2} \sum_{m \in \mathbb{N}} \mu_m |\xi_{m,0}|^2 + \frac{4 \Delta t}{\alpha} \sum_{j=0}^{N-1} \sum_{m \in \mathbb{N}} |\langle f^j, \omega_m \rangle_H|^2 \frac{1}{(1 - |\tau_{\max}|)^2} \\ &\leq \frac{1}{\alpha} \frac{2}{1 - |\tau_{\max}|^2} \|u_0\|_V^2 + \Delta t \frac{4}{\alpha (1 - |\tau_{\max}|)^2} \sum_{j=0}^{N-1} \|f^j\|_H^2 \\ &=: C_1 \|u_0\|_V^2 + \Delta t C_2 \sum_{j=0}^{N-1} \|f^j\|_H^2, \end{aligned}$$

où  $C_1$  et  $C_2$  dépendent de  $\alpha$  et  $|\tau_{\max}|$ . On complète l'estimation en utilisant la définition de  $f^j$  pour écrire

$$\Delta t \sum_{j=0}^{N-1} \|f^j\|_H^2 = \Delta t \sum_{j=0}^{N-1} \left\| \frac{1}{\Delta t} \int_{s=t^j}^{t^{j+1}} f(s) ds \right\|_H^2 \leq \Delta t \sum_{j=0}^{N-1} \frac{1}{\Delta t} \int_{s=t^j}^{t^{j+1}} \|f(s)\|_H^2 ds = \|f\|_{L^2(0,T;H)}^2.$$

D'où la stabilité pour la constante  $C = \max(C_1, C_2)$ .  $\square$

Continuons avec un résultat de convergence assez grossier, mais qui permet de percevoir l'utilité de la linéarité pour l'étude des schémas numériques.

**Lemme 30 – Un résultat de convergence pour le cas parabolique** Supposons que la solution exacte  $u \in \mathcal{C}([0, T]; V)$  avec pour constante de Lipschitz  $[u] < \infty$ , et que les hypothèses du Lemme 29 sont satisfaites. Alors le  $\theta$ -schéma converge en norme  $\|\cdot\|_V$  à l'ordre 1.

### Démonstration

Si  $u$  est une solution du problème parabolique au sens des distributions, alors (voir preuve en TD)

$$\left\langle \frac{u(t) - u(s)}{t-s}, v \right\rangle_H + a\left(\frac{1}{t-s} \int_{r=s}^t u(r) dr, v\right) = \left\langle \frac{1}{t-s} \int_{r=s}^t f(r) dr, v \right\rangle_H \quad \forall v \in V, \quad \forall 0 \leq s < t \leq T.$$

En particulier, pour  $s = t^n$  et  $t = t^{n+1}$ , on obtient

$$\left\langle \frac{u(t^{n+1}) - u(t^n)}{\Delta t}, v \right\rangle_H + a\left(\frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} u(r) dr, v\right) = \langle f^n, v \rangle_H \quad \forall v \in V, \quad \forall n \in \llbracket 0, N-1 \rrbracket.$$

Prenons la différence entre cette égalité et la définition du schéma. On pose  $w^n := u^n - u(t^n) \in V$ . Il vient

$$\left\langle \frac{w^{n+1} - w^n}{\Delta t}, v \right\rangle_H + a\left(\theta w^{n+1} + (1-\theta)w^n, v\right) = a\left(\theta u(t^{n+1}) + (1-\theta)u(t^n) - \frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} u(r) dr, v\right).$$

Notons  $g^n := \theta u(t^{n+1}) + (1-\theta)u(t^n) - \frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} u(r) dr \in V$  un terme qui ne dépend que de la solution exacte  $u(\cdot)$ . La suite  $(w^n)_{n \in \mathbb{N}}$  satisfait presque un schéma de second membre  $g^n$ , à la différence que c'est le produit scalaire  $a(\cdot, \cdot)$  qui apparaît dans le terme source et non  $\langle \cdot, \cdot \rangle_H$ . En reprenant la preuve du Lemme 29 en substituant  $a(g^n, \omega_m)$  à  $\langle f^n, \omega_m \rangle_H$ , on obtient l'existence de  $C_1$  et  $C_2$  telles que

$$\sum_{n \in \llbracket 0, N \rrbracket} \|w^n\|_V^2 \leq C_1 \|w_0\|_V^2 + \Delta t C_2 \sum_{j=0}^{N-1} a^2(g^j, g^j) \leq C_1 \times 0 + \Delta t C_2 M \sum_{j=0}^{N-1} \|g^j\|_V^2,$$

