Notes de cours
Bases des méthodes numériques (MM006)

M1 - Master mention Mathématiques et Applications

Albert Cohen et Sidi Mahmoud Kaber

3 décembre 2018

Table des matières

1 Introduction **7**
- 1.1 "Bases des méthodes numériques" . . . 7
- 1.2 Présentation du cours : les EDP et leur approximation . . . 9
- 1.3 Quelques rappels sur les matrices . . . 14
- 1.4 Rappel sur les EDO : le théorème de Cauchy Lipschitz . . . 18
- 1.5 Problèmes aux limites du second ordre en dimension 1 . . . 19

2 La méthode des différences finies **25**
- 2.1 Principe de la méthode . . . 25
- 2.2 Étude de la convergence . . . 31
- 2.3 Une brève excursion en dimension 2 . . . 37
- 2.4 La méthode des différences finies en dimension 2 . . . 42
- 2.5 Différences finies pour le problème en temps . . . 50

3 Approximation numérique de l'équation de transport **59**
- 3.1 Équation d'advection . . . 59
- 3.2 Approximation numérique de l'équation d'advection . . . 63
  - 3.2.1 Méthode des volumes finis . . . 63
  - 3.2.2 Schémas explicites à 3 et 5 points . . . 64
  - 3.2.3 Propriété des schémas . . . 66
- 3.3 Exemples de schémas pour l'équation de transport . . . 71

4 Formulation variationnelle des problèmes aux limites **75**
- 4.1 La formulation variationnelle : d'où vient-elle ? . . . 75
- 4.2 Problèmes variationnels abstraits et espaces de Hilbert . . . 78
- 4.3 Espaces de Sobolev . . . 87
- 4.4 Application au problème aux limites . . . 98

5 Les méthodes d'approximation variationnelle **103**
- 5.1 Définition et premières propriétés . . . 103
- 5.2 Forme matricielle de la méthode de Galerkin . . . 106
- 5.3 Perturbation d'une approximation variationelle . . . 107

6 La méthode des éléments finis **111**
- 6.1 Définition de la méthode dans le cas dit P_1 . . . 111
- 6.2 Analyse de la convergence dans le cas P_1 . . . 117

6.3 Eléments finis de degré et régularité plus élevés 121
6.4 Analyse générale de convergence 125
6.5 Compléments 129

7 Méthodes de bases hilbertiennes **135**
7.1 Bases hilbertiennes 135
7.2 Bases de Fourier 137
7.3 Bases polynomiales 142
7.4 Bases d'ondelettes 150

Avant-propos

Ces notes reprennent pour l'essentiel celles rédigées par Hervé Le Dret puis par Pascal Frey, Edwige Godlewski et Laurent Boudin, qui ont été en charge de ce cours avant nous. Nous les remercions de nous avoir ainsi aidés dans la préparation de ce polycopié.

Bilbiographie pour les notions prérequises :

— G. Allaire et S. M. Kaber : Algèbre linéaire numérique - Cours et exercices. Ellipses 2002
— G. Allaire, S. M. Kaber : Introduction à Scilab, exercices pratiques corrigés d'algèbre linéaire, Ellipses, 2002
— J.-P. Chancelier, F. Delebecque, C. Gomez, M. Goursat, R. Nikoukhah, S. Steer : Introduction à Scilab, Springer, 2001
— P. G. Ciarlet : Introduction à l'analyse numérique matricielle et à l'optimisation. Dunod, 2006
— M. Crouzeix et A.-L. Mignot : Analyse Numérique des Equations Différentielles, Masson, 1989
— S. Delabrière et M. Postel : Méthodes d'approximation. Equations différentielles, applications Scilab, Ellipses, 2004
— J.-P. Demailly : Analyse numérique et équations différentielles, EDP Sciences, 2006
— P. Lascaux, R. Théodor : Analyse numérique matricielle appliquée à l'art de l'ingénieur, Dunod 2000
— M. Schatzman : Analyse numérique - Une approche mathématique, Dunod 2001.

Pour la suite du cours, on pourra consulter

— G. Allaire : Analyse numérique et optimisation, Editions de l'Ecole Polytechnique, Palaiseau 20
— B. Mohammadi, J.-H. Saïac, Pratique de la simulation numérique, Dunod, 2003
— Laurent Di Menza : Analyse numérique des équations aux dérivées partielles, Cassini 2009
— F. Filbet : Analyse numérique, algorithme et étude mathématique, Dunod, 2009
— F. Hubert et J. Hubbard, Calcul scientifique 2, de la théorie à la pratique, Vuibert, 2006
— H. Le Dret, B. Lucquin, Partial differential equations : modeling, analysis and numerical approximation, Birkhaüser, 2016
— B. Lucquin : Equations aux dérivées partielles et leurs approximations, Ellipses, 2004
— B. Mohammadi, J.-H. Saïac, Pratique de la simulation numérique, Dunod, 2003
— P.-A. Raviart, J.-M. Thomas : Introduction l'analyse numérique des équations aux dérivées partielles, Dunod, 2004
— L. Sainsaulieu : Calcul scientifique, Masson enseignement des mathématiques, Dunod, 2000

Chapitre 1

Introduction

"En quelques décennies, les simulations numériques sont devenues un outil privilégié d'investigation dans les sciences et les technologies. Elles ont pour but de reproduire par le calcul le comportement d'un système décrit par un modèle, très souvent constitué d'équations aux dérivées partielles. Ces équations correspondent à la traduction mathématique de lois scientifiques. L'essor des simulations numériques renforce donc la nécessité de l'étude mathématique (analyse) de ces équations et de leur résolution numérique."

Extrait de la leçon inaugurale de Pierre-Louis Lions au Collège de France (Fayard, 2003)

1.1 "Bases des méthodes numériques"

On va présenter des méthodes numériques de façon très complète sur quelques exemples génériques simples. On n'abordera pas toutes les méthodes (comme celles de résolution de systèmes linéaires, d'optimisation, les méthodes probabilistes, statistiques, ...), mais un certain nombre de celles qui permettent le calcul de solutions approchées d'équations comportant des dérivées partielles, ce qui représente déjà un domaine très important dont nous verrons quelques exemples. Donnons les grandes lignes de la démarche.

On introduit le principe d'une méthode (prenons l'exemple de la méthode des différences finies, nous verrons aussi la méthode des éléments finis, celle des volumes finis) ; on la décrit précisément sur l'exemple générique à laquelle elle est associée (exemple : le Laplacien pour l'équation de Poisson), cet exemple étant d'abord analysé, pour s'assurer de l'existence et de l'unicité de la solution. Par exemple, pour les différences finies, le *principe* de la méthode est de calculer les valeurs approchées de la solution aux points d'une grille, en approchant les dérivées par des différences finies (nous détaillerons bien sûr plus loin ce que cela signifie) ; on écrit alors le (ou un) *schéma* qui permet le calcul d'une valeur approchée de la solution (par exemple, on définit un "Laplacien discret"), on en fait l'*analyse numérique*, en étudiant ses propriétés (consistance, stabilité, convergence, estimation de l'erreur) ; on effectue la *mise en oeuvre* en écrivant un algorithme et un programme informatique qui permet la résolution numérique, c'est à dire le calcul effectif de la solution approchée ; le programme est écrit de telle sorte que pour changer le cas d'application, il suffit de changer quelques données en entrée, cela conduit ainsi à un *code*. Le code permet alors de vérifier sur des exemples particuliers (des cas test) dont on connait la solution exacte les propriétés démontrées théoriquement (l'ordre de

convergence) : de telle manière, on assure la *validation* de la méthode. Lorsqu'il y a plusieurs méthodes ou plusieurs schémas possibles, on les compare sur des cas test. Ces grandes lignes sont les "bases" de l'étude des méthodes numériques.

Les codes utilisés dans la pratique (industrie, recherche appliquée) utilisent des méthodes numériques peut être plus élaborées mais qui ont été développées à partir d'exemple simplifiés et ont suivi le même processus de validation. Cependant, si la mise en oeuvre sur des cas réels suit même cette démarche, elle est plus compliquée et nécessite d'autres procédures complémentaires (construction du modèle mathématique, choix des variables pour la modélisation, ajustement des paramètres, discrétisation du domaine et choix du modèle numérique, traitement des échelles multiples, comparaison avec les mesures, interprétation des résultats, ...). Par ailleurs, si la taille des cas réels, le nombre de paramètres et de variables, conduit à faire exploser le nombre de calculs à effectuer, il est alors indispensable de développer des méthodes performantes que nous n'aborderons pas, permettant de mener à bien les calculs (décomposition de domaine, calcul parallèle, adaptation de maillages, réduction de modèles, bases réduites,...).

Les systèmes complexes sont modélisés le plus souvent par parties (éléments géométriques ou composants), ou éléments correspondant à un type de fonctionnement, de comportement, de phénomène physique que l'on peut différencier,... qui sont ensuite *couplés* ; chaque "brique élémentaire" suppose une validation, de même que le couplage doit être validé sur des cas test. Les équations ne sont plus des équations génériques aussi simples que celles que nous allons étudier, cependant le processus même de modélisation fait qu'elles rentrent dans de grandes catégories identifiées puisque les équations traduisent des lois "physiques" (celles du domaine d'application), des processus plus ou moins connus (certains doivent encore faire l'objet de nouvelles études approfondies). Un exemple très actuel est celui de la modélisation et la simulation numérique appliquées à différentes problématiques liées au projet de construction du réacteur expérimental ITER.

Le travail du mathématicien appliqué intervient dans le processus de modélisation (en relation avec la discipline d'application ou le métier), pour traduire le phénomène observé par des équations, dans l'analyse des équations, du comportement des solutions (par rapport au temps, leur dépendance par rapport à des paramètres, des données,...), le couplage, dans le choix des méthodes de discrétisation, l'analyse numérique de ces méthodes et aussi dans la mise en oeuvre. Il est donc utile d'illustrer toute la chaîne de cette activité sur des exemples bien compris, avant de suivre cette démarche sur les cas complexes, d'aborder des cas "industriels", ou encore d'envisager la définition de nouvelles méthodes numériques. Mais le domaine de recherche et d'applications se développe continuellement : on étudie de nouvaux contextes, de nouveaux couplages multiphysiques, de nouvelles méthodes dont le comportement est robuste par rapport à un changement d'échelle (passage à limite dans l'équation quand un paramètre devient très petit), on veut traiter les problèmes en grande dimension, on cherche à contrôler la propagation des incertitudes sur les données, introduire l'incertitude sur les modèles....

Nous allons donc présenter les principales méthodes numériques utilisées dans le contexte de l'approximation des EDP et en détailler l'étude dans le cas de la dimension 1. Nous supposerons connues les méthodes classiques d'interpolation par un polynôme, celles permettant d'approcher une dérivée en un point (différences finies), d'approcher une intégrale (formules de quadrature), ou de calculer de façon approchée la solution d'une EDO (équation différentielle ordinaire), comme la méthode d'Euler. L'algorithmique et la mise en oeuvre seront détaillées en travaux pratiques.

1.2 Présentation du cours : les EDP et leur approximation

Ce cours se propose donc d'introduire un certain nombre de méthodes d'approximation numérique des solutions de problèmes d'équations aux dérivées partielles (EDP). Il s'agit d'un exposé relativement élémentaire et les exemples traités seront essentiellement des exemples en dimension 1, ce qui relativise singulièrement le terme *dérivée partielle*, bien sûr. Les "vrais" problèmes d'EDP, lesquels sont posés en dimension d'espace 2, 3 (voire plus encore), et peuvent aussi faire intervenir une variation par rapport au temps, nécessitent des concepts et des techniques mathématiques qui se situent largement au delà des limites de ce cours. Mais pourquoi s'intéresse-t-on à l'approximation des solutions d'EDP ?

Il se trouve que la *modélisation* de très nombreux problèmes issus de la physique, de la mécanique et des sciences de l'ingénieur, font qu'ils se présentent sous forme d'EDP. Plus précisément, bon nombre de phénomènes physiques font intervenir des grandeurs qui sont, plus ou moins approximativement, solution (de systèmes) d'EDP. Donnons en quelques exemples.

Considérons un objet matériel, occupant l'adhérence d'un ouvert Ω de $\mathbb{R}^3$. Dans cet objet on observe la présence d'une densité de charge électrique, *i.e.*, une fonction f de Ω dans $\mathbb{R}$ qui donne en tout point x la densité de charge électrique au voisinage de ce point. Ceci signifie que la charge électrique contenue dans une petite boule centrée en x est approximativement égale à $f(x)$ multiplié par le volume de cette boule, à condition que la boule ne soit pas trop petite quand même, puisque la matière est en fait discontinue et qu'une boule trop petite finirait par ne contenir aucune charge ! On n'en définit pas moins la fonction f en tout point de Ω par un procédé de moyennisation qui ne nous concerne pas ici. Le potentiel électrique dans l'objet est une fonction $u\colon \bar{\Omega} \to \mathbb{R}$ dont le gradient, *i.e.*, le vecteur $\nabla u = (\partial u/\partial x_1, \partial u/\partial x_2, \partial u/\partial x_3)^T$ où x_1, x_2 et x_3 sont des coordonnées cartésiennes orthonormées choisies une fois pour toutes, donne au signe près le champ électrique qui règne dans l'objet. Ce champ peut être en principe mesuré puisqu'il se manifeste par l'apparition de forces. Supposant que la surface de Ω, c'est-à-dire en termes mathématiques, sa frontière $\partial\Omega$, est couverte d'un matériau *conducteur*, on montre en physique que le potentiel u est solution de l'*équation de Poisson*

$$\Delta u(x) = f(x) \text{ pour tout } x \in \Omega, \tag{1.1}$$

où l'opérateur Δ, le Laplacien, est défini par

$$\Delta u = \sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2},$$

et, ayant normalisé les constantes physiques à 1 pour simplifier les notations, avec la *condition aux limites* (c'est-à-dire vérifiée par u sur la frontière)

$$u(x) = 0 \text{ pour tout } x \in \partial\Omega, \tag{1.2}$$

due au fait que la surface est supposée conductrice, donc équipotentielle. Le couple formé par l'équation (1.1) et la condition aux limites (1.2) forme ce que l'on appelle un *problème aux limites*. Dans le cas où, comme ici, c'est la valeur de la fonction u qui est prescrite sur la frontière, on parle de *condition de Dirichlet*.

Il est intéressant de noter que l'interprétation électrostatique précédente du problème aux limites (1.1)–(1.2) n'est pas la seule interprétation physique que l'on puisse en donner. Ainsi,

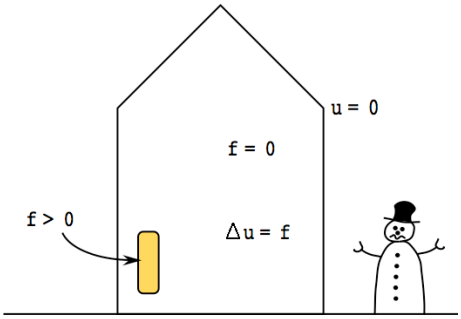


FIGURE 1.1 – Modélisation de l'équilibre thermique

si f représente une densité de sources de chaleur, alors $\theta = -u$ sera la température à l'équilibre thermique dans l'ouvert Ω quand on impose une température nulle à la frontière. La figure 1.1 représente une maison (bi-dimensionnelle et mal isolée) en hiver : les murs sont à la température extérieure 0 °C, les sources de chaleur sont nulles sauf dans le radiateur où $f > 0$. L'edp s'obtient alors en combinant deux équations de la physique : l'une traduit la loi de conservation de l'énergie (ou de la quantité de chaleur) proportionnelle à la température, l'autre relie le flux de chaleur q au gradient de température $\nabla\theta$ (loi de Fourier $q = -k\nabla\theta$, k conductivité thermique). Les constantes physiques (ici la conductivité thermique k) ont été prises égales à 1 pour simplifier l'écriture (mais aussi par ce qu'on s'y ramène par des changements simples, on étudie le *problème adimensionné*), alors que dans la modélisation physique au départ θ, par exemple, est exprimé en degré Kelvin.

De même, dans le cas où $\Omega \subset \mathbb{R}^2$, si $\partial\Omega$ représente la forme d'un cadre sur lequel on tend une membrane élastique en baudruche ou une peau de tambour, et f une densité de forces verticales que l'on applique sur cette membrane, $-u(x)$ sera le déplacement vertical que subit le point x de la membrane lorsque celle-ci atteint une position d'équilibre. Avant tout adimensionnement, le modèle issu de la mécanique des milieux élastiques s'écrit sous la forme $-div(K\nabla u) = f$, et dans le cas général, K peut dépendre de x, nous avons supposé K constant.

On conçoit bien que, dans ce type de situation, on a besoin de déterminer numériquement la fonction u. Or, pour des problèmes aux limites du type (1.1)–(1.2), si l'on sait souvent démontrer *l'existence* et *l'unicité* de solutions, grâce à une analyse mathématique adaptée, il n'existe pas en général de solution "analytique", c'est-à-dire de formule plus ou moins explicite donnant les valeurs de u dans Ω. On n'a par conséquent pas d'autre choix que de calculer des *approximations numériques* de cette solution. On distingue trois étapes :

1) La description des différentes méthodes d'approximation,

2) Leur analyse numérique : les méthodes considérées font intervenir un paramètre de discrétisation, car on ne peut en pratique calculer effectivement qu'un *nombre fini* de valeurs numériques, et il s'agit d'étudier le comportement des approximations ainsi calculées quand le paramètre de discrétisation tend vers 0, afin de les comparer à la solution exacte. En d'autres termes, on veut démontrer que l'on a bien déterminé des approximations de la solution, et cela en un sens précis.

3) L'implémentation effective des méthodes sur un ordinateur. Se posent alors des questions d'algorithmique, de complexité et de programmation, questions que nous n'aborderons pas ici.

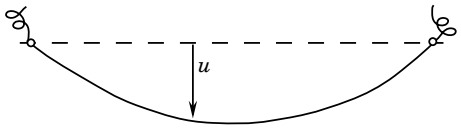


FIGURE 1.2 – Le fil élastique

Par la suite, nous nous intéresserons essentiellement aux problèmes en dimension 1 d'espace. L'analogue du problème aux limites (1.1)–(1.2) en dimension 1 est le suivant : on prend $\Omega =]0,1[$ (l'intervalle $]0,1[$ est habituellement choisi comme intervalle de référence, la généralisation à un intervalle quelconque est immédiate), on se donne $f\colon]0,1[\to\mathbb{R}$ et l'on cherche $u\colon]0,1[\to\mathbb{R}$ telle que

$$-u''(x) = f(x) \text{ dans }]0,1[\quad \text{et} \quad u(0) = u(1) = 0. \tag{1.3}$$

Notons qu'ici la frontière est formée de deux points, $\partial\Omega = \{0,1\}$. Considérant le segment Ω comme un sous-ensemble de $\mathbb{R}^2$ qui représente un fil élastique tendu entre deux points et si f modélise une densité de forces verticales que l'on applique sur ce fil divisée par la tension du fil, alors $u(x)$ sera le déplacement vertical que subit le point x du fil à l'équilibre (voir la figure 1.2). Il est très important de bien réaliser que ce *problème aux limites* est très différent du *problème de Cauchy* pour l'équation différentielle ordinaire (EDO) :

$$-u''(x) = f(x) \text{ dans }]0,1[\quad \text{et} \quad u(0) = \alpha, u'(0) = \beta. \tag{1.4}$$

Dans le problème de conditions initiales, on se donne des *valeurs initiales*, notées ici α, β, pour u et u' au *même point* 0. Il faut penser plutôt dans ce cas à une évolution d'un système après un instant initial, la variable x s'interprétant alors comme le temps, $u(x)$ étant la position à l'instant x, $u'(x)$ la vitesse et $u''(x)$ l'accélération. Par conséquent, $u(0)$ et $u'(0)$ sont respectivement la position et la vitesse initiales.

Dans le problème aux limites, on se donne au contraire la valeur de u aux *deux extrémités* de l'intervalle et la variable x s'interprète plutôt comme une variable d'espace, comme le montre l'exemple du fil élastique ci-dessus. En raison de cette différence essentielle de nature, les méthodes d'analyse des EDO (équations différentielles ordinaires) ne s'appliquent pas directement et les méthodes d'analyse numérique adaptées aux EDO comme les méthodes d'Euler progressive et rétrograde, les méthodes de Runge-Kutta, les méthodes à pas multiples, etc., ne sont adaptées à l'approximation des problèmes aux limites et ne peuvent pas être utilisées telles quelles. Pour ces derniers, il nous faudra mettre en œuvre des idées et des méthodes parfois entièrement différentes.

Notons cependant que beaucoup de problèmes physiques, ceux pour lesquels on étudie l'évolution d'une quantité u par rapport au temps à partir d'une condition initiale u_0, sont modélisés par des edp qui font intervenir à la fois la variable de temps t et la variable d'espace x comme variables indépendantes, la fonction $u = u(x,t)$ dépend de deux variables. Il faut alors se donner (en plus des conditions aux limites) la condition inititiale $u(x,t=0) = u_0(x)$; par convention on prend $t = 0$, si le temps initial est noté t_0, on se ramène simplement à 0 par changement de variable, en posant $s = t - t_0$ et $v(x,s) = u(x,t)$.

Mentionnons ainsi : *l'équation de la chaleur* en dimension 1

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f, t > 0, x \in]0,1[,$$

où encore $u = u(x,t)$, et en dimension (d'espace) supérieure

$$\frac{\partial u}{\partial t} - \Delta u = f, t > 0, x \in \Omega,$$

avec $u = u(x_1, x_2, x_3, t)$; f, la source de chaleur, peut aussi dépendre du temps. Notons que lorsque f ne dépend pas de t, et lorsque u ne dépend plus du temps, c'est à dire si on atteint un *état stationnaire*, u vérifie une équation de Poisson (1.3) ou (1.1) : l'équation de Poisson (1.3) en dimension 1 modélise donc aussi la répartition de chaleur dans un fil conducteur homogène (la conductivité thermique est constante et prise égale à 1), soumis à un rayonnement extérieur f, dont la rempérature est fixée nulle aux extrémités, une fois qu'il a atteint l'équilibre thermique. En milieu hétérogeène, la conductivité thermique est une fonction de la variable x, l'équation s'écrit alors sous la forme $-(k(x)u'(x))' = f(x),$.

L'équation de la chaleur, dite aussi équation de *diffusion*, modélise l'évolution de la température dans le domaine Ω à partir d'une condition initiale connue. Elle intervient également dans la modélisation de bien d'autres phénomènes, par exemple la diffusion d'un polluant, la migration d'espèces chimiques ou encore en finances, où une variante (l'équation de Black-Scholes) est utilisée pour modéliser le prix d'une option d'achat.

À titre d'exemple d'un autre type d'équations, nous serons aussi amenés à étudier *l'équation de transport* linéaire en dimension 1

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, t > 0, x \in \mathbb{R}, \tag{1.5}$$

où a représente une vitesse, par exemple constante, qui est supposée connue. Pour le *problème de Cauchy* (problème de condition initiale pour $x \in \mathbb{R}$ tout entier), on se donne seulement la valeur $u(x,0) = u_0(x)$ au temps initial, $t = 0$. L'unique solution est alors $u(x,t) = u_0(x - at)$, ce qui représente un phénomène de propagation à vitesse constante a (le profil u_0 est translaté avec le temps, sans modification). Attention si on regarde ce problème sur un intervalle borné, par exemple $x \in [0,1]$, on a aussi un problème aux limites, mais avec une dérivée en espace d'ordre 1 ; il faut alors aussi se donner soit $u(0,t)$ pour tout $t > 0$, si $a > 0$ soit $u(1,t)$ si $a < 0$. Nous reviendrons en plus en détail sur cette edp à la fin du cours.

L'équation des ondes, modélise aussi des phénomènes de propagation. En dimension (d'espace) 1, elle est de la forme

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = f, t > 0. \tag{1.6}$$

Lorsque on considère le problème de Cauchy, $x \in \mathbb{R}$, on vérifie assez facilement que l'équation a des solutions de la forme $u(x,t) = v(x - ct) + w(x + ct)$, où v, w sont des fonctions réelles d'une variable : on a la superposition de deux ondes qui se propagent à des vitesses opposées, $\pm c$. Pour $x \in [0,1[$, (1.6) est l'équation dite des cordes vibrantes et en dimension supérieure, elle s'écrit

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f, t > 0, x \in \Omega,$$

en dimension 2, l'équation modélise les vibrations d'une membrane élastique tendue (tambour). De nouveau, les constantes physiques ont été normalisées.

Pour ces équations (chaleur, onde), en plus des conditions aux limites du type de celles qui ont été mentionnées pour l'équation de Poisson, par exemple l'analogue de (1.2) est $u(x,t) = 0, x \in \partial\Omega$ pour tout $t > 0$ (mais la condition aux limites pourrait aussi dépendre du temps), on se donne une condition initiale $u(x,0) = u_0(x)$, ou $u(x_1,x_2,x_3,0)$, au temps $t = 0$, et aussi $\frac{\partial u}{\partial t}(x,0) = u_1(x)$ pour l'équation des ondes qui fait intervenir une dérivée partielle par rapport à t d'ordre 2. En dimension 1, a connaissance de u_0 et u_1 permet alors de déterminer les fonctions v, w et la solution du problème de Cauchy pour l'équation des ondes.

Pour l'approximation des problèmes modélisés par l'équation de la chaleur (par exemple), on utilisera les méthodes d'analyse numérique adaptées aux EDO (comme la méthode d'Euler) pour la discrétisation en temps combinées avec les méthodes pour les problèmes aux limites (en espace) que nous allons étudier. Pour l'équation de transport (1.5), nous développerons des méthodes spécifiques de type "volumes finis".

L'ordre de l'edp est l'ordre de la plus grande dérivée présente dans l'équation : par exemple l'équation des ondes est d'ordre 2 en espace et en temps, alors que l'équation de la chaleur est d'ordre deux en espace et un en temps. On utilise souvent le vocabulaire suivant pour caractériser la nature des edp d'ordre ≤ 2 : elliptique (exemple : équation de Laplace, de Poisson), parabolique (exemple : équation de la chaleur), hyperbolique (exemple : équation des ondes). Les équations ci-dessus sont linéaires : u dépend *linéairement* de la donnée f. Elles s'écrivent sous la forme $Lu = f$ où l'opérateur L est un opérateur aux dérivées partielles linéaires. Pour deux variables (avec des variables notées x, y), l'opérateur L a la forme générale suivante

$$Lu \equiv a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + a_0 u$$

avec des coefficients constants, on regarde le *symbole* $P(x,y) = ax^2 + bxy + cy^2 + dx + ey + a_0$ et le signe de $b^2 - 4ac$. Le cas dit elliptique correspond à $b^2 - 4ac < 0$, parabolique à $b^2 - 4ac = 0$, hyperbolique à $b^2 - 4ac > 0$. Un problème est *bien posé* si pour toute donnée (c'est à dire la source f et les conditions initiales et/ou aux limites), il possède une unique solution qui dépend continûment des données (ce qui suppose avoir bien précisé sur quel espace de fonctions on définit l'opérateur L). Les trois catégories précédentes correspondent à des problèmes bien posés à condition d'avoir bien défini le type de conditions initiales et/ou aux limites.

Nous commencerons par étudier le problème aux limites pour une équation elliptique en dimension 1 dans une formulation classique, c'est à dire dans des espaces usuels de fonctions $\mathcal{C}^k$, et d'abord avec les outils traditionnels, ceux déjà connus pour les EDO, ce qui est possible en dimension 1 (mais ne l'est plus en dimension supérieure), puis par des méthodes qui, elles, seront généralisables en dimension supérieure. Outre l'existence et l'unicité, il est utile de connaître les propriétes de la solution, avant de chercher à l'approcher, afin de construire des méthodes numériques qui vont essayer de conserver l'analogue discret de ces propriétés.

L'approximation numérique de ces problèmes aux limites fait intervenir la résolution de systèmes linéaires et nous faisons donc auparavant quelques rappels d'algèbre linéaire matricielle.

1.3 Quelques rappels sur les matrices

Nous donnons ici quelques rappels sur les matrices qui nous seront utiles dans la suite. D'abord, pour des raisons pratiques, on écrira souvent un vecteur (colonne) v de $\mathbb{R}^n$

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

sous la forme du transposé d'un vecteur ligne

$$v = (v_1, v_2, \ldots, v_n)^T,$$

et alors $v^T = (v_1, v_2, \ldots, v_n)$.

Pour tout entier $n \in \mathbb{N}^*$, on note $M_n(\mathbb{R})$ l'espace des matrices carrées $n \times n$ à coefficients réels. Si l'on munit $\mathbb{R}^n$ de sa base canonique, c'est-à-dire, du n-uplet de vecteurs e_i, $1 \leq i \leq n$, définis par $e_1 = (1,0,\ldots,0)^T$, $e_2 = (0,1,\ldots,0)^T$, ..., $e_n = (0,0,\ldots,1)^T$, l'espace $M_n(\mathbb{R})$ s'identifie à l'espace des endomorphismes de $\mathbb{R}^n$: si $A \in M_n(\mathbb{R})$ est la matrice de coefficients $a_{i,j}$, $A = (a_{i,j})$, $1 \leq i, j \leq n$, les vecteurs colonnes $Ae_k = (a_{1,k}, a_{2,k}, \ldots, a_{n,k})^T$, $1 \leq k \leq n$ déterminent complètement l'endomorphisme correspondant. Nous utiliserons cette identification sans vergogne et noterons $\ker A$, $\operatorname{Im} A$, Av le noyau, l'image et l'action de l'endomorphisme associé à la matrice A sur un vecteur v de $\mathbb{R}^n$. Dans ce dernier cas, il s'agit aussi du produit de A, matrice carrée $n \times n$, avec la matrice colonne (à n lignes) des coordonnées de v dans la base canonique, ce produit matriciel donnant le vecteur colonne des coordonnées de Av dans la base canonique. Ce type d'abus de langage n'est pas dangereux pour la classe d'applications que l'on a en vue, tant que l'on s'est fixé une base de $\mathbb{R}^n$ et, plus fondamentalement, que l'on sait ce que l'on fait.

L'espace $M_n(\mathbb{R})$ est lui-même un $\mathbb{R}$-espace vectoriel (de dimension n^2) pour l'addition des matrices et la multiplication par un scalaire. On peut donc parfaitement définir des normes sur cet espace, c'est-à-dire des applications $\|\cdot\| : M_n(\mathbb{R}) \to \mathbb{R}_+$ qui satisfont les trois propriétés :

$$\begin{cases} \forall A, B \in M_n(\mathbb{R}), \|A+B\| \leq \|A\| + \|B\| \text{ (inégalité triangulaire)}, \\ \forall A \in M_n(\mathbb{R}), \forall \lambda \in \mathbb{R},\ \|\lambda A\| = |\lambda| \|A\|, \\ \|A\| = 0 \Longleftrightarrow A = 0. \end{cases}$$

En plus de sa structure d'espace vectoriel, l'espace $M_n(\mathbb{R})$ est doté d'une structure multiplicative qui en fait un anneau non commutatif. En général, une norme sur $M_n(\mathbb{R})$ n'a aucune raison d'être compatible avec cette structure multiplicative. Les normes qui prennent en compte la structure multiplicative présentent un intérêt particulier.

Définition 1.3.1 *Une norme sur $M_n(\mathbb{R})$ est appelée* norme matricielle *si elle satisfait*

$$\forall A, B \in M_n(\mathbb{R}), \qquad \|AB\| \leq \|A\| \|B\|.$$

Remarque. Il n'existe aucune norme qui satisfasse la relation plus forte $\|AB\| = \|A\|\|B\|$ pour tout couple de matrices (A,B). En effet, il existe des couples (A,B) tels que $A \neq 0$, $B \neq 0$ mais $AB = 0$ (l'anneau $M_n(\mathbb{R})$ admet des diviseurs de zéro). Il suffit pour cela que l'on ait $\operatorname{Im} B \subset \ker A$. □

Après avoir posé la définition 1.3.1, encore faut-il vérifier qu'il existe des normes qui la satisfont ! En fait, nous allons voir qu'il en existe beaucoup.

Proposition 1.3.1 (et définition) *Soit $\|\cdot\|$ une norme sur $\mathbb{R}^n$. L'application*

$$\begin{cases} |||\cdot|||: M_n(\mathbb{R}) & \longrightarrow \mathbb{R}_+ \\ A & \longmapsto |||A||| = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} \end{cases}$$

définit une norme matricielle sur $M_n(\mathbb{R})$ appelée norme subordonnée *à la norme vectorielle $\|\cdot\|$.*

Démonstration. Soit A une matrice donnée. Posant, pour $v \in \mathbb{R}^n \setminus \{0\}$, $w = v/\|v\|$, on voit que

$$|||A||| = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \Big\|A\Big(\frac{v}{\|v\|}\Big)\Big\| = \sup_{w\in\mathbb{R}^n,\|w\|=1} \|Aw\|.$$

Comme la sphère unité de $\mathbb{R}^n$ est compacte (on est dans un espace de dimension finie) et que la norme choisie est une fonction continue, on en déduit d'abord que $|||A||| < +\infty$, donc l'application $A \mapsto |||A|||$ envoie bien $M_n(\mathbb{R})$ dans $\mathbb{R}_+$. Montrons dans un premier temps qu'il s'agit d'une norme. Considérons un couple de matrices (A,B) et un scalaire λ. Pour tout $v \in \mathbb{R}^n \setminus \{0\}$,

$$\frac{\|(A+B)v\|}{\|v\|} = \frac{\|Av+Bv\|}{\|v\|} \leq \frac{\|Av\|+\|Bv\|}{\|v\|} = \frac{\|Av\|}{\|v\|} + \frac{\|Bv\|}{\|v\|} \leq |||A||| + |||B|||,$$

puisque $\|\cdot\|$ satisfait l'inégalité triangulaire. Prenant le sup du membre de gauche, on obtient

$$|||A+B||| \leq |||A||| + |||B|||.$$

De même,

$$|||\lambda A||| = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|\lambda Av\|}{\|v\|} = \sup_{v\in\mathbb{R}^n\setminus\{0\}} |\lambda| \frac{\|Av\|}{\|v\|} = |\lambda|\,|||A|||.$$

Enfin, si $|||A||| = 0$, on a $Av = 0$ pour tout $v \in \mathbb{R}^n$, d'où $A = 0$ évidemment. On a donc bien affaire à une norme sur $M_n(\mathbb{R})$.

Pour montrer que c'est une norme matricielle, on remarque d'abord que pour tout $v \in \mathbb{R}^n$,

$$\|Av\| \leq |||A|||\,\|v\|, \tag{1.7}$$

ce qui découle immédiatement de la définition. On prend alors deux matrices A et B et l'on note que pour tout $v \in \mathbb{R}^n \setminus \{0\}$,

$$\|ABv\| = \|A(Bv)\| \leq |||A|||\,\|Bv\| \leq |||A|||\,|||B|||\,\|v\|,$$

d'où en divisant par $\|v\|$,

$$\frac{\|ABv\|}{\|v\|} \leq |||A|||\,|||B|||,$$

et l'on conclut en prenant le sup du membre de gauche sur tous les v dans $\mathbb{R}^n \setminus \{0\}$. □

Proposition 1.3.2 *On a également la caractérisation suivante de la norme subordonnée :*

$$|||A||| = \inf\{\mu \in \mathbb{R}_+; \forall v \in \mathbb{R}^n, \|Av\| \leq \mu\|v\|\}.$$

Démonstration. Par la remarque (1.7) de la démonstration précédente, il est clair que

$$|||A||| \geq \inf\{\mu \in \mathbb{R}_+; \forall v \in \mathbb{R}^n, \|Av\| \leq \mu\|v\|\}.$$

Soit donc $\mu \in \mathbb{R}_+$ tel que $\|Av\| \leq \mu\|v\|$ pour tout $v \in \mathbb{R}^n$. On a alors

$$|||A||| = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} \leq \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\mu\|v\|}{\|v\|} = \mu,$$

d'où

$$|||A||| \leq \inf\{\mu \in \mathbb{R}_+; \forall v \in \mathbb{R}^n, \|Av\| \leq \mu\|v\|\},$$

et le résultat. □

Remarques. i) Toute norme matricielle subordonnée satisfait $|||\,\mathrm{Id}\,||| = 1$.

ii) Il existe des normes matricielles non subordonnées, par exemple

$$|||A||| = \big(\mathrm{tr}\,(A^TA)\big)^{1/2} = \Big(\sum_{ij} |a_{ij}|^2\Big)^{1/2},$$

(le vérifier).

iii) On introduit parfois la notion de norme subordonnée en prenant le sup sur $\mathbb{C}^n \setminus \{0\}$ au lieu de $\mathbb{R}^n \setminus \{0\}$ (en complexifiant donc l'espace). Il peut arriver, même pour des matrices réelles, que les deux expressions ne coïncident pas, le sup «réel» étant toujours naturellement inférieur au sup «complexe». Cela n'est pas le cas pour les exemples que nous allons considérer plus loin. □

Proposition 1.3.3 (et définition) *Pour tout $p \in [1, +\infty]$, la fonction $\|v\|_p : \mathbb{R}^n \to \mathbb{R}_+$ définie par*

$$\begin{cases} \|v\|_p = \Big(\sum_{i=1}^n |v_i|^p\Big)^{1/p} & \text{pour } p < +\infty, \\ \|v\|_\infty = \max_{i=1,\dots,n} |v_i| & \text{pour } p = +\infty, \end{cases}$$

est une norme vectorielle sur $\mathbb{R}^n$ appelée «norme p».

Nous ne démontrerons pas ce point. Pour $p = 2$, on retrouve la norme euclidienne usuelle. On notera $|||\cdot|||_p$ la norme matricielle subordonnée à la norme p. On sait évaluer plus ou moins explicitement ces normes subordonnées pour trois valeurs de p. On note $\rho(A)$ le *rayon spectral* de la matrice A, c'est-à-dire le module de la valeur propre de A de plus grand module. C'est le rayon du plus petit disque de $\mathbb{C}$ centré en 0 qui contient toutes les valeurs propres de A.

Théorème 1.3.1 *On a*

$$\begin{aligned} |||A|||_1 &= \max_j \Big(\sum_i |a_{ij}|\Big), \\ |||A|||_2 &= \sqrt{\rho(A^TA)}, \\ |||A|||_\infty &= \max_i \Big(\sum_j |a_{ij}|\Big) \end{aligned}$$

Démonstration. Commençons par le cas $p = +\infty$. Soit v un vecteur non nul. Comme $(Av)_i = \sum_{j=1}^n a_{ij}v_j$, il vient

$$\|Av\|_\infty = \max_i \Big| \sum_{j=1}^n a_{ij}v_j \Big| \leq \max_i \Big(\sum_{j=1}^n |a_{ij}||v_j| \Big)$$

par l'inégalité triangulaire. Or pour tout indice j, $|v_j| \leq \|v\|_\infty$ par définition de la norme ∞. On en déduit donc que

$$\|Av\|_\infty \leq \Big(\max_i \sum_{j=1}^n |a_{ij}| \Big) \|v\|_\infty,$$

d'où, grâce à la proposition 1.3.2,

$$|||A|||_\infty \leq \max_i \sum_{j=1}^n |a_{ij}|.$$

Pour conclure, il suffit par conséquent de trouver un vecteur $w \in \mathbb{R}^n$, non nul, qui soit tel que $\|Aw\|_\infty = \big(\max_i \sum_{j=1}^n |a_{ij}|\big) \|w\|_\infty$. Pour cela, on note i_0 un indice de ligne qui satisfait

$$\max_i \sum_{j=1}^n |a_{ij}| = \sum_{j=1}^n |a_{i_0 j}|$$

(il en existe évidemment au moins un). On introduit alors le vecteur w de composantes

$$w_j = \frac{a_{i_0 j}}{|a_{i_0 j}|} \text{ si } a_{i_0 j} \neq 0, \quad w_j = 1 \text{ si } a_{i_0 j} = 0.$$

Dans tous les cas $|w_j| = 1$ et donc $\|w\|_\infty = 1$. De plus

$$(Aw)_{i_0} = \sum_{j=1}^n a_{i_0 j} w_j = \sum_{j=1}^n |a_{i_0 j}|,$$

par construction des w_j. On en déduit que

$$\|Aw\|_\infty \geq \sum_{j=1}^n |a_{i_0 j}| = \max_i \sum_{j=1}^n |a_{ij}|,$$

et w répond donc à la question.

Le cas $p = 1$ est analogue et est laissé en exercice. Traitons le cas $p = 2$. On a

$$\|Av\|_2^2 = (Av)^T Av = v^T A^T Av, \text{ et } \|v\|_2^2 = v^T v.$$

La matrice $A^T A$ est symétrique et positive (puisque $\|Av\|_2^2 \geq 0$). Par conséquent, elle est orthogonalement diagonalisable et ses valeurs propres λ_i sont réelles positives. On convient de les ordonner en ordre croissant, ce qui implique que $\rho(A^T A) = \lambda_n$. On note f_i une base orthornormée de vecteurs propres de $A^T A$. Décomposant v sur cette base $v = \sum_{i=1}^n v_i f_i$, on voit que $A^T Av = \sum_{i=1}^n \lambda_i v_i f_i$, d'où

$$\|v\|_2^2 = \sum_{i=1}^n v_i^2 \quad \text{et} \quad \|Av\|_2^2 = \sum_{i=1}^n \lambda_i v_i^2.$$

D'une part

$$\frac{\|Av\|_2^2}{\|v\|_2^2} = \frac{\sum_{i=1}^n \lambda_i v_i^2}{\sum_{i=1}^n v_i^2} \leq \frac{\lambda_n \sum_{i=1}^n v_i^2}{\sum_{i=1}^n v_i^2} = \lambda_n$$

d'où $|||A|||_2 \leq \sqrt{\lambda_n}$, et d'autre part

$$\|Af_n\|_2^2 = \lambda_n,$$

d'où $|||A|||_2 \geq \sqrt{\lambda_n}$, ce qui conclut la démonstration. □

Remarque. Attention aux évidences trompeuses. Il ne faut pas confondre les normes subordonnées $|||A|||_p$ et les "normes p" sur l'espace vectoriel des matrices $M_n(\mathbb{R})$ (de dimension n^2, donc isomorphe à $\mathbb{R}^{n^2}$). On voit ici que l'on a $|||A|||_1 \neq \sum_{ij} |a_{ij}|$, $|||A|||_\infty \neq \max_{ij} |a_{ij}|$ et $|||A|||_2 \neq \left(\sum_{ij} |a_{ij}|^2\right)^{1/2}$... □

1.4 Rappel sur les EDO : le théorème de Cauchy Lipschitz

L'exemple (1.3) de l'équation de Poisson en dimension un d'espace, qui contient une EDO (équation différentielle ordinaire), nous conduit à faire des rappels sur l'important théorème de Cauchy-Lipschitz pour des EDO plus générales. Nous rappelons son énoncé dans le cas où u est un vecteur de $\mathbb{R}^n$, et F continue Lipschitzienne, même si on peut le poser dans un contexte plus général (espaces de Banach) et avec des hypothèses plus faibles (F localement Lipschitzienne).

Théorème 1.4.1 *Soit $I = [t_1, t_2]$ un intervalle compact d'intérieur non vide de $\mathbb{R}$ et $F \in \mathcal{C}^0(I \times \mathbb{R}^n, \mathbb{R}^n)$ une fonction continue Lipschitzienne par rapport à la deuxième variable :*

$$\forall t \in I, \forall u, v \in \mathbb{R}^n, \|F(t,v) - F(t,u)\| \leq L\|v - u\|$$

Alors $\forall t_0 \in I, \forall u_0 \in \mathbb{R}^n$, il existe une unique fonction $u \in \mathcal{C}^1(I, \mathbb{R}^n)$ qui vérifie

$$\begin{cases} u'(t) = F(t, u(t))) \quad \textit{dans }]0,1[, \\ \\ u(t_0) = u_0. \end{cases} \tag{1.8}$$

La démonstration utilise le théorème du point fixe pour les applications contractantes.

Rappelons aussi qu'on peut ramener une EDO d'ordre p

$$u^{(p)}(t) = F(t, u, u', \ldots u^{(p-1)})$$

avec $u \in \mathcal{C}^p(I, \mathbb{R}^n)$, à une EDO du premier ordre en introduisant un nouveau vecteur inconnu $U = (u, u', \ldots u^{(p-1)})^T$ et en travaillant dans $\mathbb{R}^{np}$. Pour le problème de Cauchy, il faut se donner une condition initiale sur U c'est à dire les p valeurs $U(t_0) = (u, u', \ldots u^{(p-1)})^T(t_0)$ de u et de ses dérivées d'ordre $\leq p-1$ en t_0.

Dans l'exemple que nous allons étudier ci-dessous, l'EDO s'écrit $-u''(x) + c(x)u(x) = f(x)$ dans $I = [0,1]$), on a $p = 2$, la fonction correspondante à la notation ci-dessus est $F(x, u, u') = c(x)u - f(x)$ (F ne dépend pas de la dernière variable) et le problème de Cauchy suppose de se donner $u(0)$ et $u'(0)$. Dans le paragraphe suivant, pour étudier l'existence de solutions de l'équation de Poisson en dimension un d'espace, nous allons effectivement utiliser le théorème 1.4.1 en l'appliquant à un problème de Cauchy (1.4). Cependant, notons tout de suite qu'une autre approche sera développée ultérieurement pour étudier (1.3), puisqu'il s'agit d'un *problème aux limites* et non d'un problème de *condition initiale* (ou problème de Cauchy).