où  $M$  est la constante de continuité de  $a(\cdot, \cdot)$  dans  $V$ . Évaluons  $\|g^n\|_V^2$  : on a

$$\begin{aligned} \|\theta u(t^{n+1}) + (1-\theta)u(t^n) - \frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} u(r) dr\|_V &\leq \frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} \theta \|u(t^{n+1}) - u(r)\|_V + (1-\theta) \|u(t^n) - u(r)\|_V dr \\ &\leq \frac{1}{\Delta t} \int_{r=t^n}^{t^{n+1}} \theta [u] \Delta t + (1-\theta) [u] \Delta t dr = [u] \Delta t. \end{aligned}$$

D'où

$$\sqrt{\sum_{n \in \llbracket 0, N \rrbracket} \|w^n\|_V^2} \leq \sqrt{\Delta t C_2 M \sum_{j=0}^N ([u] \Delta t)^2} = \sqrt{C_2 T M} [u] \Delta t,$$

et on obtient une convergence à l'ordre 1 du schéma.  $\square$

# Annexe

## Démonstration du théorème de Bolzano-Weierstraß

1. Supposons que  $K$  est compact. Pour  $\varepsilon > 0$  fixé, on considère le recouvrement ouvert par les boules de rayon  $\varepsilon/2$  centrées en chacun des points de  $K$  : par compacité, il existe un sous-recouvrement fini. Donc  $K$  est totalement borné.

2. Par contraposée, [compact  $\implies$  fermé] est équivalent à [non fermé  $\implies$  non compact]. Supposons que  $C$  est non fermé, ce qui, dans un espace métrique, équivaut à l'existence d'une suite de Cauchy  $(x_n)_n \subset C$  telle que  $x_n \rightarrow x \in E$  mais  $x \notin C$ . En particulier,  $d(x, x_n) > 0$  pour tout  $n$ . La famille des

$$E_n := E \setminus \overline{\mathcal{B}}(x, d(x, x_n))$$

est une famille d'ouverts. De plus, pour tout  $y \in C$ , il existe  $n$  suffisamment grand pour que  $d(y, x) > d(x, x_n)$ , donc pour que  $y \in E_n$ . Ainsi, on a une famille d'ouverts qui recouvrent  $C$ . Mais s'il existait un nombre fini  $M$  tel que  $C \subset \bigcup_{n \in [0, M]} E_n$ , alors tout point de  $C$  se trouverait à une distance supérieure à  $d(x, x_M) > 0$  de  $x$ , et ceci contredit la convergence de  $(x_n)_n$  vers  $x$ . Donc compact implique fermé.

3. Supposons que  $K$  est totalement borné et fermé, et soit  $(x_n)_{n \in \mathbb{N}} \subset K$ . On va construire une sous-suite de Cauchy, et utiliser la fermeture de  $K$  pour conclure. Notons  $\varepsilon_m := 2^{-m}$ . Pour  $m = 0$ , recouvrons  $K$  par un nombre fini d'ouverts de diamètre inférieur à  $\varepsilon_0$ . Par l'absurde, l'un de ces ensembles contient un nombre infini de termes de la suite  $(x_n)_n$  : sinon, la suite elle-même serait finie. Notons  $(x_n^0)_n$  une telle sous-suite de  $(x_n)_n$ , et  $y^0 = x_0^0$ . Par construction, cette première extraction satisfait  $d(x_i^0, x_j^0) \leq 2^{-0}$ . En appliquant le même raisonnement à un recouvrement de diamètre  $\varepsilon_1$ , on obtient une sous-sous-suite  $(x_n^1)_n$  qui satisfait  $d(x_i^1, x_j^1) \leq 2^{-1}$  pour tout  $i, j$ , et un deuxième terme  $y_1 = x_1^1$ . Par récurrence, on construit ainsi des extractions successives  $(x_n^m)_n$  et des termes  $y_m$  qui appartiennent tous à  $K$ , et satisfont  $d(y_m, y_k) \leq 2^{-m}$  pour tout  $k \geq m$ . La suite  $(y_m)_m \subset K$  est donc de Cauchy, comme  $K$  est fermé, elle converge vers un élément  $\bar{y} \in K$ .