1.5 Problèmes aux limites du second ordre en dimension 1

On s'intéresse à l'existence et unicité de la solution d'un problème aux limites. On va considérer un problème modèle légèrement plus général que le problème du fil élastique (1.3). Les données du problème sont deux fonctions f et c continues sur l'intervalle $[0,1]$ fermé et deux nombres réels α et β. On va chercher à résoudre le problème aux limites suivant : trouver $u \in \mathcal{C}^0([0,1]) \cap \mathcal{C}^2(]0,1[)$ tel que

$$\text{(P)} \qquad \begin{cases} -u''(x) + c(x)u(x) = f(x) & \text{dans }]0,1[, \\ \\ u(0) = \alpha, \quad u(1) = \beta. \end{cases}$$

Remarque. Un petit commentaire sur les notations : u appartient à $\mathcal{C}^0([0,1]) \cap \mathcal{C}^2(]0,1[)$ signifie que u est une fonction continue sur $[0,1]$ fermé, dont la restriction à $]0,1[$ ouvert est deux fois continûment dérivable. La première condition est indispensable pour qu'imposer des conditions aux limites de Dirichlet non homogènes comme celles de la deuxième ligne du problème (P) ait un sens. La deuxième condition nous autorise à parler de dérivée seconde de u dans l'intervalle ouvert, ce qui donne un sens à la première ligne de (P). □

ATTENTION ! Contrairement au cas des EDO, il se peut que ce problème n'ait aucune solution. Ainsi, par exemple, si $c(x) = -\pi^2$, $f(x) = 1$ et $\alpha = \beta = 0$, il n'existe pas de solution au problème (P). Pour le montrer, on suppose qu'il en existe une, $u \in \mathcal{C}^0([0,1]) \cap \mathcal{C}^2(]0,1[)$, donc telle que $-u'' - \pi^2 u = 1$ dans $]0,1[$ et $u(0) = u(1) = 0$. Remarquons tout de suite que, comme $u'' = -1 - \pi^2 u$, la fonction u est en fait de classe $\mathcal{C}^2([0,1])$ (*i.e.*, sur l'intervalle fermé, et pas seulement sur l'intervalle ouvert), nous reviendrons plus en détail sur ce point plus loin. On multiplie d'abord l'équation différentielle par $\sin \pi x$, ce qui donne

$$-u''(x)\sin \pi x - \pi^2 u(x)\sin \pi x = \sin \pi x \quad \text{dans }]0,1[,$$

puis on intègre l'égalité résultante entre 0 et 1. Il vient

$$-\int_0^1 u''(x)\sin \pi x\, dx - \pi^2 \int_0^1 u(x)\sin \pi x\, dx = \int_0^1 \sin \pi x\, dx = \frac{2}{\pi}.$$

D'un autre côté, en intégrant deux fois par parties la première intégrale, ce que l'on a le droit de faire puisque $u \in \mathcal{C}^2([0,1])$, on voit que

$$-\int_0^1 u''(x)\sin \pi x\, dx = -[u'(x)\sin \pi x]_0^1 + \pi \int_0^1 u'(x)\cos \pi x\, dx.$$

Comme $\sin 0 = \sin \pi = 0$, *i.e.*, la fonction $\sin \pi x$ satisfait les conditions aux limites, le terme tout intégré s'en va. On intègre par parties une seconde fois et l'on trouve

$$-\int_0^1 u''(x)\sin \pi x\, dx = [u(x)\cos \pi x]_0^1 + \pi^2 \int_0^1 u(x)\sin \pi x\, dx = \pi^2 \int_0^1 u(x)\sin \pi x\, dx$$

puisque la fonction u satisfait les conditions aux limites. On a donc obtenu l'égalité $0 = 2/\pi$, qui est absurde. Il n'existe par conséquent pas de solution u à ce problème.

Moralité : on n'aura pas existence de solution aux problèmes aux limites pour n'importe quelles données. On remarque la différence frappante avec le problème de Cauchy pour l'EDO, lequel a toujours une solution unique pour n'importe quel choix de c et f continues.

On va néanmoins utiliser les résultats connus sur le problème de Cauchy pour donner un résultat d'existence et d'unicité pour le problème aux limites dans le cas où la fonction c ne prend que des valeurs positives (il est possible de faire mieux).

Théorème 1.5.1 *Pour toutes $f, c \in \mathcal{C}^0([0,1])$ avec $c \geq 0$ et pour tous $\alpha, \beta \in \mathbb{R}$, il existe une solution du problème (P) et une seule. Cette solution appartient à $\mathcal{C}^2([0,1])$.*

Démonstration. Remarquons pour commencer que si u est solution de (P) alors nécessairement, $u \in \mathcal{C}^2([0,1])$, *i.e.*, les dérivées première et seconde de u se prolongent par continuité à $[0,1]$ fermé. En effet, comme $u''(x) = c(x)u(x) - f(x)$ pour $x \in]0,1[$ et que c, f et u sont continues sur $[0,1]$ fermé par hypothèse, u'' se prolonge par continuité en 0 et en 1, donc est bornée sur $[0,1]$, et par conséquent en intégrant u'' à partir de $x = 1/2$ par exemple, il en va de même pour u'.

Démontrons d'abord l'*unicité.* On suppose donc que l'on a deux solutions du problème (P), u_1 et u_2. Posant $v = u_1 - u_2$, on voit facilement que v est solution du problème aux limites :

$$\begin{cases} -v''(x) + c(x)v(x) = 0 \quad \text{dans }]0,1[, \\ v(0) = v(1) = 0. \end{cases}$$

Multiplions l'équation par v et intégrons le résultat sur $]0,1[$ (un procédé qui sera utilisé de nombreuses fois dans la suite). Il vient

$$-\int_0^1 v''(x)v(x)\,dx + \int_0^1 c(x)v(x)^2\,dx = 0.$$

Intégrant la première intégrale par parties, on obtient donc

$$-[v'v]_0^1 + \int_0^1 \left(v'(x)^2\,dx + c(x)v(x)^2\right)dx = 0.$$

Comme v s'annule en $x = 0$ et en $x = 1$, $[v'v]_0^1 = 0$. La fonction $v'(x)^2 + c(x)v(x)^2$ est continue sur $[0,1]$, positive puisque c est supposée positive *par hypothèse* et d'intégrale nulle d'après le calcul qui précède. Elle est donc identiquement nulle. C'est une somme de deux termes positifs, donc chacun de ces termes est nul. En particulier $v'(x) = 0$, donc v est constante, ce qui implique que $v(x) = v(0) = 0$ pour tout $x \in [0,1]$. En d'autres termes, $u_1 = u_2$ et il y a unicité.

Remarque. On déduit aussi du calcul ci-dessus que $c(x)v(x)^2 = 0$ dans $[0,1]$. Ceci n'implique pas directement que $v = 0$ car les hypothèses n'empêchent pas c de s'annuler sur un ensemble plus ou moins grand. Il faut passer par v' et la condition aux limites pour pouvoir conclure. On voit aussi que si c prend des valeurs strictement négatives, le raisonnement ne marche plus car la fonction intégrée n'a plus de raison d'être positive et peut donc être d'intégrale nulle sans être identiquement nulle. □

Pour l'*existence*, on utilise la *méthode de tir*. Le théorème 1.4.1 de Cauchy-Lipschitz nous assure que pour tout $\alpha \in \mathbb{R}$ et tout $\lambda \in \mathbb{R}$, il existe un unique $u_\lambda \in \mathcal{C}^2([0,1])$ solution du problème de Cauchy

$$\begin{cases} -u_\lambda''(x) + c(x)u_\lambda(x) = f(x) \quad \text{dans }]0,1[, \\ u_\lambda(0) = \alpha, \quad u_\lambda'(0) = \lambda. \end{cases}$$

En effet, on peut écrire cette EDO sous forme d'un système d'EDO linéaire du premier ordre dans $\mathbb{R}^2$

$$\begin{cases} u'_\lambda(x) = -w_\lambda(x) \\ w'_\lambda(x) = -c(x)u_\lambda(x) + f(x) \end{cases}$$

soit

$$\begin{pmatrix} u_\lambda \\ w_\lambda \end{pmatrix}'(x) = A(x)\begin{pmatrix} u_\lambda \\ w_\lambda \end{pmatrix}(x) + \begin{pmatrix} 0 \\ f(x) \end{pmatrix}$$

avec la matrice $A(x) = \begin{pmatrix} 0 & -1 \\ -c(x) & 0 \end{pmatrix}$ et la condition initiale $(u_\lambda, w_\lambda)(0) = (\alpha, -\lambda)$. Notons $U = (u,w)^T$, $B(x) = (0, f(x))^T$ et $G(x,U) = A(x)U + B(x)$. La fonction G est bien Lipschitzienne par rapport à la variable U

$$\|G(x,U) - G(x,V)\| \leq |||A(x)||| \, \|U - V\|$$

pour tout norme matricielle subordonnée, et, par exemple, si on prend la norme 1, $|||A(x)||| \leq max(1, |c(x)|)$ est borné uniformément en x car c est supposée continue sur le compact $[0,1]$.

On aura manifestement construit une solution du problème (P) si l'on réussit à trouver un nombre λ tel que $u_\lambda(1) = \beta$. Comme ceci doit être possible pour tout $\beta \in \mathbb{R}$ donné, on se ramène donc à prouver que l'application

$$\begin{cases} g : \mathbb{R} \to \mathbb{R}, \\ \lambda \longmapsto u_\lambda(1), \end{cases}$$

est *surjective*. On va en fait montrer que l'application g est *affine*, c'est-à-dire qu'il existe deux constantes a et b telles que $g(\lambda) = a\lambda + b$, avec $a \neq 0$ pour la surjectivité. Si cela est vérifié, on aura $g(0) = b$ c'est à dire $u_0(1) = b$, la notation u_0 signifie que u_0 est la solution de $-u''_0 + cu_0 = f$ avec $u_0(0) = \alpha$ et $u'_0(0) = 0$.

Posons donc $b = u_0(1)$, et $v_\lambda(x) = u_\lambda(x) - u_0(x)$. Alors on a

$$\begin{cases} -v''_\lambda(x) + c(x)v_\lambda(x) = 0 \quad \text{dans }]0,1[, \\ v_\lambda(0) = 0, \quad v'_\lambda(0) = \lambda, \end{cases}$$

si bien que si l'on introduit la fonction

$$\begin{cases} l : \mathbb{R} \to \mathbb{R}, \\ \lambda \longmapsto v_\lambda(1), \end{cases}$$

on a visiblement

$$g(\lambda) = l(\lambda) + b$$

et il suffit de montrer que l est *linéaire* et *surjective*.

Soient donc deux nombres λ_1 et λ_2. La fonction $v_{\lambda_1} + v_{\lambda_2}$ satisfait clairement

$$\begin{cases} -(v_{\lambda_1} + v_{\lambda_2})''(x) + c(x)(v_{\lambda_1} + v_{\lambda_2})(x) = 0 \quad \text{dans }]0,1[, \\ (v_{\lambda_1} + v_{\lambda_2})(0) = 0, \quad (v_{\lambda_1} + v_{\lambda_2})'(0) = \lambda_1 + \lambda_2, \end{cases}$$

c'est-à-dire le problème de Cauchy correspondant aux données initiales 0 pour la fonction et $\lambda_1 + \lambda_2$ pour sa dérivée. Or le théorème de Cauchy-Lipschitz assure l'*unicité* de la solution de

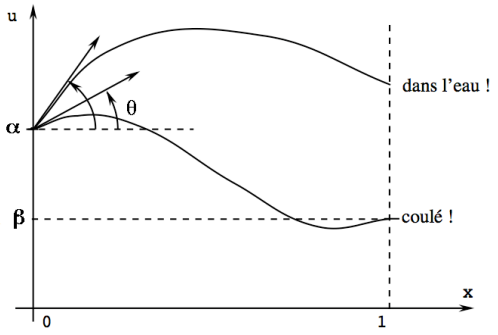


FIGURE 1.3 – Il suffit de bien viser

ce problème de Cauchy, que nous avions précédemment notée $v_{\lambda_1+\lambda_2}$. On a donc $v_{\lambda_1+\lambda_2}(x) = v_{\lambda_1}(x)+v_{\lambda_2}(x)$ pour tout $x \in [0,1]$. On montre de la même façon que $v_{\mu\lambda_1} = \mu v_{\lambda_1}$. En particulier, en prenant $x = 1$, on obtient

$$l(\lambda_1+\lambda_2) = l(\lambda_1)+l(\lambda_2) \quad \text{et} \quad l(\mu\lambda_1) = \mu l(\lambda_1),$$

i.e., l est linéaire de $\mathbb{R}$ dans $\mathbb{R}$.

Une application linéaire d'un espace vectoriel de dimension finie dans lui-même est surjective si et seulement si elle est injective. Il suffit donc de montrer que l est injective, c'est-à-dire, puisque c'est une application linéaire, montrer que son noyau $\ker l = \{\lambda \in \mathbb{R}, l(\lambda) = 0\} = \{\lambda \in \mathbb{R}, v_\lambda(1) = 0\}$ est réduit au vecteur nul. Or, si $\lambda \in \ker l$, il vient que

$$\begin{cases} -v_\lambda''(x)+c(x)v_\lambda(x) = 0 \quad \text{dans }]0,1[, \\ v_\lambda(0) = v_\lambda(1) = 0. \end{cases}$$

Comme $c \geq 0$, le même calcul que pour l'unicité montre donc que $v_\lambda(x) = 0$, ce qui implique bien sûr que $v_\lambda'(x) = 0$. Or, par définition de v_λ, on a $\lambda = v_\lambda'(0)$, par conséquent, on a montré que $\lambda = 0$, *i.e.*, $\ker l = \{0\}$, l est injective, donc surjective. □

Remarques. i) L'hypothèse $c \geq 0$ ne joue aucun rôle dans l'existence d'une solution u_λ au *problème de Cauchy* par le théorème de Cauchy-Lipschitz. Par contre, elle joue un rôle crucial dans l'unicité (qui implique ici l'existence) pour le *problème aux limites*. L'exemple $c(x) = -\pi^2$ montre d'ailleurs que l'on ne peut se passer de faire des hypothèses restrictives sur c pour pouvoir espérer un théorème d'existence (cependant l'hypothèse $c \geq 0$ n'est pas optimale).

ii) L'origine de la terminologie "méthode de tir" est claire (voir la figure 1.3) : on part de la valeur $u(0) = \alpha$ et on cherche à ajuster "l'angle de tir" (noté θ dans la figure1.3, dont la tangente est λ) pour toucher la valeur cible β en $x = 1$, la trajectoire de notre projectile fictif entre les deux points étant gouvernée par l'équation différentielle du problème (P).

Malheureusement, cette méthode est strictement monodimensionnelle et n'a aucun analogue en dimension supérieure. Les théories d'existence sont dans ce cas beaucoup plus sophistiquées et sortent largement du cadre de ce cours ; nous en donnerons simplement un exemple (dans un cas simplifié) dans le chapître suivant pour le problème aux limites en dimension 2 d'espace.

iii) La méthode s'applique également à des problèmes plus compliqués, par exemple

$$\begin{cases} u''''(x)-(a(x)u'(x))'+c(x)u(x) = f(x) \quad \text{dans}]0,1[, \\ u(0) = u(1) = u'(0) = u'(1) = 0, \end{cases}$$

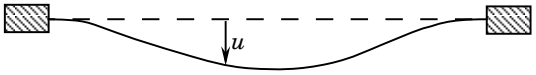


FIGURE 1.4 – Poutre encastrée en flexion

qui modélise l'équilibre d'une poutre élastique en flexion encastrée à ses extrémités sous l'action d'une force f (schématisé dans la figure 1.4). Ici, on prendra $a \geq 0$ et $c \geq 0$.

iv) On a montré l'existence et l'unicité d'une solution au problème aux limites, mais, sauf pour des données c et f très particulières, on n'a pas de formule "analytique" permettant d'évaluer numériquement $u(x)$, d'où le besoin de définir des méthodes d'approximation. □

On va maintenant montrer que si on suppose que les données f et c sont plus régulières, alors la solution u du problème aux limites est aussi plus régulière (la régularité des données se propage en un certain sens à la solution du problème). On suppose encore $c \geq 0$, condition qui nous permet d'utiliser le théorème 1.5.1.

Théorème 1.5.2 *Si f et c sont de classe $\mathcal{C}^m([0,1])$ avec $m \in \mathbb{N}$, alors u appartient à $\mathcal{C}^{m+2}([0,1])$.*

Démonstration. Par récurrence sur m.

Pour $m = 0$, le résultat est inclus dans le théorème 1.5.1.

Supposons donc le résultat vrai pour $m-1$, $m \geq 1$. Si f et c sont de classe $\mathcal{C}^m([0,1])$, alors, *a fortiori*, $f, c \in \mathcal{C}^{m-1}([0,1])$. Par hypothèse de récurrence, on a donc $u \in \mathcal{C}^{(m-1)+2}([0,1]) = \mathcal{C}^{m+1}([0,1]) \subset \mathcal{C}^m([0,1])$. Par conséquent, $u'' = cu - f \in \mathcal{C}^m([0,1])$ (car c'est une algèbre), c'est-à-dire $u \in \mathcal{C}^{m+2}([0,1])$. □

Remarque. Le résultat est correct et se démontre très simplement en dimension 1. Il est faux tel quel en dimension supérieure dans les espaces $\mathcal{C}^m(\bar{\Omega})$ (par exemple, il existe des fonctions u sur un ouvert Ω de $\mathbb{R}^2$ telles que $\Delta u \in \mathcal{C}^0(\bar{\Omega})$ mais $u \notin \mathcal{C}^2(\bar{\Omega})$). Par contre, on a des théorèmes de régularité analogues à celui de la dimension 1, mais beaucoup plus difficiles à montrer et valables dans des espaces plus compliqués comme les espaces de fonctions höldériennes $\mathcal{C}^{m,\alpha}(\bar{\Omega})$, $0 < \alpha < 1$, ou les espaces de Sobolev $H^m(\Omega)$. □

Donnons maintenant quelques résultats que l'on regroupe sous le terme générique de "principe du maximum".

Théorème 1.5.3 *Supposons que $f \geq 0$ dans $[0,1]$, $\alpha \geq 0$ et $\beta \geq 0$. Alors la solution u de* (P) *est telle que $u(x) \geq 0$ pour tout $x \in [0,1]$.*

Démonstration. On raisonne par l'absurde. Supposons qu'il existe un point $x_0 \in [0,1]$ tel que $u(x_0) < 0$. Comme $u(0) = \alpha \geq 0$ et $u(1) = \beta \geq 0$, on a nécessairement $x_0 \in]0,1[$ et comme u est continue, il existe un intervalle $[a,b] \subset [0,1]$ contenant x_0 tel que $u(x) \leq 0$ sur $[a,b]$ et $u(a) = u(b) = 0$ (ceci par le théorème des valeurs intermédiaires). Sur cet intervalle, on a donc $u''(x) = c(x)u(x) - f(x) \leq 0$, puisque $c \geq 0$ par hypothèse. En d'autres termes, u est *concave* sur cet intervalle, c'est-à-dire $u(ta + (1-t)b) \geq tu(a) + (1-t)u(b) = 0$ pour tout $t \in [0,1]$. En particulier, pour $t = (b - x_0)/(b-a)$, qui appartient bien à $[0,1]$ car $x_0 \in [a,b]$, il vient $u(x_0) \geq 0$, ce qui contredit l'hypothèse $u(x_0) < 0$. □

Remarque. Il s'agit d'un résultat de dépendance *monotone* de la solution u par rapport aux données du problème f, α, β. Il a une interprétation physique très agréable : si on tire sur le fil élastique vers le haut (ou vers le bas), le fil se déplace vers le haut (resp. vers le bas). □

On déduit du théorème précédent une borne *a priori* pour la solution du problème (P). Attention, dans le résultat qui suit, il n'y a aucune hypothèse de signe sur α, β et f.

Théorème 1.5.4 *Supposons qu'il existe* $\eta > 0$ *tel que* $c(x) \geq \eta$ *dans* $[0,1]$. *Alors la solution u de* (P) *est telle que*

$$\max_{x\in[0,1]} |u(x)| \leq \max\left\{ |\alpha|, |\beta|, \eta^{-1} \max_{x\in[0,1]} |f(x)| \right\}.$$

Démonstration. Posons $\Gamma = \max\left\{|\alpha|, |\beta|, \eta^{-1} \max_{x\in[0,1]} |f(x)|\right\}$.

i) Soit $v^-(x) = u(x) - \Gamma$. Alors $v^-(0) = u(0) - \Gamma = \alpha - \Gamma \leq \alpha - |\alpha| \leq 0$. De même, $v^-(1) = \beta - \Gamma \leq 0$. Enfin,

$$v^{-\prime\prime}(x) + c(x)v^-(x) = f(x) - \Gamma c(x) \leq f(x) - \Gamma\eta \leq f(x) - \max_{x\in[0,1]} |f(x)| \leq 0.$$

Le théorème 1.5.3 nous donne donc $v^-(x) \leq 0$ dans $[0,1]$, c'est-à-dire $u(x) \leq \Gamma$ dans $[0,1]$.

ii) Soit $v^+(x) = u(x) + \Gamma$. On voit de la même façon que $v^+(0) \geq 0$, $v^+(1) \geq 0$ et

$$v^{+\prime\prime}(x) + c(x)v^+(x) \geq 0.$$

Par conséquent, $v^+(x) \geq 0$ dans $[0,1]$, c'est-à-dire $u(x) \geq -\Gamma$ dans $[0,1]$. □

Chapitre 2

La méthode des différences finies

2.1 Principe de la méthode

La méthode des différences finies est celle, parmi les méthodes d'approximation des problèmes aux limites, qui ressemble le plus aux schémas numériques utilisés pour approcher les solutions des équations différentielles ordinaires. L'idée de base est identique : il s'agit de remplacer les dérivées qui apparaissent dans l'équation par des quotients différentiels appropriés, d'où le nom de la méthode, *différences finies* par opposition à des différences "infinitésimales" qui correspondraient aux dérivées elles-mêmes. Dans cette méthode, ce que l'on calcule effectivement n'est pas une fonction définie sur l'intervalle sur lequel on travaille, mais des approximations des valeurs que prend la solution du problème aux limites en un nombre *fini* de points de cet intervalle (ce qui est nécessaire si on veut pouvoir implémenter les algorithmes correspondants). On est bien entendu libre ensuite d'interpoler les valeurs approchées ainsi obtenues pour construire une fonction (on peut construire par exemple une fonction continue linéaire par morceaux ou une fonction plus régulière en utilisant des splines cubiques) et dessiner un beau graphe par exemple, mais il ne s'agit plus de la méthode elle-même, tout au plus d'un "post-processing" du résultat de celle-ci.

On commence donc par introduire une *grille de discrétisation uniforme* en se donnant un entier $N \geq 1$ et en posant $h = \frac{1}{N+1}$ et $x_i = ih$ pour $i = 0, 1, \ldots, N+1$, de telle sorte que les points x_i sont uniformément espacés entre eux du *pas* h, *i.e.*, $x_{i+1} - x_i = h$, avec $x_0 = 0$ et $x_{N+1} = 1$. Il y a donc N points de la grille dans l'intérieur de l'intervalle, correspondant aux indices $i = 1, \ldots, N$, voir la figure 2.1 ci-dessous. Dans la suite, on fera tendre N vers l'infini, ce qui est équivalent à faire tendre h vers 0. On va calculer des valeurs numériques notées $\bar{u}_i$, $i = 1, \ldots, N$, qui seront des approximations des valeurs exactes $u_i = u(x_i)$ d'autant meilleures que N est grand ou encore h est petit, ce que l'on démontrera ultérieurement, en supposant $\bar{u}_0 = \alpha, \bar{u}_{N+1} = \beta$ (dans la figure 2.1 on a illustré le cas $\alpha = \beta = 0$). L'idée heuristique est que, les dérivées étant par définition des limites de quotients différentiels, on ne devrait pas commettre une trop grande erreur en remplaçant celles-ci par de tels quotients, appelés traditionnellement *différences finies*.

Si φ est une fonction assez régulière sur $[0, 1]$, on note $\varphi_i = \varphi(x_i)$. Ainsi, on peut approcher la dérivée de φ en x_i, en supposant $x_{i\pm1} - x_i = \pm h$ "assez petit", par la *différence finie* décentrée

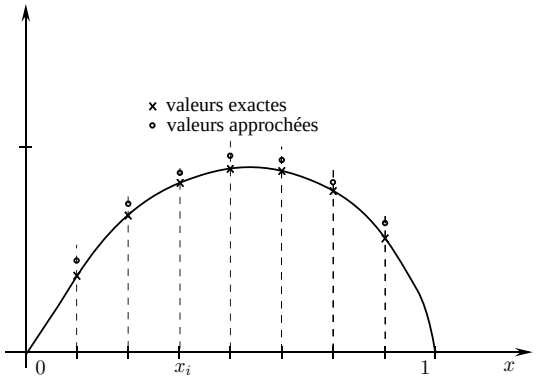


FIGURE 2.1 – Idée de la méthode

à droite

$$\varphi'(x_i) \approx \frac{\varphi(x_{i+1}) - \varphi(x_i)}{x_{i+1} - x_i} = \frac{\varphi_{i+1} - \varphi_i}{h},$$

ou bien par la différence finie décentrée à gauche

$$\varphi'(x_i) \approx \frac{\varphi(x_i) - \varphi(x_{i-1})}{x_i - x_{i-1}} = \frac{\varphi_i - \varphi_{i-1}}{h}.$$

Combinant ces deux approximations entre elles, on voit apparaître pour la dérivée seconde

$$\varphi''(x_i) \approx \frac{\varphi'(x_{i+1}) - \varphi'(x_i)}{x_{i+1} - x_i} \approx \frac{\frac{\varphi_{i+1}-\varphi_i}{h} - \frac{\varphi_i - \varphi_{i-1}}{h}}{h} = \frac{\varphi_{i+1} - 2\varphi_i + \varphi_{i-1}}{h^2}.$$

Naturellement, ici le signe $\approx$ n'a aucun sens précis. Il indique simplement une façon *a priori* raisonnable d'approcher la dérivée seconde d'une fonction en un point de la grille quand on connaît ses valeurs ponctuelles, ou des approximations de celles-ci, aux points voisins de cette même grille. On peut préciser les choses.

Théorème 2.1.1 *Soit $\varphi \in \mathcal{C}^4([0,1])$. Pour tout $i \in \{1,\dots,N\}$, il existe un nombre θ_i, avec $|\theta_i| < 1$ tel que*

$$-\varphi''(x_i) = \frac{-\varphi(x_{i+1}) + 2\varphi(x_i) - \varphi(x_{i-1})}{h^2} + \frac{h^2}{12}\varphi^{(4)}(x_i + \theta_i h).$$

Démonstration. Comme toujours pour ce type de résultats, la démonstration utilise la formule de Taylor-Lagrange. Comme φ est supposée de classe $\mathcal{C}^4$ sur $[0,1]$, on peut utiliser cette dernière jusqu'à l'ordre 4 en tout point de la grille. En particulier, pour tout $i \in \{1,\dots,N\}$, il existe un nombre $\theta_i^+ \in]0,1[$ tel que

$$\varphi(x_{i+1}) = \varphi(x_i) + h\varphi'(x_i) + \frac{h^2}{2}\varphi''(x_i) + \frac{h^3}{6}\varphi'''(x_i) + \frac{h^4}{24}\varphi^{(4)}(x_i + \theta_i^+ h).$$

De même, il existe $\theta_i^- \in]0,1[$ tel que

$$\varphi(x_{i-1}) = \varphi(x_i) - h\varphi'(x_i) + \frac{h^2}{2}\varphi''(x_i) - \frac{h^3}{6}\varphi'''(x_i) + \frac{h^4}{24}\varphi^{(4)}(x_i - \theta_i^- h).$$

Additionnant ces deux relations entre elles, il vient

$$\varphi(x_{i+1})+\varphi(x_{i-1}) = 2\varphi(x_i)+h^2\varphi''(x_i)+\frac{h^4}{24}\big(\varphi^{(4)}(x_i+\theta_i^+h)+\varphi^{(4)}(x_i-\theta_i^-h)\big).$$

Comme $\varphi^{(4)}$ est continue par hypothèse, le théorème des valeurs intermédiaires nous dit qu'il existe $y_i \in [x_i-\theta_i^-h, x_i+\theta_i^+h]$ tel que

$$\frac{1}{2}\big(\varphi^{(4)}(x_i+\theta_i^+h)+\varphi^{(4)}(x_i-\theta_i^-h)\big) = \varphi^{(4)}(y_i).$$

En effet, le terme de gauche est la moyenne des valeurs prises par $\varphi^{(4)}$ aux extrémités de l'intervalle $[x_i-\theta_i^-h, x_i+\theta_i^+h]$. Par conséquent, on voit que

$$-\varphi''(x_i) = \frac{-\varphi(x_{i+1})+2\varphi(x_i)-\varphi(x_{i-1})}{h^2}+\frac{h^2}{12}\varphi^{(4)}(y_i).$$

Pour conclure, on remarque que $y_i \in [x_i-\theta_i^-h, x_i+\theta_i^+h] \subset]x_{i-1}, x_{i+1}[$, donc $\theta_i = \frac{y_i-x_i}{h}$ est tel que $|\theta_i|<1$ et trivialement $y_i = x_i+\theta_i h$. □

Corollaire 2.1.2 *Sous les hypothèses du théorème 2.1.1, on a*

$$\max_{1\le i\le N}\left|-\varphi''(x_i)-\frac{-\varphi(x_{i+1})+2\varphi(x_i)-\varphi(x_{i-1})}{h^2}\right| \le \frac{h^2}{12}\max_{x\in[0,1]}|\varphi^{(4)}(x)|. \tag{2.1}$$

Démonstration. Immédiat d'après le théorème 2.1.1. □

Remarques. i) La quantité au membre de gauche de (2.1) s'appelle *erreur de consistance* de la méthode (vecteur de $\mathbb{R}^N$ évalué en norme $\|.\|_\infty$), comme pour les schémas d'approximation des EDO.

ii) Si φ est un polynôme de degré inférieur ou égal à 3, l'erreur de consistance est nulle.

iii) Si l'on suppose seulement φ de classe $\mathcal{C}^3$, on peut uniquement dire que l'erreur de consistance est en $O(h)$ car alors la formule de Taylor-Lagrange n'est valable que jusqu'à l'ordre trois. De même, si φ est seulement $\mathcal{C}^2$, alors l'erreur de consistance tend vers 0 quand h tend vers 0, mais pas plus *a priori*. □

Appliquons ces résultats à la solution u du problème aux limites (P). On note u_h le vecteur $(u_1,u_2,\ldots,u_N)^T$ de $\mathbb{R}^N$ de composantes $u_1=u(x_1)$, $u_2=u(x_2)$, …, $u_N=u(x_N)$ (attention à cette notation traditionnelle qui manque un peu de cohérence ; h et N sont liés par la relation $(N+1)h=1$ donc en particulier la dimension du vecteur $u_h=(u_1,\ldots,u_N)^T$ dépend de h).

Corollaire 2.1.3 *Supposons que la solution u du problème aux limites soit de classe $\mathcal{C}^4$ sur $[0,1]$. Il existe alors des points $y_i \in]x_{i-1},x_{i+1}[$ tels que le vecteur u_h est solution du système suivant :*

$$A_h u_h = b_h+\varepsilon_h(u), \tag{2.2}$$

avec

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2+c_1h^2 & -1 & 0 & \dots & \dots & \dots & 0 \\ -1 & 2+c_2h^2 & -1 & 0 & & & \\ 0 & -1 & \ddots & \ddots & & & \\ \vdots & & \ddots & \ddots & \ddots & & \\ \vdots & & 0 & -1 & 2+c_ih^2 & -1 & 0 \\ \vdots & & & & \ddots & \ddots & \ddots \\ \vdots & & & 0 & -1 & 2+c_{N-1}h^2 & -1 \\ 0 & & \dots & & 0 & -1 & 2+c_Nh^2 \end{pmatrix} \tag{2.3}$$

où $c_i = c(x_i)$,

$$b_h = \begin{pmatrix} f_1 + \frac{\alpha}{h^2} \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N + \frac{\beta}{h^2} \end{pmatrix} \tag{2.4}$$

où $f_i = f(x_i)$ *et*

$$\varepsilon_h(u) = -\frac{h^2}{12}\begin{pmatrix} u^{(4)}(y_1) \\ u^{(4)}(y_2) \\ \vdots \\ u^{(4)}(y_N) \end{pmatrix}. \tag{2.5}$$

Démonstration. C'est presque immédiat. En effet, en chaque point x_i, $1 \le i \le N$, on a par l'équation différentielle

$$-u''(x_i) + c(x_i)u(x_i) = f(x_i).$$

Il suffit de remplacer $-u''(x_i)$ par les expressions déduites du théorème 2.1.1. Il faut distinguer trois cas, suivant que $i = 1$, $2 \le i \le N-1$ ou $i = N$.

- Le cas $i = 1$. Dans ce cas, il vient

$$f(x_1) = -u''(x_1) + c(x_1)u(x_1) = \frac{-u(x_2) + 2u(x_1) - u(x_0)}{h^2} + c(x_1)u(x_1) + \frac{h^2}{12}u^{(4)}(y_1).$$

Comme $u(x_0) = u(0) = \alpha$ est connu par la condition aux limites, on le passe au second membre et l'on obtient donc

$$\frac{-u_2 + 2u_1}{h^2} + c_1u_1 = f_1 + \frac{u(x_0)}{h^2} - \frac{h^2}{12}u^{(4)}(y_1) = f_1 + \frac{\alpha}{h^2} - \frac{h^2}{12}u^{(4)}(y_1),$$

ou encore

$$\frac{1}{h^2}[(2+c_1h^2)u_1 - u_2] = f_1 + \frac{\alpha}{h^2} - \frac{h^2}{12}u^{(4)}(y_1).$$

- Le cas $i = N$. De façon analogue

$$f(x_N) = \frac{-u(x_{N+1}) + 2u(x_N) - u(x_{N-1})}{h^2} + c(x_N)u(x_N) + \frac{h^2}{12}u^{(4)}(y_N).$$

Comme $u(x_{N+1}) = u(1) = \beta$, on obtient donc

$$\frac{1}{h^2}[-u_{N-1} + (2 + c_N h^2)u_N] = f_N + \frac{\beta}{h^2} - \frac{h^2}{12}u^{(4)}(y_N).$$

- Le cas $2 \leq i \leq N-1$. Ici rien de spécial,

$$f(x_i) = \frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} + c(x_i)u(x_i) + \frac{h^2}{12}u^{(4)}(y_i),$$

est exactement la $i$ème ligne du système linéaire annoncé. □

Remarques. i) Les valeurs de la solution aux points de la grille satisfont exactement le système (2.2) : il n'y a aucune approximation ici. Bien sûr, il n'y a pas de miracle : on ne peut pas résoudre ce système pour calculer effectivement ces valeurs, puisque le second membre contient un terme $\varepsilon_h(u)$ inconnu ! Il sera néanmoins utile dans l'analyse de la convergence de la méthode.

ii) Par contre, la matrice A_h de dimension $N \times N$ et le vecteur b_h sont connus.

iii) On note $\|\cdot\|_{N,\infty}$ la norme ∞ sur $\mathbb{R}^N$ (définie dans la proposition 1.3.3) en faisant explicitement apparaître N, lequel est destiné à tendre vers l'infini. Alors $\|\varepsilon_h(u)\|_{N,\infty}$ est l'erreur de consistance calculée sur u et satisfait donc, en appliquant le corollaire 2.1.2,

$$\|\varepsilon_h(u)\|_{N,\infty} \leq \frac{h^2}{12} \max_{x\in[0,1]} |u^{(4)}(x)| \to 0 \text{ quand } N \to +\infty, \tag{2.6}$$

puisque $h = \frac{1}{N+1}$. □

Idée de la méthode des différences finies : Puisque $\varepsilon_h(u)$ est de toutes façons petit quand h est petit (supposant u suffisamment régulière, bien sûr), on décide de l'enlever du second membre du système linéaire et on considère alors le problème *discret* suivant

$$(S_h) \quad \begin{cases} \text{Trouver } \bar{u}_h \in \mathbb{R}^N \text{ tel que} \\ \\ A_h \bar{u}_h = b_h. \end{cases}$$

Il s'agit donc de résoudre un système de N équations linéaires à N inconnues qui sont les composantes du vecteur $\bar{u}_h$, de matrice A_h et de second membre b_h connus.

Plusieurs questions se posent :

1) La matrice A_h est elle inversible ? Si ce n'est pas le cas, on n'a aucune chance de calculer ce nouveau vecteur $\bar{u}_h$.

2) En supposant que ce soit le cas, en général on aura $\bar{u}_h \neq u_h$, *i.e.*, $(\bar{u}_h)_i \neq u(x_i)$, $u(x_i)$ étant une des valeurs qui nous intéressent, on commet une *erreur* qu'il faut pouvoir estimer (cette estimation est liée à ce qu'on appelle le *conditionnement* de la matrice A_h). A-t-on alors $\bar{u}_h - u_h \to 0$ en un sens raisonnable et à quelle vitesse en fonction de h (ou de N) ? En d'autres termes, a-t-on ainsi construit des approximations des valeurs de la solution aux points de la grille, et quelle est la qualité de ces approximations ?

Définition 2.1.1 *On dit que la méthode est*
i) convergente si

$$\max_{1\le i\le N} |(u_h)_i - (\bar{u}_h)_i| \to 0 \text{ quand } N \to +\infty.$$

ii) d'ordre p si

$$\max_{1\le i\le N} |(u_h)_i - (\bar{u}_h)_i| \le C(u)h^p,$$

où $C(u)$ est une constante qui ne dépend que de u.

Donc, si la méthode est convergente, alors on a bien obtenu des approximations (ici uniformes, mais on pourrait utiliser d'autres normes que la norme $\|\cdot\|_{N,\infty}$) des valeurs exactes et ces approximations convergent d'autant plus vite que la méthode est d'ordre élevé.

En résumé, il s'agit maintenant d'effectuer ce que l'on appelle l'*analyse numérique* de la méthode : est-elle bien définie, est-elle convergente et de quel ordre ?

Commençons par traiter la première question, à savoir est-il bien raisonnable de vouloir calculer le vecteur $\bar{u}_h$.

Théorème 2.1.4 *Si $c(x) \ge 0$ sur $[0,1]$, la matrice A_h est symétrique, définie positive, donc inversible.*

Démonstration. Il est évident que A_h est tridiagonale symétrique, quel que soit le signe de c. Supposons maintenant que $c \ge 0$. Soit $v \in \mathbb{R}^N \setminus \{0\}$. Nous devons évaluer $v^T A_h v$; on calcule

$$h^2 v^T A_h v = h^2 \sum_{i,j} a_{ij} v_i v_j = \sum_{i=1}^{N} (2 + c_i h^2) v_i^2 - 2 \sum_{i=1}^{N-1} v_i v_{i+1} \ge 2 \sum_{i=1}^{N} v_i^2 - 2 \sum_{i=1}^{N-1} v_i v_{i+1},$$

puisque $c_i = c(x_i) \ge 0$. Par conséquent, en réarrangeant ces deux dernières sommes, il vient

$$\begin{aligned} h^2 v^T A_h v \ &\ge v_1^2 + (v_1^2 - 2v_1v_2 + v_2^2) + (v_2^2 - 2v_2v_3 + v_3^2) \\ &\qquad + \cdots + (v_{N-1}^2 - 2v_{N-1}v_N + v_N^2) + v_N^2 \\ &= v_1^2 + (v_1 - v_2)^2 + (v_2 - v_3)^2 + \cdots + (v_{N-1} - v_N)^2 + v_N^2 \ge 0. \end{aligned}$$

La matrice A_h est donc positive. De plus, si $v^T A_h v = 0$, on voit que nécessairement

$$v_1 = 0, v_1 - v_2 = 0, v_2 - v_3 = 0, \ldots, v_{N-1} - v_N = 0 \text{ et } v_N = 0,$$

c'est-à-dire en fait $v = 0$. Elle est donc définie positive. On en déduit immédiatement qu'elle est inversible, car

$$v \in \ker A_h \Leftrightarrow A_h v = 0 \Rightarrow v^T A_h v = 0 \Leftrightarrow v = 0,$$

donc $\ker A_h = \{0\}$. □

Corollaire 2.1.5 *Pour tout f et pour tout N, il existe un unique vecteur $\bar{u}_h \in \mathbb{R}^N$ solution du problème aux différences finies (S_h).*

Remarque. Il est intéressant de noter que c'est la même hypothèse de signe sur c qui assure l'existence de la solution du problème discret et l'existence de la solution du problème aux limites lui-même. □

2.2 Étude de la convergence

Dans la suite, on suppose toujours que la solution u est de classe $\mathcal{C}^4$ sur $[0,1]$. On va mesurer l'écart entre la solution discrète $\bar{u}_h$ et les valeurs de la solution continue aux points de la grille u_h en utilisant la norme uniforme $\|\bar{u}_h - u_h\|_{N,\infty}$. C'est cette norme que nous avons utilisé dans la définition 2.1.1 de la convergence. La notation $\|.\|_{N,\infty}$, avec l'indice N met en valeur le fait que nous sommes dans l'espace vectoriel $\mathbb{R}^N$. Toutes les normes sur $\mathbb{R}^N$ sont équivalentes mais elles ne sont pas toutes bien adaptées à l'étude de la convergence, puisque N est amené à tendre vers l'infini ; notons que la norme $\|u_h\|_{N,\infty}$ garde bien un sens alors que, par exemple, $\|u_h\|_{N,1} \to \infty$ quand $N \to \infty$. De plus il est possible d'évaluer la norme $|||A_h^{-1}|||_{N,\infty}$ comme nous allons le voir, et c'est elle qui va nous permettre d'estimer l'erreur $\bar{u}_h - u_h$.

Proposition 2.2.1 *On a*

$$\|\bar{u}_h - u_h\|_{N,\infty} \leq |||A_h^{-1}|||_{N,\infty} \left(\frac{h^2}{12} \max_{[0,1]} |u^{(4)}(x)| \right).$$

Démonstration. Écrivons les systèmes linéaires respectivement satisfaits par $\bar{u}_h$ et u_h : par (2.2)

$$\begin{aligned} A_h \bar{u}_h &= b_h, \\ A_h u_h &= b_h + \varepsilon_h(u). \end{aligned}$$

Soustrayant ces deux relations entre vecteurs, on obtient

$$A_h(\bar{u}_h - u_h) = -\varepsilon_h(u) \Longleftrightarrow \bar{u}_h - u_h = -A_h^{-1}\varepsilon_h(u)$$

puisque A_h est inversible. Prenant les normes N,∞ de ces vecteurs, on obtient par définition des normes subordonnées

$$\|\bar{u}_h - u_h\|_{N,\infty} \leq |||A_h^{-1}|||_{N,\infty} \|\varepsilon_h(u)\|_{N,\infty}.$$

Le résultat se déduit alors immédiatement de (2.6) (par application du corollaire 2.1.2). □

L'étude de la convergence se ramène donc maintenant à étudier le comportement de la quantité $|||A_h^{-1}|||_{N,\infty}$ (qui ne dépend plus de u) en fonction de N ou h.

Notons que de façon semblable à ce que l'on fait dans l'étude des schémas d'approximation numérique pour les équations différentielles ordinaires, l'erreur est constituée de deux morceaux, d'une part l'erreur de consistance $\|\varepsilon_h(u)\|_{N,\infty}$, que l'on a déjà estimée, et d'autre part cette quantité $|||A_h^{-1}|||_{N,\infty}$ qui ne dépend pas de la solution et que l'on peut appeler constante de *stabilité*.

L'estimation de $|||A_h^{-1}|||_{N,\infty}$ est liée aux propriétés de la matrice A_h qui est issue de la modélisation d'un problème physique. La démonstration demande un certain nombre d'étapes. On commence par définir la notion de matrice *monotone*.

Définition 2.2.1 *i) On introduit une relation d'ordre partiel sur $\mathbb{R}^N$ en posant*

$$v \overrightarrow{\geq} 0 \text{ si et seulement si } \forall i,\ v_i \geq 0.$$

ii) De même pour les matrices, on dit que

$$A \overrightarrow{\geq} 0 \text{ si et seulement si } \forall i, j,\ A_{ij} \geq 0.$$

iii) On dit qu'une matrice A est monotone si elle est inversible et si $A^{-1} \overrightarrow{\geq} 0$.

Attention, la relation d'ordre sur les matrices ainsi définie n'a rien à voir avec celle définie sur les matrices symétriques à l'aide des formes quadratiques : il existe des matrices qui sont positives au sens des formes quadratiques mais pas positives au sens présent et inversement (en trouver des exemples).

Donnons une autre caractérisation de la monotonie d'une matrice, plus pratique que la définition.

Lemme 1 *Soit A une matrice $N \times N$. Elle est monotone si et seulement si quand on a un vecteur v tel que $Av \overrightarrow{\geq} 0$, alors cela implique que $v \overrightarrow{\geq} 0$.*

Démonstration. On procède par condition nécessaire et condition suffisante.

• Condition nécessaire. Soit A une matrice monotone. On se donne un vecteur $v \in \mathbb{R}^N$ tel que $Av \overrightarrow{\geq} 0$, c'est-à-dire $(Av)_i \geq 0$ pour tout indice $1 \leq i \leq N$. Naturellement, $v = A^{-1}(Av)$, ce qui se lit en composantes sous la forme

$$v_i = \sum_{j=1}^{N} (A^{-1})_{ij} (Av)_j.$$

Or $(A^{-1})_{ij} \geq 0$ puisque A est monotone, il vient donc $v_i \geq 0$, c'est-à-dire $v \overrightarrow{\geq} 0$.

• Condition suffisante. Soit A une matrice telle que $Av \overrightarrow{\geq} 0$ implique $v \overrightarrow{\geq} 0$. Montrons tout d'abord qu'elle est inversible. Pour cela, soit w un élément du noyau de A, donc tel que $Aw = 0$. Comme évidemment, $Aw = 0 \overrightarrow{\geq} 0$, on en déduit que $w \overrightarrow{\geq} 0$. De même, $-w$ appartient au noyau et donc $-w \overrightarrow{\geq} 0$. Par conséquent, $w_i = 0$ pour tout i et le noyau est réduit au vecteur nul.

Notons b_j le $j$ème vecteur-colonne de la matrice A^{-1}. Ceci signifie que $A^{-1} e_j = b_j$ où e_j est le $j$ème vecteur de base. En d'autres termes, $Ab_j = e_j$, avec bien sûr $e_j \overrightarrow{\geq} 0$. Par conséquent, on en déduit que $b_j \overrightarrow{\geq} 0$ pour tout j, ce qui implique immédiatement que $A^{-1} \overrightarrow{\geq} 0$. □

Continuons par une propriété qui nous sera utile dans une démonstration plus loin.

Lemme 2 *Soit A et B deux matrices monotones, avec $B \overrightarrow{\geq} A$. Alors*

$$A^{-1} \overrightarrow{\geq} B^{-1}$$

et

$$|||A^{-1}|||_{N,\infty} \geq |||B^{-1}|||_{N,\infty}.$$

Démonstration. On remarque que pour tout couple de matrices A et B inversibles, on a l'identité :

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

En effet, $A^{-1} - B^{-1} = A^{-1}(BB^{-1}) - (A^{-1}A)B^{-1}$. Si les matrices sont monotones, et si de plus $B - A \overrightarrow{\geq} 0$ ceci implique $A^{-1} - B^{-1} \overrightarrow{\geq} 0$ car un produit de matrices positives est visiblement positif (il suffit d'écrire la définition). On a donc montré que

$$A^{-1} \overrightarrow{\geq} B^{-1}$$

Considérons maintenant deux matrices telles que $B \overrightarrow{\geq} A \overrightarrow{\geq} 0$. Ceci signifie simplement que $b_{ij} \geq a_{ij} \geq 0$ pour tous i, j. Par conséquent,

$$|||B|||_{N,\infty} = \max_i \sum_j |b_{ij}| = \max_i \sum_j b_{ij} \geq \max_i \sum_j a_{ij} = \max_i \sum_j |a_{ij}| = |||A|||_{N,\infty}.$$

On applique alors ce résultat à $A^{-1} \overrightarrow{\geq} B^{-1} \overrightarrow{\geq} 0$. □

Revenons à la matrice A_h.

Lemme 3 *La matrice A_h est monotone.*

Démonstration. On utilise la caractérisation donnée dans le lemme 1. Soit $v \in \mathbb{R}^N$ tel que $A_h v \overrightarrow{\geq} 0$, c'est-à-dire

$$\begin{cases} (A_h v)_k = \frac{1}{h^2}(-v_{k-1} + (2 + c_k h^2) v_k - v_{k+1}) \geq 0 \text{ pour } 2 \leq k \leq N-1, \\ (A_h v)_1 = \frac{1}{h^2}((2 + c_1 h^2) v_1 - v_2) \geq 0, \\ (A_h v)_N = \frac{1}{h^2}(-v_{N-1} + (2 + c_N h^2) v_N) \geq 0. \end{cases}$$

On choisit un indice p de $1, \ldots N$ tel que $v_p \leq v_k$ pour tout $1 \leq k \leq N$ (il existe évidemment un tel p). Distinguons trois cas.

• Si $p = 1$, alors $v_2 \geq v_1$ ou encore $v_1 - v_2 \leq 0$ et, d'après la deuxième ligne des inégalités précédentes,

$$(1 + c_1 h^2) v_1 \geq (1 + c_1 h^2) v_1 + v_1 - v_2 = h^2 (A_h v)_1 \geq 0.$$

Par conséquent, en divisant par $1 + c_1 h^2 > 0$, on en déduit que $v_1 \geq 0$. Comme v_1 réalise le minimum de toutes les composantes de v, $v_k \geq v_1 \geq 0$ pour tout k.

• Si $p = N$, on fait le même raisonnement en utilisant la troisième ligne.

• Enfin, si $2 \leq p \leq N-1$, alors $v_p \leq v_{p-1}$ et $v_p \leq v_{p+1}$, donc

$$c_p h^2 v_p \geq c_p h^2 v_p + v_p - v_{p-1} + v_p - v_{p+1} = h^2 (A_h v)_p \geq 0.$$

On distingue deux sous-cas.

♢ Si $c_p > 0$, alors on divise par c_p et l'on obtient $v_p \geq 0$. On conclut comme précédemment.

♡ Si $c_p = 0$, alors l'inégalité ci-dessus donne

$$0 \geq v_p - v_{p-1} + v_p - v_{p+1} \geq 0,$$

ce qui, compte tenu du fait que $v_p - v_{p-1} \leq 0$ et $v_p - v_{p+1} \leq 0$ implique que $v_{p-1} = v_p = v_{p+1}$. Dans ce cas, on recommence le raisonnement avec v_{p-1} autant de fois que nécessaire (jusqu'à, si tous les c_i sont nuls, arriver à v_1 où l'on utilise le premier cas). □

Remarque. La monotonie de la matrice A_h est l'analogue discret de la monotonie du problème aux limites (cf. théorème 1.5.3). En effet, par définition (2.4) de b_h, si les hypothèses du théorème 1.5.3 sont satisfaites, alors $b_h \overrightarrow{\geq} 0$. Si $b_h \overrightarrow{\geq} 0$ et $A_h \bar{u}_h = b_h$, on en déduit que $\bar{u}_h \overrightarrow{\geq} 0$. On parle donc de *principe du maximum discret*. □

On peut maintenant donner le résultat de convergence.

Théorème 2.2.1 *(de Gerschgorin). Supposons que c et f sont de classe $\mathcal{C}^2$. Alors on a la majoration d'erreur :*

$$\max_{1\le i\le N} |(\bar{u}_h)_i - u(x_i)| \le \frac{h^2}{96} \max_{0\le x\le 1} |u^{(4)}(x)|.$$

Démonstration. D'après le théorème de régularité (théorème 1.5.2), si c et f sont $\mathcal{C}^2$, alors u est $\mathcal{C}^4$. On peut donc utiliser la proposition 2.2.1. Il nous reste alors à estimer la quantité $|||A_h^{-1}|||_{N,\infty}$. Introduisons la matrice

$$A_{0h} = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & & \\ 0 & -1 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & -1 & 2 & -1 \\ 0 & \dots & & 0 & -1 & 2 \end{pmatrix} \tag{2.7}$$

qui correspond au cas où $c = 0$. On sait déjà que $A_h^{-1} \overrightarrow{\ge} 0$ et que $A_{0h}^{-1} \overrightarrow{\ge} 0$. De plus,

$$A_h - A_{0h} = \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & c_N \end{pmatrix} \overrightarrow{\ge} 0,$$

car $c_i = c(x_i) \ge 0$. En appliquant maintenant le lemme 2, on obtient alors

$$A_{0h}^{-1} \overrightarrow{\ge} A_h^{-1} \overrightarrow{\ge} 0.$$

et aussi

$$|||A_h^{-1}|||_{N,\infty} \le |||A_{0h}^{-1}|||_{N,\infty},$$

et l'on s'est ramené à traiter le cas $c = 0$.

Il s'agit donc d'estimer la norme N,∞ d'une matrice A_{0h}^{-1} positive. Or, si $B \overrightarrow{\ge} 0$, il est facile de voir que $|||B|||_{N,\infty} = \|Be\|_{N,\infty}$ où e est le vecteur $e = (1,1,\dots,1)^T$. En effet, $(Be)_i = \sum_j b_{ij} = \sum_j |b_{ij}|$ puisque $b_{ij} \ge 0$. Soit donc

$$\bar{u}_{0h} = A_{0h}^{-1} e \Longleftrightarrow A_{0h}\bar{u}_{0h} = e.$$

Le vecteur $\bar{u}_{0h}$ n'est donc autre que la solution du problème aux différences finies associé au problème aux limites particulier

$$\begin{cases} -u_0''(x) = 1 \text{ dans }]0,1[, \\ u_0(0) = u_0(1) = 0. \end{cases}$$

Or il est facile de calculer la solution exacte de ce problème : c'est la fonction $u_0(x) = \frac{1}{2}x(1-x)$. Il se trouve que c'est un polynôme du second degré et, par conséquent $u_0^{(4)} = 0$, son erreur de

consistance est nulle, $\varepsilon_h(u_0) = 0$. On déduit alors du corollaire 2.1.3 que le vecteur $(u_0)_h$ associé est solution *du même* système linéaire que le vecteur $\bar{u}_{0h}$. D'où l'expression de celui-ci :

$$(\bar{u}_{0h})_i = u_0(x_i) = \frac{1}{2} ih(1-ih).$$

Dans tous les cas,

$$\|\bar{u}_{0h}\|_{N,\infty} \leq \max_{x\in[0,1]} \left\{ \frac{1}{2} x(1-x) \right\} = \frac{1}{8}.$$

On obtient le théorème en combinant toutes les estimations obtenues. □

Remarques. i) On a montré que la méthode des différences finies est convergente et d'ordre 2 (pour la norme N,∞). Les valeurs calculées se placent donc dans un voisinage du graphe de la solution exacte dont l'épaisseur est en $O(h^2)$.

ii) On peut montrer que la convergence n'est pas plus rapide en général, *i.e.*, il existe une donnée $f \in \mathcal{C}^2$ telle que les solutions exactes et approchées correspondantes vérifient l'estimation $\|\bar{u}_h - u_h\|_{N,\infty} = Ch^2(1+\delta(h))$ avec $\delta(h) \to 0$ quand $h \to 0$ et $C > 0$.

iii) L'estimation d'erreur dépend de u, qui est inconnue, par l'intermédiaire de sa dérivée quatrième. Elle n'est donc pas fonction explicite des données du problème, et ne donne pas d'indication quantitative sur l'erreur commise. On dit qu'il s'agit d'une estimation *a priori*.

iv) On a obtenu une estimation en norme $\|.\|_{N,\infty}$. On peut obtenir un résultat dans une autre norme. Rappelons que les normes sur un espace de dimension finie sont équivalentes, mais les constantes qui interviennent dans les inégalités entre les normes peuvent dépendre de la dimension de l'espace. □

Pour obtenir un résultat de convergence qu'on interprète en terme de norme L^2, on va utiliser une autre technique. La norme euclidienne de $\mathbb{R}^N$ n'est pas intéressante telle quelle car quand $h \to 0$, alors $N \to \infty$ et le nombre de termes de la somme augmente. Nous allons d'abord définir une *norme L^2 discrète* qui elle a un sens quand $h = \frac{1}{N+1} \to 0$.

Définition 2.2.2 *Pour tout $v \in \mathbb{R}^{N+2}$, $v = (v_0, ..., v_{N+1})^T$, on définit la norme L^2 discrète $\|.\|_{2,\Delta}$ par*

$$\|v\|_{2,\Delta}^2 = h(\frac{1}{2}v_0^2 + \sum_{i=1}^{N} v_i^2 + \frac{1}{2}v_{N+1}^2)$$

On vérife facilement que c'est bien une norme sur $\mathbb{R}^{N+2}$. C'est la norme dans $L^2(0,1)$ de la fonction constante par morceaux v_Δ égale à v_i sur l'intervalle $](i-1/2)h, i+1/2)h]$, $i = 1, ..N$, et v_0 sur $[0, h/2]$, v_N sur $]1-h/2, 1]$. La notation avec le symbole Δ signifie traduit simplement le côté *discret*, on aurait pu noter avec l'indice h.

On aura besoin du résultat suivant.

Proposition 2.2.2 *Pour tout $v \in \mathbb{R}^{N+2}$, tel que $v_0 = v_{N+1} = 0$,*

$$\|v\|_{2,\Delta}^2 \leq \frac{1}{2} \sum_{i=0}^{N} \frac{|v_{i+1} - v_i|^2}{h}.$$

Démonstration. Pour évaluer

$$\|v\|_{2,\Delta}^2 = h\sum_{i=1}^{N} |v_i|^2,$$

on commence par écrire l'identité

$$v_i = v_0 + \sum_{k=1}^{i} (v_k - v_{k-1})$$

qui est l'analogue discret de l'intégrale de la dérivée. On obtient par l'inégalité de Cauchy-Schwarz, puisque $v_0 = 0$,

$$v_i^2 = (\sum_{k=1}^{i} |v_k - v_{k-1}|)^2 \leq i \sum_{k=1}^{i} |v_k - v_{k-1}|^2.$$

En sommant ces inégalités

$$\sum_{i=1}^{N} |v_i|^2 \leq \sum_{i=0}^{N} (i \sum_{k=1}^{i} |v_k - v_{k-1}|^2)$$

on majore le second membre par $\sum_{i=0}^{N}(i\sum_{k=1}^{N} |v_k - v_{k-1}|^2) = (\sum_{k=1}^{N} |v_k - v_{k-1}|^2)\frac{N(N+1)}{2}$. Il suffit alors de se rappeler que $h = \frac{1}{N+1}$.

Remarque. Le terme $\frac{v_{i+1}-v_i}{h}$ sur $]x_i, x_{i+1}[$ peut s'interpréter comme une dérivée discrète sur l'intervalle $]x_i, x_{i+1}[$. L'inégalité de la proposition 2.2.2 correspond en fait à une version discrète de l'inégalité de Poincaré que nous verrons au chapitre 3.

Remarquons aussi que la démontration n'utilise en fait qu'une partie de l'hypothèse, à savoir $v_0 = 0$; on peut faire une démonstration similaire en supposant seulement $v_{N+1} = 0$. □

Jusque là, u_h était un vecteur de $\mathbb{R}^N$ on peut définir un vecteur de $\mathbb{R}^{N+2}$ noté $\hat{u}_h$ en ajoutant les valeurs en x_0, x_{N+1} :

$$\hat{u}_h = (\alpha, u_1, .., u_N, \beta)^T.$$

Remarque. On sait que quand $h \to 0$, on a la convergence

$$\|\hat{u}_h\|_{2,\Delta} = h(\frac{1}{2}u_0^2 + \sum_{i=1}^{N} u_i^2 + \frac{1}{2}u_{N+1}^2) \to \int_0^1 u(x)^2 dx$$

il s'agit en effet d'une formule de quadrature de type Newton-Cotes (formule des trapèzes composée) dont on sait qu'elle converge pour u continue. □

De même pour $\bar{u}_h$, solution du problème (S_h), on pose $\hat{\bar{u}}_h = (\alpha, \bar{u}_1, ...\bar{u}_N, \beta)^T$. On peut maintenant donner un résultat de convergence en norme L^2 discrète.

Théorème 2.2.2 *Supposons que c et f sont de classe $\mathcal{C}^2$, $c \geq 0$, et soit u la solution du problème aux limites (P), et $\bar{u}_h \in \mathbb{R}^N$, la solution du problème discret (S_h). Alors il existe une constante $C > 0$ (qui ne dépend que de $u^{(4)}$, pas de N), telle que*

$$\|\hat{u}_h - \hat{\bar{u}}_h\|_{2,\Delta} \leq Ch^2.$$

Démonstration. Le vecteur "erreur" $\hat{u}_h - \hat{\bar{u}}_h$ vérifie les hypothèses de la proposition précédente et $\|\hat{u}_h - \hat{\bar{u}}_h\|_{2,\Delta} = (h\sum_{i=1}^N |(\bar{u}_h)_i - u(x_i)|^2)^{1/2}$.

Posons $v = (v_i), v_i = (\bar{u}_h)_i - u(x_i), i = 1,..,N$, et $v_0 = v_{N+1} = 0$, alors on a déjà vu que $A_h v = \varepsilon_h(u)$, où $\varepsilon_h(u)_i = -h^2 u^{(4)}(y_i)/12$. Le calcul fait dans la démonstration du théorème 2.1.4 donne

$$h^2 v^T \varepsilon_h(u) = h^2 v^T A_h v \geq v_1^2 + (v_1 - v_2)^2 + (v_2 - v_3)^2 + \cdots + (v_{N-1} - v_N)^2 + v_N^2 = \sum_{i=1}^N (v_i - v_{i-1})^2.$$

Par l'inégalité de Cauchy-Schwarz, le terme de gauche est lui même majoré par

$$h^2 (\sum_i v_i^2)^{1/2} (\sum_i |(\varepsilon_h(u))_i|^2)^{1/2},$$

et en appliquant la proposition 2.2.2, on obtient une minoration du terme de droite, qui conduit à

$$h(\sum_{i=1}^N v_i^2) \leq \frac{1}{2} \sum_{i=0}^N \frac{|v_{i+1} - v_i|^2}{h} \leq \frac{h}{2} (\sum_i v_i^2)^{1/2} (\sum_i |(\varepsilon_h(u))_i|^2)^{1/2}.$$

On peut diviser les deux membres par $h(\sum_i v_i^2)^{1/2}$, puis élever au carré ; cela donne

$$\sum_i v_i^2 \leq \frac{1}{4} \sum_{i=1}^N |(\varepsilon_h(u))_i|^2 \leq \frac{1}{4} N \max_{0 \leq x \leq 1} |u^{(4)}(x)|^2 (\frac{h^2}{12})^2,$$

d'où le résultat final. □

Le résultat du théorème 2.2.2 implique alors que $(\hat{\bar{u}}_h)_\Delta$ (fonction constante par morceaux construite à partir des valeurs $(\bar{u}_h)_i$ tend vers u dans $L^2(0,1)$. En effet, si on note $(\hat{u}_h)_\Delta$ la fonction constante par morceaux construite à partir des valeurs $(u_h)_i$, on peut écrire

$$\|u - (\hat{\bar{u}}_h)_\Delta\|_{L^2} \leq \|u - (\hat{u}_h)_\Delta\|_{L^2(0,1)} + \|(\hat{u}_h)_\Delta - \hat{\bar{u}}_h)_\Delta\|_{L^2(0,1)},$$

le dernier terme du membre de droite est exactement $\|\hat{u}_h - \hat{\bar{u}}_h\|_{2,\Delta}$ et il tend vers 0 avec h d'après le résultat précédent, le premier tend également vers 0 (on approche une fonction continue par une fonction en escalier sur $[0,1]$).

2.3 Une brève excursion en dimension 2

À ce niveau du cours, on ne peut pas dire grand-chose des problèmes aux limites en dimension supérieure à 1. Néanmoins, un certain nombre de propriétés peuvent être démontrées de façon élémentaire dans des cas particuliers.

On considérera donc ici un carré ouvert du plan $\Omega =]0,a[\times]0,a[$, avec $a > 0$. On note les coordonnées x et y. Soient $F : \Omega \to \mathbb{R}$ et $g : \partial\Omega \to \mathbb{R}$ deux applications continues. Le problème aux limites va consister à chercher une fonction $u : \bar{\Omega} \to \mathbb{R}$ appartenant à $\mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ telle que :

$$\left\{ \begin{array}{l} -\Delta u = F \text{ dans } \Omega, \\ u = g \text{ sur } \partial\Omega. \end{array} \right. \tag{2.8}$$

Théorème 2.3.1 *(Principe du maximum). Soit $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ telle que $-\Delta u \geq 0$ dans Ω et $u \geq 0$ sur $\partial\Omega$. Alors $u \geq 0$ dans tout $\bar{\Omega}$.*

Démonstration. Soit $v \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ une fonction qui atteint un minimum relatif en un point $(x_0, y_0) \in \Omega$ *intérieur*. L'application $v_{y_0} :]0, a[\to \mathbb{R}, t \mapsto v(t, y_0)$ admet donc en particulier un minimum relatif en $t = x_0$, qui est point intérieur d'un intervalle où cette fonction est de classe $\mathcal{C}^2$. Par la formule de Taylor-Lagrange, on en déduit que $\frac{d^2 v_{y_0}}{dt^2}(x_0) \geq 0$ (raisonner par l'absurde). Par définition des dérivées partielles, ceci n'est autre que $\frac{\partial^2 v}{\partial x^2}(x_0, y_0) \geq 0$. De même, on montre que $\frac{\partial^2 v}{\partial y^2}(x_0, y_0) \geq 0$. Additionnant ces deux inégalités, on obtient

$$\Delta v(x_0, y_0) \geq 0$$

en tout point de minimum relatif intérieur de v.

Soit maintenant $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ telle que $-\Delta u \geq 0$ dans Ω et $u \geq 0$ sur $\partial\Omega$. On raisonne par l'absurde et l'on suppose que u prend des valeurs strictement négatives dans $\bar{\Omega}$. Comme $\bar{\Omega}$ est compact et u est continue, elle atteint son minimum absolu, qui est donc une valeur strictement négative, en un point (x_0, y_0) de $\bar{\Omega}$. Ce point ne peut se trouver sur la frontière du carré $\partial\Omega$ où u est supposée positive, donc $(x_0, y_0) \in \Omega$. Posons $-M = u(x_0, y_0) = \min_{\bar{\Omega}} u < 0$.

On introduit alors une fonction auxiliaire

$$u_\varepsilon(x, y) = u(x, y) - \varepsilon(x^2 + y^2) \quad \text{avec} \quad 0 < \varepsilon < \frac{M}{2a^2}.$$

La fonction u_ε est continue sur le compact $\bar{\Omega}$. Elle y atteint donc son minimum absolu en un point (x_1, y_1). En (x_0, y_0), on a

$$u_\varepsilon(x_0, y_0) = -M - \varepsilon(x_0^2 + y_0^2) < -M,$$

donc $u_\varepsilon(x_1, y_1) \leq u_\varepsilon(x_0, y_0) < -M$. De plus, si $(x, y) \in \partial\Omega$ on a $u(x, y) \geq 0$, et par conséquent,

$$u_\varepsilon(x, y) \geq -\varepsilon(x^2 + y^2) \geq -2\varepsilon a^2 > -M.$$

Comme $u_\varepsilon(x_1, y_1) < u_\varepsilon(x, y)$ pour tout $(x, y) \in \partial\Omega$, on en déduit que $(x_1, y_1) \notin \partial\Omega$, ou encore $(x_1, y_1) \in \Omega$.

Appliquons la première remarque à la fonction u_ε au point (x_1, y_1). Il vient que

$$-\Delta u_\varepsilon(x_1, y_1) \leq 0.$$

D'un autre côté, comme $-\Delta u \geq 0$ dans Ω,

$$-\Delta u_\varepsilon(x, y) = -\Delta u(x, y) + \varepsilon\Delta(x^2 + y^2) = -\Delta u(x, y) + 4\varepsilon > 0.$$

En particulier, en $(x_1, y_1) \in \Omega$, on obtient

$$-\Delta u_\varepsilon(x_1, y_1) > 0,$$

contradiction.

Par conséquent, u ne peut pas prendre de valeurs strictement négatives dans $\bar{\Omega}$, ce qui était notre hypothèse de départ, soit en d'autres termes, $u \geq 0$ dans $\bar{\Omega}$. □

Remarques. i) On ne peut pas faire le raisonnement par l'absurde directement sur u. En effet, celui-ci conduit alors à $-\Delta u(x_0,y_0) \leq 0$ et $-\Delta u(x_0,y_0) \geq 0$, ce qui n'est pas une contradiction mais implique seulement que $-\Delta u(x_0,y_0) = 0$, et on ne va pas plus loin.

ii) Le raisonnement précédent montre qu'une fonction u telle que $-\Delta u \geq 0$ atteint son minimum absolu sur le bord de l'ouvert. De même, une fonction u telle que $-\Delta u \leq 0$ atteint son maximum absolu sur le bord. □

Corollaire 2.3.2 *Soit $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ telle que $-\Delta u = 0$ dans Ω et $u = 0$ sur $\partial\Omega$. Alors $u = 0$ dans $\bar{\Omega}$.*

Démonstration. On applique le théorème 2.3.1 à u et à $-u$. □

On déduit immédiatement du corollaire 2.3.2 un résultat d'*unicité*.

Théorème 2.3.3 *Le problème aux limites (2.8) admet au plus une solution.*

Démonstration. Soient u_1 et u_2 deux solutions du problème aux limites (2.8). On pose $v = u_1 - u_2$. Il vient $\Delta v = \Delta u_1 - \Delta u_2 = F - F = 0$ dans Ω d'une part, et $v = g - g = 0$ sur $\partial\Omega$ d'autre part. D'après le corollaire 2.3.2, on a donc $v = 0$. □

Remarque. Les raisonnements précédents sont encore valables quand Ω est un ouvert borné quelconque de $\mathbb{R}^n$, pour toute valeur de n, et pas seulement pour un carré en dimension 2 (le vérifier). □

La question de l'*existence* d'une solution au problème (2.8) est nettement plus délicate. Donnons en un cas particulier avec $F = 0$.

Proposition 2.3.1 *On suppose que $g(0,y) = g(a,y) = g(x,a) = 0$ et $g(x,0) = f(x)$ où $f \in \mathcal{C}^1([0,a])$, avec $f(0) = f(a) = 0$. Alors il existe une solution u du problème (2.8) avec $F = 0$.*

Démonstration. On commence par prolonger f de $[0,a]$ à $[-a,a]$ par imparité, puis à $\mathbb{R}$ tout entier par $2a$-périodicité. La fonction $\tilde{f}$ ainsi construite est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et $2a$-périodique (le vérifier). Par conséquent, sa série de Fourier converge normalement vers $\tilde{f}$ par le théorème de Dirichlet sur les séries de Fourier. En d'autres termes,

$$\forall x \in [0,a], f(x) = \sum_{k=1}^{\infty} b_k \sin\left(\frac{k\pi x}{a}\right),\ b_k = \frac{2}{a}\int_0^a f(s)\sin\left(\frac{k\pi s}{a}\right) ds \text{ et } \sum_{k=1}^{\infty} |b_k| < +\infty.$$

On va construire u à l'aide d'une série de fonctions. Pour cela, on pose

$$u_k(x,y) = b_k \sin\left(\frac{k\pi x}{a}\right) h_k(y),$$

et l'on va déterminer la fonction d'une seule variable h_k pour que $\Delta u_k = 0$. Comme

$$\Delta u_k(x,y) = b_k \sin\left(\frac{k\pi x}{a}\right)\left[h_k''(y) - \frac{k^2\pi^2}{a^2} h_k(y)\right],$$

il suffit pour cela que

$$h_k''(y) - \frac{k^2\pi^2}{a^2} h_k(y) = 0 \Leftrightarrow h_k(y) = A\cosh\left(\frac{k\pi y}{a}\right) + B\sinh\left(\frac{k\pi y}{a}\right).$$

Choisissons les constantes A et B pour que $h_k(0) = 1$ et $h_k(1) = 0$. Il vient

$$h_k(y) = \frac{\sinh\left[k\pi(1-\frac{y}{a})\right]}{\sinh k\pi}.$$

Cette fonction est visiblement positive sur $[0,a]$, décroissante, d'où $0 \le h_k(y) \le 1$ pour $y \in [0,a]$.

On en déduit que la série $\sum_k u_k$ converge *normalement* sur $\bar{\Omega}$. En effet,

$$\forall (x,y) \in \bar{\Omega}, \left| b_k \sin\left(\frac{k\pi x}{a}\right) h_k(y)\right| \le |b_k| \text{ et la série } \sum_k |b_k| \text{ est convergente.}$$

On pose donc $u = \sum_k u_k$. Chaque fonction u_k est continue sur $\bar{\Omega}$, donc la somme de leur série u est aussi une fonction continue sur $\bar{\Omega}$, puisqu'elle converge normalement. De plus, en remplaçant dans la série les valeurs de x et y appropriées, on voit aisément que

$$u(0,y) = u(a,y) = u(x,a) = 0, u(x,0) = f(x)$$

sur $\partial\Omega$. La fonction u satisfait donc les conditions aux limites. Il reste à montrer qu'elle est de classe $\mathcal{C}^2$ dans Ω et que son Laplacien est identiquement nul dans cet ouvert. Pour cela, on doit utiliser les théorèmes de dérivations de séries de fonctions. Commençons par la dérivée partielle $\partial u/\partial x$, dont il s'agit de démontrer l'existence.

On a

$$\frac{\partial u_k}{\partial x}(x,y) = b_k \frac{k\pi}{a}\cos\left(\frac{k\pi x}{a}\right) h_k(y)$$

d'où

$$\left|\frac{\partial u_k}{\partial x}(x,y)\right| \le |b_k| \frac{k\pi}{a} h_k(y).$$

Pour $y = 0$, $h_k(y) = 1$ et l'on voit apparaître la série majorante $\sum_k k|b_k|$ sur laquelle *on ne sait rien*. Cette série peut être convergente comme elle peut être divergente. Donc on ne peut rien conclure en $y = 0$ par cette méthode. On exclut donc cette situation en prenant un nombre $a > \eta > 0$ et en se restreignant à l'intervalle $[\eta, a]$ dans la variable y. On remarque d'abord que la série des $|b_k|$ est convergente. Ceci implique en particulier que $M = \max_{k\in\mathbb{N}^*} |b_k| < +\infty$. Comme la fonction h_k est positive décroissante, il vient

$$\forall x \in [0,a], \forall y \in [\eta, a], \left|\frac{\partial u_k}{\partial x}(x,y)\right| \le M\frac{k\pi}{a} h_k(\eta) = M\frac{k\pi}{a}\frac{\sinh\left[k\pi(1-\frac{\eta}{a})\right]}{\sinh k\pi}.$$

On va raisonner par équivalents (il s'agit de séries à termes *positifs*, donc les équivalents sont permis).

$$k\sinh\left[k\pi(1-\frac{\eta}{a})\right] \sim \frac{k}{2} e^{k\pi(1-\frac{\eta}{a})}, \ \sinh k\pi \sim \frac{1}{2}e^{k\pi} \text{ quand } k \to +\infty,$$

donc

$$M\frac{k\pi}{a} h_k(\eta) \sim \frac{M\pi}{a} k e^{-\frac{k\pi\eta}{a}} = \frac{M\pi}{a} k \left(e^{-\frac{\pi\eta}{a}}\right)^k.$$

Comme $\eta > 0$, on voit que $e^{-\frac{\pi\eta}{a}} < 1$, donc la série $\sum_k k e^{-\frac{k\pi\eta}{a}}$ est *convergente*. Par conséquent, la série $\sum_k \frac{\partial u_k}{\partial x}$ converge *normalement* sur l'ensemble $[0,a]\times[\eta,a]$. D'après les théorèmes de dérivation des séries de fonctions, on en déduit que la fonction u est continûment dérivable par rapport à x sur $[0,a]\times[\eta,a]$, avec $\partial u/\partial x = \sum_k \partial u_k/\partial x$. Comme ceci est vrai pour tout $\eta > 0$, on a le même résultat sur $[0,a]\times]0,a]$, donc, par restriction, sur Ω. On procède de même pour $\partial u/\partial y$ (le faire) ce qui montre que $u \in \mathcal{C}^1(\Omega)$.

On recommence le même raisonnement pour toutes les dérivées secondes $\partial^2 u/\partial x^2$, $\partial^2 u/\partial x\partial y$ et $\partial^2 u/\partial y^2$ (le faire) et l'on obtient que $u \in \mathcal{C}^2(\Omega)$ avec

$$\frac{\partial^2 u}{\partial x^2} = \sum_{k=1}^{\infty} \frac{\partial^2 u_k}{\partial x^2}, \quad \frac{\partial^2 u}{\partial x\partial y} = \sum_{k=1}^{\infty} \frac{\partial^2 u_k}{\partial x\partial y}, \quad \frac{\partial^2 u}{\partial y^2} = \sum_{k=1}^{\infty} \frac{\partial^2 u_k}{\partial y^2}$$

dans Ω. Additionnant entre elles les deux expressions extrêmes, on obtient

$$\Delta u = \sum_{k=1}^{\infty} \Delta u_k = 0 \text{ dans } \Omega,$$

puisque $\Delta u_k = 0$ par construction. □

Remarque. On ne peut pas faire l'économie de la démonstration de la convergence de la série et de celle de ses dérivées. En effet, c'est par des séries de fonctions aussi régulières que l'on veut que l'on construit des exemples de fonctions continues mais dérivables nulle part. Donc la dérivabilité de la somme d'une série de fonctions n'a rien d'évident, pas plus que la dérivation terme à terme de cette série. □

Théorème 2.3.4 *Soit g continue sur $\partial\Omega$ et de classe $\mathcal{C}^1$ sur chaque côté du carré. Il existe une solution au problème (2.8) avec $F = 0$.*

Démonstration. On peut trouver quatre nombres $\alpha,\beta,\gamma,\delta$ tels que la fonction

$$\bar{g}(x,y) = g(x,y) - (\alpha + \beta x + \gamma y + \delta xy)$$

s'annule aux quatre sommets du carré (c'est de l'interpolation de Lagrange Q_1 dans $\mathbb{R}^2$, c'est à dire on interpole g par un polynôme de deux variables de degré ≤ 1 par rapport à chaque variable, qui prend les mêmes valeurs que g aux quatre sommets ; l'ensemble de ces polynômes forme un espace vectoriel de dimension 4 dont $1, x, y, xy$ est une base). On écrit ensuite $\bar{g}$ comme la somme de ses restrictions à chaque côté : $\bar{g} = \bar{g}_1 + \bar{g}_2 + \bar{g}_3 + \bar{g}_4$, où chaque $\bar{g}_i$ est bien continue et comme à la proposition 2.3.1 (mais sur chaque côté). Leur correspondent quatre solutions $\bar{u}_1, \ldots, \bar{u}_4$ du problème aux limites.

On pose alors $u(x,y) = \sum_{i=1}^4 \bar{u}_i(x,y) + \alpha + \beta x + \gamma y + \delta xy$. Visiblement, $u \in \mathcal{C}^0(\bar{\Omega}) \cap \mathcal{C}^2(\Omega)$ et

$$\begin{cases} u(x,y) = \sum_{i=1}^4 \bar{g}_i(x,y) + \alpha + \beta x + \gamma y + \delta xy = g(x,y) \text{ sur } \partial\Omega, \\ \Delta u(x,y) = \sum_{i=1}^4 \Delta\bar{u}_i(x,y) + \Delta(\alpha + \beta x + \gamma y + \delta xy) = 0 \text{ dans } \Omega, \end{cases}$$

donc u est solution du problème aux limites. □

Remarques. i) Il est essentiel pour cette démonstration que $F = 0$ et que Ω soit un carré (ou un rectangle) en dimension 2.

ii) La somme partielle $\sum_{k=1}^n u_k$ donne une approximation calculable de u, à condition de savoir calculer – ou au moins approcher – les coefficients de Fourier de la donnée au bord g. □

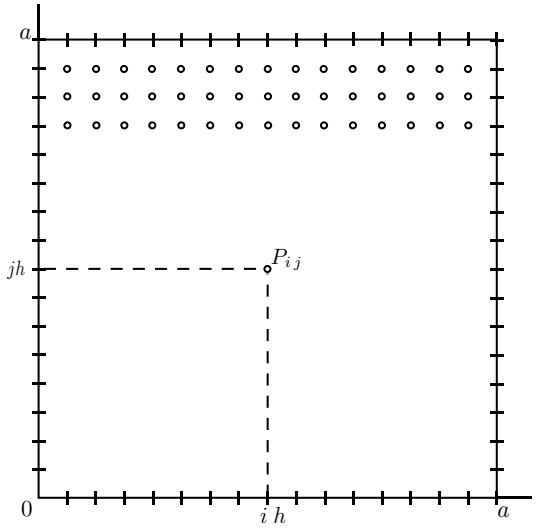


FIGURE 2.2 – Grille de discrétisation 2D

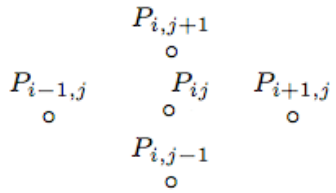


FIGURE 2.3 – les points intervenant pour le calcul du Laplacien discret à 5 points

2.4 La méthode des différences finies en dimension 2

Donnons un aperçu de ce que l'on peut faire en différences finies sur le problème précédent. On se donne un entier $N \geq 1$, on pose encore $h = \frac{a}{N+1}$ et l'on construit la grille de discrétisation composée des points ou *nœuds* $P_{ij} = (ih, jh)$ pour $0 \leq i \leq N+1$ et $0 \leq j \leq N+1$. Il y a donc au total $(N+2)^2$ points dans cette grille, dont N^2 situés à l'intérieur ($1 \leq i \leq N$ et $1 \leq j \leq N$) et $4N+4$ situés sur le bord ($i = 0$ ou $i = N+1$ ou $j = 0$ ou $j = N+1$). Si φ est une fonction définie sur $\bar{\Omega}$, on écrira indifféremment $\varphi(x,y) = \varphi(P)$ avec $P = (x,y)$. De façon analogue à la dimension un, on introduit une discrétisation du Laplacien.

Définition 2.4.1 *Soit $\varphi \in \mathcal{C}^0(\bar{\Omega})$. On appelle Laplacien discret à 5 points de φ la quantité*

$$\Delta_h \varphi(P_{ij}) = \frac{1}{h^2}[-4\varphi(P_{ij}) + \varphi(P_{i-1,j}) + \varphi(P_{i+1,j}) + \varphi(P_{i,j-1}) + \varphi(P_{i,j+1})]$$

définie en tout point P_{ij} ($1 \leq i \leq N$ et $1 \leq j \leq N$) intérieur à la grille.

Remarque. La terminologie est claire, le Laplacien discret à 5 points utilise les quatre plus proches voisins du point considéré sur la grille (voir la figure 2.3). Il existe des variantes à 9 points, etc. □

Estimons l'erreur de consistance.

Théorème 2.4.1 *Soit* $\varphi \in \mathcal{C}^4(\bar{\Omega})$. *Alors* $\forall i,j \in [1,N]$,

$$|\Delta\varphi(P_{ij}) - \Delta_h\varphi(P_{ij})| \leq \frac{h^2}{12}\max_{\bar{\Omega}}\Big(\Big|\frac{\partial^4\varphi}{\partial x^4}\Big| + \Big|\frac{\partial^4\varphi}{\partial y^4}\Big|\Big).$$

Noter la distinction entre $\Delta\varphi$ qui est une fonction définie sur Ω, dont on prend ici la valeur en un point de la grille et $\Delta_h\varphi$ qui est défini seulement sur la grille.

Démonstration. Par définition de ce qu'est une dérivée partielle, si l'on introduit la fonction $\theta_1^{ij}(t) = \varphi(ih+t, jh)$, qui est définie et de classe $\mathcal{C}^4$ sur un voisinage de zéro en t (lequel contient au moins l'intervalle $]-h,h[$, vues les valeurs de i et j), on a $\frac{\partial^k\varphi}{\partial x^k}(ih+t,jh) = \frac{d^k\theta_1^{ij}}{dt^k}(t)$ pour $0 \leq k \leq 4$. De même avec $\theta_2^{ij}(s) = \varphi(ih, jh+s)$, $\frac{\partial^k\varphi}{\partial y^k}(ih,jh+s) = \frac{d^k\theta_2^{ij}}{ds^k}(s)$. En particulier,

$$\Delta\varphi(P_{ij}) = (\theta_1^{ij})''(0) + (\theta_2^{ij})''(0).$$

Or, les mêmes calculs de développements de Taylor-Lagrange qu'en dimension 1 montrent que

$$h^2(\theta_1^{ij})''(0) = \theta_1^{ij}(-h) - 2\theta_1^{ij}(0) + \theta_1^{ij}(h) + \frac{h^4}{12}(\theta_1^{ij})^{(4)}(t_{ij})$$

pour un certain $t_{ij} \in]-h,h[$ et

$$h^2(\theta_2^{ij})''(0) = \theta_2^{ij}(-h) - 2\theta_2^{ij}(0) + \theta_2^{ij}(h) + \frac{h^4}{12}(\theta^{ij})^{(4)}(s_{ij})$$

pour un certain $s_{ij} \in]-h,h[$. Par définition des fonctions θ_1^{ij} et θ_2^{ij}, il est facile de voir que $\theta_1^{ij}(0) = \varphi(P_{ij})$, $\theta_1^{ij}(-h) = \varphi(P_{i-1,j})$, $\theta_1^{ij}(h) = \varphi(P_{i+1,j})$, $\theta_2^{ij}(0) = \varphi(P_{ij})$, $\theta_2^{ij}(-h) = \varphi(P_{i,j-1})$ et $\theta_1^{ij}(h) = \varphi(P_{i,j+1})$. En remplaçant et en sommant, on obtient donc :

$$\Delta\varphi(P_{ij}) = \Delta_h\varphi(P_{ij}) + \frac{h^2}{12}\Big(\frac{\partial^4\varphi}{\partial x^4}(ih+t_{ij},jh) + \frac{\partial^4\varphi}{\partial y^4}(ih,jh+s_{ij})\Big),$$

et l'on conclut immédiatement à l'aide des valeurs absolues. □

On voit donc apparaître une erreur de consistance en $O(h^2)$. Par le même cheminement qu'en dimension 1, on introduit la méthode des différences finies de la façon suivante. Notons $\Omega_h = \{P_{ij}, 1 \leq i \leq N, 1 \leq j \leq N\}$; il s'agit des points de la grille intérieurs. Posons également $(\bar{\Omega})_h = \{P_{ij}, 0 \leq i \leq N+1, 0 \leq j \leq N+1\}$; c'est l'ensemble de tous les points de la grille, y compris ceux du bord. On va donc chercher une fonction *définie sur la grille* et qui ne prend qu'un nombre fini de valeurs $\bar{u}_h \colon (\bar{\Omega})_h \to \mathbb{R}$, telle que

$$\boxed{\begin{cases} -\Delta_h\bar{u}_h(P_{ij}) = F(P_{ij}) \ \text{ sur } \Omega_h, \\ \bar{u}_h(P_{ij}) = g(P_{ij}) \ \text{ sur } (\bar{\Omega})_h \setminus \Omega_h. \end{cases}} \tag{2.4}$$

Il n'y a bien sûr aucune difficulté à appliquer le Laplacien discret, défini initialement pour une fonction continue sur $\bar{\Omega}$, à une fonction *discrète* définie seulement sur $(\bar{\Omega})_h$. On vérifie aisément en regardant les indices que toutes les valeurs dont on a besoin pour calculer le Laplacien discret sont bien à notre disposition dans ce cas.

On voit sur la formulation 2.4 qu'en fait seules les valeurs de $\bar{u}_h$ aux points intérieurs sont inconnues, ses valeurs sur les points du bord étant données par la condition aux limites (comme en dimension 1). Il s'agit donc d'un problème à N^2 inconnues (les valeurs de $\bar{u}_h$ sur Ω_h) et N^2 équations linéaires (les valeurs de $-\Delta_h \bar{u}_h$ sur Ω_h).

À la différence de la dimension 1, ces inconnues et équations ne sont pas arrangées naturellement comme les composantes d'un vecteur — en fait, elles sont arrangées naturellement comme les composantes d'une matrice avec deux indices — et la forme matricielle du problème aux différences finies n'apparaît pas directement à l'œil nu comme précédemment.

Pour faire apparaître cette forme matricielle, il faut réarranger abstraitement les valeurs de $\bar{u}_h$, c'est-à-dire en fait les points de la grille, en une seule colonne. En d'autres termes, on doit *numéroter* les nœuds. En dimension 1, la question ne se posait pas puisqu'une numérotation naturelle s'imposait, donnée par les indices des points de la grille. Si l'on avait eu l'esprit compliqué, on aurait pu tout aussi bien choisir alors une autre numérotation, c'est-à-dire effectuer une permutation des indices, c'est-à-dire un changement de base qui permute les vecteurs de base. La matrice qui en aurait résulté aurait été obtenue à partir de la matrice initiale par changement de base par une matrice de permutation. Elle aurait en particulier perdu ses agréables propriétés d'être tridiagonale et symétrique (en passant, que peut-on dire de la monotonie ?), propriétés qui sont très utiles lorsqu'il est temps d'appliquer effectivement des méthodes de résolution de systèmes linéaires. C'était idiot, on ne l'a pas fait. Ici, on n'a pas le choix. Il n'y a pas de numérotation naturelle qui saute aux yeux. On va en parachuter une qui marche bien au sens où elle fournit une matrice symétrique et dont les éléments non nuls sont relativement concentrés autour de la diagonale, deux conditions favorables à l'utilisation, dans une étape ultérieure, d'algorithmes efficaces de résolution de systèmes linéaires.

On convient donc de numéroter les nœuds de gauche à droite et de bas en haut, comme sur la figure 2.4. On se convainc sans grand peine que le numéro du point P_{ij} est donné par $(j-1)N+i$. On définit donc le vecteur $\bar{u}_h$ par ses composantes :

$$(\bar{u}_h)_k = \bar{u}_h(P_{ij}) \quad \text{pour } k = (j-1)N+i.$$

Attention à l'abus de notation : on ne distingue pas la fonction $\bar{u}_h$ définie sur la grille du vecteur $\bar{u}_h$, bien que leur identification passe par une numérotation des nœuds largement arbitraire.

Les indices k ainsi définis varient entre 1 et N^2, donc $\bar{u}_h \in \mathbb{R}^{N^2}$. La bijection entre $\{1,\dots,N\} \times \{1,\dots,N\}$ et $\{1,\dots,N^2\}$ que l'on vient juste d'introduire admet pour inverse l'application $k \mapsto \left(k - \left[\frac{k-1}{N}\right]N, \left[\frac{k-1}{N}\right]+1\right)$ où $[t]$ désigne la partie entière de t (le vérifier).