4. Soit  $K$  satisfaisant (c). Pour  $\varepsilon > 0$  fixé, recouvrons  $K$  par les boules de rayon  $\varepsilon$  centrées en chaque point de  $K$ . Supposons qu'il n'existe pas de sous-recouvrement fini : on pioche alors  $x_0 \in K$  au hasard, puis  $x_1 \in K \setminus \mathcal{B}(x_0, \varepsilon)$ , puis  $x_2 \in K \setminus \bigcup_{i=0}^1 \mathcal{B}(x_i, \varepsilon)$ ... Par récurrence, on obtient une suite  $(x_n)_{n \in \mathbb{N}}$  telle que  $d(x_i, x_n) \geq \varepsilon$  pour tout  $i \in [0, n-1]$ , donc par symétrie, pour tous  $i, n \in \mathbb{N}$ . Cette suite ne peut pas admettre de sous-suite convergente, ce qui contredit (c). Donc  $K$  est totalement borné. De plus, toute suite de Cauchy de  $K$  est en particulier une suite, donc admet une sous-suite convergente dans  $K$  : comme la limite d'une suite de Cauchy est unique, on en déduit que toute la suite converge.

5. Soit  $K$  satisfaisant (b). On considère  $(r_n)_{n \in \mathbb{N}} \subset \mathbb{R}^+$  une suite qui tend vers 0. Pour chaque  $n$ , il existe un recouvrement fini de  $K$  par des ouverts de diamètre  $r_n$ . On choisit un  $x$  dans chacun de ces ouverts, et on concatène ces choix en une suite dénombrable, qui satisfait la séparabilité.

6. On considère l'union sur  $n \in \mathbb{N}_*$  des familles de boules de rayon  $1/n$  centrées en chacun des points de la partie dénombrable dense. Pour montrer que cette famille d'ouverts est dénombrable, on applique le même raisonnement que pour la dénumérabilité des rationnels. De plus, pour tout ouvert  $G \subset E$  et tout point  $x \in G$ , il existe  $\varepsilon > 0$  tel que  $\mathcal{B}(x, \varepsilon) \subset G$ , et  $y$  dans la partie dénombrable dense tel que  $d(x, y) \leq \varepsilon/2$ . En prenant la boule de rayon  $\varepsilon/4$  centrée en  $y$ , on a bien le caractère de base.

7. Soit  $(\mathcal{O}_\alpha)_\alpha$  un recouvrement ouvert d'un ensemble  $K$  totalement borné, donc séparable, et une base dénombrable  $(V_i)_{i \in \mathbb{N}}$ . Chaque  $x \in C$  est contenu dans un  $\mathcal{O}_\alpha$ , pour lequel on choisit un certain  $V_k \subset \mathcal{O}_\alpha$ . L'union sur  $x$  de ces choix reste un sous-ensemble de la base, donc dénombrable.

8. Soit  $K$  satisfaisant (c), et un recouvrement ouvert de  $K$ , que l'on peut supposer dénombrable, noté  $(\mathcal{O}_n)_n$ .

Supposons par l'absurde qu'il n'existe pas de sous-recouvrement fini. On pioche  $x_0 \in K \cap \mathcal{O}_0$  au hasard. Comme  $K$  n'est pas contenu dans  $\mathcal{O}_0$ , il existe  $x_1 \in K \setminus \mathcal{O}_0$ . Soit  $n_1$  le plus petit entier tel que  $x_1 \in \mathcal{O}_{n_1}$ . Par récurrence, on construit une suite  $(x_k, \mathcal{O}_{n_k})_n$  telle que  $n_k \geq k$ ,  $x_k \in \mathcal{O}_{n_k}$  et  $x_k \in K \setminus \bigcup_{j=0}^{n_k-1} \mathcal{O}_j$ .

Par (c), il existe une sous-suite convergente  $(x_{k_\ell})_{\ell \in \mathbb{N}}$  vers un certain  $\bar{x} \in K$ . Soit  $m$  tel que  $\bar{x} \in \mathcal{O}_m$  : comme  $\mathcal{O}_m$  est ouvert, il existe  $\varepsilon > 0$  tel que  $\mathcal{B}(\bar{x}, \varepsilon) \subset \mathcal{O}_m$ , et par convergence, il existe  $\bar{\ell}$  suffisamment grand pour que  $x_{k_\ell} \in \mathcal{O}_m$  pour tout  $\ell \geq \bar{\ell}$ . C'est absurde pour  $k_\ell \geq m$ . Donc il existe un sous-recouvrement fini.