Nous pouvons maintenant écrire la forme matricielle de la méthode associée à la numérotation

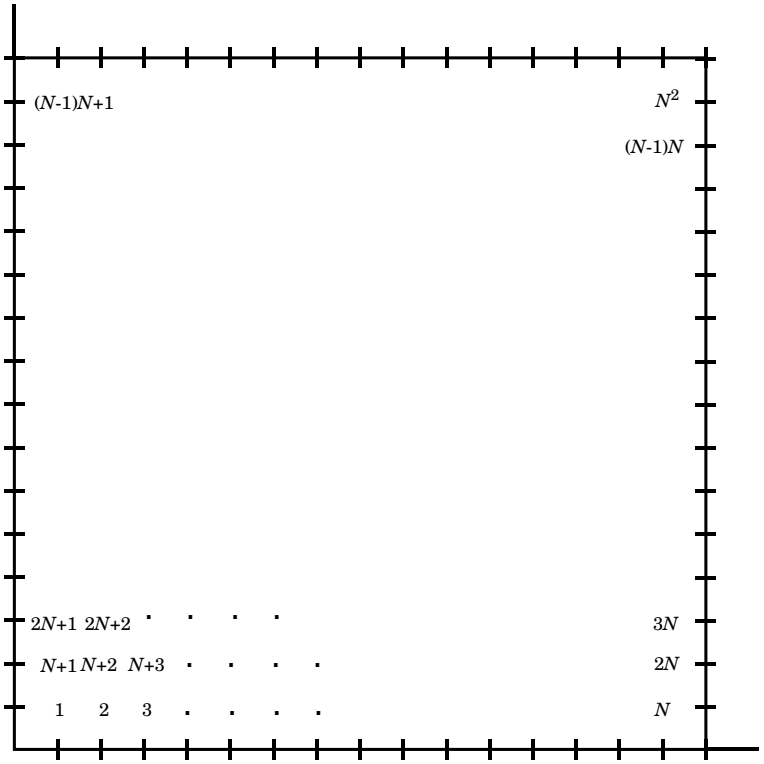


FIGURE 2.4 – Numérotation des noeuds

choisie. Pour cela, on introduit la matrice $N \times N$

$$T_4 = \begin{pmatrix} 4 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & & \vdots \\ 0 & -1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -1 & 4 & -1 \\ 0 & \dots & \dots & 0 & -1 & 4 \end{pmatrix}$$

et l'on note I la matrice identité $N \times N$.

On va définir la matrice A_h du système linéaire, c'est une matrice $N^2 \times N^2$ que l'on peut écrire *par blocs* $N \times N$, en utilisant les blocs T_4 et I, ainsi que le bloc 0 correspondant à la matrice $N \times N$ nulle. De même, on peut écrire le second membre b_h par blocs, on fait apparaître ci-dessous les deux premiers blocs (sur N blocs) et le début du troisième.

Proposition 2.4.1 *Le vecteur $\bar{u}_h \in \mathbb{R}^{N^2}$ est solution du système linéaire suivant :*

$$A_h \bar{u}_h = b_h,$$

avec

$$A_h = \frac{1}{h^2}\begin{pmatrix} T_4 & -I & 0 & \dots & \dots & 0 \\ -I & T_4 & -I & 0 & & \vdots \\ 0 & -I & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -I & T_4 & -I \\ 0 & \dots & \dots & 0 & -I & T_4 \end{pmatrix}$$

et

$$b_h = \begin{pmatrix} F_1 + \frac{1}{h^2}(g(P_{1,0}) + g(P_{0,1})) \\ F_2 + \frac{1}{h^2} g(P_{2,0}) \\ F_3 + \frac{1}{h^2} g(P_{3,0}) \\ \vdots \\ F_N + \frac{1}{h^2}(g(P_{N,0}) + g(P_{N+1,1})) \\ F_{N+1} + \frac{1}{h^2} g(P_{0,2}) \\ F_{N+2} \\ \vdots \\ F_{2N} + \frac{1}{h^2} g(P_{N+1,2}) \\ F_{2N+1} + \frac{1}{h^2} g(P_{0,3}) \\ F_{2N+2} \\ \vdots \end{pmatrix}$$

où $F_k = F(P_{ij})$ pour $k = (j-1)N + i$.

Démonstration. Il faut distinguer plusieurs cas, suivant que l'indice de ligne k considéré correspond à un point dont les quatre voisins sont dans Ω_h, un point dont seuls trois voisins sont dans Ω_h ou un point dont seuls deux voisins sont dans Ω_h.

♣ Quatre voisins dans Ω_h. Ce sont les points P_{ij} avec $2 \le i \le N-2$ et $2 \le j \le N-2$. Ils correspondent aux indices $k \in \cup_{l=1}^{N-2}\{lN+2, lN+3, \dots, (l+1)N-1\}$. (Par exemple, pour $l = 1$, ceci donne $k = N+2, N+3, \dots, 2N-1$. Ensuite on saute directement à $2N+2$ et ainsi de suite)

En un tel point P_{ij}, de numéro $k = (j-1)N + i$, les points voisins sont tous numérotés et ont pour numéro respectif

$$\begin{aligned} \#P_{i-1,j} &= (j-1)N + (i-1) = k-1, \\ \#P_{i+1,j} &= (j-1)N + (i+1) = k+1, \\ \#P_{i,j-1} &= ((j-1)-1)N + i = k-N, \\ \#P_{i,j+1} &= ((j+1)-1)N + i = k+N. \end{aligned}$$

On obtient donc pour le problème discret la ligne

$$\frac{1}{h^2}[-(\bar{u}_h)_{k-N} - (\bar{u}_h)_{k-1} + 4(\bar{u}_h)_k - (\bar{u}_h)_{k+1} - (\bar{u}_h)_{k+N}] = F(P_{ij}) = F_k.$$

♠ Trois voisins dans Ω_h. Ceci peut se produire de quatre façons : $j = 1, i = 2, 3, \dots, N-1$; $j = N, i = 2, 3, \dots, N-1$; $i = 1, j = 2, 3, \dots, N-1$ ou $i = N, j = 2, 3, \dots, N-1$. Ces quatre façons correspondent respectivement à $k = 2, 3, \dots, N-1$; $k = (N-1)N+2, (N-1)N+3, \dots, N^2 -$

1 ; $k = N+1, 2N+1, \ldots, (N-2)N+1$ et $k = 2N, 3N, \ldots, (N-1)N$. Regardons la première possibilité (les autres sont laissées en exercice). Le point $P_{i,j-1} = P_{i,0}$ est situé sur le bord. On doit donc passer la valeur de $\bar{u}_h$ correspondante, qui est donnée par la condition aux limites, dans le membre de droite, ce qui donne

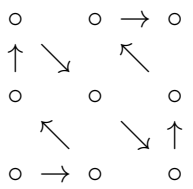
$$\frac{1}{h^2}[-(\bar{u}_h)_{k-1} + 4(\bar{u}_h)_k - (\bar{u}_h)_{k+1} - (\bar{u}_h)_{k+N}] = F_k + \frac{1}{h^2} g(P_{k,0}).$$

$\diamondsuit$ Deux voisins dans Ω_h. Ceci se produit aux quatre coins : $P_{1,1}$, $P_{N,1}$, $P_{1,N}$ et $P_{N,N}$, donc pour $k = 1$, $k = N$, $k = (N-1)N+1$ et $k = N^2$. Regardons le cas $k = 1$. Il vient

$$\frac{1}{h^2}[4(\bar{u}_h)_1 - (\bar{u}_h)_2 - (\bar{u}_h)_{1+N}] = F_1 + \frac{1}{h^2}[g(P_{1,0}) + g(P_{0,1})].$$

Il faut enfin recoller tous les morceaux. C'est un peu fastidieux, il faut utiliser une grande feuille de papier, mais on arrive à la forme matricielle annoncée. □

Remarque. Pour se convaincre de l'importance du choix de la numérotation, il peut être amusant d'écrire entièrement la matrice pour $N = 3$ par exemple (donc une matrice 9×9) pour la numérotation proposée, puis pour une autre numérotation, *a priori* tout aussi naturelle, celle que l'on utilise pour montrer que $\mathbb{N}^2$ est en bijection avec $\mathbb{N}$:



□

Proposition 2.4.2 *La matrice A_h est symétrique, définie positive et monotone.*

Démonstration. On montre que A_h est symétrique, définie positive comme en dimension un. Pour démontrer qu'elle est monotone, on passe par un *principe du maximum discret.*

Soit $v_h \in \mathbb{R}^{N^2}$ un vecteur de composantes $(v_h)_k$. On identifie ce vecteur à une fonction $v_h \colon \Omega_h \to \mathbb{R}$ en posant $v_h(P_{ij}) = (v_h)_k$ pour $k = (j-1)N + i$. D'après ce qui précède, si on étend v_h par 0 sur $(\bar{\Omega})_h$, alors on a

$$(A_h v_h)_k = -\Delta_h v_h(P_{ij}) \quad \text{pour } k = (j-1)N + i$$

(l'extension par 0 est cruciale pour les points de Ω_h qui ont des voisins sur le bord).

Supposons maintenant que $A_h v_h \overrightarrow{\geq} 0$. Ceci signifie simplement que $-\Delta_h v_h(P_{ij}) \geq 0$ pour $(i,j) \in \{1, \ldots, N\}^2$. Si v_h atteint son minimum sur $(\bar{\Omega})_h \setminus \Omega_h$, alors immédiatement, $v_h \overrightarrow{\geq} 0$ et on a terminé. Supposons donc que v_h atteint son minimum sur Ω_h en un point P_{ij} tel que $v_h(P_{ij}) \leq v_h(P_{mn})$ pour tous m et n. En ce point, on a

$$\begin{aligned} -\Delta_h v_h(P_{ij}) &= \tfrac{1}{h^2}[4v_h(P_{ij}) - v_h(P_{i-1,j}) - v_h(P_{i+1,j}) - v_h(P_{i,j-1}) - v_h(P_{i,j+1})] \\ &\leq 0, \end{aligned}$$

d'où par l'hypothèse $A_h v_h \overrightarrow{\geq} 0$,

$$-\Delta_h v_h(P_{ij}) = 0$$

et donc

$$v_h(P_{ij}) = \min_{(\bar{\Omega})_h} v_h = v_h(P_{i-1,j}) = v_h(P_{i+1,j}) = v_h(P_{i,j-1}) = v_h(P_{i,j+1}).$$

On recommence de proche en proche, et on finit par atteindre un point du bord, où $v_h = 0$. Donc, $\min_{(\bar{\Omega})_h} v_h = 0$, soit $v_h \overrightarrow{\geq} 0$ dans ce cas aussi (en fait, il est facile de voir que dans ce cas, v_h est identiquement nulle). □

Remarque. Une conséquence immédiate de la proposition 2.4.2 est que la méthode des différences finies est encore bien posée. □

Théorème 2.4.2 *On a* $|||A_h^{-1}|||_\infty \leq \frac{a^2}{2}$.

Démonstration. On procède encore par principe du maximum discret. Soit $F_h \in \mathbb{R}^{N^2}$ et $v_h \in \mathbb{R}^{N^2}$ non nul tels que $A_h v_h = F_h$. Identifiant v_h avec une fonction sur $(\bar{\Omega})_h$ nulle au bord comme précédemment, ceci est équivalent à $-\Delta_h v_h = F_h$ dans Ω_h.

Introduisons la fonction discrète $z_h(P_{ij}) = \frac{h^2}{4}(i^2 + j^2)$. On vérifie par un petit calcul que $-\Delta_h z_h = -1$ dans Ω_h. Posons alors

$$w_h^+ = \|-\Delta_h v_h\|_\infty z_h - v_h,$$

où on laisse en évidence le signe pour se souvenir que $-\Delta_h v_h = F_h$. Prenant le Laplacien discret de w_h^+, il vient, pour tout $P_{i,j} \in \Omega_h$

$$-\Delta_h w_h^+(P_{i,j}) = -\|-\Delta_h v_h\|_\infty \Delta_h z_h(P_{i,j}) + \Delta_h v_h(P_{i,j}) = \Delta_h v_h(P_{i,j}) - \|-\Delta_h v_h\|_\infty \leq 0.$$

Le raisonnement de la proposition précédente s'applique et montre que w_h^+ atteint son maximum sur $(\bar{\Omega})_h \setminus \Omega_h$. Mais $v_h = 0$ sur $(\bar{\Omega})_h \setminus \Omega_h$, donc pour tout $P_{i,j} \in \Omega_h$

$$w_h^+(P_{ij}) \leq \|-\Delta_h v_h\|_\infty \max_{(\bar{\Omega})_h \setminus \Omega_h} z_h = \frac{a^2}{2}\|-\Delta_h v_h\|_\infty.$$

Par conséquent,

$$v_h(P_{i,j}) = \|-\Delta_h v_h\|_\infty z_h(P_{i,j}) - w_h^+(P_{i,j}) \geq -w_h^+(P_{i,j}) \geq -\frac{a^2}{2}\|-\Delta_h v_h\|_\infty.$$

Utilisant de la même façon la fonction $w_h^- = \|-\Delta_h v_h\|_\infty z_h + v_h$, on montre que

$$v_h(P_{i,j}) \leq \frac{a^2}{2}\|-\Delta_h v_h\|_\infty.$$

Par conséquent,

$$\|v_h\|_\infty = \max_{P_{i,j} \in \Omega_h} |v_h(P_{i,j})| \leq \frac{a^2}{2}\|-\Delta_h v_h\|_\infty.$$

En termes matriciels, ceci s'écrit encore

$$\|A_h^{-1}F_h\|_\infty \leq \frac{a^2}{2}\|F_h\|_\infty,$$

d'où le résultat en divisant cette inégalité par $\|F_h\|_\infty$. □

Remarque. Noter la profonde analogie entre les raisonnements de principe du maximum discret et les raisonnements de principe du maximum continu. □

Nous avons étudié la consistance grâce au théorème 2.4.1 et la stabilité par l'estimation de $|||A_h^{-1}|||_\infty$. Nous pouvons maintenant énoncer le théorème de convergence.

Théorème 2.4.3 *Supposons que $u \in \mathcal{C}^4(\bar{\Omega})$, alors*

$$\max_{i,j}|u(P_{ij}) - \bar{u}_h(P_{ij})| \leq \frac{a^2h^2}{24}\max_{\bar{\Omega}}\Big(\Big|\frac{\partial^4 u}{\partial x^4}\Big| + \Big|\frac{\partial^4 u}{\partial y^4}\Big|\Big).$$

Démonstration. On note u_h (resp. F_h) le vecteur de composantes $u(P_{ij}), 1 \leq i,j \leq N$ (resp. $F(P_{ij})$), et $\varepsilon_h = -\Delta_h u_h - F_h = -\Delta_h u_h + (\Delta u)_h$. On prolonge u_h et $\bar{u}_h$ sur le bord $(\bar{\Omega})_h \setminus \Omega_h$ par les conditions aux limites discrétisées $g(P_{i,j})$. On sait déjà que

$$\|\varepsilon_h\|_\infty \leq \frac{h^2}{12}\max_{\bar{\Omega}}\Big(\Big|\frac{\partial^4 u}{\partial x^4}\Big| + \Big|\frac{\partial^4 u}{\partial y^4}\Big|\Big).$$

Mais

$$\begin{aligned} -\Delta_h \bar{u}_h &= F_h, \\ -\Delta_h u_h &= F_h + \varepsilon_h, \end{aligned}$$

d'où

$$\begin{cases} -\Delta_h(u_h - \bar{u}_h) = \varepsilon_h \text{ dans } \Omega_h, \\ u_h - \bar{u}_h = 0 \text{ sur } (\bar{\Omega})_h \setminus \Omega_h. \end{cases}$$

Comme $u_h - \bar{u}_h$ s'annule au bord, on peut donc réécrire ceci matriciellement sous la forme

$$A_h(u_h - \bar{u}_h) = \varepsilon_h \Longrightarrow \|u_h - \bar{u}_h\|_\infty \leq |||A_h^{-1}|||_\infty \|\varepsilon_h\|_\infty.$$

Ceci termine la démonstration. □

Remarques. i) On a fait apparaître (pour une fois) le facteur d'échelle a pour faire remarquer que $\frac{a^2h^2}{24} = \frac{a^4}{24(N+1)^2}$ est un facteur qui croît *très* vite avec a à nombre total de points fixé. On peut donc s'attendre à une rapide dégradation de la précision sur de grands ouverts, à nombre de points fixé.

ii) On peut étendre ces idées en dimension 3, 4 ou plus sans difficulté de principe. Toutefois, les matrices se compliquent et surtout leur taille est de l'ordre de N^3, N^4 et ainsi de suite. Cette augmentation de taille exponentielle par rapport à la dimension devient vite prohibitive du point de vue pratique.

iii) Attention à l'hypothèse de régularité $u \in \mathcal{C}^4(\bar{\Omega})$. Comme on l'a déjà indiqué en passant, elle ne va pas de soi en dimension 2... □

Faisons maintenant une autre excursion vers la méthode des différences finies pour le problème dépendant du temps. La méthode est la même mais il faut l'adapter à un problème de nature différente.

2.5 Différences finies pour le problème en temps

On considère maintenant le problème dépendant du temps

$$\begin{cases} \partial_t u - \mu \partial_{xx} u = f(x,t) \quad x \in]0,1[, 0 < t \leq T, \\ \\ u(0,t) = 0, \quad u(1,t) = 0, \quad 0 \leq t \leq T, \\ u(x,0) = u_0(x), x \in]0,1[, \end{cases} \tag{2.9}$$

c'est l'équation de la chaleur, ou équation de diffusion. La fonction f et la condition initiale u_0 sont données et $\mu > 0$ est une constante donnée. On suppose que la donnée initiale et les données au bord sont compatibles, c'est à dire que la donnée initiale vérifie les conditions aux limites $u_0(0) = u_0(1) = 0$. On ne cherche pas ici à étudier cette edp dans le cadre général (avec des données générales, par exemple des conditions aux limites non constantes, i.e. dépendant du temps, $u(0,t) = \alpha(t), u(1,t) = \beta(t)$) mais à comprendre comment construire un schéma aux différences finies pour en approcher la solution. On va se placer dans un cas particulier où il est assez simple de construire une solution. Pour pouvoir faire l'analyse numérique d'une méthode, il est en effet utile de savoir que la solution qu'elle est censée approcher existe dans un certain espace, et est unique. Dans le cas $f = 0$, on peut effectivement monter le résultat d'existence et d'unicité suivant (qui utilise un développement en série de Fourier).

Proposition 2.5.1 *Soit $u_0 \in \mathcal{C}^2([0,1])$ vérifiant $u_0(0) = u_0(1) = 0$. Le problème (2.9) avec $f = 0$ a une solution $u \in \mathcal{C}^0([0,1] \times [0,T]) \cap \mathcal{C}^1(]0,1[\times]0,T])$, $\partial_{xx} u \in \mathcal{C}^0([0,1]\times]0,T])$ et une soution ayant une telle régularité est unique.*

Pour l'approximation par différences finies de la solution de (2.9), en plus de la grille uniforme en espace, on introduit une grille en temps : on se donne un entier $M > 0$ et on pose $\Delta t = \frac{T}{M}$, Δt est le pas de temps, et $t_n = n\Delta t$, $0 \leq n \leq M$. Les points de la grille espace-temps sont donc les points (x_j, t_n), $0 \leq j \leq N+1, 0 \leq n \leq M$. Pour une fonction $\phi(x,t)$ quelconque définie sur $[0,1] \times [0,T]$, on posera $\phi_j^n = \phi(x_j, t_n)$. On cherche à calculer une solution approchée qu'on notera plutôt v_j^n que $\overline{u}_j^n$ de la solution exacte $u_j^n = u(x_j, t_n)$ en ces points de grille grâce à un schéma aux différences finies. On garde la même discrétisation pour le "Laplacien" $\partial_{xx} u$, mais suivant la formule utilisée pour approcher la dérivée en temps, on obtient des schémas différents. Le premier schéma est le suivant :

$$\begin{cases} \dfrac{v_j^{n+1} - v_j^n}{\Delta t} - \mu \dfrac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{h^2} = f_j^n, \\ \qquad\qquad 1 \leq j \leq N, 0 \leq n \leq M, \\ v_0^n = 0, \quad v_{N+1}^n = 0, 0 \leq n \leq M, \\ \\ v_j^0 = u_0(x_j), \quad 0 \leq j \leq N+1, \end{cases} \tag{2.10}$$

où $f_j^n = f(x_j, t_n)$. Il est obtenu en prenant l'équation exacte en (x_j, t_n), en discrétisant le Laplacien par différences finies comme on l'a vu au début de ce chapitre et en remplaçant la dérivée en temps par une différence finie progressive

$$\partial_t u(x_j, t_n) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t}.$$

C'est un schéma explicite en temps : si on pose

$$r = \mu \frac{\Delta t}{h^2}, \tag{2.11}$$

l'équation aux différences du schéma (2.10) s'écrit

$$v_j^{n+1} = (1-2r)v_j^n + r(v_{j+1}^n + v_{j-1}^n) + \Delta t f_j^n, \tag{2.12}$$

qui est une formule *explicite* permettant de calculer v_j^{n+1} à partir des valeurs $v_{j+1}^n, v_j^n, v_{j-1}^n$ connues au temps t_n, sans nécessité d'inverser une fonction ou un système.

Posons pour $0 \leq n \leq M$, $v_h^n = (v_1^n, v_2^n, ..., v_N^n)^T \in \mathbb{R}^N$, $b_h^n = (f_1^n, f_2^n, ..., f_N^n)^T \in \mathbb{R}^N$ et définissons la matrice tridiagonale

$$Q_1(r) = \begin{pmatrix} 1-2r & r & 0 & \dots & \dots & 0 \\ r & 1-2r & r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & r & 1-2r & r \\ 0 & \dots & \dots & 0 & r & 1-2r \end{pmatrix}$$

la formule (2.12) écrite pour $j = 1, ..., N$ donne la relation

$$v_h^{n+1} = Q_1(r)v_h^n + \Delta t b_h^n,$$

et permet le calcul des vecteurs v_h^n, $0 < n \leq M$ à partir d'une donnée initiale qu'on a prise exacte $v_h^0 = u_h^0$ vecteur de $\mathbb{R}^N$ de composantes $(u_0(x_j), 1 \leq j \leq N)$.

Si au contraire, on considère une différence finie régressive (qu'on écrit alors au temps t_{n+1})

$$\partial_t u(x_j, t_{n+1}) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t},$$

en prenant l'équation exacte en (x_j, t_{n+1}), on obtient le deuxième schéma

$$\begin{cases} \dfrac{v_j^{n+1} - v_j^n}{\Delta t} - \mu \dfrac{v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}}{h^2} = f_j^{n+1}, \\ \qquad\qquad 1 \leq j \leq N, 0 \leq n \leq M, \\ v_0^n = 0, \quad v_{N+1}^n = 0, 0 \leq n \leq M \\ \\ v_j^0 = u_0(x_j), \quad 0 \leq j \leq N+1, \end{cases} \tag{2.13}$$

qui, lui, est un schéma *implicite*. On peut écrire l'équation aux différences du schéma (2.13) avec la notation (2.11) sous la forme

$$(1+2r)v_j^{n+1} - r(v_{j+1}^{n+1} + v_{j-1}^{n+1}) = v_j^n + \Delta t f_j^{n+1}, 1 \leq j \leq N, \tag{2.14}$$

et on a un système de matrice tridiagonale à résoudre pour calculer les N composantes du vecteur $v_h^{n+1} = (v_j^{n+1})$, $1 \leq j \leq N$. On définit la matrice $N \times N$

$$Q_2(r) = \begin{pmatrix} 1+2r & -r & 0 & \dots & \dots & 0 \\ -r & 1+2r & -r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -r & 1+2r & -r \\ 0 & \dots & \dots & 0 & -r & 1+2r \end{pmatrix}.$$

La formule (2.14) écrite pour $j = 1, ..., N$ donne le système linéaire

$$Q_2(r) v_h^{n+1} = v_h^n + \Delta t b_h^n.$$

On peut prendre une différence finie centrée,

$$\partial_t u(x_j, t_{n+1/2}) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t},$$

on prend alors l'équation exacte en $(x_j, t_{n+1/2})$ puis on approche les valeurs $u(x_j, t_{n+1/2})$ dans

$$\partial_{xx} u(x_j, t_{n+1/2}) \sim \frac{u(x_{j+1}, t_{n+1/2}) - 2u(x_j, t_{n+1/2}) + u(x_{j-1}, t_{n+1/2})}{h^2}$$

par des moyennes $u(x_j, t_{n+1/2}) \sim \frac{1}{2}(u(x(j, t_n) + u(x_j, t_{n+1}))$ et on obtient un troisième schéma

$$\begin{cases} \dfrac{v_j^{n+1} - v_j^n}{\Delta t} - \mu \dfrac{(v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}) + (v_{j+1}^n - 2v_j^n + v_{j-1}^n}{2h^2} = f_j^{n+1/2} \\ \qquad\qquad 1 \leq j \leq N, 0 \leq n \leq M \\ v_0^n = 0, \quad v_{N+1}^n = 0, 0 \leq n \leq M \\ \\ v_j^0 = u_0(x_j), \quad 0 \leq j \leq N+1, \end{cases} \tag{2.15}$$

où $f_j^{n+1/2} = f(x_j, (n+1/2)\Delta t)$ ou on pourrait aussi prendre l'approximation $\frac{1}{2}(f_j^n + f_j^{n+1})$. Ce schéma (appelé schéma de Crank-Nicolson) est également implicite. En introduisant encore deux matrices $N \times N$ tridiagonales, disons Q_3, Q_4 (le faire), on obtient une relation de la forme

$$Q_3 v_h^{n+1} = Q_4 v_h^n + \Delta t f_h^{n+1/2}.$$

Remarque. Dans les trois exemples, on vérifie que les schémas obtenus correspondent à avoir utilisé le schéma d'Euler (explicite ou implicite) ou celui de Crank-Nicolson pour l' EDO

$$\begin{cases} \dfrac{d\underline{u}_h}{dt}(t) + \mu A_{0h} \underline{u}_h(t) = f_h(t) \quad 0 < t \leq T \\ \\ \underline{u}_h(0) = u_h^0 \end{cases} \tag{2.16}$$

où A_{0h} est définie par (2.7), les N composantes de $\underline{u}_h(t)$ approchent les $u(x_i,t)$ et u_h^0 est le vecteur de composantes $u_0(x_j)$. Ce système d'équations différentielles (2.16) est celui qui résulte de la semi-discrétisation en espace par la méthode des différences finies en espace de la section 2.1, le temps t étant considéré comme un "paramètre", on parle aussi de "méthode des lignes". On discrétise ensuite en temps, en calculant un vecteur v_h^n qui approche le vecteur $\underline{u}_h(t_n)$.

On pourrait aussi commencer par semi-discrétiser le problème en temps, x étant considéré comme un "paramètre". Par exemple si on utilise la méthode d'Euler implicite, on obtient

$$\left\{\begin{array}{l} \dfrac{\underline{u}^{n+1}(x)-\underline{u}^n(x)}{\Delta t}-\mu\dfrac{d^2\underline{u}^{n+1}}{dx^2}(x)=f^{n+1}(x) \quad x\in]0,1[,\ 0\le n\le M-1 \\ \underline{u}^n(0)=\underline{u}^n(1)=0,\ 0\le n\le M \\ \underline{u}^0(x)=u_0(x),\ x\in]0,1[, \end{array}\right. \tag{2.17}$$

où $f^n(x)=f(x,t_n)$ et la fonction $\underline{u}^n(x)$ approche $u(x,t_n)$. On utilise ensuite la méthode des différences finies en espace pour (2.17), elle consiste à prendre l'équation aux points x_j, en remplaçant la dérivée exacte par une différence finie comme à la section 2.1. On obtient alors le schéma 2 (2.13) où v_h^n approche le vecteur $\underline{u}^n(x_i)$.

Cette remarque va aider à comprendre comment on pourra relier l'ordre en temps et en espace de la méthode résultant des deux discrétisations, aux ordres respectifs de chaque méthode de semi discrétisation. Cela permet aussi de comprendre comment on peut généraliser en utilisant d'autres méthodes de discrétisation en temps (par exemple une méthode de Runge-Kutta) ou en espace (par exemple une méthode d'éléments finis). □

Pour les schémas 2 et 3, on vérifie que les matrices Q_2 et Q_3 sont inversibles (pourquoi ? faire le lien avec la matrice A_h de (2.3)). Les trois schémas précédents peuvent donc tous se mettre sous la forme générale

$$v_h^{n+1}=Qv_h^n+\Delta t g_h^n,\ 0\le n\le M-1, \tag{2.18}$$

où $Q=Q(r)$ est une matrice $N\times N$ connue, et g_h^n un vecteur de $\mathbb{R}^N$ connu à partir des données du problème, et $M\Delta t\le T$. La condition initiale est $v_h^0=u_h^0$, i.e. $v_h^0=(u_1^0,...,u_N^0)^T$, où $u_j^0=u^0(x_j)$. Il s'agit d'un schéma à deux niveaux en temps (seules les valeurs en t_n interviennent pour le calcul au temps t_{n+1}), on dit aussi schéma à un pas de temps.

Notons que si les conditions aux limites dans (2.9) ne sont pas 0 mais $u(0,t)=\alpha(t), u(1,t)=\beta(t)$, ces valeurs une fois discrétisées en $\alpha(t_n),\beta(t_n)$ interviendront dans les composantes du terme au second membre g_h^n.

Introduisons pour ces schémas les notions de consistance, stabilité et convergence. Pour cela, on considère la solution exacte $u_j^n=u(x_j,t_n)$ aux points de la grille et le vecteur associé $u_h^n=(u_1^n,u_2^n,...,u_N^n)^T\in\mathbb{R}^N$. Alors u_h^n ne vérifie pas le schéma (sinon on saurait calculer la solution exacte en n'importe quel point) mais on peut écrire

$$u_h^{n+1}=Qu_h^n+\Delta t g_h^n+\Delta t\tau_h^n,$$

ce qui définit le vecteur τ_h^n, c'est l'*erreur de consistance* locale

$$\tau_h^n=\frac{1}{\Delta t}(u_h^{n+1}-Qu_h^n-\Delta t g_h^n).$$

Enfin, on considère une norme vectorielle $\|\ .\|$ sur $\mathbb{R}^N$ et la norme matricielle associée $|||.|||$.

Définition 2.5.1 *On dit que le schéma (2.18) est*
i) consistant avec l'edp (2.9) si pour toute solution u de (2.9)

$$\sup_{m;m\Delta t\leq T}\|\tau_h^m\|\to 0 \text{ quand } \Delta t,h\to 0.$$

ii) La méthode est d'ordre (p,q) (p en espace et q en temps) si, pour toute solution u suffisamment régulière de (2.9) , il existe une constante $C=C(u)>0$ telle que

$$\sup_{m;m\Delta t\leq T}\|\tau_h^m\|\leq C(h^p+\Delta t^q).$$

iii) Le schéma est stable (pour la norme $\|\ .\|$) s'il existe une constante K_0 (qui peut dépendre de T) telle que

$$\sup_{m;m\Delta t\leq T}|||Q^m|||\leq K_0.$$

iv) Le schéma est convergent si

$$\sup_{m;m\Delta t\leq T}\|u_h^m-v_h^m\|\to 0 \text{ quand } \Delta t,h\to 0.$$

Remarques. i) Attention aux *notations* : dans τ_h^n, v_h^n, qui sont des vecteurs de $\mathbb{R}^N$, n est un indice de temps ; dans Q^m, où $Q=Q(r)$ est une matrice $N\times N$, pour m il s'agit de la puissance $Q^2=QQ$, ... Si on utilisait une méthode à pas variable, dans laquelle $\Delta t_n\equiv t_{n+1}-t_n$ n'est pas nécessairement constant, la matrice Q dans (2.18) dépendrait de n, soit $Q^{(n)}$, et on aurait alors un produit $Q^{(m)}Q^{(m-1)}...Q^{(1)}$ à la place de la puissance Q^m.

ii) La régularité demandée pour la consistance est celle de la proposition 2.5.1 ; pour estimer l'ordre, on a besoin de supposer que la solution est plus régulière pour pouvoir faire des développements limités, voir la proposition 2.5.2 ci-dessous. Bien qu'on n'ait pas donné de résultat général de régularité de la solution par rapport aux données (qui serait un analogue du théorème 1.5.2 dans le cas de l'équation de la chaleur), cette régularité peut effectivement être atteinte et découler d'hypothèses de régularité sur les données.

iii) Même si on ne l'a pas rappelé, la norme utilisée dans $\mathbb{R}^N$ doit être telle que les quantités continuent à avoir un sens quand $N\to\infty$, par exemple la norme 'infini' ou la norme 2 discrète introduite dans la définition 2.2.2 ; notons que le facteur h change la valeur de la norme mais pas la norme matricielle associée. □

Théorème 2.5.1 *Si le schéma est consistant et stable, alors il est convergent.*

Démonstration. On introduit l'erreur $e_h^n=u_h^n-v_h^n, 0\leq n\leq M=T/\Delta t$. Alors $e_h^0=0$ par choix de la condition initiale, et des deux relations

$$u_h^{n+1}=Qu_h^n+\Delta t g_h^n+\Delta t\tau_h^n,\ v_h^{n+1}=Qv_h^n+\Delta t g_h^n$$

on obtient par différence, pour $n\leq M-1$,

$$e_h^{n+1}=Qe_h^n+\Delta t\tau_h^n,$$

et en itérant le processus, puisque $e_h^0 = 0$, on obtient

$$e_h^{n+1} = \Delta t \sum_{k=0}^{n} Q^k \tau_h^{n-k}.$$

On déduit

$$\|e_h^{n+1}\| \leq \Delta t \sum_{k=0}^{n} |||Q^k||| \, \|\tau_h^{n-k}\|.$$

Si la méthode est stable

$$\|e_h^{n+1}\| \leq K_0 (n+1) \Delta t \sup_{0 \leq k \leq n} \|\tau_h^{n-k}\|,$$

et donc, comme cela est valable pour tout $n \leq M-1$, avec $M\Delta t \leq T$

$$\sup_{m; m\Delta t \leq T} \|e_h^m\| \leq K_0 T \sup_{n; n\Delta t \leq T} \|\tau_h^n\|,$$

et si la méthode est consistante, le second membre tend vers 0 quand $\Delta t, h \to 0$. □

Remarques. i) En calquant la démonstration précédente on montrerait que si on a une perturbation δ_h^0 sur la donnée initiale et que le schéma calcule une suite de valeurs w_h^n solution d'un schéma perturbé

$$w_h^{n+1} = Q w_h^n + \Delta t g_h^n + \Delta t \delta_h^n,$$

avec $w_h^0 = \eta_h^0 + \delta_h^0$, si le schéma est stable, la différence $\|w_h^n - v_h^n\|$ reste bornée en fonction des perturbations $\|\delta_h^0\|$ et $\max_{n \leq M} \|\delta_h^n\|$, ce qui explique le terme de *stabilité*.

ii) Le principe "stabilité + consistance implique convergence" est très général dans les méthodes de discrétisation. □

Appliquons ce résultat aux schémas 1 et 2. Il faut vérifier la consistance et la stabilité ; commençons par étudier le **schéma 1**.

Proposition 2.5.2 *Supposons que la solution u du problème (2.9) vérifie : $u \in \mathcal{C}^0([0,1] \times [0,T] \cap \mathcal{C}^1(]0,1[\times]0,T])$, et $\frac{\partial^4 u}{\partial x^4}, \frac{\partial^2 u}{\partial t^2} \in \mathcal{C}^0([0,1]\times]0,T])$. Alors, pour le schéma (2.10), $u_j^n = u(x_j, t_n)$ $1 \leq j \leq N, n \leq M-1$, satisfait*

$$u_j^{n+1} = (1-2r)u_j^n + r(u_{j+1}^n + u_{j-1}^n) + \Delta t f_j^n + \Delta t \tau_j^n,$$

avec

$$\sup_{m; m\Delta t \leq T} \|\tau_h^m\|_{N,\infty} \leq C(h^2 + \Delta t)$$

où la constante C dépend de $\max_{x \in [0,1], t \in [0,T]} |\frac{\partial^4 u}{\partial x^4}|$, $\max_{x \in [0,1], t \in [0,T]} |\frac{\partial^2 u}{\partial t^2}|$.

Démonstration. L'erreur de consistance τ_j^n vérifie

$$u_j^{n+1} = u_j^n + \mu \frac{\Delta t}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t f_j^n + \Delta t \tau_j^n,$$

donc

$$\tau_j^n = \frac{1}{\Delta t}\left(u_j^{n+1} - u_j^n\right) - \mu \frac{1}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) - f_j^n$$

$$= \frac{1}{\Delta t}\big(u(x_j,t_{n+1}) - u(x_j,t_n)\big) - \mu\frac{1}{h^2}\Big(u(x_{j+1},t_n) - 2u(x_j,t_n) + u(x_{j-1},t_n)\Big) - f(x_j,t_n).$$

On utilise de nouveau la formule de Taylor-Lagrange. Pour une fonction ϕ supposée de classe $\mathcal{C}^2$ sur $[0,T]$, on peut écrire : pour tout $n \in \{0,\dots,M-1\}$, il existe un nombre $\theta^{(n)} \in]0,1[$ tel que

$$\phi(t_{n+1}) = \phi(t_n) + \Delta t\phi'(t_n) + \frac{\Delta t^2}{2}\phi''(t_n + \theta^{(n)}\Delta t),$$

on applique le résultat à $\phi(t) = u(x_j,t)$, et alors $\phi'(t_n) = \partial_t u(x_j,t_n)$. De même on applique le théorème 2.1.1 à $\varphi(x) = u(x,t_n)$ et alors $\varphi''(x_j) = \partial_{xx}u(x_j,t_n)$. Comme u est solution de (2.9),

$$\partial_t u(x_j,t_n) - \mu\partial_{xx}u(x_j,t_n) = f_j^n$$

donc

$$\tau_j^n = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x_j,t_n + \theta^{(n)}\Delta t) - \mu\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j + \theta_j h, t_n),$$

pour des $\theta_j, \theta^{(n)} \in]-1,1[$, d'où le résultat. On a donc montré que le schéma est d'ordre 2 en espace et 1 en temps comme on s'y attendait puisqu'il résulte de l'utilisation d'un schéma de différences finies en espace d'ordre 2 (associé au Laplacien discret) et d'un schéma de différences finies en temps d'ordre 1 (la méthode d'Euler). □

Pour la stabilité, on a le résultat suivant.

Proposition 2.5.3 *Supposons $0 < r \leq 1/2$. Alors le schéma est stable pour la norme $|||.|||_\infty$ et pour $|||.|||_2$. Plus précisément $|||Q_1(r)|||_\infty = 1$ et $|||Q_1(r)|||_2 < 1$.*

Démonstration. Considérons d'abord la norme $|||.|||_\infty$. Si $0 < r \leq 1/2$, alors les coefficients de $Q_1(r)$ sont positifs, et par le théorème 1.3.1, $|||Q_1(r)|||_\infty = 1$, ce qui entraîne la stabilité pour cette norme. Pour la norme $|||.|||_2$, comme Q_1 est symétrique, le même théorème donne $|||Q_1|||_2 = \max_j |\lambda_j(Q_1)|$, or les valeurs propres λ_j de Q_1 peuvent se calculer et on trouve

$$\lambda_j(Q_1) = 1 - 4r\sin^2\Big(\frac{j\pi}{2(N+1)}\Big),\ 1 \leq j \leq N.$$

Si $0 < r \leq 1/2$, on a $|\lambda_j(Q_1)| < 1$, et la méthode est stable. □

Sous la condition de stabilité $0 < r \leq 1/2$, c'est à dire si le pas de temps vérifie

$$0 < \Delta t \leq h^2/2\mu, \tag{2.19}$$

la méthode est donc convergente et

$$\sup_{m;m\Delta t \leq T} \|u_h^m - v_h^m\|_p \leq c(h^2 + \Delta t),$$

pour $p = 2$ et $p = \infty$, où c ne dépend que de u.

D'après l'estimation de l'erreur de consistance, on voit que la méthode est d'ordre 2 (en h) et 1 (en Δt). donc l'erreur est finalement en $O(h^2)$ et n'est pas détériorée par l'ordre 1 en temps (on raisonne avec $\mu > 0$ fixé, qui n'est a priori pas très petit devant h). Cette estimation suppose cependant une condition qui limite le pas de temps, ici (2.19), on dit que la méthode est *conditionnellement stable*, c'est la conséquence de son caractère explicite.

Schéma 2. On écrit maintenant que la solution exacte vérifie

$$(1+2r)u_j^{n+1} - r(u_{j+1}^{n+1} + u_{j-1}^{n+1}) = u_j^n + \Delta t f_j^{n+1} + \Delta t \overline{\tau}_j^n,$$

ce qui définit l'erreur de consistance $\overline{\tau}_j^n$ et on a encore

$$|\overline{\tau}_j^n| \leq C(h^2 + \Delta t),$$

où la constante C dépend de $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^4 u}{\partial x^4}|$, $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^2 u}{\partial t^2}|$. Si on définit alors la matrice Q par $Q = Q_2^{-1}(r)$, et si on pose $\tau_h^n = Q\overline{\tau}_h^n$, $g_h^n = Qb_h^{n+1}$, le schéma peut se mettre sous la forme générale (2.18).

La stabilité résulte de la proposition suivante.

Proposition 2.5.4 *Le schéma (2.13) est stable pour la norme* $|||.|||_\infty$ *et pour* $|||.|||_2$. *Plus précisément* $|||Q|||_\infty \leq 1$ *et* $|||Q|||_2 < 1$.

Démonstration. Considérons d'abord la norme $|||.|||_\infty$. Par définition de Q, pour $v, w \in \mathbb{R}^N$,

$$w = Qv \Leftrightarrow v = Q_2 w.$$

On peut alors écrire

$$w_i = \frac{r}{1+2r}(w_{i-1} + w_{i+1}) + \frac{v_i}{1+2r},\ 1 \leq i \leq N.$$

Soit i l'indice de $\{1,..N\}$ tel que $\|w\|_\infty = |w_i|$, alors

$$\|w\|_\infty = |w_i| \leq \frac{r}{1+2r} 2\|w\|_\infty + \frac{\|v\|_\infty}{1+2r}$$

ce qui entraine

$$\|Qv\|_\infty = \|w\|_\infty \leq \|v\|_\infty$$

pour tout $v \in \mathbb{R}^N$, et donc $|||Q|||_\infty \leq 1$ ce qui implique la stabilité pour cette norme. Pour la norme $|||.|||_2$, Q_2 comme Q_1 est symétrique, et on peut aussi calculer ses valeurs propres

$$\lambda_j(Q_2) = 1 + 4r\sin^2(\frac{j\pi}{2(N+1)}),\ 1 \leq j \leq N.$$

Le même théorème donne $|||Q|||_2 = \max_j |\lambda_j(Q)|$, or les valeurs propres de Q sont les inverses des $\lambda_j(Q_2)$, donc les valeurs $(1+4r\sin^2(\frac{j\pi}{2(N+1)}))^{-1}$, $1 \leq j \leq N$. On a $|\lambda_j(Q)| < 1$, donc $|||Q|||_2 \leq 1$ et la méthode est stable pour cette norme. □

Le schéma 2 qui est implicite est (inconditionnellement) stable, il est aussi consistant, donc il est convergent.

Dans les deux "excursions" précédentes, pour lesquelles on a introduit une nouvelle variable (y en espace ou t en temps) au problème aux limites en dimension un, la même *méthode des différences finies* a été utilisée, mais les schémas en résultant sont sensiblement différents, parce que les problèmes continus sont de nature différente, la variable temps (pour le problème d'évolution) ne jouant pas le même rôle que la variable d'espace (pour le problème aux limites), et aussi l'ordre des opérateurs de dérivation (∂_t et ∂_{yy}) n'est pas le même.

Notons aussi que si les notions de consistance, de stabilité et de convergence sont essentiellement les mêmes, il faut néanmoins adapter la définition précise au contexte du problème considéré (voir par exemple les définitions 2.1.1 et 2.5.1).

Chapitre 3

Approximation numérique de l'équation de transport

L'objet de ce chapitre est d'étudier sur un exemple simple, l'équation de transport (ou d'advection) linéaire, un autre type de méthodes d'approximation numériques, les méthodes de *volumes finis*. Avec des hypothèses simplifiées, elles ont des points communs avec les méthodes de différences finies, mais se prêtent à d'autres généralisations. Nous verrons ensuite, dans les chapîtres ultérieurs, la méthode des *éléments finis*, qui suppose d'avoir au préalable introduit l'approximation variationnelle ; elle sera bien adaptée à l'étude et l'approximation des problèmes aux limites du second ordre que nous reprendrons à ce moment là.

3.1 Équation d'advection

Le problème de Cauchy pour l'équation d'advection s'écrit

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0,\ x \in \mathbb{R}, t > 0 \\ u(x,0) = u_0(x),\ x \in \mathbb{R}, \end{cases} \tag{3.1}$$

où a, qui représente la vitesse de propagation, est une constante donnée. Si u est une quantité par unité de longueur, le terme $Q = au$ représente un débit, quantité par unité de temps. La donnée initiale, u_0 est supposée connue. Dans l'écriture de l'équation aux dérivées partielles, les variables x,t semblent jouer le même rôle, mais dans l'interpétation, les variables de temps t et d'espace x ne jouent pas le même rôle. On regarde l'évolution de la solution avec le temps à partir d'une condition initiale donnée. De ce fait, l'analyse de l'équation continue et les méthodes d'approximation vont respecter cette différence. En particulier, on fait une hypothèse implicite que l'évolution en temps est continue, alors qu'il n'est pas nécessaire de supposer la continuité en espace de u_0. La solution peut être cherchée au sens classique, dans l'espace $\mathcal{C}^1(\mathbb{R} \times (0,\infty))$. On peut aussi chercher une solution pour laquelle les dérivées sont des dérivées au sens faible.

Soit une fonction test $\mathcal{C}^1$ en les variables x,t ; on dit qu'elle est à support compact dans $\mathbb{R} \times [0,\infty)$, si elle est la restriction à $\mathbb{R} \times [0,\infty)$ d'une fonction à support compact dans un ouvert de la forme $\mathbb{R} \times]-\varepsilon,\infty), \varepsilon > 0$, on notera $\mathcal{C}^1_c(\mathbb{R} \times [0,\infty))$ l'ensemble correspondant. On introduit la notion de solution faible en suivant une démarche analogue à celle que nous suivrons

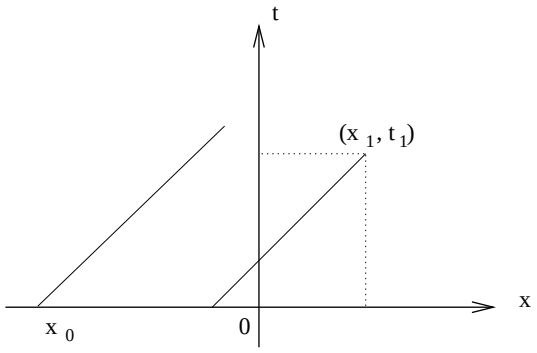


FIGURE 3.1 – Caractéristiques dans le plan (x,t)

pour définir la formulation variationnelle d'un problème aux limites, démarche adaptée à cette nouvelle équation. On suppose d'abord qu'on a une solution classique, on multiplie l'équation (3.1) par une fonction test et on intègre sur le domaine $\mathbb{R}\times[0,\infty)$. On remarque ensuite qu'on peut étendre la formule à des fonctions moins régulières, et pour des dérivées définies au sens faible (nous détaillerons cette notion plus tard). Dans la définition qui suit, on vérifie qu'il n'est pas nécessaire de se placer dans l'espace de Sobolev H^1 (espace que nous définirons pour la formulation variationnelle) pour que l'intégrale ait un sens.

Définition 3.1.1 *Soit $u_0 \in L^\infty(\mathbb{R})$ donnée. Une solution faible de (3.1) est une fonction $u \in L^\infty(\mathbb{R}\times(0,\infty))$ qui vérifie pour toute fonction test $\varphi \in \mathcal{C}^1_c(\mathbb{R}\times[0,\infty))$*

$$\int_0^\infty \int_{\mathbb{R}} \{u\frac{\partial\varphi}{\partial t} + au\frac{\partial\varphi}{\partial x}\}(x,t)dxdt + \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0.$$

On montre que ce problème a une solution faible unique. Nous n'étudierons pas plus en détail ces aspects théoriques généraux.

Si u_0 est donnée dans $\mathcal{C}^1(\mathbb{R})$, une solution faible est une solution au sens classique. La solution est simplement

$$u(x,t) = u_0(x-at). \tag{3.2}$$

Il est facile de vérifier que la fonction (3.2) est effectivement solution. Si l'on suppose la solution régulière, il n'est pas difficile de montrer que c'est la seule. En effet, une propriété importante des équations de transport est que la solution est constante sur les *caractéristiques*. Une caractéristique est par définition une droite $t \to X(t;x_0,t_0)$ passant par un point donné (x_0,t_0), définie par

$$\begin{cases} \dfrac{dX}{dt} = a \\ X(t_0) = x_0. \end{cases} \tag{3.3}$$

Lorsque $t_0 = 0$, on résout pour $t > 0$, le point $(x_0,0)$ est *le pied* de la caractéristique, l'équation de la droite est $x - x_0 = a(t-t_0)$, donc $X(t;x_0,t_0) = x_0 + a(t-t_0)$. On peut aussi considérer une caractéristique rétrograde $X(t;x_1,t_1)$ passant par un point disons (x_1,t_1), c'est à dire pour $t < t_1$ lorsque $t_1 > 0$ est donné (voir la figure 3.1, notons que pour l'étude des problèmes de transport, on fait souvent une représentation de ce type dans le plan (x,t)). Le pied de la caractéristique est $X(0;x_1,t_1)$.

Proposition 3.1.1 *Soit $u \in \mathcal{C}^1(\mathbb{R}\times[0,\infty))$ une solution de (3.1). La restriction de u à une droite caractéristique est constante.*

Démonstration. Soit u une fonction $\mathcal{C}^1$et $\tilde{u}(t) = u(X(t),t)$ sa restriction à une caractéristique. La variation de $\tilde{u}(t)$ est donnée par

$$\frac{d\tilde{u}}{dt}(t) = \Big\{\frac{\partial u}{\partial x}\frac{dX}{dt} + \frac{\partial u}{\partial t}\Big\}(X(t),t) = \Big\{\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}\Big\}(X(t),t)$$

donc si u est solution de (3.1), $\tilde{u}' = 0$ et

$$\tilde{u}(t) = \tilde{u}(t_0) = u(X(t_0;x_0,t_0),t_0) = u(x_0,t_0) = u_0(x_0),$$

si on choisit $t_0 = 0$, donc u est constant sur la caractéristique et égal à la valeur de u au pied de la caractéristique, valeur qui est donnée par la condition initiale. □

On peut alors construire la solution à partir de u_0 en propageant les valeurs initiales le long des caractéristiques ce qui se traduit par $u(x,t) = u_0(X(0;x,t))$, ce qui donne (3.2). La condition initiale est simplement transportée à la vitesse a, vers la droite si $a > 0$, vers la gauche si $a < 0$, elle est stationnaire si $a = 0$. La solution vérifie de façon évidente les estimations

$$\|u(.,t)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}, \text{ et } \|u(.,t)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})}.$$

Si on s'intéresse maintenant au problème posé dans un intervalle borné en espace, par exemple $x \in [0,L]$, on a un problème de condition initiale (u_0 est donnée dans $L^\infty(0,L)$) et aux limites. Notons que comme l'opérateur de dérivation en espace est d'ordre 1, intuitivement on ne peut se fixer des conditions aux limites aux deux extrémités, 0 et L, mais seulement à une des deux. Effectivement, si $a > 0$, pour que le problème soit bien posé il faut se donner $u(0,t) = g(t)$ et cela suffit alors à déterminer u dans la bande $[0,L]\times(0,\infty)$ (voir figure 3.2, à gauche). On utilise encore que u est constante sur les caractéristiques dans la bande $[0,L]\times(0,\infty)$. On vérifie que la formule que l'on obtient pour $u(x,t)$, pour $t > 0$, est la suivante

$$u(x,t) = \begin{cases} u_0(x-at),\ at < x \leq L, \\ g(t-\frac{x}{a}),\ 0 \leq x < \min(at,L). \end{cases}$$

Par contre si $a < 0$, il faut se donner $u(L,t) = h(t)$, et alors pour $t > 0$,

$$u(x,t) = \begin{cases} u_0(x-at),\ 0 \leq x < L+at, \\ h(t-\frac{x-L}{a}),\ \max(0,L+at) < x \leq L. \end{cases}$$

On peut aussi se donner des conditions aux limites périodiques $u(0,t) = u(L,t)$, ce qui signifie que "ce qui rentre est ce qui sort". Par exempe si $a > 0$, la condition initiale permet de connaître $u(x,t) = u_0(x-at)$, $at < x \leq L, 0 < t \leq \frac{L}{a}$, donc de connaître ce qui sort en $x = L$, à savoir $u(L,t) = u_0(L-at)$ pour $0 < t \leq \frac{L}{a}$. La condition de périodicité donne donc la condition sur le bord $x = 0$, $h(t) = u_0(L-at)$ pour $0 < t \leq \frac{L}{a}$, ce qui permet alors de construire $u(x,t)$ dans toute la bande $[0,L]\times(0,\infty)$ par périodicité. On peut aussi le comprendre au sens suivant : si u_0 est périodique de période L sur $\mathbb{R}$, $u(x,t) = u_0(x-at)$ reste périodique en espace et on peut

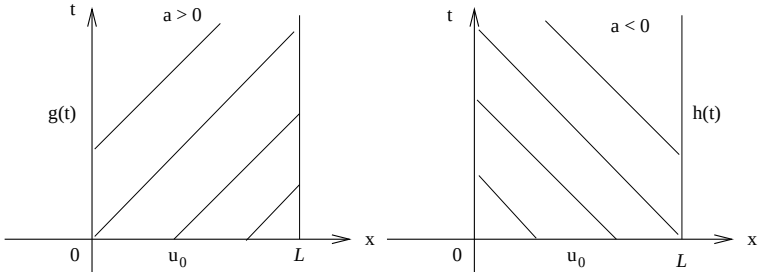


FIGURE 3.2 – Problème posé dans un intervalle $[0,L]$

encore limiter la résolution à $x \in [0,L]$. Elle est aussi périodique en temps de période L/a car $u(x,t+L/a) = u_0(x-at-L) = u_0(x-at) = u(x,t)$.

L'équation (3.4) ayant une solution explicite, on peut se demander pourquoi introduire des méthodes pour l'approcher. La première réponse est que ces méthodes pourront se généraliser (plus ou moins directement) aux équations de transport non linéaires

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0,\ x \in \mathbb{R}, t > 0 \\ u(x,0) = u_0(x), x \in \mathbb{R}, \end{cases} \tag{3.4}$$

où f est une fonction de $\mathcal{C}^2(\mathbb{R})$ donnée, appelée *flux*, on pourra penser à l'exemple classique de l'équation de Burgers pour lequel $f(u) = u^2/2$. Dans le cas linéaire le flux est $f(u) = au$, on retrouve l'équation (3.1). Les équations linéaires servent alors de cas test pour valider les méthodes utilisées dans des cas plus généraux des équations nonlinéaires, voire des systèmes d'edp nonlinéaires, car une première idée pour commencer à étudier ces équations est de les linéariser autour d'un état constant, même si cette approche ne suffit bien sûr pas à comprendre les phénomènes non linéaires.

Remarque. La définition des caractéristiques peut s'étendre au cas des équations de transport avec une vitesse non constante (a dépend de (x,t))

$$\partial_t u + a(x,t)\partial_x u = 0,$$

les caractéristiques définies par (3.3) ne sont plus des droites mais des courbes puisque la pente $dX/dt = a(X(t),t)$ n'est plus constante. Leur existence locale lorsque a est assez régulière (Lipschitz en variable x, uniformément en t) est assurée par le théorème de Cauchy-Lipschitz, on note $X(t;x_0,t_0)$ l'absisse au temps t du point de la caractéristique passant par x_0 au temps t_0. La solution u est encore constante sur une caractéristique. Cette approche permet aussi de résoudre le cas où il y a un terme source (un second membre) $\partial_t u + a(x,t)\partial_x u = S(x,t)$, la variation du u sur une caractéristique est connue, et on intègre donc pour obtenir la valeur de u en tenant compte de la condition initiale

$$u(x,t) = u_0(X(0;x,t)) + \int_0^t S(X(s;x,t),s)ds,$$

le point $X(0;x,t) = x_0$ est le pied de la caractéristique rétrograde passant par un point (x,t) (la caractéristique par (x,t) coupe l'axe $t=0$ en ce point d'abscisse x_0).

Cette approche s'applique aussi au cas des équations nonlinéaires qui est beaucoup plus compliqué. Ce sont encore des équations de transport, mais avec une vitesse qui dépend de l'état u. En effet, supposons que u soit une solution $\mathcal{C}^1$ de (3.4). On peut alors écrire

$$\frac{\partial u}{\partial t} + f'(u)\frac{\partial u}{\partial x} = 0.$$

Notons que cette écriture n'a pas de sens si u est une solution discontinue, alors qu'on peut donner un sens faible à la notion de solution de (3.4), comme dans la définition 3.1.1 ci-dessus, en remplaçant au par $f(u)$. Dans la définition (3.3) des caractéristiques, on remplace a par $f'(u(X(t),t))$ où u est supposée être une solution $\mathcal{C}^1$ de (3.4)

$$\frac{dX}{dt}(t) = f'(u(X(t),t)).$$

On montre encore que la solution u est nécessairement constante le long d'une caractéristique (le faire) et que les caractéristiques sont donc des droites. Mais la vitesse de propagation le long de ces droites (donc leur pente) dépend de l'état u et ces droites ne sont plus parallèles (sauf dans les zones où u est constante). Les difficultés arrivent si ces droites se coupent, mais nous n'aborderons pas ce problème ici. □

3.2 Approximation numérique de l'équation d'advection

3.2.1 Méthode des volumes finis

Nous allons décrire les principes de la méthode des volumes finis, utilisée ici pour la discrétisation en espace. Pour commencer, on s'intéresse donc à la semi-discrétisation en espace et t est un "paramètre" ($t \in [0,T]$). Comme ce sera le cas pour la méthode des éléments finis, on se donne un recouvrement de $\mathbb{R}$ (ou de $[0,L]$) par des cellules. Ici, on notera C_j les cellules qui sont des intervalles deux à deux disjoints, $C_j = (x_{j-1/2}, x_{j+1/2})$, $j \in \mathbb{Z}$, et on cherche à approcher la solution u sur chaque cellule C_j. En dimension 1 d'espace, les cellules sont effectivement de même nature que celles d'un maillage élement fini, en dimension supérieure les contraintes imposées au maillage "volume fini" peuvent être moins fortes que celles imposées dans la méthode d'éléments finis.

On pose $|C_j| = x_{j+1/2} - x_{j-1/2}$, et on définit $u_j(t) = \frac{1}{|C_j|}\int_{C_j} u(x,t)dx$ la moyenne de $u(.,t)$ sur une cellule. On cherche à approcher $u_j(t)$ par une valeur calculable (à t fixé), notée $v_j(t)$. Pour obtenir le schéma, l'idée de la méthode des volumes finis est d'intégrer l'équation dont on approche la solution sur une cellule, notant $Q = au$,

$$\int_{C_j} \Big\{\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}Q\Big\}(x,t)dx = 0$$

ensuite on calcule (ou on approche) chaque terme. Le premier terme donne

$$\int_{C_j} \frac{\partial u}{\partial t}(x,t)dx = \frac{d}{dt}\int_{C_j} u(x,t)dx = |C_j|\frac{du_j}{dt}(t)$$

(l'interversion entre la dérivation et l'intégrale peut se justifier, nous ne détaillons pas ce passage) et le deuxième

$$\int_{C_j} \frac{\partial}{\partial x} Q(x,t)dx = Q(x_{j+1/2},t) - Q(x_{j-1/2},t)$$

donc

$$|C_j|\frac{du_j}{dt}(t) + Q(x_{j+1/2},t) - Q(x_{j-1/2},t) = 0.$$

On ne veut faire intervenir pour le calcul effectif qu'une seule inconnue par cellule, donc que les valeurs $v_j(t) \sim u_j(t)$ (ou les valeurs aux centres des mailles, c'est à dire en x_j), on ne veut pas introduire les valeurs aux interfaces $x_{j+1/2}$ comme nouvelles inconnues. On approche donc le "flux exact" $Q_{j+1/2} = au(x_{j+1/2},t)$ (resp. $Q_{j-1/2} = au(x_{j-1/2},t)$) à travers la frontière $x_{j+1/2}$ (resp. $x_{j-1/2}$) de C_j par "un flux approché", ou flux numérique, fonction (notée g) des deux valeurs de chaque côté :

$$Q_{j+1/2} = au(x_{j+1/2},t) \sim g_{j+1/2} = g(u_j(t), u_{j+1}(t))$$

(respectivement, on approche $q_{j-1/2} = au(x_{j-1/2},t) \sim g(u_{j-1}(t), u_j(t))$ avec la même fonction g) ce qui donne

$$|C_j|\frac{du_j}{dt}(t) + g(u_j(t), u_{j+1}(t)) - g(u_{j-1}(t), u_j(t)) \sim 0.$$

L'ensemble des valeurs à calculer $v_j(t)$ va donc satisfaire le système d'équations différentielles ordinaires obtenu en remplaçant le signe "approche" par le signe égal

$$|C_j|\frac{dv_j}{dt}(t) + g(v_j(t), v_{j+1}(t)) - g(v_{j-1}(t), v_j(t)) = 0.$$

Toutes les propriétés du schéma découlent du choix du flux numérique g, et nous donnerons des exemples un peu plus loin. Il est naturel de supposer que lorsque $v_j = v_{j+1}$, on retombe sur le flux exact av_j, ce qui se traduit par la propriété

$$\forall v \in \mathbb{R},\ g(v,v) = av. \tag{3.5}$$

Cette propriété s'appelle la *consistance* du flux approché avec le flux exact.

Pour simplifier, nous avons pris une fonction g de deux variables, le flux $Q_{j+1/2}$ à l'interface $x_{j+1/2}$ est approché par une fonction des valeurs de chaque côté (des cellules C_j, C_{j+1}), mais il est possible de faire aussi intervenir des valeurs dans des cellules voisines, C_{j-1}, C_{j+2} par exemple.

Pour pouvoir calculer effectivement une approximation de la solution, il faut rajouter un schéma de discrétisation en temps.

3.2.2 Schémas explicites à 3 et 5 points

Rajoutons maintenant une discrétisation en temps de ces EDO par une méthode classique comme par exemple la méthode d'Euler explicite. Soit Δt un pas de temps et $t_n = n\Delta t, n \in \mathbb{N}$

(en fait $n \leq \frac{T}{\Delta t}$). On note v_j^n l'approximation de $u_j^n \equiv \frac{1}{\Delta x}\int_{C_j} u(x,t_n)dx$ que l'on veut calculer. On obtient le schéma explicite

$$\frac{|C_j|}{\Delta t}(v_j^{n+1} - v_j^n) + g(v_j^n, v_{j+1}^n) - g(v_{j-1}^n, v_j^n) = 0.$$

Cette formule est écrite pour tout $n \in \mathbb{N}$ (en fait $n \leq \frac{T}{\Delta t}$) et $j \in \mathbb{Z}$, dans la pratique pour un nombre fini de valeurs j, soit parce qu'on travaille sur un intervalle borné (il faut alors discrétiser les conditions aux limites qui sont données en tenant compte du signe de a), soit parce qu'on se donnera une condition initiale à support compact, et un intervalle suffisamment grand pour ne pas avoir atteint la frontière au temps T. À partir des valeurs initiales v_j^0 que l'on va prendre égales aux moyennes de la donnée initiale

$$v_j^0 = \frac{1}{\Delta x}\int_{C_j} u_0(x)dx, \tag{3.6}$$

on peut calculer pour tout n l'ensemble des valeurs $v_j^n, j \in \mathbb{Z}$. Si on avait utilisé un schéma d'Euler implicite, cela aurait naturellement conduit à une méthode implicite. On peut aussi utiliser d'autres schémas de discrétisation en temps, par exemple utilisant deux niveaux t_n, t_{n-1} au lieu d'un (v_j^{n+1} dépendra alors de valeurs v_j^n et aussi de v_j^{n-1}, voir par exemple le schéma dit *saute-mouton*), ou des méthodes de Runge-Kutta.

Supposons pour simplifier le maillage uniforme, $|C_j| = x_{j+1/2} - x_{j-1/2} = \Delta x$, et $x_j = j\Delta x$. On pose

$$\lambda = \frac{\Delta t}{\Delta x}, \tag{3.7}$$

le schéma devient

$$v_j^{n+1} = v_j^n - \lambda(g(v_j^n, v_{j+1}^n) - g(v_{j-1}^n, v_j^n)). \tag{3.8}$$

Cette formulation du schéma, où on rend apparent un incrément du flux, est appelée *formulation conservative*. On peut aussi utiliser une formulation générale

$$v_j^{n+1} = H(v_{j-1}^n, v_j^n, v_{j+1}^n), \tag{3.9}$$

où on a introduit une fonction H 'solution discrète', notons que la fonction H doit vérifier

$$H(v,v,v) = v$$

(ce qui correspond au fait que les états constants sont propagés exactement par l'opérateur 'solution exacte', c'est à dire : $u_0 = cte \Rightarrow u(x,t) = u_0 = cte$). Un schéma de type (3.9) permet de calculer de façon explicite v_j^{n+1} à partir de trois valeurs, $v_{j-1}^n, v_j^n, v_{j+1}^n$, on dit que c'est un schéma *à trois points*. On peut envisager des schémas à plus de points, par exemple 5 points,

$$v_j^{n+1} = H(v_{j-2}^n, v_{j-1}^n, v_j^n, v_{j+1}^n, v_{j+2}^n),$$

ce qui correspond à prendre un flux numérique g fonction de quatre variables, qui prend en compte deux valeurs de chaque côté de l'interface $g_{j+1/2}^n = g(v_{j-1}^n, v_j^n, v_{j+1}^n, v_{j+2}^n)$, comme dit précedemment ; la consistance du flux se traduit par $g(v,v,v,v) = av$. On doit encore avoir

$H(v,v,v,v,v) = v$, cette condition est vérifiée dès que H est associée à un flux numérique g.

Remarque. La présentation que l'on a faite peut s'étendre au cas d'un flux non linéaire $Q = f(u)$. Dans ce cas il est très important de pouvoir écrire un schéma (3.9) sous la forme dite *conservative* (3.8). La consistance se traduit alors par $g(u,u) = f(u)$, pour toute valeur u (dans le cas d'un schéma à 3 points). □

Dans le cas linéaire où $f(u) = au$, le flux numérique $g^n_{j+1/2} = g(v^n_j, v^n_{j+1})$ (pour un schéma à 3 points, $g^n_{j+1/2} = g(v^n_{j-1}, v^n_j, v^n_{j+1}, v^n_{j+2})$ pour un schéma à 5 points), et donc la fonction H vont être souvent une combinaison linéaire des valeurs v^n_j. Si c'est le cas, le schéma s'écrira sous la forme

$$v^{n+1}_j = \sum_{\ell=-k}^{+k} c_\ell v^n_{j+\ell} \tag{3.10}$$

avec $k = 1$ pour un schéma à 3 points, $k = 2$ pour un schéma à 5 points, ... et des coefficients constants c_ℓ qui dépendent de λ et a. Nous allons étudier plus précisément ces schémas linéaires.

3.2.3 Propriété des schémas

L'analyse numérique des schémas, en étudiant la consistance et la stabilité, permet d'étudier leur convergence et, lorsqu'ils convergent, d'estimer l'erreur entre u^n_j valeur exacte de la moyenne de la solution sur une maille au temps t_n (dans la suite du paragraphe, on prendra en fait la valeur au centre de la maille $u(x_j, t_n)$) et v^n_j la valeur calculée par le schéma. Comme nous l'avons vu dans le chapitre sur les différences finies, il faut supposer des propriétés de stabilité et de consistance (et préciser l'ordre).

Notons tout de suite que dans les schémas explicites pour l'équation d'advection, l'ordre de grandeur des pas d'espace Δx et de temps Δt n'est pas le même que pour le problème en temps (équation de la chaleur) de la section 2.5, qui concernait un opérateur différentiel en espace d'ordre 2. Le paramètre $r = \mu\frac{\Delta t}{h^2}$ défini en (2.11) devait vérifier une condition de stabilité du type $0 < r \leq 1/2$ (voir la proposition 2.5.3). C'est

$$c = \lambda a = \frac{a\Delta t}{\Delta x} \tag{3.11}$$

parfois appelé *nombre de Courant*, qui va jouer ce rôle pour les méthodes explicites d'approximation de l'équation d'advection.

Définition 3.2.1 *On dit qu'un schéma est stable pour la norme* $\|.\|_q, 1 \leq q \leq \infty$ *si la suite* $v^n = (v^n_j)_{j\in\mathbb{Z}}$ *vérifie pour tout* $n \geq 0$

$$\|v^n\|_q \leq \|v^0\|_q.$$

Les normes les plus utilisées sont les normes usuelles $\|.\|_1$, $\|.\|_2$ et $\|.\|_\infty$ dont on rappelle qu'elles sont définies pour une suite $v = (v_i)$ par $\|v\|_1 = \Sigma|u_i|$, $\|v\|_2 = (\Sigma u_i^2)^{1/2}$ et $\|v\|_\infty = sup|v|_i$. En fait, et comme nous l'avions vu dans la section 2.2 du chapitre 2, il est plus cohérent d'introduire à une suite une fonction constante par morceaux, et de considérer les normes dans les espaces de fonctions L^q. On va donc aussi considérer la fonction constante par morceaux

$$v_\Delta(x,t) = v^n_j, x \in C_j, t \in [t_n, t_{n+1}[$$

et alors la stabilité s'exprime facilement en terme de norme de v_Δ dans les espaces L^q, et on parlera de façon équivalente de stabilité L^q, en particulier on a

$$\|v_\Delta(.,t_n)\|_{L^2(\mathbb{R})} = \sqrt{\Delta x}\|v^n\|_2,\ \|v_\Delta(.,t_n)\|_{L^\infty(\mathbb{R})} = \|v^n\|_\infty.$$

Il y a des conditions suffisantes de stabilité L^∞ très simples à vérifier.

Proposition 3.2.1 *Un schéma linéaire (3.10) qui peut être associé à un flux numérique et dont les coefficients vérifient $c_\ell \geq 0$, $-k \leq \ell \leq k$, est stable dans L^∞.*

Démonstration. On a vu que si H est associé à un flux, alors $H(u,...,u) = u$, d'où

$$\sum_{\ell=-k}^{k} c_\ell = 1. \tag{3.12}$$

Si les coefficients c_ℓ sont de plus positifs, v_j^{n+1} est une combinaison convexe des v_j^n donc

$$\min_{\ell=-k,...k}(v_{j+\ell}^n) \leq v_j^{n+1} \leq \max_{\ell=-k,...k}(v_{j+\ell}^n)$$

ce qui implique la stabilité L^∞. □

Un tel schéma linéaire est dit *monotone*. En effet, H est alors une fonction croissante de chacun de ses arguments et si les conditions initiales vérifient $v_j^0 \geq w_j^0, \forall j$, alors $v_j^n \geq w_j^n, \forall j$, pour tout n. On verra des exemples de tels schémas ci-dessous (schéma décentré, schéma de Lax-Friedrichs). Notons que puisque les coefficients dépendent de λ et a, la condition sur les coefficients (positivité et somme égale à un) est vérifiée sous une hypothèse de *stabilité* qui concerne λ et a (voir (3.15)).

Pour un schéma linéaire, la stabilité L^2 peut se caractériser grâce à la transformation de Fourier.

Proposition 3.2.2 *Condition de Von Neumann. Un schéma (3.10) est stable dans L^2 si et seulement si le coefficient*

$$h(\xi) = \sum_{\ell=-k}^{+k} c_\ell \exp(i\ell\xi\Delta x) \tag{3.13}$$

vérifie

$$|h(\xi)| \leq 1, \forall \xi \in \mathbb{R}. \tag{3.14}$$

Le coefficient h s'appelle coefficient d'amplification *du schéma.*

Démonstration. Le schéma (3.10) peut aussi s'écrire en terme de fonction v_Δ sous la forme

$$v_\Delta(x,t_{n+1}) = \sum_{\ell=-k}^{+k} c_\ell v_\Delta(x+\ell\Delta x,t_n).$$

Il suffit de vérifier que les deux membres coincident sur chaque intervalle C_j. On prend la transformée de Fourier en espace de cette identité (transformée définie par $\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} v(x)e^{-ix\xi}dx$), on obtient,

$$\hat{v}_\Delta(\xi,t_{n+1}) = \sum_{\ell=-k}^{+k} c_\ell \exp(i\ell\xi\Delta x)\hat{v}_\Delta(\xi,t_n) = h(\xi)\hat{v}_\Delta(\xi,t_n),$$

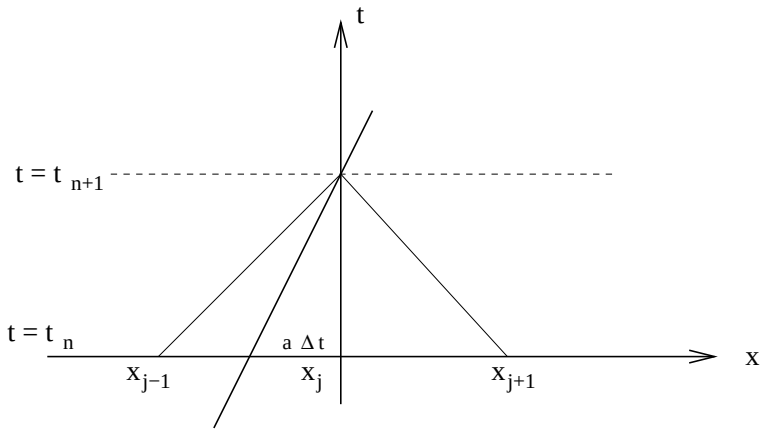


FIGURE 3.3 – Cône de dépendance numérique et caractéristique

où h est définie par (3.13). Il est donc clair que la condition de Von Neumann (3.14) implique que

$$\|\hat{v}_\Delta(.,t_{n+1})\|_{L^2(\mathbb{R})} \leq \|\hat{v}_\Delta(.,t_n)\|_{L^2(\mathbb{R})}.$$

La transformée de Fourier étant une isométrie de $L^2(\mathbb{R})$, on en déduit

$$\|v_\Delta(.,t_{n+1})\|_{L^2(\mathbb{R})} \leq \|v_\Delta(.,t_n)\|_{L^2(\mathbb{R})}$$

et le schéma est bien stable dans L^2. On montre que cette condition est aussi nécéssaire (nous admettrons ce résultat qui est lié à l'évaluation de la norme de l'opérateur linéaire sur L^2 défini par $\varphi \to \varphi h$ pour $h \in L^\infty$). □

La stabilité d'un schéma *explicite* fait nécessairement intervenir une *condition de stabilité* qui se traduit par une inégalité de la forme

$$0 \leq |a|\frac{\Delta t}{\Delta x} \leq 1 \tag{3.15}$$

ou plus généralement $0 \leq |a|\frac{\Delta t}{\Delta x} \leq \mu$ avec $\mu \leq 1$ pour un schéma à 3 points, $\mu \leq k$ pour un schéma à $2k+1$ points. Une telle condition de stabilité (3.15) s'appelle *condition CFL* (pour Courant-Friedrichs-Lewy) et cette terminologie est très utilisée. La quantité $|a|\Delta t$ est la distance parcourue pendant un pas de temps à la vitesse a. Or la solution exacte vérifie

$$u(x_j,t_{n+1}) = u(x_j - a\Delta t, t_n).$$

La condition de stabilité $0 < |c| \leq 1$ dit que cette distance ne doit pas excéder une maille, pour un schéma à trois points (2 mailles pour un schéma à 5 points,...). Cela exprime ainsi le fait que le cône de dépendance numérique doit contenir la caractéristique rétrograde passant par (x_j,t_{n+1}) (voir la figure 3.3). Le *cône de dépendance numérique* au point (x_j,t_{n+1}) est l'ensemble des points de la grille qui interviennent pour le calcul de v_j^{n+1}.

Remarque. Lorsqu'on programme un schéma explicite, on fixe un maillage, par exemple uniforme, donc un pas Δx, et on ajuste le pas de temps Δt pour qu'une condition CFL (éventuellement plus restrictive que (3.15)) soit respectée. On peut bien sûr ensuite raffiner le maillage

(par exemple passer de $\Delta x = 10^{-2}$ à $\Delta x = 10^{-3}$, mais il faut alors diminuer le pas de temps en conséquence. □

L'autre définition importante, qui permet de caractériser la précision d'un schéma, est celle de l'ordre.

Définition 3.2.2 *On dit qu'un schéma (3.9) est d'ordre p, si pour toute solution assez régulière de l'équation (3.1), pour $\Delta x = \frac{\Delta t}{\lambda}$, et λ constant*

$$\frac{1}{\Delta t}\big(u(x,t+\Delta t) - H(u(x-\Delta x,t),u(x,t),u(x+\Delta x,t))\big) = O(\Delta t^p). \tag{3.16}$$

La définition pour un schéma à 5 points est tout à fait analogue, on met à la place $H(u(x-2\Delta x,t),...,u(x+2\Delta x,t))$ dans le deuxième terme du membre de gauche. L'hypothèse de régularité est faite pour pouvoir faire des développements limités qui sont utilisés pour évaluer le premier membre de (3.16) appelé *erreur de consistance*, et est noté E. Le fait que λ soit constant implique que Δx et Δt tendent vers 0 à la même vitesse. Ainsi, dans l'estimation de l'erreur de consistance, on peut ne faire apparaître que le pas de temps, à condition de remplacer Δx par $\frac{\Delta t}{\lambda}$. Nous admettrons le résultat important suivant qui concerne les schémas conservatifs, et est valable y compris dans le cas non linéaire, pour des schémas à $2k+1$ points.

Théorème 3.2.1 *Un schéma (3.9) qui est conservatif et associé à un flux numérique consistant est d'ordre au moins 1.*

Dans le cas linéaire, cela donne le résultat suivant, où c est le nombre de Courant (3.11).

Proposition 3.2.3 *Les coefficients c_ℓ d'un schéma (3.10) associé à un flux consistant vérifient*

$$\sum_{\ell=-k}^{\ell=+k} c_\ell = 1 \text{ et } \sum_{\ell=-k}^{\ell=+k} \ell c_\ell = -c$$

et le schéma est d'ordre au moins 1. Si de plus $\sum_{\ell=-k}^{\ell=+k} \ell^2 c_\ell = c^2$, il est d'ordre au moins deux.

Démonstration. Le fait qu'un schéma (3.10) soit associé à un flux consistant se traduit exactement par les deux premières identités $\sum_{\ell=-k}^{\ell=+k} c_\ell = 1$ et $\sum_{\ell=-k}^{\ell=+k} \ell c_\ell = -c$. La première a déjà été vue (*cf.* 3.12). Pour la deuxième, détaillons le cas $k=1$. Un schéma linéaire à 3 points s'écrit

$$v_j^{n+1} = c_{-1}v_{j-1}^n + c_0 v_j^n + c_1 v_{j+1}^n.$$

S'il est conservatif, nécessairement $\forall u, H(u,u,u) = u \;\Rightarrow\; c_{-1}+c_0+c_1 = 1$. On peut alors l'écrire

$$v_j^{n+1} = v_j^n - \Big((c_{-1}v_j - c_1 v_{j+1}^n) - (c_{-1}v_{j-1} - c_1 v_j^n)\Big)$$

donc sous forme conservative avec le flux $g(u,v) = (1/\lambda)(c_{-1}u - c_1 v)$ et le flux est consistant si $c_{-1} - c_1 = c = \lambda a$. Dans le cas général, la démonstration est similaire.

On effectue alors des développements de Taylor, supposons pour simplifier $u \in \mathcal{C}^3(\mathbb{R}\times[0,T])$. Pour le développement de $u(x,t+\Delta t)$ (resp. $u(x+\ell\Delta x,t)$) au points (x,t), il existe θ (resp. θ_ℓ) $\in]0,1[$ tels que

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}\Delta t^2 u_{tt}(x,t) + \frac{1}{6}\Delta t^3 u_{ttt}(x,t+\theta\Delta t)$$

$$u(x+\ell\Delta x,t) = u(x,t) + \ell\Delta x u_x(x,t) + \frac{1}{2}(\ell\Delta x)^2 u_{xx}(x,t) + \frac{1}{6}(\ell\Delta x)^3 u_{xxx}(x+\theta_\ell\Delta x,t).$$

On remplace ensuite dans la définition de l'erreur de consistance. Si u est solution régulière de (3.1), on a $u_t + au_x = 0$, et aussi $u_{tt} = (-au_x)_t = -a(u_t)_x = a^2 u_{xx}$. Alors, si $\sum_{\ell=-k}^{\ell=+k} c_\ell = 1$ et $\sum_{\ell=-k}^{\ell=+k} \ell c_\ell = -c$, on obtient en posant $\lambda = \Delta t/\Delta x$

$$E(x,t) = \frac{\Delta t}{2\lambda^2}\big((-\sum_{\ell=-k}^{\ell=k} \ell^2 c_\ell + c^2)u_{xx}\big) + O(\Delta t)^2$$

d'où le résultat. □

Corollaire 3.2.2 *Il existe un seul schéma linéaire à 3 points conservatif, avec un flux numérique consistant, et d'ordre deux.*

Démonstration. Un schéma linéaire à 3 points s'écrit

$$v_j^{n+1} = c_{-1}v_{j-1}^n + c_0 v_j^n + c_1 v_{j+1}^n.$$

On vient de voir que s'il est conservatif $c_{-1} + c_0 + c_1 = 1$ et si le flux est consistant, $c_{-1} - c_1 = c = \lambda a$. Un tel schéma peut donc s'écrire en fonction d'un seul paramètre qu'on prend égal à $q \equiv c_{-1} + c_1 = 1 - c_0$. Il s'écrit alors

$$v_j^{n+1} = v_j^n - \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{q}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n), \tag{3.17}$$

il est d'ordre deux si $c_{-1} + c_1 = c^2$, c'est à dire $q = \lambda^2 a^2$, et c'est le seul schéma à 3 points d'ordre 2. □

Le coefficient q s'appelle *coefficient de viscosité*. Le schéma obtenu pour $q = c^2$ s'appelle schéma de Lax-Wendroff. La formule (3.17) permet de montrer le résultat de stabilité suivant.

Proposition 3.2.4 *Un schéma linéaire (3.10) à 3 points conservatif et consistant est stable dans L^2 si et seulement si son coefficient de viscosité q vérifie*

$$(\lambda a)^2 \leq q \leq 1. \tag{3.18}$$

Démonstration. Le coefficient d'amplification du schéma (3.17) est donné par

$$h(\xi) = 1 - \frac{c}{2}(\exp(i\xi\Delta x) - \exp(-i\xi\Delta x)) + \frac{q}{2}(\exp(i\xi\Delta x) - 2 + \exp(-i\xi\Delta x))$$

$$= 1 - q + q\cos(\xi\Delta x) - ic\sin\xi\Delta x$$

d'où,

$$h(\xi) = 1 - q(1 - \cos(\xi\Delta x)) - 2ic\cos(\xi\Delta x/2)\sin(\xi\Delta x/2).$$

On calcule le module de $h(\xi)$ et, en posant $y = (\sin(\xi\Delta x/2))^2$, on obtient

$$|h(\xi)|^2 = (1-2qy)^2 + 4c^2 y(1-y).$$

La condition de Von Neumann (3.14) est vérifiée si

$$0 \leq (1-2qy)^2 + 4c^2y(1-y) \leq 1, \forall y \in [0,1].$$

Pour que ces inégalités soient vérifiées, il faut et suffit que

$$-qy + q^2y^2 + c^2y(1-y) \leq 0, \forall y \in [0,1],$$

donc pour $y \in]0,1]$

$$-q + q^2y + c^2(1-y) \leq 0.$$

Si on fait tendre y vers 0, on obtient la condition $c^2 \leq q$ et si on fait tendre y vers 1, on obtient $q^2 \leq q$. La première inégalité suppose $q > 0$, la deuxième donne alors $q \leq 1$. En résumé, on obtient l'encadrement du coefficient q : $(\lambda a)^2 \leq q \leq 1$. Réciproquement, si cette condition est vérifiée, alors

$$c^2 - q + (q^2 - c^2)y \leq 0 \text{ pour } y = 0 \text{ et } y = 1$$

donc pour tout $y \in [0,1]$ car on a une expression affine en y. On peut alors remonter les calculs, ce qui donne (3.14). □

Donnons maintenant des exemples de schémas linéaires.

3.3 Exemples de schémas pour l'équation de transport

Pour tous les schémas qui suivent, on se donne la condition initiale $v_j^0 = u_j^0 \equiv \frac{1}{\Delta x}\int_{C_j} u_0(x)dx$, et on écrit la formule permettant de calculer v_j^{n+1} à partir des v_j^n.

Exemple 1 : le schéma décentré. Lorsque $a > 0$, on prend simplement $g_{j+1/2}^n = g(v_j^n, v_{j+1}^n) = av_j^n$, c'est à dire le flux exact venant de gauche (car quand a est positif, le mouvement se fait vers la droite, donc la solution exacte vérifie $u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$ et $x_j - a\Delta t$ est à gauche de x_j). On obtient la formule

$$v_j^{n+1} = v_j^n - \lambda a(v_j^n - v_{j-1}^n).$$

Au contraire, quand $a < 0$, on décentre vers la droite, $g_{j+1/2}^n = g(v_j^n, v_{j+1}^n) = av_{j+1}^n$,

$$v_j^{n+1} = v_j^n - \lambda a(v_{j+1}^n - v_j^n).$$

En anglais, le schéma est dit *upwind* : on tient compte du sens du vent. Si on ne précise pas le signe de a, on peut écrire

$$v_j^{n+1} = v_j^n - \lambda(a_-(v_{j+1}^n - v_j^n) + a_+(v_j^n - v_{j-1}^n)),$$

où $a_- = \min(a,0) = a$ si $a \leq 0$, 0 sinon et $a_+ = \max(a,0) = a$ si $a \geq 0$, 0 sinon. On a alors $a = a_+ + a_-, |a| = a_+ - a_-$, d'où la formule unique

$$v_j^{n+1} = v_j^n - \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda|a|}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n).$$

Proposition 3.3.1 *Sous la condition CFL (3.15) le schéma décentré est stable dans L^∞ et stable dans L^2.*

Démonstration. Posons $c = \lambda a$. Si $0 \leq c \leq 1$, on peut aussi écrire

$$v_j^{n+1} = (1-c)v_j^n + cv_{j-1}^n,$$

et v_j^{n+1} apparaît comme combinaison convexe des valeurs (v_{j-1}^n, v_j^n), ce qui entraîne

$$\min(v_{j-1}^n, v_j^n) \leq v_j^{n+1} \leq \max(v_{j-1}^n, v_j^n)$$

et a fortiori

$$\|v^{n+1}\|_{\ell^\infty} \leq \|v^n\|_{\ell^\infty}$$

Le cas $-1 \leq c \leq 0$ se traite de façon identique.

Les formules peuvent aussi s'interpréter comme le fait qu'on a utilisé l'interpolation linéaire pour approcher $u(x_j, t_{n+1})$. La solution exacte vérifie $u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$. Pour $a \geq 0$ par exemple, le point $x_j - a\Delta t$ est dans l'intervalle $[x_{j-1}, x_j]$ si $c \leq 1$, on utilise les valeurs en x_{j-1}, et x_j, v_{j-1}^n approche $u(x_j - \Delta x, t_n)$ et v_{j-1}^n approche $u(x_j, t_n)$. Le polynôme d'interpolation linéaire, $p_1(x) = v_j^n + (x - x_j)\frac{v_j^n - v_{j-1}^n}{\Delta x}$ est alors pris au point $x_j - a\Delta t = x_j - c\Delta x$.

Le coefficient de viscosité du schéma décentré est $q = |c|$ qui vérifie bien la condition (3.18) sous la condition CFL (3.15). Le schéma est d'ordre 1. □

Exemple 2 : le schéma centré. Donnons sa formule bien qu'il ne soit pas utilisé car il est instable

$$v_j^{n+1} = v_j^n - \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n).$$

Il correspond encore à un flux $g_{j+1/2}^n = \frac{a}{2}(v_{j+1}^n + v_j^n)$, on a prix la moyenne des flux exacts de part et d'autre de l'interface. On ne peut plus écrire v_j^{n+1} comme combinaison linéaire convexe des valeurs $(v_{j-1}^n, v_j^n, v_{j+1}^n)$, même si la somme de ces coefficients vaut 1, il y en a toujours un qui est négatif (sauf si $a = 0$ bien sûr, mais ce cas a peu d'intérêt). Regardons par exemple quand $a > 0$ le cas de la condition initiale suivante : $v_j^0 = 1$ pour $j < 0$, $v_j^0 = 0$ pour $j \geq 0$. On calcule facilement $v_j^1 = v_j^0$ pour $j \neq 0, 1$, et $v_{-1}^1 = 1 + c/2, v_0^1 = c/2$. Si $c > 0$, on a $\|v^1\|_\infty = 1 + c/2 > \|v^0\|_\infty$. Cela ne prouve pas que $\|v^n\|_\infty$ explose mais montre qu'il n'y a pas décroissance de la norme. Regardons la condition initiale $v_j^0 = \sin(2\pi j\Delta x)$, alors $v_j^1 = \sin(2\pi j\Delta x)(1 - c\cos(2\pi\Delta x)) = (1 - c\cos(2\pi\Delta x))v_j^0$, par suite, par linéarité, $v_j^2 = (1 - c\cos(2\pi\Delta x))v_j^1 = (1 - c\cos(2\pi\Delta x))^2 v_j^0$, et par récurrence, si $c < 0$, et $\Delta x < 1/4$ on voit que $\|v^n\|_\infty$ explose. On montre qu'il n'est pas stable dans L^2 car son coefficient de viscosité q est nul et ne vérifie donc pas (3.18).

Exemple 3 : le schéma de Lax-Friedrichs. Il est donné par la formule

$$v_j^{n+1} = \frac{v_{j+1}^n + v_{j-1}^n}{2} - \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n).$$

On vérifie que ce schéma est associé au flux

$$g_{j+1/2}^n = \frac{a}{2}(v_{j+1}^n + v_j^n) - \frac{1}{2\lambda}(v_{j+1}^n - v_j^n)$$

c'est à dire $g(u,v) = a(u+v)/2 - (1/2\lambda)(v-u)$.

Proposition 3.3.2 *Sous la condition (3.15) le schéma de Lax-Friedrichs est stable dans L^∞ et stable dans L^2.*

Démonstration. On peut encore écrire

$$v_j^{n+1} = \frac{1}{2}(1-c)v_{j+1}^n + \frac{1}{2}(1+c)v_{j-1}^n$$

et sous la condition (3.15), on a encore une combinaison convexe des valeurs $u_{j\pm1}^n$. On conclut comme pour le schéma décentré. Le coefficient de viscosité du schéma de Lax-Friedrichs est $q = 1$, borne supérieure admissible dans (3.18). Le schéma est d'ordre 1. □

Remarque. Le schéma décentré et le schéma de Lax-Friedrichs sont des exemples de schémas d'ordre un très utilisés. Ils ont des extensions au cas non linéaire. Pour le schéma de Lax-Friedrichs, la formule est immédiate, posant $f_j = f(v_j)$, le schéma est simplement

$$v_j^{n+1} = \frac{v_{j+1}^n + v_{j-1}^n}{2} - \frac{\lambda}{2}(f_{j+1}^n - f_{j-1}^n).$$

Pour le schéma décentré, la généralisation est immédiate lorsque $f' > 0$ ou $f' < 0$, par contre elle nécessite beaucoup plus de travail si f' peut changer de signe, donc lorsqu'il y a un point *sonique* (un état $\overline{u}$ est dit sonique si $f'(\overline{u}) = 0$). Ce vocabulaire vient de la dynamique des gaz, dont la la modélisation conduit à des systèmes non linéraires qui rentrent dans un cadre généralisant l'équation (3.4) ; en un point sonique la vitesse du gaz est égale à la 'vitesse du son'). □

Exemple 4 : le schéma de Lax-Wendroff. Il est donné par la formule

$$v_j^{n+1} = v_j^n - \frac{\lambda a}{2}(v_{j+1}^n - v_{j-1}^n) + \frac{\lambda^2 a^2}{2}(v_{j+1}^n - 2v_j^n + v_{j-1}^n)$$

ce schéma est associé au flux

$$g_{j+1/2}^n = \frac{a}{2}(v_{j+1}^n + v_j^n) - \lambda\frac{a^2}{2}(v_{j+1}^n - v_j^n)$$

c'est à dire $g(u,v) = a(u+v)/2 - (\lambda a^2/2)(v-u)$. C'est le seul schéma à trois points linéaire qui soit d'ordre 2. On peut écrire

$$v_j^{n+1} = v_j^n(1-c^2) + \frac{c}{2}(c-1)v_{j+1}^n + \frac{c}{2}(c+1)v_{j-1}^n.$$

Ce n'est pas une combinaison convexe (la somme des coefficients vaut 1, mais ils ne peuvent être tous les trois positifs) et le schéma ne fait pas décroître la norme dans L^∞ et n'est pas monotone.

Cette formule montre que le schéma est obtenu par interpolation quadratique. En effet les polynômes $p_{-1} = \frac{x}{2}(1-x), p_0 = (1-x^2), p_1 = \frac{x}{2}(1+x)$ sont les polynômes de base de l'interpolation de Lagrange P_2 aux points $-1, 0, 1$, le polynôme $p_j^n(x) = v_j^n p_0 + v_{j+1}^n p_1 + v_{j-1}^n p_{-1}$ prend les valeurs v_j^n, $v_{j\pm1}^n$, en $0, \pm1$, $p_j^n(-c)$ est la valeur prise par ce polynôme en $-c = -a\Delta t/\Delta x$. Par un changement de variable affine, qui envoie l'intervalle $[-1,1]$ sur $[x_{j-1}, x_{j+1}]$, on voit que v_j^{n+1} est la valeur prise en $(x_j - a\Delta t)$ par le polynôme d'interpolation qui prend respectivement les valeurs v_j^n, $v_{j\pm1}^n$ aux points $x_j, x_{j\pm1}$.

Le schéma de Lax-Wendroff est stable dans L^2 sous condition CFL (3.15) car son coefficient de viscosité est $q = c^2$, borne inférieure admissible dans les inégalités (3.18).

□

Chapitre 4

Formulation variationnelle des problèmes aux limites

4.1 La formulation variationnelle : d'où vient-elle ?

La formulation "classique" des problèmes aux limites que nous avons utilisée jusqu'à maintenant n'est pas suffisante, à la fois pour obtenir des résultats théoriques d'existence et d'unicité et pour définir des méthodes d'approximation plus générales. Cela sera d'autant plus vrai que l'on travaillera en dimension supérieure à 1, mais nous n'aborderons pas ces questions dans ce cours.

Reprenons notre problème modèle, avec condition aux limites de Dirichlet homogènes pour simplifier : trouver $u \in \mathcal{C}^0([0,1]) \cap \mathcal{C}^2(]0,1[)$ telle que

$$\text{(P)} \qquad \begin{cases} -u''(x) + c(x)u(x) = f(x) \text{ dans }]0,1[, \\ u(0) = u(1) = 0. \end{cases}$$

On a vu en utilisant la méthode de tir que, si f et $c \geq 0$ sont continues sur l'intervalle $[0,1]$, alors le problème admet une solution et une seule, laquelle appartient en fait à $\mathcal{C}^2([0,1])$. On va maintenant exprimer le problème (P) différemment.

Théorème 4.1.1 *Soit u la solution du problème (P). Alors pour toute fonction $v \in \mathcal{C}^1([0,1])$ telle que $v(0) = v(1) = 0$, u vérifie*

$$\int_0^1 (u'(x)v'(x) + c(x)u(x)v(x))\,dx = \int_0^1 f(x)v(x)\,dx.$$

Réciproquement, si $u \in \mathcal{C}^2([0,1])$ satisfait

$$\text{(fv)} \qquad \begin{cases} \int_0^1 (u'(x)v'(x) + c(x)u(x)v(x))\,dx = \int_0^1 f(x)v(x)\,dx \quad \forall v \in \mathcal{C}^1([0,1]),\ v(0) = v(1) = 0, \\ u(0) = u(1) = 0, \end{cases}$$

alors u est solution de (P).

Démonstration. i) Sens direct. Soit u la solution du problème (P). On a donc $-u'' + cu = f$ dans $[0,1]$. Soit $v \in \mathcal{C}^1([0,1])$ telle que $v(0) = v(1) = 0$. On multiplie l'équation différentielle par v

et on intègre le résultat sur $[0,1]$. Il vient

$$-\int_0^1 u''v + \int_0^1 cuv = \int_0^1 fv.$$

Or $u'' \in \mathcal{C}^0([0,1])$ et $v \in \mathcal{C}^1([0,1])$, donc on a le droit d'intégrer la première intégrale par parties :

$$-\int_0^1 u''v = -[u'v]_0^1 + \int_0^1 u'v'.$$

Comme v satisfait les conditions de Dirichlet homogènes, le terme $[u'v]_0^1$ disparaît et l'on obtient bien

$$\int_0^1 (u'v' + cuv) = \int_0^1 fv$$

pour toute fonction $v \in \mathcal{C}^1([0,1])$ telle que $v(0) = v(1) = 0$.

ii) Réciproque. Par hypothèse, u vérifie les conditions aux limites de (P). Ensuite, le même calcul mené en sens inverse montre que si $u \in \mathcal{C}^2([0,1])$ satisfait (fv) pour toute fonction $v \in \mathcal{C}^1([0,1])$ telle que $v(0) = v(1) = 0$, alors

$$-\int_0^1 u''v + \int_0^1 cuv - \int_0^1 fv = \int_0^1 (-u'' + cu - f)v = 0.$$

Or on sait (ou on devrait savoir) que si une fonction $g \in \mathcal{C}^0([0,1])$ est telle que pour tout $v \in \mathcal{C}^1([0,1])$ avec $v(0) = v(1) = 0$, $\int_0^1 gv = 0$, alors nécessairement, $g = 0$.

Juste au cas où on ne sait pas, voici la preuve de cette propriété fort utile. On raisonne par contradiction en supposant que g n'est pas identiquement nulle. Il existe donc un point $x_0 \in [0,1]$ tel que $g(x_0) = c \neq 0$ et par continuité de g on peu supposer que $x_0 \in]0,1[$. Supposons par exemple que $c > 0$. Toujours par continuité de g, ceci implique qu'il existe un intervalle $I = [x_0 - \delta, x_0 + \delta]$ contenu dans $]0,1[$ avec $\delta > 0$, tel que $g(x) \geq c/2$ pour tout $x \in I$. On considère à présent une fonction $v \in \mathcal{C}^1([0,1])$ dont le support est contenu dans I et telle que $\int_I v = 1$. Il est facile de construire une telle fonction en partant d'une fonction $w \in \mathcal{C}^1(\mathbb{R})$, dont le support est contenu dans $[-1,1]$ et telle que $\int_{\mathbb{R}} w = 1$, puis en posant $v(x) = \delta^{-1} w(\delta^{-1}(x - x_0))$. On a alors

$$\int_0^1 gv = \int_I gv \geq \frac{c}{2} \int_I v = \frac{c}{2} > 0,$$

ce qui est une contradiction. Le cas $c < 0$ se traite de la même manière.

Pour en revenir à la formulation variationnelle, on applique le résultat précédent à la fonction continue $g = -u'' + cu - f$. □

Au vu du théorème 4.1.1, on est naturellement amené à poser la définition suivante.

Définition 4.1.1 *Soit l'espace vectoriel* $\mathcal{V} = \{v \in \mathcal{C}^1([0,1]), v(0) = v(1) = 0\}$. *On appelle* formulation variationnelle *du problème (P) le problème :*

$$\text{(FV)} \qquad \begin{cases} \text{trouver } u \in \mathcal{V} \\ a(u,v) = \ell(v), \qquad \forall v \in \mathcal{V}, \end{cases}$$

où la forme bilinéaire $a\colon \mathcal{V}\times\mathcal{V}\to\mathbb{R}$ *est donnée par*

$$a(u,v)=\int_0^1 (u'v'+cuv),$$

et la forme linéaire $\ell\colon \mathcal{V}\to\mathbb{R}$ *est donnée par*

$$\ell(v)=\int_0^1 fv.$$

Remarque 4.1.1 *On vient de voir que le problème (FV) admet au moins une solution, à savoir la solution de (P). On n'est pas encore sûr que c'est la seule ! En effet, il pourrait très bien y en avoir une qui soit* $\mathcal{C}^1([0,1])$ *mais pas* $\mathcal{C}^2([0,1])$ *(d'après le théorème 4.1.1, toute solution* $\mathcal{C}^2([0,1])$ *est solution de (P)). On verra plus loin que ce n'est pas le cas.*

Remarque 4.1.2 *On a remplacé une équation différentielle posée dans* $]0,1[$ *par une infinité d'équations dans* $\mathbb{R}$*. Noter que* $\mathcal{V}$ *est un espace vectoriel de dimension infinie. Les éléments de* $\mathcal{V}$ *sont appelés* fonctions-test *(elles servent à "tester" l'équation différentielle).*

Réglons d'abord la question de l'unicité.

Proposition 4.1.1 *Le problème (FV) admet une solution unique.*

Démonstration. On sait déjà que la solution de (P) est solution de (FV). Reste à montrer l'unicité. Soient donc $u_1,u_2\in\mathcal{V}$ deux solutions de (FV) :

$$\forall v\in\mathcal{V},\quad \begin{cases} a(u_1,v)=l(v),\\ a(u_2,v)=l(v),\end{cases}$$

d'où en soustrayant membre à membre, et en posant $w=u_1-u_2$,

$$\forall v\in\mathcal{V},\quad a(w,v)=a(u_1,v)-a(u_2,v)=0$$

en utilisant la bilinéarité de a. Comme $w\in\mathcal{V}$, on a le droit de prendre comme fonction-test $v=w$. Ceci donne

$$\int_0^1 (|w'|^2+c|w|^2)=0,$$

d'où, par un raisonnement déjà effectué plusieurs fois, $w'=0$, c'est-à-dire $w=\text{Cte}=w(0)=0$. □

Dans la suite, on ommetra souvent le signe $\forall$ en écrivant simplement

$$a(u,v)=\ell(v),\quad v\in V.$$

Nous terminons cette section en présentant les formulations variationnelles pour la même edp mais avec d'autres conditions aux limites.

i) Conditions aux limites de Dirichlet non homogènes $u(0)=\alpha$ et $u(1)=\beta$.
On ne peut pas définir $\mathcal{V}$ comme l'espace des fonctions $\mathcal{C}^1([0,1])$ telles que $v(0)=\alpha$ et $v(1)=\beta$

car ce n'est pas un espace vectoriel. On se ramène au problème de Dirichlet homogène en posant $\tilde{u} = u - g$ où g est une fonction de $\mathcal{C}^1([0,1])$ qui prends les valeurs prescrites $g(0) = \alpha$ et $g(1) = \beta$ (on dit que g est un *relèvement* de ces conditions limites). On peut par exemple prendre la fonction affine $g(x) = \alpha(1-x) - \beta x$. Ainsi $\tilde{u} = u - g$ vérifie les conditions de Dirichlet homogènes et on vérifie qu'elle est solution d'un problème de type (P) : trouver $\tilde{u} \in \mathcal{V}$ telle que

$$a(\tilde{u}, v) = \tilde{\ell}(v), \quad v \in \mathcal{V}.$$

avec $\tilde{\ell}(v) = \ell(v) - a(g, v)$. Si $\tilde{u}$ est l'unique solution de ce problème, on reviens à u en posant $u = \tilde{u} + g$. Un exercice instructif consiste vérifier que la solution u ainsi obtenue ne dépend pas du choix du relèvement g.

ii) Conditions aux limites de Neumann $u'(0) = \gamma$ et $u'(1) = \delta$.
Nous n'avons pas abordé ce problème aux limites au chapitre 1 mais il peut être résolu par une méthode du tir. Indiquons qu'il faut supposer $c > 0$ en un point x_0 pour montrer l'existence d'une solution. Pour la formulation variationnelle, il faut changer d'espace de fonctions-test : pour les conditions aux limites de Neumann il n'y a aucune raison d'imposer $v(0) = v(1) = 0$, on prend $\mathcal{V} = \mathcal{C}^1([0,1])$. Par intégration par partie de l'équation, on obtient la formulation variationelle : trouver $u \in \mathcal{V}$ telle que

$$a(u, v) = \tilde{\ell}(v), \quad v \in \mathcal{V}.$$

dans laquelle la forme de a est la même que pour les conditions de Dirichlet et $\tilde{\ell}(v) = \int_0^1 fv + \delta v(1) - \gamma v(0)$. On remarque qu'on n'impose pas aux fonctions tests de vérifier les conditions limites de Neumann (on voit que ces conditions n'apparaissent pas naturellement sur la fonction test quand on intègre par parties). On pourra vérifier que réciproquement, lorsque $u \in \mathcal{C}^2([0,1])$, cette formulation variationnelle permet bien de récupérer l'équation *et* les conditions aux limites, donc la solution du problème aux limites.

On a montré l'existence de la solution de (FV) en se raccrochant à l'existence de la solution de (P), obtenue par la méthode de tir. Cette méthode n'étant pas susceptible de beaucoup de généralisations, il convient de voir si l'on peut démontrer directement l'existence pour (FV), *sans passer* par (P). Pour cela, il faut prendre un peu de recul et reconsidérer l'affaire sous un angle abstrait.

4.2 Problèmes variationnels abstraits et espaces de Hilbert

Quelques rappels de vocabulaire pour commencer. Soit H un espace vectoriel sur $\mathbb{R}$. On appelle *produit scalaire* sur H toute forme bilinéaire, symétrique, définie, positive. On notera $(\cdot,\cdot)_H$ un tel produit scalaire. Il n'est pas difficile de montrer que l'application $u \mapsto (u,u)_H^{1/2}$ définit une norme sur H, dite norme *préhilbertienne*. Un espace H muni d'un produit scalaire (et de la norme qui va avec) est dit *préhilbertien*. Dans le cas où H est en outre de dimension finie, on parle d'espace *euclidien* et de norme *euclidienne*. Si H est *complet* pour la norme associée au produit scalaire, *i.e.*, si toute suite de Cauchy pour cette norme est convergente, on dit que H muni de son produit scalaire est un *espace de Hilbert*. La norme est alors appelée

norme *hilbertienne*. Comme tout espace de dimension finie est complet pour n'importe quelle norme, ce qui découle immédiatement du fait que $\mathbb{R}$ lui-même est complet en identifiant H à $\mathbb{R}^n$ pour un certain n à l'aide d'une base, tout espace euclidien est un espace de Hilbert. Le cas non trivial est bien entendu celui de la dimension infinie, où le caractère complet ou non de l'espace prend toute son importance.

Le cadre des espaces de Hilbert se révele particulièrement utile pour étudier les formulations variationelles de type (FV). Avant de revenir à ces formulations, on va rappeler quelques propriétés fondamentales des espaces de Hilbert.

On a tout d'abord le théorème de projection sur un convexe fermé. On rappelle qu'un convexe d'un espace vectoriel est un sous-ensemble C de cet espace tel que pour tout couple de points de cet ensemble, le segment qui joint ces deux points reste dans l'ensemble, c'est-à-dire

$$\lambda u + (1-\lambda)v \in C, \qquad u, v \in C,\ \lambda \in [0,1].$$

Théorème 4.2.1 *Soit H un espace de Hilbert et C un convexe fermé non vide de H. Alors, pour tout $u \in H$, il existe un unique $v \in C$ tel que*

$$\|u-v\|_H = \inf_{w\in C} \|u-w\|_H. \tag{4.1}$$

De plus, cet élément v est caractérisé par les conditions suivantes

$$\begin{cases} v \in C, \\ (u-v, w-v)_H \leq 0,\ \forall w \in C. \end{cases} \tag{4.2}$$

On note $v = P_C(u)$ et on l'appelle *projection orthogonale de u sur C*.

Démonstration. 1) *Existence*. Posons $d = \inf_{w\in C} \|u-w\|_H$ et soit $(w_n)_{n\in\mathbb{Z}} \in C$ une suite d'éléments de C telle que $d_n = \|u-w_n\|_H \to d$ quand $n \to +\infty$ (une telle suite, dite *suite minimisante*, existe par définition de la borne inférieure). Appliquons l'identité du parallélogramme

$$\left\|\frac{x+y}{2}\right\|_H^2 + \left\|\frac{x-y}{2}\right\|_H^2 = \frac{1}{2}(\|x\|_H^2 + \|y\|_H^2), \tag{4.3}$$

à $x = u-w_n$ et $y = u-w_m$. Il vient

$$\left\|\frac{w_n-w_m}{2}\right\|_H^2 + \left\|u - \frac{w_n+w_m}{2}\right\|_H^2 = \frac{1}{2}(\|u-w_n\|_H^2 + \|u-w_m\|_H^2) = \frac{1}{2}(d_n^2+d_m^2).$$

Or C est convexe et comme w_n et w_m appartiennent à C, leur milieu $\frac{w_n+w_m}{2}$ appartient aussi à C. Par conséquent,

$$\left\|u - \frac{w_n+w_m}{2}\right\|_H^2 \geq d^2.$$

On en déduit

$$\left\|\frac{w_n-w_m}{2}\right\|_H^2 \leq \frac{1}{2}(d_n^2+d_m^2) - d^2.$$

Comme d_n et d_m tendent vers d quand n et m tendent vers l'infini, pour tout $\varepsilon > 0$, on peut donc trouver un entier n_0 tel que pour tous $n \geq n_0$ et $m \geq n_0$, $\frac{1}{2}(d_n^2+d_m^2) - d^2 \leq \varepsilon^2$, soit

$$\left\|\frac{w_n-w_m}{2}\right\|_H \leq \varepsilon,$$

c'est-à-dire que la suite w_n est une *suite de Cauchy*. Or H est un espace de Hilbert. Il est donc complet (c'est là que la complétude intervient crucialement) et cette suite de Cauchy est convergente, il existe $v \in H$ tel que $w_n \to v$ dans H (*i.e.*, $\|w_n - v\|_H \to 0$).

On utilise maintenant le fait que C est fermé pour en déduire que $v \in C$. De plus, par continuité de la norme, $\|u-v\|_H = \lim_{n\to\infty} \|u-w_n\|_H = d$, donc v répond à la question.

2) *Équivalence entre (4.1) et (4.2).* Soit $v \in C$ vérifiant (4.1). On se donne $w \in C$ quelconque et l'on pose $z = (1-t)v + tw$ pour $t \in]0,1]$. Comme C est convexe, $z \in C$. Donc, par (4.1),

$$\|u-v\|_H \leq \|u-z\|_H = \|u-(1-t)v-tw\|_H = \|u-v-t(w-v)\|_H.$$

Élevant les deux membres de cette inégalité au carré et développant le carré scalaire du membre de droite, on obtient

$$\|u-v\|_H^2 \leq \|u-v\|_H^2 - 2t(u-v,w-v)_H + t^2\|w-v\|_H^2,$$

soit

$$2t(u-v,w-v)_H \leq t^2\|w-v\|_H^2.$$

On divise cette inégalité par $t > 0$, on fait tendre t vers 0 et on obtient (4.2).

Réciproquement, soit $v \in C$ vérifiant (4.2). Pour tout $w \in C$, on a

$$\|u-w\|_H^2 = \|u-v+v-w\|_H^2 = \|u-v\|_H^2 + 2(u-v,v-w)_H + \|v-w\|_H^2 \geq \|u-v\|_H^2,$$

et v satisfait (4.1).

3) *Unicité*. Soient $v_1 \in C$ et $v_2 \in C$ satisfaisant (4.2). On prend $w = v_2$ pour le premier et $w = v_1$ pour le second et l'on additionne les inégalités correspondantes. Il vient,

$$(u-v_1,v_2-v_1)_H + (u-v_2,v_1-v_2)_H \leq 0$$

c'est-à-dire

$$(v_2-v_1,v_2-v_1)_H = \|v_2-v_1\|_H^2 \leq 0.$$

Donc $v_1 = v_2$. □

Le point $v = P_C(u)$ réalise donc par (4.1) le minimum de la distance de u aux points de C. La caractérisation (4.2) permet de comprendre pourquoi on parle de projection orthogonale, *cf* figure 4.1.

Remarque 4.2.1 *Si l'espace H n'est pas complet, il n'existe pas forcément de projection orthogonale. Un espace qui n'est pas complet est un espace qui contient des «trous», des endroits où des suites de Cauchy aimeraient bien converger, mais ne le peuvent pas car leur limite manque à l'espace. En particulier, il se peut que la borne inférieure des distances d'un point à un convexe fermé ne soit pas atteinte, car elle tombe justement dans un «trou» de l'espace. On peut boucher ces trous de façon systématique. Plus précisément, on rappelle que pour tout espace métrique X, il existe un espace métrique $\bar{X}$, unique à homéomorphisme près, tel que $\bar{X}$ soit complet et qu'il existe une injection continue de X dans $\bar{X}$ dont l'image est dense. L'espace $\bar{X}$ s'appelle le complété de X. L'injection continue permet d'identifier X à un sous-ensemble dense de $\bar{X}$.*

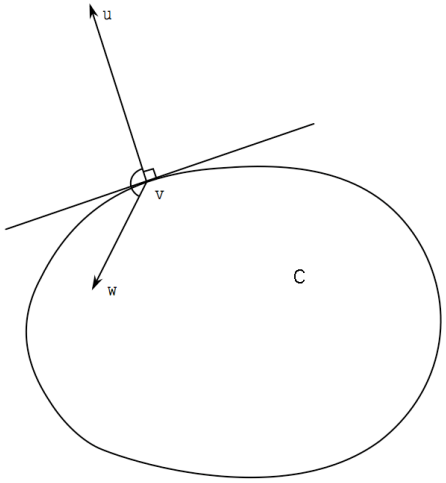


FIGURE 4.1 – Projection orthogonale sur un convexe fermé

On construit $\bar{X}$ à partir de X de la même façon que l'on construit $\mathbb{R}$ à partir de $\mathbb{Q}$: on regarde l'ensemble des suites de Cauchy sur X, on dit que deux suites x_n et y_n sont équivalentes si la distance $d(x_n, y_n)$ tend vers 0 et on définit $\bar{X}$ comme le quotient de l'ensemble des suites de Cauchy par cette relation d'équivalence. La distance entre deux classes d'équivalence est définie comme étant la limite de la distance entre deux représentants de ces classes (cette suite étant de Cauchy dans $\mathbb{R}$). L'injection continue est définie en prenant la classe d'équivalence des suites constantes. Dans le cas où X est un espace vectoriel normé, $\bar{X}$ est un espace de Banach. Dans le cas où X est un espace vectoriel préhilbertien, $\bar{X}$ est un espace de Hilbert. Dans le cas d'un espace préhilbertien X, la projection orthogonale d'un point sur un convexe fermé C existera donc toujours sur le complété $\bar{C}$ du convexe, mais pas forcément dans l'espace de départ. C'est pourquoi il est crucial dans ces questions d'avoir affaire à un espace (préhilbertien) complet, i.e. de Hilbert. On a une remarque analogue si H est un espace de Hilbert, mais C est un convexe non fermé : la projection orthogonale d'un point existe dans l'adhérence $\bar{C}$ de C, qui coïncide avec son complété dans ce cas. Enfin, si C est compact mais pas convexe, il existe bien un point qui minimise la distance, mais il n'y a pas unicité en général (faire un dessin). De plus, l'interprétation géométrique en termes de projection orthogonale ne tient plus. Attention : si C est seulement fermé, il n'existe pas forcément de point qui minimise la distance (en dimension infinie).

L'application la plus importante du théorème de projection dans le cadre de ce cours est celle où le convexe M sur lequel on projette est un sous-espace vectoriel fermé, c'est à dire un sous-espace Hilbertien quand on l'équipe du même produite scalaire que celui de H.

Corollaire 4.2.2 *Si M est un sous-espace vectoriel fermé de H, alors $v = P_M(u)$ est caractérisé par*

$$\begin{cases} v \in M, \\ (u-v, w)_H = 0, \ \forall w \in M. \end{cases} \tag{4.4}$$

De plus, l'application P_M est linéaire continue de H dans H. On note ainsi $P_M u = P_M(u)$.

Démonstration. D'après (4.2), on a pour tout $z \in M$

$$(u-v, z-v)_H \leq 0.$$

Or M est un espace vectoriel, donc pour tout $w \in M$, $z = v+w \in M$, d'où

$$(u-v, w)_H \leq 0.$$

Mais encore, comme M est un espace vectoriel, si $w \in M$, $-w \in M$, d'où

$$(u-v, w)_H \geq 0.$$

On en déduit immédiatement (4.4) et réciproquement. On en déduit également que l'application P_M est linéaire. En effet, étant donnés u_1, $u_2 \in H$ et $\lambda \in \mathbb{R}$, on a

$$\lambda(u_1 - P_M(u_1), w)_H = 0, \quad (u_2 - P_M(u_2), w)_H = 0,$$

pour tout $w \in M$, d'où en additionnant ces deux relations et par bilinéarité du produit scalaire,

$$(\lambda u_1 + u_2 - (\lambda P_M(u_1) + P_M(u_2)), w)_H = 0,$$

pour tout $w \in M$. Or cette relation caractérise $P_M(\lambda u_1 + u_2)$. Par conséquent,

$$P_M(\lambda u_1 + u_2) = \lambda P_M(u_1) + P_M(u_2).$$

Enfin, on note que

$$(u - P_M(u), P_M(u))_H = 0 \Longrightarrow \|P_M(u)\|_H^2 = (u, P_M(u))_H \leq \|u\|_H \|P_M(u)\|_H,$$

par l'inégalité de Cauchy-Schwarz. Donc

$$\|P_M(u)\|_H \leq \|u\|_H \Longleftrightarrow \|P_M\|_{\mathcal{L}(H,H)} \leq 1.$$

Donc P_M est continu. En fait sa norme d'application linéaire continue est égale à 1 car il est facile de voir que P_M est un projecteur, *i.e.* $P_M \circ P_M = P_M$. □

Remarque 4.2.2 *Le corollaire ci-dessus implique que la décomposition*

$$u = P_M u + (u - P_M u) = P_M u + (I - P_M)u,$$

est orthogonale et que $(I-P_M)$ est l'opérateur de projection orthogonale sur $M^\perp$, le supplémentaire orthogonal de M. Par Pythagore, on a

$$\|u\|_H^2 = \|P_M u\|_H^2 + \|u - P_M u\|_H^2. \tag{4.5}$$

On rappelle qu'une application linéaire L allant d'un espae de Banach X dans un autre espace de Banach Y est continue si et seulement si il existe une constante finie C_L telle que

$$\|Lu\|_X \leq C_L \|u\|_X,$$

pour tout $u \in X$. En particulier une forme linéaire ℓ sur X est continue si et seulement si il existe finie C_ℓ telle que

$$|\ell(u)| \leq C_\ell \|u\|_X, \quad u \in X.$$

L'ensemble des formes linéaires continue sur X est appelé le *dual topologique* de X. C'est aussi un espace vectoriel de Banach muni de la norme

$$\|\ell\|_{X'} = \max\{\ell(u) \ : \ \|u\|_X = 1\}.$$

Il ne faut pas confondre X' avec le *dual algébrique* de X qui contient toutes les formes linéaires sur X et est noté X^*. Ce dernier est moins utilisé en pratique car on s'intéresse principalement aux formes linéaires continues.

De la même manière, on peut vérifier qu'une forme bilinéaire $a : X \times Y \to \mathbb{R}$ est continue si et seulement si il existe une constante C_a telle que

$$|a(u,v)| \leq C_a \|u\|_X \|v\|_X, \qquad (u,v) \in X \times Y.$$

En particulier, une forme bilinéaire sur X est continue si et seulement si il existe une constante C_a telle que

$$|a(u,v)| \leq C_a \|u\|_X \|v\|_X, \qquad u,v \in X.$$

Lemme 4 *(Théorème de représentation de Riesz) Soit H un espace de Hilbert et ℓ une forme linéaire continue sur H. Alors il existe un unique $u_0 \in H$ qui représente ℓ au sens où*

$$\ell(v) = (u_0, v)_H,$$

pour tout $v \in H$.

Démonstration. Soit $M = \ker \ell$. Comme ℓ est continue, c'est un sous-espace vectoriel fermé de H. Deux cas se présentent :

• Soit $M = H$, c'est-à-dire $\ell = 0$. Dans ce cas, $u_0 = 0$ est la solution qui s'impose.

• Soit $M \neq H$ et alors il existe $v_1 \in H$ tel que $v_1 \notin M$. Comme $P_M(v_1) \in M$, on a donc $v_1 \neq P_M(v_1)$ et l'on peut définir

$$v_0 = \frac{v_1 - P_M(v_1)}{\|v_1 - P_M(v_1)\|_H}.$$

Il est clair que $\|v_0\|_H = 1$ et que $(v_0, w)_H = 0$ pour tout $w \in M$ par définition de la projection orthogonale sur M (v_0 est un vecteur unitaire orthogonal à M). Par ailleurs, comme $v_1 \notin M$, de même, $v_0 \notin M$, d'où $\ell(v_0) \neq 0$. Pour tout $v \in H$, on pose alors

$$\lambda = \frac{\ell(v)}{\ell(v_0)} \text{ et } w = v - \lambda v_0.$$

On voit immédiatement que $w \in M$. En effet, $\ell(w) = \ell(v - \lambda v_0) = \ell(v) - \lambda \ell(v_0) = 0$ par définition de λ. On a donc ainsi décomposé tout vecteur $v \in H$ en somme orthogonale $v = w + \lambda v_0$. Par conséquent

$$(v, v_0)_H = (w, v_0)_H + \lambda \|v_0\|_H^2 = \frac{\ell(v)}{\ell(v_0)},$$

ou, en d'autres termes,

$$\ell(v) = \ell(v_0)(v, v_0)_H = (v, u_0)_H$$

avec $u_0 = \ell(v_0)v_0$ et l'existence du vecteur u_0 est prouvée. L'unicité est immédiate. □

Remarque 4.2.3 *On a montré que le noyau d'une forme linéaire continue non nulle est un sous-espace vectoriel fermé qui admet un supplémentaire orthogonal de dimension* 1. *C'est donc un hyperplan fermé. De façon plus générale, le noyau d'une forme linéaire non nulle est toujours un hyperplan, c'est à dire un sous-espace de codimension* 1 *(il admet un supplémentaire de dimension* 1*), mais si la forme linéaire n'est pas continue, cet hyperplan n'est pas fermé. Il n'admet pas de supplémentaire orthogonal. En fait, on montre que tout hyperplan non fermé est dense dans l'espace. En dimension finie, toute forme linéaire est continue et la question de fermé ou pas fermé ne se pose pas : tout hyperplan (tout sous-espace vectoriel en fait) est fermé. La dimension d'un hyperplan est égale à la dimension de l'espace moins* 1.

Remarque 4.2.4 *Soit H' le dual topologique de H espace de Hilbert. Alors l'application $\ell \mapsto u_0$ donnée par le théorème de Riesz (on dit aussi de Riesz-Fréchet) est linéaire, bijective et isométrique. La linéarité est évidente, par unicité du vecteur qui représente une forme linéaire. La bijectivité est essentiellement l'objet du théorème (en effet, réciproquement, toute forme linéaire de la forme $v \mapsto (v, u_0)_H$ est trivialement continue par l'inégalité de Cauchy-Schwarz). Pour montrer qu'il s'agit d'une isométrie, on note que*

$$\|\ell\|_{H'} = \sup_{v \neq 0} \frac{|\ell(v)|}{\|v\|_H} = \sup_{v \neq 0} \frac{|(v, u_0)_H|}{\|v\|_H} \leq \|u_0\|_H,$$

encore par Cauchy-Schwarz, avec égalité pour $v = u_0$, d'où $\|\ell\|_{H'} = \|u_0\|_H$. Cette isométrie permet d'identifier H et H' (si l'on veut, on n'est pas obligé de le faire systématiquement), et montre en passant que H' est aussi un espace de Hilbert.

Nous revenons à présent sur les problèmes variationnels de la forme (FV) que l'on peut généraliser ainsi.

Définition 4.2.1 *Un problème variationnel abstrait est défini par la donnée d'un espace vectoriel H, d'une forme bilinéaire a sur H et d'une forme linéaire ℓ sur H. Il consiste à trouver un élément $u \in H$ tel que*

$$\text{(PVA)} \quad a(u, v) = \ell(v), \quad v \in H.$$

Il s'agit bien d'une abstraction de la formulation de la définition 4.1.1. On va montrer un théorème d'existence et d'unicité pour les problèmes variationnels abstraits. La propriété suivante joue un rôle fondamental.

Définition 4.2.2 *Soit H un espace de Hilbert muni de sa norme $\|\cdot\|_H$. On dit qu'une forme bilinéaire a sur H est coercive (ou elliptique) s'il existe une constante $\alpha > 0$ telle que pour tout $u \in H$,*

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

Notons qu'une forme bilinéaire coercive est automatiquement définie, positive. Par contre, la réciproque est fausse en dimension infinie.

Théorème 4.2.3 *(théorème ou lemme de Lax-Milgram) Soit un problème variationnel abstrait (PVA) pour lequel*

i) l'espace H est un espace de Hilbert,

ii) la forme linéaire ℓ est continue : $|\ell(v)| \leq C_\ell \|v\|_H$ pour tout $v \in H$.

iii) la forme bilinéaire a est continue : $|a(u,v)| \leq C_a \|u\|_H \|v\|_H$ pour tout $u, v \in H$.

iv) la forme bilinéaire a est coercive : $a(v,v) \geq \alpha \|v\|^2$ pour tout $v \in H$, avec $\alpha > 0$.

Alors le problème variationnel abstrait admet une solution u et une seule. Cette solution vérifie l'estimation dite a-priori*,*

$$\|u\|_H \leq \frac{C_\ell}{\alpha}. \tag{4.6}$$

Démonstration du théorème 4.2.3 dans le cas symétrique. Pour l'unicité on suppose qu'il existe deux solutions u_1 et u_2. En posant $w = u_1 - u_2$, et en prenant la différence des équations vérifiées par les deux solutions, on voit que

$$a(w,v) = 0, \qquad v \in H.$$

en prenant $v = w$, et utilisant la propriété de coercivité (iv), on trouve

$$\alpha \|w\|_H^2 \leq a(w,w) = 0,$$

et par conséquent $w = 0$ ce qui montre l'unicité. L'estimation a-priori (4.6) se démontre en remarquant que si u est solution, on a

$$\alpha \|u\|_H^2 = a(u,u) = \ell(u) \leq C_\ell \|u\|_H,$$

et en divisant par u (il n'y a rien à prouver dans le cas $u = 0$).

Pour l'existence, on donne la preuve dans le cas où la forme bilinéaire a est symétrique. Puisqu'elle est aussi définie positive, elle définit un produit scalaire sur H noté $(u,v)_a = a(u,v)$. La norme préhilbertienne sur H associée à ce produit scalaire est notée $\|u\|_a = (u,u)_a^{1/2}$. Les propriétés de continuité et de coercivité de a montrent que

$$\alpha \|v\|_H^2 \ \leq \|v\|_a^2 \leq C_a \|v\|_H^2,$$

pour tout $v \in H$, et par conséquent les normes $\|\cdot\|_a$ et $\|\cdot\|_H$ sont équivalentes. On en déduit immédiatement que H est complet pour la norme $\|\cdot\|_a$, c'est-à-dire est un espace de Hilbert pour cette norme également, et que la forme linéaire ℓ est continue pour la nouvelle norme. Le problème variationel peut ainsi s'écrire : trouver u tel que

$$(u,v)_a = \ell(v), \qquad v \in H,$$

et le théorème de Riesz nous assure qu'il existe une unique solution $u \in H$. □

Remarque 4.2.5 *Le théorème s'applique naturellement en dimension finie. Dans ce cas, il suffit que la matrice de la forme bilinéaire soit définie positive pour avoir la coercivité (le démontrer). Il est facile de démontrer que ceci implique son inversibilité et par conséquent l'existence et l'unicité. Cette preuve fonctionne aussi lorsque a n'est pas symmétrique. La démonstration est plus élaborée dans le cas où H est de dimension infinie et a n'est pas symmétrique.*

Les problèmes variationnels abstraits associés à une forme bilinéaire symétrique ont une formulation équivalente en termes de *problèmes de minimisation.*

Théorème 4.2.4 *Soit H un espace de Hilbert, a une forme bilinéaire symétrique, continue, coercive et ℓ une forme linéaire continue. On introduit la fonction $J\colon H \to \mathbb{R}$ par*

$$J(v) = \frac{1}{2}a(v,v) - \ell(v).$$

Alors $u \in H$ est solution du problème (PVA) si et seulement si u minimise J sur H, c'est-à-dire

$$J(u) = \min_{v\in H} J(v).$$

Démonstration. Notons d'abord que pour tous $v, w \in H$, on a

$$\begin{aligned} J(v+w) &= \tfrac{1}{2}a(v+w,v+w) - \ell(v+w) \\ &= \tfrac{1}{2}a(v,v) + \tfrac{1}{2}a(v,w) + \tfrac{1}{2}a(w,v) + \tfrac{1}{2}a(w,w) - \ell(v) - \ell(w) \\ &= \tfrac{1}{2}a(v,v) - \ell(v) + a(v,w) - \ell(w) + \tfrac{1}{2}a(w,w) \ \ (a \text{ symétrique}) \\ &= J(v) + a(v,w) - \ell(w) + \tfrac{1}{2}a(w,w). \end{aligned}$$

Soit alors $u \in H$ la solution du problème (PVA). Pour tout $w \in H$, on a donc $a(u,w) - \ell(w) = 0$ et il vient

$$J(u+w) = J(u) + \frac{1}{2}a(w,w) \geq J(u)$$

pour tout $w \in H$. Or H est un espace vectoriel, donc tout $v \in H$ s'écrit sous la forme $v = u + w$, avec $w = v - u \in H$. On a ainsi montré que

$$J(u) \leq J(v), \quad v \in H,$$

c'est à dire $J(u) = \min_{v\in H} J(v)$.

Réciproquement, si $u \in H$ est tel que $J(u) = \min_{v\in H} J(v)$, alors pour tout $v \in H$ et tout $t \in \mathbb{R}$, $J(u+tv) \geq J(u)$. En développant comme plus haut, on remarque que

$$P(t) = J(u+tv) = J(u) + (a(u,v) - \ell(v))t + \frac{1}{2}a(v,v)t^2,$$

est un polynôme de degré 2 qui atteint son minimum en $t = 0$. Ceci entraine que $P'(0) = 0$, autrement dit

$$a(u,v) - \ell(v) = 0.$$

Puisque $v \in H$ est arbitraire, on obtient que u est solution du problème (PVA). □

Remarque 4.2.6 *Le théorème de Lax-Milgram montre l'existence et l'unicité de la solution de (PVA), ce qui entraîne l'existence et l'unicité du minimum de J sur H. Il est possible de démontrer directement, sans passer par le problème (PVA), que le problème de minimisation admet une solution et une seule sous les hypothèses du théorème de Lax-Milgram. L'étude générale des problèmes de minimisation fait partie de ce qu'on appelle le calcul des variations : toute variation autour d'une solution produit un élément dont le J est plus grand que celui de la solution. Il s'agit d'un domaine très vaste qui a de nombreuses applications dans divers domaines où interviennent des problèmes d'équations aux dérivées partielles. C'est de là que vient le terme "variationnel".*

Remarque 4.2.7 *Le calcul de $J(u+v)$ montre que la différentielle $J'(u) = dJ_u \in H'$ de fonction quadratique J est donnée par la forme linéaire*

$$v \mapsto J'(u)v = a(u,v) - \ell(v)$$

ce qui montre que $u \mapsto J'(u)$ est affine. Elle s'annule si et seulement si u est solution de (PVA).

Remarque 4.2.8 *Si a n'est plus symétrique, alors le (PVA) garde encore une solution unique par Lax-Milgram, mais il n'y a plus de problème de minimisation associé. On peut encore écrire la fonction J, le problème de minimisation correspondant aura bien une solution et une seule, mais cette solution est solution du (PVA) correspondant à la partie symétrique de la forme bilinéaire de départ.*

4.3 Espaces de Sobolev

Les espaces bien adaptés au type de produit scalaire qui apparaît naturellement dans les problèmes variationnels associés aux problèmes aux limites doivent prendre en compte le caractère intégrable des fonctions, puisque la norme fait intervenir des intégrales. Rappelons quelques définitions.

Définition 4.3.1 *On note $L^2(0,1)$, ou $L^2(]0,1[)$, l'espace vectoriel des (classes de) fonctions mesurables sur $]0,1[$, de carré intégrable,*

$$L^2(0,1) = \left\{ v \text{ mesurable}, \int_0^1 |v(x)|^2\, dx < +\infty \right\}.$$

Muni de la norme $\|v\|_{L^2(0,1)} = \left(\int_0^1 |v|^2\right)^{1/2}$, associée au produit scalaire $(u,v)_{L^2(0,1)} = \int_0^1 uv$, c'est un espace de Hilbert.

On rappelle qu'une fonction est dite mesurable si l'image réciproque d'un ouvert est un sous-ensemble mesurable de $[0,1]$ (pour la mesure de Lebesgue). On sait alors calculer l'intégrale de Lebesgue $\int_0^1 v^2$, qui appartient *a priori* à $[0,+\infty]$, et pour un élément de $L^2(0,1)$, cette intégrale est dans $[0,+\infty[$. Il y a une petite subtilité : comme l'intégrale ne voit pas les ensembles de mesure nulle, c'est-à-dire si v est nulle sauf sur un ensemble de mesure nulle (on dit que v est nulle *presque partout*), alors $\int_0^1 |v|^2 = 0$, si l'on considérait vraiment l'espace des fonctions mesurables de carré intégrable, alors l'intégrale du carré ne produirait qu'une semi-norme

sur cet espace. Pour fabriquer une norme, on introduit une relation d'équivalence sur l'ensemble des fonctions mesurables — l'égalité presque partout, *i.e.*, deux fonctions sont équivalentes si elles sont égales sauf sur un ensemble de mesure nulle — on prend l'espace quotient par cette relation, la norme passe au quotient et l'espace quotient devient un espace vectoriel normé. En toute rigueur, L^2 (ainsi que tous les espaces L^p) est donc un espace de classes d'équivalence de fonctions. Dans la pratique, on y pense le plus souvent comme à un espace de fonctions ordinaires. Il y a toutefois des situations dans lesquelles il faut faire attention à la distinction entre fonction et classe d'équivalence.

Notation : Dans la suite, on notera systématiquement $]0,1[=\Omega$ et $[0,1]=\bar{\Omega}$. En particulier, on notera $L^2(\Omega)$ pour $L^2(0,1)$, $\mathcal{C}^1(\bar{\Omega})$ pour $\mathcal{C}^1([0,1])$, etc. De façon générale, on omet de mentionner le domaine de définition de la fonction en notant sa norme

$$\|v\|_{L^2} = \|v\|_{L^2(\Omega)} = \|v\|_{L^2(0,1)}$$

et de même le produit scalaire $(u,v)_{L^2}$. Il en sera de même pour d'autres normes.

On a vu que le problème aux limites est équivalent, sous des hypothèses de continuité des fonctions c et f, à un problème variationnel abstrait posé sur l'espace $\mathcal{V} = \{v \in \mathcal{C}^1(\bar{\Omega}), v(0) = v(1) = 0\}$, avec la forme bilinéaire $a(u,v) = \int_0^1 (u'v' + cuv)$ et la forme linéaire $\ell(v) = \int_0^1 fv$. Peut-on appliquer le résultat d'existence abstrait à ce problème variationnel ? Nous avons donc besoin d'un produit scalaire sur $\mathcal{V}$ pour lequel $\mathcal{V}$ soit préhilbertien, a soit continue et coercive et ℓ soit continue. Définissons donc

$$(u,v)_{H^1} = \int_\Omega (u'v' + uv).$$

C'est visiblement un produit scalaire sur $\mathcal{V}$. La norme préhilbertienne associée est la norme H^1 définie par

$$\|v\|_{H^1} = \Big(\int_\Omega [(v')^2 + v^2]\Big)^{1/2} = \Big(\|v\|_{L^2}^2 + \|v'\|_{L^2}^2\Big)^{1/2}.$$

On définit la *semi-norme* H^1 par

$$|v|_{H^1} = \|v'\|_{L^2},$$

et on note en particulier que

$$\|v\|_{H^1}^2 = \|v\|_{L^2}^2 + |v|_{H^1}^2.$$

Alors ℓ et a sont clairement continues pour cette norme, en utilisant l'inégalité de Cauchy-Schwarz, et si de plus $c(x) \geq \eta > 0$ (hypothèse trop forte comme nous le verrons plus tard) la forme a est visiblement coercive. Le seul point délicat est la question de savoir si $\mathcal{V}$ est complet pour cette norme. La réponse, négative comme on va le voir, montre que $\mathcal{V}$ n'est pas le bon espace pour travailler avec la formulation variationnelle.

Proposition 4.3.1 *L'espace $\mathcal{V}$ n'est pas complet pour la norme $\|\cdot\|_{H^1}$.*

Démonstration. Il faut fabriquer une suite de Cauchy dans $\mathcal{V}$ qui ne soit pas convergente. Pour cela on peut considèrer une fonction la fonction

$$u(x) = 1 - |1 - 2x|,$$

qui est nulle en 0 et 1. La dérivée de u admet une discontinuité en $\frac{1}{2}$ et donc $u \notin \mathcal{V}$.

On considère alors des versions régularisées de u que l'on peut définir par exemple de la manière suivante : on considère une fonction φ symmétrique, positive et de classe $\mathcal{C}^1$ sur $\bar{\Omega}$ et telle que

$$\varphi(-1) = \varphi(1) = 0, \quad \varphi'(-1) = 2, \quad \varphi(1) = -2.$$

Un exemple d'une telle fonction est $\varphi(x) = (x+1)(1-x)$. Pour $n \geq 2$, on définit ensuite u_n par

$$u_n(x) = u(x), \quad |x - \frac{1}{2}| \geq 1/n,$$

et

$$u_n(x) = u(\frac{1}{2} - \frac{1}{n}) + n^{-1}\varphi(n(x - \frac{1}{2})).$$

On vérifie aisément que u_n ainsi définie est de classe $\mathcal{C}^1$, nulle en 0 et 1 et donc appartient à $\mathcal{V}$. On note que lorsque $|x - \frac{1}{2}| \leq c$ avec $c \geq \frac{1}{n}$, on a

$$1 - c \leq u_n(x) \leq 1,$$

Ceci permet de montrer que

$$\|u_n - u\|_{L^2}^2 \leq \int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}+\frac{1}{n}} |u_n - u|^2 \leq \frac{2}{n}\frac{1}{n^2},$$

et par conséquent u_n converges vers u dans $L^2(\Omega)$. D'autre part, en remarquant que $-2 \leq u_n'(x) \leq 2$ pour tout $x \in [0,1]$, on trouve que pour tout $m \geq n$,

$$\|u_n - u_m\|_{H^1}^2 = \int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}+\frac{1}{n}} \left(|u_n - u_m|^2 + |u_n' - u_m'|^2 \right) \leq \frac{2}{n}(\frac{1}{n^2} + 16),$$

ce qui montre que $(u_n)_{n\geq 2}$ est une suite de cauchy pour la norme H^1.

Si $\mathcal{V}$ était complet elle convergerait dans cette norme vers une limite $u^* \in \mathcal{V}$. Puisque $\|v\|_{L^2} \leq \|v\|_{H^1}$, on voit que u^* serait aussi la limite de la suite $(u_n)_{n\geq 2}$ pour la norme L^2 et par conséquence $u^* = u$, ce qui est impossible puisque $u \notin \mathcal{V}$. □

Cette démonstration indique que le complété de $\mathcal{V}$ pour le produit scalaire $(\cdot,\cdot)_{H^1}$ doit d'une certaine façon contenir des fonctions qui ne sont pas dérivables au sens usuel. Il faut noter que l'espace $\mathcal{V}$ est par contre bien complet pour sa norme naturelle

$$\|v\|_{\mathcal{C}^1([0,1])} = \max_{x\in[0,1]} |v(x)| + \max_{x\in[0,1]} |v'(x)|.$$

Mais, cette norme n'est pas hilbertienne (le montrer). Une situation similaire se rencontre si on cherche à utiliser la norme L^2 sur l'espace $\mathcal{C}^0(\bar{\Omega})$ des fonctions continues sur $\bar{\Omega}$.

On va à présent chercher à construire un espace qui, au contraire de $\mathcal{V}$ est complet pour la norme H^1. Au vu de la définition de cette norme, on est tenté de définir l'espace comme celui des fonctions $u \in L^2(\Omega)$ telle que $u' \in L^2(\Omega)$ mais cette définition n'a pas de sens car on ne sais pas bien définir la dérivée d'une fonction de $L^2(\Omega)$. Nous allons voir comment on peut contourner cette difficulté.

Définition 4.3.2 *On note $\mathcal{D}(\Omega)$ l'espace vectoriel des fonctions de classe $\mathcal{C}^\infty$ sur Ω à support compact*

$$\mathcal{D}(\Omega) = \{\varphi \in \mathcal{C}^\infty(\Omega),\ supp\ \varphi := \overline{\{x \in \Omega\ :\ v(x) \neq 0\}}\ est\ compact\ dans\ \Omega\}.$$

Ceci signifie que le support de φ est inclus dans un intervalle *fermé* $[a,b]$ avec $0 < a < b < 1$. En d'autres termes, on est sûr que φ est nulle sur $[0,a]$ et sur $[b,1]$. Attention, l'espace $\mathcal{D}(\Omega)$ n'est pas de façon naturelle un espace vectoriel normé (on peut bien sûr mettre des normes dessus, mais elles n'auront pas de bonnes propriétés). Sa topologie naturelle est nettement plus compliquée qu'une topologie d'espace vectoriel normé, et nous la passerons volontiers sous silence. Il contient beaucoup de fonctions, en voici un exemple : $\varphi(x) = 0$ si $0 \leq x \leq a$ ou $b \leq x \leq b$ et $\varphi(x) = e^{1/(x-a)(x-b)}$ si $a < x < b$. Nous admettrons le résultat de densité fondamental suivant, dont la démonstration ressort de la théorie de la mesure et de l'intégration.

Proposition 4.3.2 *L'espace $\mathcal{D}(\Omega)$ est dense dans $L^2(\Omega)$, i.e., pour tout $v \in L^2(\Omega)$, il existe une suite $\varphi_n \in \mathcal{D}(\Omega)$ telle que $\|\varphi_n - v\|_{L^2} \to 0$ quand $n \to +\infty$.*

Définition 4.3.3 *On dit qu'une fonction $v \in L^2(\Omega)$ admet une* dérivée faible *dans $L^2(\Omega)$ s'il existe $g \in L^2(\Omega)$ telle que, pour tout $\varphi \in \mathcal{D}(\Omega)$,*

$$\int_\Omega v\varphi' = -\int_\Omega g\varphi,$$

et l'on note alors $g = v'$.

Remarque 4.3.1 *On dit également que v admet une dérivée au sens des distributions, ou au sens de $\mathcal{D}'(\Omega)$, et que cette dérivée appartient à $L^2(\Omega)$. La notion de dérivée faible s'inscrit ainsi de manière plus générale dans la théorie des distributions. Comme nous n'aurons pas besoin de cette théorie ici, nous garderons la terminologie la plus brève.*

La notion de dérivée faible généralise la notion de dérivée usuelle pour des fonctions qui ne sont pas dérivables au sens où leurs quotients différentiels n'ont pas forcément de limite en certains points de l'intervalle de définition. C'est le sens de la proposition suivante.

Proposition 4.3.3 *i) Si $v \in L^2(\Omega)$ admet une dérivée faible, alors celle-ci est unique.*
ii) Si $v \in \mathcal{C}^1(\bar{\Omega})$, alors v admet une dérivée faible et celle-ci coïncide avec sa dérivée au sens usuel.

Démonstration. Montrons d'abord l'unicité de la dérivée faible. Soit $v \in L^2(\Omega)$ qui admet deux dérivées faibles g_1 et g_2. Pour tout $\varphi \in \mathcal{D}(\Omega)$, on a donc

$$\int_\Omega v\frac{d\varphi}{dx} = -\int_\Omega g_1\varphi = -\int_\Omega g_2\varphi.$$

En postant $w = g_1 - g_2 \in L^2(\Omega)$, on a donc

$$\int_\Omega w\varphi = 0.$$

Or, $\mathcal{D}(\Omega)$ est *dense* dans $L^2(\Omega)$. Il existe donc une suite $\varphi_n \in \mathcal{D}(\Omega)$ telle que $\varphi_n \to w$ quand $n \to +\infty$ au sens de $L^2(\Omega)$. Donc

$$0 = \int_\Omega w\varphi_n \to \int_\Omega w^2 = \|w\|_{L^2}^2 \text{ quand } n \to +\infty.$$

En effet, $\varphi \to \int_\Omega w\varphi$ est continu pour la norme L^2 par Cauchy-Schwarz. Par conséquent, $w = 0$, i.e. $g_1 = g_2$ et on a unicité de la dérivée faible.

Soit maintenant $v \in \mathcal{C}^1(\bar{\Omega})$ (que l'on identifie à un sous-espace de $L^2(\Omega)$ comme on l'a fait plus haut pour $\mathcal{D}(\Omega)$). Comme φ est indéfiniment dérivable, nous pouvons intégrer par parties,

$$\int_\Omega v\varphi' = [v\varphi]_0^1 - \int_\Omega v'\varphi.$$

Comme φ est à support compact, en particulier $\varphi(0) = \varphi(1) = 0$ et par conséquent $[v\varphi]_0^1 = 0$. De plus, $v' \in \mathcal{C}^0(\bar{\Omega})$ est de carré intégrable et appartient à $L^2(\Omega)$. Par conséquent, on voit que v admet une dérivée faible et que cette dérivée faible est égale à sa dérivée usuelle v'. □

Définition 4.3.4 *On appelle* espace de Sobolev *et on note $H^1(\Omega)$ l'ensemble des fonctions v de $L^2(\Omega)$ qui admettent une dérivée faible dans $L^2(\Omega)$.*

Remarque 4.3.2 *La notion de dérivée faible s'étend naturellement en dimension supérieure de la manière suivante : si Ω est un ouvert de $\mathbb{R}^d$ une fonction $f \in L^2(\Omega)$ admet une dérivée partielle faible dans la direction x_i si et seulement si il existe une fonction $g_i \in L^2(\Omega)$ telle que $\int_\Omega f\frac{\partial\varphi}{\partial x_i} = -\int g_i\varphi$ pour tout $\varphi \in \mathcal{D}(\Omega)$. On pose alors $\frac{\partial f}{\partial x_i} = g_i$, et on note $H^1(\Omega)$ l'ensemble des fonctions v de $L^2(\Omega)$ qui admettent des dérivée partielles faible dans $L^2(\Omega)$ dans toutes les directions $(x_1,\ldots,x_d)$.*

Remarque 4.3.3 *D'après la proposition précédente, on a $\mathcal{C}^1(\bar{\Omega}) \subset H^1(\Omega) \subset L^2(\Omega)$.*

Théorème 4.3.1 *Muni de la norme*

$$\|v\|_{H^1} = \Big(\int_\Omega [|v|^2 + |v'|^2]\Big)^{1/2},$$

l'espace $H^1(\Omega)$ est un espace de Hilbert.

Démonstration. Il est clair que $H^1(\Omega)$ est un espace vectoriel et que la norme $\|\cdot\|_1$ est une norme préhilbertienne. Il s'agit donc de vérifier que $H^1(\Omega)$ est complet pour cette norme. On va pour cela s'appuyer sur le fait que $L^2(\Omega)$ lui-même est complet (c'est dans ce dernier résultat que se trouve tout le travail difficile en fait, que nous admettrons, voir en annexe pour sa preuve).

Soit u_n une suite de Cauchy dans $H^1(\Omega)$. Pour tout $\varepsilon > 0$, il existe donc un entier n_0 tel que pour tous $m,n \geq n_0$, $\|u_m - u_n\|_{H^1} \leq \varepsilon$. Comme $\|u_m - u_n\|_{L^2} \leq \|u_m - u_n\|_{H^1}$, on en déduit d'abord que u_n est aussi de Cauchy dans $L^2(\Omega)$. Comme $L^2(\Omega)$ est complet, il existe donc $u \in L^2(\Omega)$ tel que $\|u_n - u\|_{L^2} \to 0$ quand $n \to +\infty$. De même, comme $\|u'_m - u'_n\|_{L^2} \leq \|u_m - u_n\|_{H^1}$, la suite u'_n est de Cauchy dans $L^2(\Omega)$ et il existe $g \in L^2(\Omega)$ tel que $\|u'_n - g\|_{L^2} \to 0$ quand $n \to +\infty$.

Vérifions que g est bien la dérivée faible de u, afin de montrer que $u \in H^1(\Omega)$. Comme $u_n \in H^1(\Omega)$, on a par définition

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad \int_\Omega u_n \varphi' = -\int_\Omega u_n' \varphi.$$

Comme φ et φ' appartiennent à $\mathcal{D}(\Omega)$ et donc à $L^2(\Omega)$, on peut passer à la limite et obtenir

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad \int_\Omega u \varphi' = -\int_\Omega g \varphi.$$

Donc u admet une dérivée faible $u' = g$, c'est-à-dire $u \in H^1(\Omega)$. De plus,

$$\|u_n - u\|_{H^1}^2 = \|u_n - u\|_{L^2}^2 + \|u_n' - g\|_{L^2}^2 \to 0 \text{ quand } n \to +\infty,$$

et la suite u_n converge bien vers $u \in H^1(\Omega)$ au sens de la norme de $H^1(\Omega)$. Cet espace est par conséquent complet. □

Donnons quelques propriétés élémentaires de l'espace $H^1(\Omega)$. La proposition 4.3.1 nous indique que l'inclusion $\mathcal{C}^1(\bar{\Omega}) \subset H^1(\Omega)$ est stricte, puisque sa preuve montre que la fonction $v(x) = 1 - |2x - 1|$ appartient à $H^1(\Omega)$ alors qu'elle n'appartient pas à $\mathcal{C}^1(\bar{\Omega})$. Ce dernier exemple, une fonction affine par morceaux, se généralise aisément.

Proposition 4.3.4 *Toute fonction u continue et de classe $\mathcal{C}^1$ par morceaux sur $\bar{\Omega}$ appartient à $H^1(\Omega)$. Sa dérivée faible est une fonction continue par morceaux qui coïncide avec la dérivée de u sur chaque intervalle ouvert où u est $\mathcal{C}^1$.*

Démonstration. Soit u une fonction continue de classe $\mathcal{C}^1$ par morceaux. Ceci signifie que l'on a des points $0 = x_0 < x_1 < x_2 < \cdots < x_p < x_{p+1} = 1$ et que la restriction de u à chaque intervalle fermé $[x_i, x_{i+1}]$ est de classe $\mathcal{C}^1$ sur ce même intervalle. De plus, les limites de u à gauche et à droite en chaque point x_i coïncident avec la valeur de u en ce point. Notons $g_i \in \mathcal{C}^0([x_i, x_{i+1}])$ la dérivée de la restriction de u à $[x_i, x_{i+1}]$. En définissant la fonction g par

$$g(x) = g_i(x) \text{ quand } \quad x \in]x_i, x_{i+1}[,$$

et en ne la définissant pas pour $x = x_i$, on voit que $g \in L^2(\Omega)$ (une fonction de L^2 n'a pas besoin d'être définie sur un ensemble de mesure nulle). Montrons que g est bien la dérivée faible de u. Soit $\varphi \in \mathcal{D}(\Omega)$ quelconque. On a

$$\int_\Omega u \varphi' = \sum_{i=0}^{p} \int_{x_i}^{x_{i+1}} u \varphi' = \sum_{i=0}^{p} [u\varphi]_{x_i}^{x_{i+1}} - \sum_{i=0}^{p} \int_{x_i}^{x_{i+1}} g_i \varphi,$$

car la restriction de u à $[x_i, x_{i+1}]$ étant de classe $\mathcal{C}^1$, on a le droit d'intégrer par parties sur chacun de ces intervalles. Il est clair, parce qu'on a supposé u continue, que $\sum_{i=0}^{p} [u\varphi]_{x_i}^{x_{i+1}} = [u\varphi]_0^1$ et que $[u\varphi]_0^1 = 0$, puisque φ est à support compact. Par conséquent, on a

$$\int_\Omega u \varphi' = -\sum_{i=0}^{p} \int_{x_i}^{x_{i+1}} g_i \varphi = -\int_0^1 g \varphi,$$

d'où $u \in H^1(\Omega)$ et $u' = g$. □

L'hypothèse de continuité globale de u est essentielle la preuve de la proposition ci-dessus. En particulier une fonction de classe $\mathcal{C}^1$ par morceaux mais sans raccords continus n'appartiendra pas à l'espace $H^1(\Omega)$. Prenons par exemple la fonction

$$u = \chi_{[0,1/2[},$$

où χ_E désigne la fonction indicatrice d'un ensemble E. La fonction u vaut donc 1 sur $[0,1/2[$ et 0 sur $[1/2,1]$. Si elle admettait une dérivée faible $g = u'$, celle-ci vérifierait pour toute fonction $\varphi \in \mathcal{D}(]0,1/2[)$

$$\int_0^{1/2} g\varphi = -\int_0^{1/2} \varphi' = -[\varphi]_0^{1/2} = 0,$$

et par conséquent sa restriction à $]0,1/2[$ serait nulle. De la même manière sa restriction à $]1/2,1[$ serait nulle, ce qui montre que $g = 0$ (au sens presque partout). On aurait alors

$$0 = \int_\Omega g\varphi = -\int_\Omega u\varphi' = -\int_0^{1/2} \varphi' = -\varphi(1/2),$$

pour tout $\varphi \in \mathcal{D}(\Omega)$ et on aboutit à une contradiction en prenant une fonction φ telle que $\varphi(1/2) \neq 0$. Cet exemple montre en particulier que l'inclusion $H^1(\Omega) \subset L^2(\Omega)$ est stricte, et qu'il ne faut surtout pas confondre la dérivée faible avec la dérivée presque partout.

Montrons maintenant que $H^1(\Omega)$ réalise le complété de $\mathcal{C}^1(\bar{\Omega})$. On commence par un petit lemme fort utile.

Lemme 5 *Si $w \in H^1(\Omega)$ est tel que $w' = 0$, alors w est une fonction constante.*

Démonstration. Soit $w \in H^1(\Omega)$ telle que $w' = 0$, ceci signifie, par définition de la dérivée faible, que

$$\int_\Omega w\varphi' = 0, \qquad \varphi \in \mathcal{D}(\Omega).$$

Choisissons $\psi_0 \in \mathcal{D}(\Omega)$ de façon à ce que $\int_\Omega \psi_0 = 1$. Alors pour tout $\theta \in \mathcal{D}(\Omega)$, la fonction $\Theta = \theta - (\int_\Omega \theta(x)\,dx)\psi_0 \in \mathcal{D}(\Omega)$ et est telle que $\int_\Omega \Theta = 0$. Par conséquent, $\varphi(x) = \int_0^x \Theta(t)\,dt$ est $\mathcal{C}^\infty$ et à support compact (introduire des intervalles qui contiennent les supports de θ et ψ). De plus, $\varphi' = \Theta$ et en reportant dans l'intégrale $\int_\Omega w\varphi'$, on obtient

$$\int_\Omega w\theta - \Big(\int_\Omega \theta\Big)\Big(\int_\Omega w\psi_0\Big) = 0, \qquad \theta \in \mathcal{D}(\Omega).$$

Posant alors $c = \int_\Omega w\psi_0 \in \mathbb{R}$, on a montré que

$$\int_\Omega (w-c)\theta = 0, \qquad \theta \in \mathcal{D}(\Omega).$$

On utilise encore une fois la densité de $\mathcal{D}(\Omega)$ dans $L^2(\Omega)$ pour en déduire que $w - c = 0$. □

Remarque 4.3.4 *Attention, il ne s'agit pas du résultat classique analogue, puisqu'on parle ici de dérivée faible et c'est la définition de la dérivée faible qu'il faut utiliser. Néanmoins, le résultat est le même.*

Proposition 4.3.5 *L'espace $\mathcal{C}^1(\bar{\Omega})$ est dense dans $H^1(\Omega)$.*

Démonstration. Soit $u \in H^1(\Omega)$. Comme $\mathcal{D}(\Omega)$ est dense dans $L^2(\Omega)$, il existe une suite φ_n dans $\mathcal{D}(\Omega)$ telle que $\varphi_n \to u'$ dans $L^2(\Omega)$. On pose $v_n(x) = \int_0^x \varphi_n(t)\,dt$. Il est clair que $v_n \in \mathcal{C}^1(\bar{\Omega})$ (en fait ces fonctions sont indéfiniment dérivables) et que $v_n' \to u'$ dans $L^2(\Omega)$. Montrons que la suite v_n est de Cauchy dans $L^2(\Omega)$.

Pour cela, on prend deux indices n et m et l'on compare v_n et v_m :

$$(v_n - v_m)(x) = \int_0^x (\varphi_n - \varphi_m)(t)\,dt.$$

Prenant les valeurs absolues et élevant au carré, il vient immédiatement (par l'inégalité de Cauchy-Schwarz)

$$|(v_n - v_m)(x)|^2 \leq \int_\Omega |\varphi_n - \varphi_m|^2 = \|\varphi_n - \varphi_m\|_{L^2}^2.$$

Intégrant alors entre 0 et 1, on obtient,

$$\|v_n - v_m\|_{L^2} \leq \|\varphi_n - \varphi_m\|_{L^2}.$$

Mais la suite φ_n converge dans $L^2(\Omega)$ par hypothèse. C'est donc une suite de Cauchy dans L^2. L'inégalité précédente montre alors que v_n est aussi une suite de Cauchy.

Comme l'espace $L^2(\Omega)$ est complet, il existe donc $v \in L^2(\Omega)$ tel que $v_n \to v$ dans $L^2(\Omega)$. Par le même raisonnement de passage à la limite dans les intégrales que celui déjà fait au théorème 4.3.1, on voit que $v \in H^1(\Omega)$ avec $v' = u'$. Par le lemme précédent, il existe une constante $c \in \mathbb{R}$ telle que $u = v + c$. On déduit de ce qui précède que la suite de terme $u_n := v_n + c \in \mathcal{C}^1(\bar{\Omega})$ converge dans $H^1(\Omega)$ vers u. □

Remarque 4.3.5 *La preuve ci-dessus montre en fait que l'espace $\mathcal{C}^\infty(\overline{\Omega})$ est dense dans $H^1(\Omega)$ puisqu'on a en fait par construction $u_n \in \mathcal{C}^\infty(\overline{\Omega})$.*

Encore une inclusion, plus exactement une injection continue : nous allons montrer que toute fonction u de $H^1(\Omega)$ a un représentant continu, qu'on note encore u et que l'application $u \longmapsto u$ est une application (linéaire) continue de $H^1(\Omega)$ muni de $\|.\|_{H^1}$ dans $\mathcal{C}^0(\bar{\Omega})$ muni de sa norme naturelle

$$\|v\|_{\mathcal{C}^0} = \max_{x \in \bar{\Omega}} |v(x)|.$$

On rappelle au passage que $L^\infty(\Omega)$ est l'ensemble des classes d'équivalence de fonctions mesurables sur $\Omega =]0,1[$ qui contiennent un représentant borné. On peut y penser comme à des fonctions mesurables bornées presque partout. Le nombre

$$\|v\|_{L^\infty(\Omega)} = \text{esssup}_{x \in \Omega} |v(x)| = \inf\{\lambda \in \mathbb{R}_+, |v(x)| \leq \lambda \text{ presque partout}\}$$

est appelé *sup essentiel* de la fonction c définit une norme sur $L^\infty(\Omega)$ qui en fait un espace de Banach. Bien sûr, $|v(x)| \leq \|c\|_{L^\infty(\Omega)}$ presque partout. Mais dans le cas où v est continue sur $\bar{\Omega}$, le sup essentiel et le max sur $\bar{\Omega}$ coïncident. On pourra donc écrire

$$\|v\|_{L^\infty} = \|v\|_{\mathcal{C}^0}.$$

Proposition 4.3.6 *On a $H^1(\Omega) \hookrightarrow \mathcal{C}^0(\bar{\Omega})$.*

Démonstration. On raisonne par densité des fonctions régulières. Pour $v \in \mathcal{C}^1(\bar{\Omega})$, on a pour tout $x, y \in \bar{\Omega}$,

$$v(x) = v(y) + \int_y^x v'(t)dt, \tag{4.7}$$

et par conséquent

$$|v(x)|^2 \leq 2|v(y)|^2 + 2|\int_y^x v'(t)dt|^2 \leq 2|v(y)|^2 + 2\int_0^1 |v'(t)|^2 dt,$$

en utilisant l'inégalité $(a+b)^2 \leq 2a^2 + 2b^2$ et Cauchy-Schwarz. En intégrant sur y entre 0 et 1, on trouve

$$|v(x)|^2 \leq 2(\|v\|_{L^2}^2 + \|v'\|_{L^2}^2) = 2\|v\|_{H^1}^2.$$

Cette inégalité étant uniforme par rapport à x, on en déduit que

$$\|v\|_{L^\infty} \leq \sqrt{2}\|v\|_{H^1}. \tag{4.8}$$

Soit maintenant $u \in H^1(\Omega)$ et soit $v_n \in \mathcal{C}^1(\bar{\Omega})$ une suite qui converge vers u dans $H^1(\Omega)$, laquelle existe d'après la proposition 4.3.5. C'est donc une suite de Cauchy et en appliquant l'inégalité précédente, il vient

$$\|v_n - v_m\|_{L^\infty} \leq \sqrt{2}\|v_n - v_m\|_{H^1},$$

c'est-à-dire que v_n est aussi une suite de Cauchy dans $\mathcal{C}^0(\bar{\Omega})$. Cet espace est un espace de Banach pour sa norme naturelle et il existe donc $v \in \mathcal{C}^0(\bar{\Omega})$ tel que $v_n \to v$ dans $\mathcal{C}^0(\bar{\Omega})$, c'est-à-dire uniformément. En particulier, comme Ω est un intervalle borné de longueur 1,

$$\|v_n - v\|_{L^2}^2 \leq \|v_n - v\|_{L^\infty}^2,$$

par conséquent, v_n tend aussi vers v dans $L^2(\Omega)$. Or on sait déjà que v_n tend vers u dans $L^2(\Omega)$, d'où $u = v$ presque partout et la classe d'équivalence de u contient une fonction continue v. On a ainsi construit une injection de $H^1(\Omega)$ dans $\mathcal{C}^0(\bar{\Omega})$. De plus cette injection est continue, car appliquant l'inégalité (4.8) à v_n, on peut passer à la limite dans les deux membres, par convergence $\mathcal{C}^0(\bar{\Omega})$ à gauche et convergence $H^1(\Omega)$ à droite, ce qui donne

$$|u\|_{L^\infty} \leq \sqrt{2}\|u\|_1, \qquad u \in H^1(\Omega), \tag{4.9}$$

où l'on n'a pas distingué entre u et son représentant continu v.

Notons que l'on peut aussi passer à la limite dans les deux membres de l'identité (4.7) qui est donc aussi vérifiée pour toute fonction $v \in H^1(\Omega)$. □

Remarque 4.3.6 *Ce raisonnement est typique des raisonnements par densité : on établit une relation pour des fonctions régulières, en utilisant des propriétés de ces fonctions régulières — ici d'être égales à l'intégrale de leur dérivée — puis on passe à la limite dans la relation en question.*

Remarque 4.3.7 *Le fait que l'injection soit continue implique qu'une suite qui converge dans H^1 converge aussi uniformément. L'injection continue $H^1 \hookrightarrow \mathcal{C}^0$ n'est valable qu'en dimension 1. Elle est fausse pour les espaces de Sobolev H^1 que l'on définit en dimension supérieure.*

Nous pouvons maintenant parler de condition de Dirichlet pour des fonctions H^1.

Définition 4.3.5 *On pose $H_0^1(\Omega) = \{v \in H^1(\Omega), v(0) = v(1) = 0\}$. C'est un sous-espace vectoriel fermé de $H^1(\Omega)$.*

En effet, $H_0^1(\Omega)$ est bien défini en considérant la fonction continue qui représente v (si v est seulement L^2, la quantité $v(0)$ n'a aucun sens, puisque l'on peut modifier les valeurs d'une fonction de L^2 sur un ensemble de mesure nulle sans modifier la classe d'équivalence). De plus si $v_n \in H_0^1(\Omega)$ converge vers $v \in H^1(\Omega)$, alors elle converge aussi uniformément et donc ponctuellement en 0 et en 1. Donc $0 = v_n(0) \to v(0)$ et $0 = v_n(1) \to v(1)$, *i.e.*, $v \in H_0^1(\Omega)$ qui est bien fermé.

Proposition 4.3.7 *L'espace $H_0^1(\Omega)$ est l'adhérence de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$. C'est aussi l'adhérence, donc le complété, de l'espace $\mathcal{V}$ introduit au début de cette section.*

Démonstration. On commence par noter que si $u \in H_0^1(\Omega)$ alors $\int_\Omega u' = 0$. Ceci provient de l'identité (4.7) qui est vérifiée par les fonctions de $H^1(\Omega)$ et qu'on peut appliquer avec $x = 0$ et $y = 1$.

On reprend alors la suite $\varphi_n \to u'$ dans $L^2(\Omega)$ de la proposition 4.3.5, et la fonction ψ_0 du lemme 5. Comme $\int_\Omega \varphi_n \to \int_\Omega u' = 0$, on a aussi que $\theta_n = \varphi_n - (\int_\Omega \varphi_n)\psi_0 \to u'$ dans $L^2(\Omega)$. De plus, $\int_\Omega \theta_n(x) = 0$ et donc $\Theta_n(x) = \int_0^x \theta_n(t)\,dt$ est telle que $\Theta_n \in \mathcal{D}(\Omega)$ et il existe $c \in \mathbb{R}$ tel que $\Theta_n \to u + c$ dans $H^1(\Omega)$ quand $n \to \infty$. Comme la convergence est de plus uniforme, il s'ensuit que $0 = \Theta_n(0) \to u(0) + c = c$, d'où $c = 0$, et $\Theta_n \to u$ ce qui établit le premier résultat.

On remarque enfin que $\mathcal{D}(\Omega) \subset \mathcal{V} \subset H_0^1(\Omega)$ pour en déduire immédiatement que $\mathcal{V}$ est dense dans $H_0^1(\Omega)$. □

Remarque 4.3.8 *En dimension supérieure, on définit parfois directement $H_0^1(\Omega)$ comme l'adhérence de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$. La question de savoir en quel sens une fonction de $H_0^1(\Omega)$ s'annule sur le bord de Ω est délicate dès que les fonctions de $H^1(\Omega)$ ne sont plus continues.*

On rappelle la notation $|u|_{H^1} = \|u'\|_{L^2}$ pour la semi-norme H^1. La propriété suivante est très importante.

Théorème 4.3.2 *Pour tout $u \in H_0^1(\Omega)$, on a l'inégalité de Poincaré*

$$\|u\|_{L^2} \leq |u|_{H^1}. \tag{4.10}$$

La semi-norme $|\cdot|_{H^1}$ définit une norme sur $H_0^1(\Omega)$, équivalente à la norme $\|\cdot\|_{H^1}$.

Démonstration. Pour $u \in H^1(\Omega)$, on a $u(x) - u(0) = \int_0^x u'(s)\,ds$. En particulier, si $u \in H_0^1(\Omega)$, il vient

$$u(x) = \int_0^x u'(s)\,ds,$$

d'où

$$|u(x)|^2 \leq |\int_0^x u'(s)\,ds|^2 \leq \int_0^x |u'(s)|^2\,ds \leq \int_\Omega |u'(s)|^2\,ds = |u|_1^2.$$

Intégrant cette inégalité entre 0 et 1, on obtient (4.10). On en déduit immédiatement que

$$|u|_{H^1}^2 = \|u\|_{L^2}^2 + |u|_{H^1}^2 \leq 2|u|_{H^1}^2, \qquad u \in H_0^1(\Omega),$$

d'où l'équivalence des normes,

$$|u|_{H^1} \leq \|u\|_{H^1} \leq \sqrt{2}|u|_{H^1}, \qquad u \in H_0^1(\Omega),$$

qui démontre le théorème. □

Remarque 4.3.9 *Si l'on travaille sur $H_0^1(\Omega)$, on peut donc au choix utiliser la norme $\|\cdot\|_{H^1}$ ou la norme équivalente définie par*

$$\|v\|_{H_0^1} := |v|_{H^1} = \|v'\|_{L^2}.$$

Remarque 4.3.10 *L'inégalité de Poincaré est fausse sur $H^1(\Omega)$. En effet, si u est une fonction constante non nulle, $\|u\|_0 > 0$ mais $|u|_1 = 0$. Par contre, elle reste vraie, comme le montre clairement la démonstration, si l'on considère l'espace des fonctions de $H^1(\Omega)$ qui s'annulent en $x = 0$ (ou en $x = 1$). Elle se généralise facilement au cas où Ω est intervalle de taille finie quelconque, sous la forme*

$$\|u\|_{L^2} \leq C|u|_{H^1},$$

où on peut prendre pour C la longueur de l'intervalle Ω.

Donnons, pour être complet, une dernière propriété d'injection continue.

Proposition 4.3.8 *On a $H^1(\Omega) \hookrightarrow \mathcal{C}^{1/2}(\bar{\Omega})$ l'espace des fonctions hölderiennes d'exposant $1/2$. De plus, l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte.*

Démonstration. On sait déjà que $H^1(\Omega) \hookrightarrow \mathcal{C}^0(\bar{\Omega})$. Il faut montrer qu'en outre les fonctions de $H^1(\Omega)$ sont höldériennes d'exposant $1/2$. Or, pour $u \in H^1(\Omega)$,

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

d'où

$$|u(x) - u(y)| \leq |x-y|^{1/2}\|u'\|_{L^2},$$

par l'inégalité de Cauchy-Schwarz, donc u est höldérienne d'exposant $1/2$. De plus,

$$\max_{x \neq y \in \bar{\Omega}} \frac{|u(x) - u(y)|}{|x-y|^{1/2}} \leq \|u'\|_{L^2},$$

d'où

$$\|u\|_{\mathcal{C}^{1/2}(\bar{\Omega})} = \|u\|_{\mathcal{C}^0(\bar{\Omega})} + \max_{x\neq y\in\bar{\Omega}} \frac{|u(x)-u(y)|}{|x-y|^{1/2}} \leq C\|u\|_{H^1}$$

et l'injection est continue.

On dit qu'une application linéaire est compacte si elle transforme les bornés en ensembles relativement compacts. Par linéarité, cela revient à affirmer que l'image de la boule unité est relativement compacte. Soit B la boule unité de $H^1(\Omega)$. Pour tout $u \in B$, d'après ce qui précède,

$$|u(x)-u(y)| \leq |x-y|^{1/2}, \qquad x,y \in \Omega,$$

donc B est une partie équicontinue de $\mathcal{C}^0(\bar{\Omega})$, bornée. Par le théorème d'Ascoli, on en déduit que B est relativement compacte dans $\mathcal{C}^0(\bar{\Omega})$, d'où aussi dans $L^2(\Omega)$ par l'injection continue triviale $\mathcal{C}^0(\bar{\Omega}) \hookrightarrow L^2(\Omega)$. Ceci nous montre que l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ (ainsi que dans $\mathcal{C}^0(\bar{\Omega})$) est compacte. □

Remarque 4.3.11 *L'injection $H^1(\Omega) \hookrightarrow \mathcal{C}^{1/2}(\bar{\Omega})$ n'est valable qu'en dimension 1, mais la compacité de l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ reste vraie en dimension supérieure, pour Ω borné et de frontière régulière (en un sens à préciser). C'est le très important théorème de Rellich. Il implique que si l'on a une suite bornée dans $H^1(\Omega)$, alors on peut en extraire une sous-suite qui converge dans $L^2(\Omega)$ (et uniformément en dimension 1).*

Remarque 4.3.12 *Ici encore les inclusions sont strictes : $H^1(\Omega) \neq \mathcal{C}^{1/2}(\bar{\Omega})$, en effet, $\sqrt{x} \in \mathcal{C}^{1/2}(\bar{\Omega})$ mais $\sqrt{x} \notin H^1(\Omega)$. De plus, on ne peut pas faire mieux : $H^1(\Omega) \not\subset \mathcal{C}^{\alpha}(\bar{\Omega})$ pour tout $\alpha > 1/2$.*

4.4 Application au problème aux limites

Appliquons les notions que l'on vient d'introduire au problème variationnel associé au problème aux limites.

Théorème 4.4.1 *Pour tout $c \in L^\infty(\Omega)$, $c \geq 0$ et tout $f \in L^2(\Omega)$, le problème variationnel : trouver $u \in H^1_0(\Omega)$ tel que*

$$\int_\Omega (u'v' + cuv) = \int_\Omega fv, \quad v \in H^1_0(\Omega), \tag{4.11}$$

admet une solution et une seule. De plus, on a l'estimation $\|u\|_{H^1} \leq 2\|f\|_{L^2}$.

Démonstration. Il faut vérifier que les hypothèses du théorème de Lax-Milgram sont bien satisfaites. Cette fois-ci, on travaille bien dans un espace de Hilbert $V = H^1_0(\Omega)$. La forme bilinéaire et la forme linéaire sont continues :

$$|a(u,v)| \leq \max(1, \|c\|_{L^\infty})\|u\|_{H^1}\|v\|_{H^1},$$
$$|\ell(v)| \leq \|f\|_{L^2}\|v\|_{H^1},$$

par diverses applications de l'inégalité de Cauchy-Schwarz. La forme bilinéaire est bien coercive. En effet, comme c est positive, pour tout $v \in H_0^1(\Omega)$,

$$\int_\Omega [(v')^2 + cv^2] \geq |v|_{H^1}^2 \geq \frac{1}{2}\|v\|_{H^1}^2$$

par l'équivalence des normes. Donc on a existence et unicité de la solution. Pour l'estimation de cette dernière, on prend comme fonction-test $v = u$, ce qui donne

$$\frac{1}{2}\|u\|_{H^1}^2 \leq \int_\Omega [(u')^2 + cu^2] = \int_\Omega fu \leq \|f\|_{L^2}\|u\|_{H^1},$$

et l'on conclut en divisant par $\|u\|_{H^1}$ (quand $u \neq 0$, sinon il n'y a rien à montrer). □

Remarque 4.4.1 *Cette dernière inégalité signifie que l'application linéaire qui à f associe u est continue de $L^2(\Omega)$ dans $H_0^1(\Omega)$.*

Remarque 4.4.2 *Quand c et f sont en outre continues sur $[0,1]$, on sait qu'il existe une solution classique du problème aux limites qui est de classe C^2. Par unicité elle coïncide avec la solution variationelle u.*

On a ainsi résolu directement un problème variationnel, qui redonne la solution du problème aux limites quand les données sont bien régulières. Mais on a obtenu bien plus : on peut maintenant résoudre pour le même prix le problème variationnel avec des fonctions c éventuellement discontinues (mais toujours bornées) et des fonctions f éventuellement discontinues et non bornées. On a donc considérablement généralisé le champ d'applications possible. Le revers de la médaille est que l'on sait pas encore dans ce cas s'il y a un problème aux limites raisonnable associé au problème variationnel. On remarque que pour la formulation variationnelle considérée, les conditions aux limites sont naturellement vérifiées par la solution puisqu'on la cherche dans $H_0^1(\Omega)$. Il faut encore donner un sens faible à la dérivée u''.

Pour traiter cette dernière question, on a besoin d'autres espaces de Sobolev.

Définition 4.4.1 *Soit $m \in \mathbb{N}$. On définit par récurrence sur $k \geq 1$ l'espace de Sobolev*

$$H^m(\Omega) = \{v \in H^1(\Omega) \ : \ v' \in H^{m-1}(\Omega)\},$$

en posant par convention $H^0(\Omega) = L^2(\Omega)$.

Il est clair, d'après cette définition, qu'une fonction de $H^m(\Omega)$ admet m dérivées faibles successives (*i.e.*, pour $k \geq 2$, v' admet une dérivée faible, noté v'', qui admet elle-même une dérivée faible, notée v''' ou $v^{(3)}$, etc. Ces dérivées sont caractérisées par la relation

$$\int_\Omega v^{(l)}\varphi = (-1)^l \int_\Omega v\varphi^{(l)}, \quad \varphi \in \mathcal{D}(\Omega).$$

On note aussi que l'injection de $H^1(\Omega)$ dans $\mathcal{C}^0(\Omega)$ entraine par récurrence celle de $H^m(\Omega)$ dans $\mathcal{C}^{m-1}(\Omega)$: les $m-1$ premières dérivées sont donc des dérivées classiques.

Nous reviendrons sur ces espaces dans un chapitre suivant. Muni de la norme

$$\|v\|_{H^m} = \Big(\sum_{l=0}^{m} \|u^{(l)}\|_{L^2}^2\Big)^{1/2},$$

$H^m(\Omega)$ est un espace de Hilbert (la démonstration est identique au cas $m = 1$). On note par ailleurs

$$|v|_{H^m} = \|u^{(m)}\|_{L^2},$$

la semi-norme associée à cet espace.

Proposition 4.4.1 *On se donne $c \in L^\infty(\Omega)$, $c \geq 0$, et $f \in L^2(\Omega)$. Soit $u \in H_0^1(\Omega)$ la solution du problème variationnel (4.11). Alors $u \in H^2(\Omega)$ et*

$$-u'' + cu = f \text{ au sens de } L^2(\Omega),$$

donc en particulier presque partout.

Démonstration. Comme $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, on peut prendre dans (4.11) des fonctions-test de $\mathcal{D}(\Omega)$. Ceci donne

$$\forall \varphi \in \mathcal{D}(\Omega), \qquad \int_\Omega u'\varphi' = -\int_\Omega (cu - f)\varphi.$$

Or $f \in L^2(\Omega)$ d'une part et $u \in L^2(\Omega)$ et $c \in L^\infty(\Omega)$ d'autre part, d'où $cu \in L^2(\Omega)$ puisque

$$\int_\Omega (cu)^2 \leq \|c\|_{L^\infty(\Omega)}^2 \int_\Omega u^2.$$

Donc, $g = cu - f \in L^2(\Omega)$ et on reconnaît la définition de la dérivée faible pour u' : u' a une dérivée faible u'' qui coincide avec $cu - f$. On a ainsi montré que $u \in H^2(\Omega)$ et $u'' = cu - f$, donc u vérifie le problème aux limites $-u'' + cu = f$, et $u(0) = u(1) = 0$ puisque $u \in H_0^1(\Omega)$. □

Remarque 4.4.3 *Attention : il ne s'agit pas d'une intégration par partie au sens usuel. On ne fait qu'utiliser la définition des dérivées faibles. Bien sûr, si les données sont régulières, la solution l'est aussi et les dérivées faibles coïncident avec les dérivées usuelles.*

Remarque 4.4.4 *On peut bien entendu aller plus loin en régularité, en établissant par récurrence le résultat suivant : si $f \in H^m(\Omega)$ et $c \in \mathcal{C}^m(\overline{\Omega})$, alors la solution $u \in H_0^1(\Omega)$ du problème variationnel appartient à $H^{m+2}(\Omega)$. Pour cela il suffit de montrer que si*

$$v \in H^m(\Omega) \quad \text{et} \quad c \in \mathcal{C}^m(\overline{\Omega}) \implies cv \in H^m(\Omega),$$

avec la règle de Leibniz qui s'applique aux dérivées faibles (exercice). La formulation variationnelle nous montre alors que la dérivée faible de u appartient à $H^{m+1}(\Omega)$.

Remarque 4.4.5 *Les conditions aux limites de Dirichlet apparaissent par l'intermédiaire de l'espace $H_0^1(\Omega)$. Dans le cas d'un problème aux limites avec des conditions de Neumann $u'(0) = \gamma$ et $u'(1) = \delta$, nous avons déjà indiqué à la fin de la section 4.1 que ce ne serait pas le cas. Dans ce cas la formulation variationnelle correspondante est posée dans l'espace $V = H^1(\Omega)$ tout entier, la forme linéaire a reste la même, et il faut ajouter la condition $c \geq \beta$, où β est une constante strictement positive, pour qu'elle soit coercive sur $H^1(\Omega)$. Enfin on modifie la forme linéaire ℓ en la remplacant par $\tilde{\ell}(v) = \int_\Omega fv + \delta v(1) - \gamma v(0)$. Il est là aussi possible de montrer que si $f \in L^2(\Omega)$, alors la solution u de la formulation variationelle appartient à $H^2(\Omega)$ et vérifie le problème aux limites $-u'' + cu = f$ avec conditions de Neumann.*

Pour clore ce chapitre, rappelons aussi que la solution de (4.11) minimise la fonction

$$J(v) = \frac{1}{2}\int_\Omega (v')^2 - \int_\Omega fv$$

sur $H_0^1(\Omega)$, c'est à dire $J(u) = \inf J(v)$ borne inférieure prise sur tous les $v \in H_0^1(\Omega)$.

Annexe : complétude de $L^2(\Omega)$

Montrons rapidement que $L^2(\Omega)$ est complet. On se donne une suite de Cauchy u_n dans $L^2(\Omega)$. Pour tout $\varepsilon > 0$, il existe n_0 tel que pour tout $n, m \geq n_0$, $\|u_n - u_m\|_{L^2} \leq \varepsilon$. Prenant $\varepsilon = 2^{-k}$ pour $k \in \mathbb{N}$, on extrait donc une sous-suite u_{n_k} telle que $\|u_{n_k} - u_{n_{k+1}}\|_{L^2} \leq 2^{-k}$. On considère la fonction

$$g(x) = |u_{n_0}| + \sum_{k=0}^{\infty} |u_{n_{k+1}} - u_{n_k}|(x).$$

Cette fonction est bien définie comme série à termes positifs et prend ses valeurs dans $[0, +\infty]$. De plus, c'est une limite ponctuelle de fonctions mesurables, elle est donc mesurable. Comme

$$\|g\|_{L^2} \leq \|u_{n_0}\|_{L^2} + \sum_{k=0}^{\infty} 2^{-k} \leq \|u_{n_0}\|_{L^2} + 2,$$

g appartient en fait à $L^2(\Omega)$. En particulier g est finie presque partout.

Par conséquent, il existe un ensemble de mesure nulle N tel que $g(x) < +\infty$ si $x \notin N$ et pour x en dehors de cet ensemble la série

$$u(x) = u_{n_0} + \sum_{k=0}^{\infty} (u_{n_{k+1}} - u_{n_k})(x)$$

est absolument convergente. Elle définit donc une fonction presque partout, mesurable et, comme $|u(x)| \leq g(x)$, appartenant à $L^2(\Omega)$. Pour conclure, il suffit de montrer que $u_{n_p} \to u$ dans $L^2(\Omega)$ quand $p \to +\infty$. Or $u_{n_p} - u \to 0$ presque partout, et de plus, $|u_{n_p} - u| \leq |g|$ avec g dans $L^2(\Omega)$. Le théorème de convergence dominée de Lebesgue montre donc que $\|u_{n_p} - u\|_{L^2} \to 0$ quand $p \to +\infty$. La propriété de suite de Cauchy entraîne finalement que $\|u_n - u\|_{L^2} \to 0$ quand $n \to +\infty$. □

On a incidemment montré la réciproque partielle du théorème de convergence dominée de Lebesgue, à savoir que de toute suite convergente dans $L^2(\Omega)$, on peut extraire une sous-suite qui converge presque partout et qui est dominée par une fonction de $L^2(\Omega)$.

Chapitre 5

Les méthodes d'approximation variationnelle

Les formulations variationnelles se prêtent très naturellement à la définition de méthodes d'approximation, c'est-à-dire de réduction à une suite de problèmes en dimension finie que l'on peut effectivement résoudre sur ordinateur. L'objet de ce bref chapitre est d'établir les propriétés communes à toutes ces méthodes du point de vue abstrait.

5.1 Définition et premières propriétés

On considère donc un problème variationnel abstrait bien posé, à savoir on se donne un espace de Hilbert V, a une forme bilinéaire continue coercive et ℓ une forme linéaire continue sur V et on cherche $u \in V$ tel que

$$\text{(PVA)} \qquad a(u,v) = \ell(v), \qquad v \in V$$

Définition 5.1.1 *Une méthode d'approximation variationnelle consiste en la donnée d'un sous-espace vectoriel $V_n \subset V$ de dimension finie. On cherche alors $u_n \in V_n$ solution du problème discret*

$$a(u_n, v_n) = \ell(v_n), \qquad v_n \in V_n. \tag{5.1}$$

On parle également de *méthode de Galerkin*. Comme $V_n \subset V$, on dit qu'il s'agit d'une approximation *conforme.* On peut définir aussi des méthodes d'approximation non conformes pour lesquelles $V_n \not\subset V$ (nous n'en rencontrerons pas ici). Comme on le montre plus loin, l'intérêt de cette méthode est que le calcul de u_n se ramène à la résolution d'un système linéaire de taille $d_n \times d_n$ où on a posé

$$d_n := \dim(V_n).$$

La méthode de Galerkin a donc pour objectif d'approcher la solution exacte $u \in V$ par la solution calculable $u_n \in V_n$. Afin d'améliorer la précision, on considère typiquement une suite d'espaces $(V_n)_{n\geq 1}$ dont la dimension d_n augmente avec n. Dans certains cas, les sous-espaces de dimension finie seront emboîtés, $V_n \subset V_m$ pour tout $m \geq n$. Dans d'autres cas, cette propriété ne sera pas forcément vérifiée. On pourra aussi avoir exactement $d_n = n$, mais cela ne sera pas toujours le cas.

Théorème 5.1.1 *Le problème discret (5.1) admet une solution u_n et une seule.*

Démonstration. Tout sous-espace vectoriel de dimension finie est complet, donc ici V_n est un espace de Hilbert. Les formes bilinéaire a et linéaire ℓ sont naturellement continues et a reste coercive sur V_n. On peut appliquer le théorème de Lax-Milgram dans V_n. □

On sait exactement quand une méthode d'approximation variationnelle est convergente. On peut même préciser quantitativement l'erreur d'approximation. C'est en fait *l'estimation fondamentale* des méthodes d'approximation variationnelle, exprimé par le résultat fondamental suivant.

Théorème 5.1.2 *(Lemme de Cea) Il existe une constante C, indépendante de n et de u, telle que*

$$\|u-u_n\|_V \leq C \min_{v_n \in V_n} \|u-v_n\|_V, \tag{5.2}$$

avec $C = C_a/\alpha$, où C_a et α sont les constantes de continuité et de coercivité de la forme bilinéaire a. Dans le cas où la forme bilinéaire a est symmétrique, on obtient le même résultat avec $C = (C_a/\alpha)^{1/2}$.

Démonstration. La formulation variationelle indique en particulier que $a(u,v_n) = \ell(v_n)$ pour tout $v_n \in V_n$, et par soustraction avec l'équation $a(u_n,v_n) = \ell(v_n)$, on trouve

$$a(u-u_n,v_n) = 0.$$

Par ailleurs, d'après la V-ellipticité et la continuité de a, on a

$$\begin{aligned}\alpha\|u-u_n\|_V^2 &\leq a(u-u_n,u-u_n) = a(u-u_n,u-v_n+v_n-u_n)\\ &= a(u-u_n,u-v_n)+a(u-u_n,v_n-u_n)\\ &= a(u-u_n,u-v_n) \leq C_a\|u-u_n\|_V\|u-v_n\|_V,\end{aligned}$$

car $v_n - u_n \in V_n$. D'où en divisant par $\|u-u_n\|_V$ (dans le cas où cette quantité n'est pas nulle, sinon il n'y a rien à démontrer), il vient

$$\forall v_n \in V_n, \qquad \|u-u_n\|_V \leq \frac{C_a}{\alpha}\|u-v_n\|_V$$

d'où l'inégalité désirée, avec $C = C_a/\alpha$.

Dans le cas où a est symétrique, la propriété

$$a(u-u_n,v_n) = 0, \qquad v_n \in V_n$$

s'interprète en disant que u_n est la projection orthogonale de u sur le sous-espace vectoriel fermé V_n pour le produit scalaire $(\cdot,\cdot)_a = a(\cdot,\cdot)$. On a par conséquent, pour tout $v_n \in V_n$,

$$\alpha\|u-u_n\|_V^2 \leq \|u-u_n\|_a^2 \leq \|u-v_n\|_a^2 \leq C_a\|u-v_n\|_V^2,$$

ce qui entraîne l'estimation (5.2) avec la constante $C = (C_a/\alpha)^{1/2}$ □

Remarque 5.1.1 *La quantité $u-u_n$ est l'*erreur *commise par la méthode d'approximation variationnelle. Cette erreur en norme V est majorée par une constante qui ne dépend que du problème variationnel abstrait, et non pas du choix de méthode, multipliée par la distance de la solution u au sous-espace V_n. On se ramène donc à estimer cette distance. Il s'agit alors d'un problème d'approximation. Les estimations de ces erreurs d'approximation vont elles bien sûr dépendre fortement du choix de méthode, et c'est là que celles-ci vont commencer à se distinguer entre elles.*

Remarque 5.1.2 *Puisque $u_n \in V_n$ on a aussi trivialement l'inégalité*

$$\min_{v_n \in V_n} \|u-v_n\|_V \leq \|u-u_n\|_V,$$

ce qui montre que les quantités $\|u-u_n\|_V$ et $\min_{v_n\in V_n}\|u-v_n\|_V$ sont équivalentes, indépendamment du choix de l'espace V_n.

On obtient ainsi comme conséquence immédiate une condition sur les espaces V_n pour la convergence de la méthode de Galerkin.

Corollaire 5.1.3 *Soit $u \in V$ la solution du problème variationnel abstrait, et u_n celle du problème discret. Alors $u_n \to u$ dans V si et seulement si pour tout $v \in V$, il existe une suite (w_n) de V telle que $\forall n \in \mathbb{N}$, $w_n \in V_n$ et telle que $w_n \to v$ dans V quand $n \to \infty$.*

Une méthode d'approximation variationnelle est donc sûre de converger si et seulement si les sous-espaces V_n deviennent «assez gros» quand n tend vers l'infini, au sens où l'on peut approcher tout élément de V arbitrairement près au sens de la norme de V par un élément de V_n pour n assez grand.

Dans l'exemple qui nous intéresse $V = H_0^1(\Omega)$, $c \in L^\infty(0,1)$, $c \geq 0$ et $f \in L^2(0,1)$,

$$a(u,v) = \int_\Omega (u'v' + cuv), \quad \ell(v) = \int_\Omega fv,$$

on peut donner deux exemples traditionnels de sous espace de dimension finie V_n.

Le premier exemple, que nous allons développer dans le chapitre suivant, conduit à la méthode des éléments finis P_1. On prend une subdivision $0 = x_0^{(n)} < x_1^{(n)} < \ldots < x_n^{(n)} < x_{n+1}^{(n)} = 1$ de l'intervalle $[0,1]$, et on prend pour V_n les fonctions continues, affines par morceaux sur chaque intervalle $[x_i^{(n)}, x_{i+1}^{(n)}]$, nulles en 0 et 1. C'est un sous espace de $H_0^1(\Omega)$, de dimension finie $d_n = n$. On pose $h_n = \max_i(x_{i+1}^{(n)} - x_i^{(n)})$, on suppose que $h_n \to 0$ quand $n \to \infty$. Dans ce cas, les espaces ne sont pas nécessairement emboités (les points $x_i^{(n+1)}$ de la subdivision d'indice $n+1$ ne contiennent pas nécessairement les points $x_j^{(n)}$).

Dans le deuxième exemple, à l'opposé, on prend pour V_n les fonctions p_n globalement polynômiales de degré inférieur ou égal à n, sur $[0,1]$, nulles en 0 et 1. Un tel polynôme s'écrit $p_n = (1-x)xq_{n-2}$ où q_{n-2} est un polynôme de degré inférieur ou égal à $n-2$, et l'espace correspondant V_n est de dimension $n-1$. On a dans ce cas $V_n \subset V_{n+1}$. Ce choix conduit aux *méthodes spectrales* qui seront abordées dans le chapitre 7 dans le cadre général de méthodes fondées sur des bases hilbertiennes.

5.2 Forme matricielle de la méthode de Galerkin

En vue de l'implémentation effective sur ordinateur de la méthode de Galerkin, il convient de réinterpréter celle-ci en termes de systèmes linéaires. Pour cela, on doit d'abord se donner une base de V_n, soit w_i, $i = 1, \ldots, d_n$ où on a noté $\dim V_n = d_n$. La solution du système discret se décompose sur cette base sous la forme $u_n = \sum_{i=1}^{d_n} \lambda_i w_i$, et calculer u_n est équivalent à calculer ses composantes λ_i dans la base choisie w_i. On notera

$$\bar{u}_n = (\lambda_1, \ldots, \lambda_{d_n})^t$$

le vecteur colonne de $\mathbb{R}^{d_n}$ correspondant.

Théorème 5.2.1 *Le vecteur $\bar{u}_n \in \mathbb{R}^{d_n}$ est l'unique solution du système linéaire*

$$A_n \bar{u}_n = b_n$$

avec A_n la matrice de coefficients $(A_n)_{ij} = a(w_j, w_i)$ et b_n le vecteur de composantes $(b_n)_i = \ell(w_i)$.

Démonstration. On remarque que la solution discrete u_n est caractérisée par les n équations

$$a(u_n, w_i) = \ell(w_i), \quad i = 1, \ldots, d_n,$$

car pour tout $v_n \in V_n$, on peut écrire $v_n = \sum_{i=1}^{d_n} \mu_i w_i$. En sommant les équations multipliées par μ_i on obtient ainsi $a(u_n, v_n) = \ell(v_n)$.

L'équation i se développe suivant

$$\sum_{j=1}^{d_n} \lambda_j a(w_i, w_j) = \ell(w_i) = (b_n)_i,$$

et ceci montre que $\bar{u}_n$ est solution du système annoncé. □

Une fois choisie une base de V_n, on se ramène donc finalement à la construction de la matrice A_n, parfois appelée *matrice de rigidité* et à celle du second membre b_n, puis à la résolution effective du système linéaire obtenu. Cette matrice a automatiquement de bonnes propriétés.

Proposition 5.2.1 *La matrice A_n est définie positive (et par conséquent inversible).*

Démonstration. Soit $v_n = \sum_{i=1}^{d_n} \mu_i w_i$ un élément quelconque de V_n et $\bar{v}_n$ le vecteur de $\mathbb{R}^{d_n}$ associé. On a

$$\bar{v}_n^T A_n \bar{v}_n = \sum_{i,j=1}^{d_n} \mu_i \mu_j a(w_j, w_i) = a(v_n, v_n) \geq \alpha \|v_n\|_V^2.$$

Donc $\bar{v}_n^T A_n \bar{v}_n \geq 0$ et $\bar{v}_n^T A_n \bar{v}_n = 0$ implique $v_n = 0$, donc bien sûr $\bar{v}_n = 0$. □

Remarque 5.2.1 *Si de plus la forme bilinéaire a est symétrique, alors la matrice A_n est aussi symétrique. On pourra dans ce cas utiliser des méthodes de résolution de systèmes linéaires adaptées aux matrices symétriques, définies positives (comme la méthode de Cholesky, par exemple). Ceci est à comparer aux matrices produites par la méthode des différences finies qui n'avaient ces même bonnes propriétés que par accident finalement, et encore pour un bon choix de numérotation des nœuds. Rien de tel ici.*

Notons enfin une dernière propriété des méthodes d'approximation variationnelles dans le cas symétrique.

Proposition 5.2.2 *Lorsque la forme bilinéaire a est de plus symétrique, le problème variationnel discret est équivalent au problème de minimisation quadratique : minimiser la fonction*

$$J(\bar{v}_n) = \frac{1}{2}\bar{v}_n^T A_n \bar{v}_n - b_n^T \bar{v}_n$$

sur $\mathbb{R}^{d_n}$.

Démonstration. Évident (le faire !). On a $J'(\bar{u}_n).v_n = \bar{u}_n^T A_n \bar{v}_n - b_n^T \bar{v}_n$. □

Pour la résolution effective du système linéaire, on pourra donc également appliquer des méthodes de descente : gradient à pas optimal, mais surtout gradient conjugué et ses multiples variantes. Ces méthodes calculent la solution du système avec une erreur qui décroit au cours des itérations et qu'il faut prendre en compte en plus de l'erreur d'approximation $\|u-u_n\|_V$ de la méthode de Galerkin.

Remarque 5.2.2 *On rappelle en particulier que la rapidité de convergence des méthode itératives est intimement liée au nombre de conditionnement, qui dans le cas d'une matrice A symmétrique définie positive est donné par*

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},$$

où $\lambda_{\max}(A)$ *et* $\lambda_{\min}(A)$ *désignent la plus grande et plus petite valeur propre. Prenons par exemple la descente de gradient à pas fixe* τ *appliquée à J. L'itération* $k \to k+1$ *a la forme*

$$\bar{u}_n^{k+1} = \bar{u}_n^k + \tau(b_n - A_n \bar{u}_n^k).$$

L'erreur $\bar{e}_n^k := \bar{u}_n - \bar{u}_n^k$ *vérifie* $\bar{e}_n^k = (I-\tau A_n)\bar{e}_n^k = \cdots = (I-\tau A_n)^k \bar{e}_n^0$. *On pourra vérifier (exercice) que le pas qui minimise la norme matricielle* $\|I-\tau A_n\|_{\ell^2\to\ell^2}$ *est* $\tau = 2(\lambda_{\min}(A_n)+\lambda_{\max}(A_n))^{-1}$, *et que la vitesse de convergence de l'erreur est alors en* $O(\rho^k)$, *où* $\rho = \frac{\kappa(A_n)-1}{\kappa(A_n)+1}$. *On voit ainsi que cette vitesse se détériore lorsque* $\kappa(A_n)$ *est très grand.*

5.3 Perturbation d'une approximation variationelle

On a vu que l'assemblage du système linéaire nécessite le calcul des éléments de matrice $a(w_j, w_i)$ et des fonctionnelles $\ell(w_i)$. En pratique ces éléments sont donnés par des intégrales : par exemple dans le cas du problème aux limites (P), on a

$$a(w_j,w_i) = \int_\Omega (w_j'(x)w_i'(x) + c(x)w_j(x)w_i(x)dx \quad \text{et} \quad \ell(w_i) = \int_\Omega f(x)w_i(x)dx.$$

Dans certains cas, ces intégrales sont calculables de manière exacte. Le plus souvent, il est nécessaire de les approcher par de formules de quadratures (rectangles, trapezes, Simpson, Gauss-Legendre...).

De façon générale, de telles quadratures approchent pour toute fonction v une intégrale $I(w) = \int_\Omega w(x)dx$ par une formule discrète

$$Q(w) = \sum_{j=1}^{p} \omega_j w(y_j),$$

où les y_j sont des points fixés sur Ω et les ω_j sont des poids réels fixés. Le choix des points et des poids est important pour assurer la précision de la quadrature, c'est à dire la petitesse de l'erreur $|I(w) - Q(w)|$ pour certaines classes de fonctions w que l'on souhaite intégrer.

Notre objectif n'est pas ici de rentrer dans les détails de telles méthodes mais de comprendre ce que l'erreur de quadrature induit sur le calcul de la solution discrète. Une manière de faire cette analyse est de remarquer que toutes les méthodes de quadrature employées reviennent à *perturber* les formes a et ℓ et les remplacer par des approximations $\tilde{a}$ et $\tilde{\ell}$. Par exemple, avec une quadrature du type ci-dessus, on a

$$\tilde{\ell}(v) = \sum_{j=1}^{p} \omega_j v(y_j) f(y_j).$$

Cela signifie que l'on résoud en fait le probleme discret perturbé suivant : trouver $\tilde{u}_n \in V_n$ tel que

$$\tilde{a}(\tilde{u}_n, v_n) = \tilde{\ell}(v_n), \quad v_n \in V_n. \tag{5.3}$$

La solution $\tilde{u}_n$ sera différente de la solution de Galerkin u_n, et il n'est même pas aquis que cette nouvelle solution soit bien définie. Afin de comprendre cela, on fait les hypothèse suivantes sur les perturbations induites par la formule de quadrature, et on obtient un contrôle de la déviation entre u_n et $\tilde{u}_n$. C'est l'objet du résultat suivant, appelé parfois *Lemme de Strang*.

Théorème 5.3.1 *On suppose qu'il existe deux constantes positives ε_1 et ε_2 telles que*

$$|\ell(v_n) - \tilde{\ell}(v_n)| \leq \varepsilon_1 \|v_n\|_V, \quad v_n \in V_n, \tag{5.4}$$

et

$$|a(v_n, w_n) - \tilde{a}(v_n, w_n)| \leq \varepsilon_2 \|v_n\|_V \|w_n\|_V, \quad v_n, w_n \in V_n. \tag{5.5}$$

On suppose aussi que $\varepsilon_2 < \alpha$ où α est la constante de coercivité de a. Alors, il existe une unique solution $\tilde{u}_n \in V_n$ de (5.3)*, et celle-ci vérifie*

$$\|u_n - \tilde{u}_n\|_V \leq \frac{\varepsilon_1 + \varepsilon_2 \frac{C_\ell}{\alpha}}{\alpha - \varepsilon_2} \tag{5.6}$$

où la constante C dépend de C_1, C_2, et α.

Démonstration. Par l'inégalité triangulaire, les hypothèses (5.4) et (5.5) entrainent

$$|\tilde{\ell}(v_n)| \leq (C_\ell + \varepsilon_1)\|v_n\|_V, \quad v_n \in V_n,$$

ainsi que

$$|\tilde{a}(v_n, w_n)| \leq (C_a + \varepsilon_2)\|v_n\|_V \|w_n\|_V, \quad v_n, w_n \in V_n.$$

On a aussi

$$|\tilde{a}(v_n, v_n)| \geq (\alpha - \varepsilon_2)\|v_n\|_V^2, \quad v_n \in V_n.$$

Les hypothèse de Lax-Milgram sont donc vérifiées sur V_n avec les constantes de continuité et de coercivité perturbées $C_{\tilde{\ell}} = C_\ell + \varepsilon_1$, $C_{\tilde{a}} = C_A + \varepsilon_2$ et $\tilde{\alpha} = \alpha - \varepsilon_2 > 0$, ce qui assure l'existence et l'unicité de la solution $\tilde{u}_n$.

Pour controler la distance entre u_n et $\tilde{u}_n$ on fait la différence des équations qui les définissent et on obtient, pour tout $v_n \in V_n$,

$$a(u_n, v_n) - \tilde{a}(\tilde{u}_n, v_n) = \ell(v_n) - \tilde{\ell}(v_n).$$

Ceci peut aussi s'écrire

$$\tilde{a}(u_n - \tilde{u}_n, v_n) = \ell(v_n) - \tilde{\ell}(v_n) + \tilde{a}(u_n, v_n) - a(u_n, v_n).$$

Les hypothèse de perturbations permettent de majorer les termes de droite, ce qui donne

$$\tilde{a}(u_n - \tilde{u}_n, v_n) \leq \varepsilon_1 \|v_n\|_V + \varepsilon_2 \|u_n\|_V \|v_n\|_V.$$

En prenant $v_n = u_n - \tilde{u}_n$ et en utilisant l'ellipticité de $\tilde{a}$ on obtient

$$\tilde{\alpha}\|u_n - \tilde{u}_n\|_V \leq \varepsilon_1 + \varepsilon_2 \|u_n\|_V \leq \varepsilon_1 + \varepsilon_2 \frac{C_\ell}{\alpha},$$

où la deuxième inégalité provient de l'estimation a-priori qur u_n. On obtient ainsi l'estimation (5.6) annoncée □

En résumé, on peut dire que toutes les méthodes d'approximation variationnelles (conformes) partagent la même structure abstraite, et notamment l'estimation d'erreur fondamentale (5.2). Tout l'art ensuite va résider dans le choix des espaces de dimension finie, qui devront approcher le mieux possible la solution u, puis dans le choix d'une base de V_n qui produise des matrices faciles à calculer et à résoudre (par exemple des matrices aussi creuses que possible). A l'erreur d'approximation entre u et u_n viennent éventuellement s'ajouter les erreurs supplémentaires liées aux perturbations dues aux quadratures, ainsi que l'erreur d'itération lorsqu'on ne résoud pas le système linéaire de manière exacte.

Chapitre 6

La méthode des éléments finis

La méthode des éléments finis est peut-être la plus importante des méthodes d'approximation variationnelle. Nous allons la décrire en détail dans le contexte de l'approximation variationnelle du problème aux limites du second ordre avec conditions de Dirichlet, en dimension un d'espace. Mais elle est d'une grande souplesse et est très souvent employée dans la pratique dans des contextes très variés (en particulier, pas seulement pour les problèmes aux limites du second ordre, pas seulement pour les conditions de Dirichlet et pas seulement en dimension un d'espace !).

6.1 Définition de la méthode dans le cas dit P_1

Pour la méthode des éléments finis, plutôt que d'indexer les sous-espaces vectoriels sur lesquels on va travailler par un entier naturel n, on les indexe traditionnellement par un réel h, qui représente un pas d'espace (en dimension 2, la mesure d'une surface, en dimension 3 d'un volume). Pour l'analyse de convergence, on conviendra que h prend ses valeurs dans une suite qui tend vers 0.

Pour construire la "méthode des éléments finis P_1", on se donne un entier $m \geq 1$ et un *maillage* de l'intervalle $]0,1[$, c'est-à-dire un découpage de l'intervalle en sous-intervalles, appelés *éléments*, délimités par des sommets $0 = x_0 < x_1 < x_2 < \cdots < x_m < x_{m+1} = 1$. Les sommets ne sont pas forcément équidistants. On pose $h = \max_i |x_{i+1} - x_i|$ et on l'appelle le *pas du maillage*.

Il faut noter que bien que ces données soient de même nature que dans le cas de la méthode des différences finies, on emploie ici un vocabulaire différent. En effet, dans la méthode des différences finies, on s'intéressait à des valeurs prises sur les points de la grille. Ici, on va s'intéresser à ce qui se passe dans les éléments, *i.e.*, entre les sommets du maillage. Cette différence de vocabulaire pour désigner des objets analogues n'est qu'un artefact de la dimension un. En dimension deux, par exemple, la méthode des éléments finis va fonctionner sur des maillages formés d'éléments triangulaires ou quadrangulaires, alors que la méthode des différences finies travaillera toujours sur une grille formée de points.

On introduit donc l'espace

$$V_h = \{v \in \mathcal{C}^0([0,1]), v_{|[x_i,x_{i+1}]} \text{ est affine}, v(0) = v(1) = 0\}$$

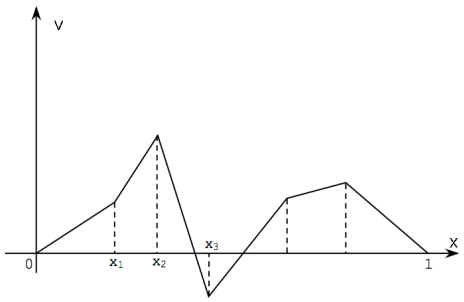


FIGURE 6.1 – Un élément de V_h

des fonctions continues, affines par morceaux sur le maillage et satisfaisant les conditions de Dirichlet. On l'appelle espace d'éléments finis. D'après les résultats du chapitre III, on voit que $V_h \subset V = H_0^1(\Omega)$. Attention quand même à la notation qui est ambiguë : l'espace V_h ne dépend pas uniquement de h, mais bien du maillage tout entier. Deux maillages de même pas h vont donner naissance à deux espaces V_h différents. Nous conserverons quand même cette notation traditionnelle. Enfin les points x_i dépendent évidemment du maillage au sens où on sera amené à prendre des suites de maillages, dont la finesse qui tend vers 0, et alors on explicitera la dépendance des points $x_i = x_i^n$ pour le n-ème maillage de la suite, mais pas lorsqu'on travaille avec un maillage donné, lorsqu'il n'y en a qu'un. Par exemple, dans le cas de subdivision uniformes on a

$$x_i^n = \frac{i}{n+1},$$

et $h_n = \frac{1}{n+1}$. Attention aussi au fait que V désigne désormais l'espace $H_0^1(\Omega)$.

Proposition 6.1.1 *L'espace V_h est de dimension finie m. Le problème variationnel discret associé, trouver $u_h \in V_h$ tel que*

$$\forall v_h \in V_h, \quad \int_\Omega (u_h' v_h' + c u_h v_h) = \int_\Omega f v_h$$

admet une solution et une seule.

Démonstration. Il est évident que V_h est un sous-espace vectoriel de V. Considérons l'application $L\colon V_h \to \mathbb{R}^m$ qui à v associe le vecteur $(v(x_i))_{i=1,\dots,m}$. C'est visiblement une application linéaire injective, car $L(v) = 0$ implique que $v(x_i) = 0$ pour tout $i = 0, 1, \dots, m+1$, et comme v est affine sur chaque intervalle $[x_i, x_{i+1}]$, elle s'annule sur cet intervalle, donc sur $[0,1]$ tout entier. Elle est aussi surjective, car pour n'importe quel vecteur $(\lambda_i) \in \mathbb{R}^m$, il existe une fonction de V_h qui prend la valeur λ_i en x_i (il suffit d'interpoler linéairement sur chaque intervalle). L'application L est donc un isomorphisme entre V_h et $\mathbb{R}^m$. Ces deux espaces ont par conséquent même dimension, c'est-à-dire que $\dim V_h = m$.

L'existence et l'unicité de u_h découlent alors de l'étude abstraite effectuée dans le chapitre précèdent. □

Remarque 6.1.1 *La solution $u \in V$ du problème variationnel continu est également solution du problème aux limites (cf. proposition 4.4.1). Il faut bien noter qu'il n'en va pas de même*

pour u_h. Le problème variationnel résolu par u_h sur V_h n'admet en effet aucune interprétation en termes d'équation différentielle.

On va maintenant exhiber une base de V_h bien adaptée pour la suite.

Théorème 6.1.1 *Les fonctions $w_i \in V_h$, $1 \leq i \leq m$, définies par $w_i(x_j) = \delta_{i,j}$, forment une base de V_h. De plus, on a pour tout $v \in V_h$,*

$$v(x) = \sum_{i=1}^{m} v(x_i) w_i(x).$$

Démonstration. Remarquons d'abord que $w_i \in V_h$ est bien définie. C'est en effet l'image par L^{-1} du $i$ème vecteur de la base canonique de $\mathbb{R}^m$ (le symbole de Kronecker $\delta_{i,j}$ vaut 1 si $j = i$, 0 sinon). La famille des w_i étant l'image d'une base par un isomorphisme est donc une base de V_h.

Soit $v \in V_h$, on note μ_i ses composantes dans la base (w_i). Pour tout $x \in [0,1]$, on a donc

$$v(x) = \sum_{i=1}^{m} \mu_i w_i(x).$$

En particulier, en $x = x_j$, on obtient

$$v(x_j) = \sum_{i=1}^{m} \mu_i w_i(x_j) = \sum_{i=1}^{m} \mu_i \delta_{i,j} = \mu_j,$$

d'où l'expression annoncée des composantes. □

Les fonctions de bases w_i sont parfois appelées *fonctions chapeau*, étant donnée l'allure de leur graphe. On utilise aussi le terme de base *nodale*, chaque fonction w_i étant naturellement associée à un noeud x_i qui est ici un sommet du maillage. La base $(w_i)_{i=1,\ldots,m}$ a une propriété particulière : les composantes d'une fonction de V_h dans cette base sont les valeurs que prend la fonction aux noeuds. Une fois résolu le système linéaire, on aura donc directement des informations exploitables sur la solution discrète en lisant ses composantes, sans autre calcul supplémentaire. Dans le langage des éléments finis, ces valeurs sont appelées *degrés de liberté.*

Remarque 6.1.2 *De façon plus générale, la méthode des éléments finis part d'une partition d'un domaine Ω en fermés d'intérieurs disjoints. En dimension 1 il s'agit d'intervalles, en dimension 2 de triangles ou de quadrangles, en dimension 3 de prismes, tétraèdres ou polyèdres. Chacun de ces fermés, avec ses degrés de liberté, est un élément fini. Les fonctions de l'espace de dimension finie V_h sont polynomiales par morceaux et globalement continues, donc H^1. Cet espace est muni d'une base nodale indexée par les degrés de libertés, qui généralise la base des fonctions chapeau.*

Si v est une fonction continue quelconque et qui s'annule en 0 et 1, on peut l'approcher de manière naturelle par une fonction de V_h en définissant l'unique fonction de V_h qui admet les mêmes valeurs que v aux points du maillage : c'est l'interpolation de Lagrange.

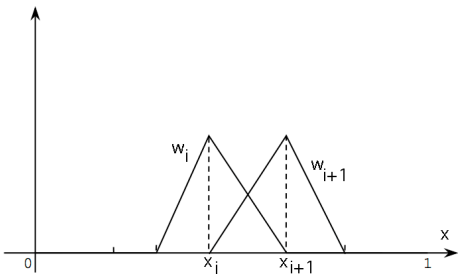


FIGURE 6.2 – Deux fonctions de base successives

Définition 6.1.1 *On appelle fonction interpolée de v dans V_h la fonction $\Pi_h v$ définie par :*

$$\Pi_h v \in V_h \quad et \quad \Pi_h v(x_i) = v(x_i), 1 \leq i \leq m.$$

Il est facile de voir que Π_h est un opérateur linéaire (appelé interpolant ou opérateur d'interpolation) que l'on peut exprimer par

$$\Pi_h v = \sum_{i=1}^{m} v(x_i) w_i.$$

On voit aussi que l'image de Π_h est V_h et que $\Pi_h v = v$ pour tout $v \in V_h$. En ce sens Π_h est un projecteur (qu'il ne faut pas confondre avec le projecteur orthogonal pour un produit scalaire).

Pour aborder les questions de convergence, il faut considérer une suite de maillages dont le pas tend vers 0. Les espaces V_h correspondants ne sont pas forcément emboîtés, sauf si les maillages successifs sont obtenus par *raffinement* les uns des autres.

Proposition 6.1.2 *Soit $(x_i^n)_{i=1,\ldots,m_n}$ une suite de maillages avec $h_n = \max |x_{i+1}^n - x_i^n| \to 0$ quand $n \to +\infty$. Soit u_{h_n} l'approximation variationelle de u dans V_{h_n}. Alors $\|u_{h_n} - u\|_{H^1} \to 0$ quand $n \to +\infty$.*

Démonstration. On a vu qu'il suffit de montrer que l'on peut approcher tout élément v de $V = H_0^1(\Omega)$ par une suite $v_n \in V_{h_n}$. Comme $\mathcal{D}(\Omega)$ est dense dans $H_0^1(\Omega)$, il suffit d'approcher en norme H^1 n'importe quel élément de v de $\mathcal{D}(\Omega)$ par une suite $v_n \in V_{h_n}$ (par un argument d'extraction diagonale, exercice !). On va plus précisément montrer que si $v \in \mathcal{D}(\Omega)$ alors son interpolation

$$v_n = \Pi_{h_n} v,$$

converge vers v en norme H^1.

On montre tout d'abord la convergence en norme L^2. Pour cela il suffit de remarquer que pour tout $x \in [x_i^n, x_{i+1}^n]$ on a d'une part par le théorème des accroissements finis

$$|v(x) - v(x_i^n)| \leq M_1 |x - x_i^n| \leq M_1 h_n,$$

où $M_1 = \|v'\|_{L^\infty}$, et d'autre part puisque v_n est affine sur cet intervalle

$$|v_n(x) - v_n(x_i^n)| \leq |v_n(x_{i+1}^n) - v_n(x_i^n)| = |v(x_{i+1}^n) - v(x_i^n)| \leq M_1 h_n.$$

Il vient par conséquent

$$|v(x) - v_n(x)| \leq 2M_1 h_n$$

ce qui montre la convergence uniforme de v_n vers v, et entraîne la convergence en norme L^2.

On montre ensuite la convergence en norme L^2 de v_n' vers v'. La dérivée faible v_n' est donnée sur chaque intervalle $]x_i^n, x_{i+1}^n[$ par la valeur constante

$$v_n'(x) = \frac{v(x_{i+1}^n) - v(x_i^n)}{x_{i+1}^n - x_i^n}.$$

Le théorème des accroissements finis indique qu'il existe $y_i^n \in]x_i^n, x_{i+1}^n[$ tel que cette valeur coincide avec $v'(y_i^n)$. Pour $x \in]x_i^n, x_{i+1}^n[$, on a donc

$$|v_n'(x) - v'(x)| = |v'(y_i^n) - v'(x)| \leq M_2 |y_i^n - x| \leq M_2 h_n,$$

où $M_2 = \|v''\|_{L^\infty}$. Ceci montre la convergence uniforme de v_n' vers v', et entraîne la convergence en norme L^2. Au final on a montré que v_n converge vers v en norme H^1. □

Donnons la forme matricielle de la méthode des éléments finis dans la base des w_i (à maillage fixé). On a vu que $(A_h)_{ij} = a(w_j, w_i) = \int_\Omega (w_j' w_i' + c w_j w_i)$. On remarque d'abord que dès que $|i - j| > 1$, $w_i w_j = w_i' w_j' = 0$ car dans ce cas les supports des fonctions de base sont disjoints ou leur intersection est réduite à un point. On va donc obtenir une matrice au plus tridiagonale. Comme elle est symétrique, il suffit de considérer le cas $i \leq j$, c'est-à-dire soit $i = j$, soit $i + 1 = j$. Supposons pour simplifier que le maillage est uniforme, *i.e.*, $x_i = ih$, avec $h = \frac{1}{m+1}$. On a alors

$$w_i(x) = \begin{cases} 0 & \text{si } x \notin [x_{i-1}, x_{i+1}], \\ \frac{x - x_{i-1}}{h} & \text{si } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1} - x}{h} & \text{si } x \in [x_i, x_{i+1}], \end{cases}$$

et

$$w_i'(x) = \begin{cases} 0 & \text{si } x \notin [x_{i-1}, x_{i+1}], \\ \frac{1}{h} & \text{si } x \in]x_{i-1}, x_i[, \\ -\frac{1}{h} & \text{si } x \in]x_i, x_{i+1}[. \end{cases}$$

Pour $i = j$, il vient donc

$$\int_\Omega (w_i')^2 = \int_{x_{i-1}}^{x_{i+1}} \frac{1}{h^2} = \frac{2}{h}$$

et

$$\int_\Omega c(x)(w_i)^2 = \frac{1}{h^2} \Big[\int_{x_{i-1}}^{x_i} c(x)(x - x_{i-1})^2 + \int_{x_{i-1}}^{x_{i+1}} c(x)(x_{i+1} - x)^2 \Big].$$

Dans le cas où c est une fonction constante, on obtient

$$\int_\Omega c(w_i)^2 = \frac{2ch}{3}.$$

Pour $i + 1 = j$, toujours à cause des supports, $w_i w_{i+1} = w_i' w_{i+1}' = 0$ si $x \notin]x_i, x_{i+1}[$. On obtient alors

$$\int_\Omega w_i' w_{i+1}' = \int_{x_i}^{x_{i+1}} (-\frac{1}{h^2}) = -\frac{1}{h}$$

et

$$\int_\Omega c(x)w_i w_{i+1} = \frac{1}{h^2}\int_{x_i}^{x_{i+1}} c(x)(x_{i+1}-x)(x-x_i).$$

Dans le cas où c est une fonction constante, il vient

$$\int_\Omega c w_i w_{i+1} = \frac{ch}{6}.$$

On obtient par conséquent une matrice A_h de la forme (pour c constante)

$$A_h = \frac{1}{h}\begin{pmatrix} 2+\frac{2ch^2}{3} & -1+\frac{ch^2}{6} & 0 & \dots & 0 \\ -1+\frac{ch^2}{6} & 2+\frac{2ch^2}{3} & -1+\frac{ch^2}{6} & \dots & 0 \\ \ddots & \ddots & \ddots & \dots & \vdots \\ 0 & \dots & 0 & -1+\frac{ch^2}{6} & 2+\frac{2ch^2}{3} \end{pmatrix}$$

Pour le second membre, il faut évaluer les intégrales $\int_\Omega f w_i$. Quand $f = \overline{f}$ est une fonction constante, ces intégrales valent $h\overline{f}$. On remarque que quand la fonction c est constante, la matrice obtenue pour le maillage uniforme ressemble beaucoup à celle de la méthode des différences finies. Cela dit, il n'est pas difficile d'écrire la matrice du système linéaire pour un maillage non uniforme.

Remarque 6.1.3 *Dans le cas où c ou f ne sont pas constantes, il se peut que l'on sache calculer exactement les diverses intégrales qui apparaissent dans le système linéaire, mais la règle générale est que l'on ne puisse pas y arriver. Il faut alors recourir à des méthodes d'intégration numérique pour obtenir des valeurs numériques pour les coefficients de la matrice et pour le second membre. On introduit ainsi une approximation supplémentaire, qu'il faut justifier à part (nous ne traiterons pas cette question ici, disons simplement que l'on peut le faire). Il faut bien sûr que cette approximation soit d'ordre suffisamment élevé pour ne pas trop perturber l'erreur commise par la méthode des éléments finis elle-même. Notons pour l'anecdote qu'il arrive ainsi, pour certains problèmes et certaines intégrations numériques, que l'on retombe exactement sur une matrice différences finies.*

Remarque 6.1.4 *On a choisi des fonctions de base qui produisent une matrice tridiagonale. En dimension supérieure, de bonnes numérotations des éléments d'un maillage produisent des matrices* creuses *dont les éléments non nuls sont regroupés autour de la diagonale (mais c'est plus compliqué qu'en dimension un).*

Pour terminer cette section, arrêtons nous sur le conditionnement de la matrice A_h donnée ci-dessus dans le cas de sommets équidistants $x_i = ih = 1/m$ et c constant. Dans ce cas il est facile d'exhiber la base de vecteur propres qui est donnée par les vecteurs s_k pour $k = 1,\dots,m$, dont les coordonnées sont données par

$$s_{k,j} = \sin(\pi k j h).$$

On voit en effet que

$$(A_h s_k)_j = \frac{1}{h}\Big((2+\frac{2ch^2}{3})s_{k,j} - (1-\frac{ch^2}{6})(s_{k,j-1}+s_{k,j+1})\Big) = (\frac{1}{h}-\frac{ch}{6})(2s_{k,j}-s_{k,j-1}-s_{k,j+1}) + chs_{k,j}.$$

En appliquant l'identité

$$2\sin(x)-\sin(x-t)-\sin(x+t)=(2-2\cos t)\sin(x)=4\sin(t/2)^2\sin(x),$$

à $x=\pi kjh$ et $t=\pi kh$, on trouve ainsi

$$A_h s_k=\lambda_k s_k,\quad \lambda_k=ch+\Big(\frac{1}{h}-\frac{ch}{6}\Big)4\sin(\pi kh/2)^2.$$

Lorsque h devient petit, la valeur minimale est $\lambda_1\approx ch+\frac{4}{h}\sin(\pi h/2)^2\sim(\pi^2+c)h$ et la valeur maximale est $\lambda_m\approx\frac{4}{h}$.

Ceci nous indique que lorsque $h\to 0$ on a

$$\kappa(A_h)\approx\frac{4}{(\pi^2+c)h^2}=O(h^{-2})$$

Une telle augmentation du nombre de conditionnement de la matrice de rigidité lorsque $h\to 0$ est typique de la méthode des éléments finis. Ceci constitue un problème lorsque l'on désire calculer la solution avec une grande précision puisque l'on est alors amené à prendre une maille h très petite, et le système à résoudre devient ainsi mal conditionné. En particulier, comme observé dans la remarque 5.2.2, les méthodes itératives classiques convergeront plus lentement. Ceci justifie le développement de méthodes dites de *préconditionnement* telles que les méthodes multigrilles qui dépassent le cadre de ce cours.

6.2 Analyse de la convergence dans le cas P_1

Nous avons déjà vu (Proposition 6.1.2) que la méthode d'éléments finis P_1 converge. Nous souhaitons estimer l'erreur en fonction de la taille du maillage. Pour analyser la convergence d'une méthode d'éléments finis, dont on rappelle que c'est un cas particulier de méthode d'approximation variationnelle, on sait qu'il convient d'étudier la distance de la solution à l'espace d'éléments finis, c'est-à-dire que l'on se ramène à un problème d'approximation. L'idée générale est très simple : si on connaît une projection $P_h: V\to V_h$, alors évidemment, $\inf_{v_h\in V_h}\|u-v_h\|_V\leq\|u-P_hu\|_V$, et si cette projection est bien choisie, alors on saura calculer des estimations de la quantité $\|u-P_hu\|_V$ en fonction de u et de h. (Attention, une projection P_h est un opérateur linéaire de V sur V_h, qui vérifie $P_h\circ P_h=Id$ mais qui n'a aucune raison d'être la projection orthogonale).

Dans le cas des éléments finis, l'opérateur de projection naturel est l'interpolant Π_h associée aux degrés de liberté que nous avons déjà utilisée dans la démonstration de la proposition 6.1.2. On rappelle qu'en une dimension $V=H^1_0(\Omega)\subset\mathcal{C}^0(\bar\Omega)$, ce qui permet de définir $\Pi_h v$ pour tout fonction $v\in V$ puisque les valeurs ponctuelles de v en x_i sont bien définies sans ambiguïté.

Les espaces H^1 sont donc des espaces de nature *locale* au sens où la norme est donnée par une intégrale sur Ω qui peut se décomposer en sous-domaines. Il est en particulier facile de voir que pour $v\in H^1(\Omega)$ et pour tout intervalle ouvert $T\subset\Omega$, la restriction de v à T, notée $v_{|T}$, appartient à $H^1(T)$, avec $(v_{|T})'=(v')_{|T}$. Dans la suite, on ne distinguera plus les restrictions des fonctions à un intervalle des fonctions elles-mêmes, pour alléger la notation. On notera aussi les éléments du maillages par

$$T_i=]x_i,x_{i+1}[.$$

Proposition 6.2.1 *Pour tout $v \in H^1(\Omega)$,*

$$\|v\|_{H^1(\Omega)} = \Big(\sum_{i=0}^{N} \|v\|^2_{H^1(T_i)}\Big)^{1/2}.$$

Démonstration. En effet, on a

$$\|v\|^2_{H^1(\Omega)} = \int_\Omega (v^2 + (v')^2) = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} (v^2 + (v')^2) = \Big(\sum_{i=0}^{N} \|v\|^2_{H^1(T_i)}\Big).$$

ce qui prouve le résultat □

Ce résultat est bien entendu aussi valable pour les normes $H^k(\Omega)$. Il nous suffit par conséquent d'estimer les quantités $\|u - \Pi_h u\|_{H^1(T_i)}$. On note que sur T_i, la fonction $\Pi_h u$ est affine et coïncide avec l'interpolé affine $\Pi_i u$ aux points x_i et x_{i+1}.

Théorème 6.2.1 *(Erreur d'interpolation locale) Soit $h_i = |T_i| = x_{i+1} - x_i$, soit $v \in H^2(T_i)$ et $\Pi_i v \in P_1$ son interpolée affine en x_i et x_{i+1}. On a alors*

$$\|(v - \Pi_i v)(x)\|_{L^2(T_i)} \leq h_i^2 \|v''\|_{L^2(T_i)},$$

et

$$\|(v' - (\Pi_i v)')(x)\|_{L^2(T_i)} \leq h_i \|v''\|_{L^2(T_i)},$$

Démonstration. Notons $g = v - \Pi_i v$. La fonction g s'annule donc aux points x_i et x_{i+1} et sa dérivée seconde (au sens faible) est donnée par

$$g'' = v'' \in L^2(T_i).$$

Nous allons montrer que $h_i g = \tilde{g}$ où

$$\tilde{g}(x) = \int_{x_i}^{x} (x - x_{i+1})(t - x_i) v''(t) dt + \int_{x}^{x_{i+1}} (t - x_{i+1})(x - x_i) v''(t) dt.$$

Pour cela, on remarque d'abord que $\tilde{g}(x_i) = \tilde{g}(x_{i+1}) = 0$. On dérive ensuite h en écrivant

$$\tilde{g}(x) = (x - x_{i+1}) \int_{x_i}^{x} (t - x_i) v''(t) dt + (x - x_i) \int_{x}^{x_{i+1}} (t - x_{i+1}) v''(t) dt,$$

et en appliquant la règle de Leibniz. Les termes issus de la dérivation des intégrales s'annulent et on obtient

$$\tilde{g}'(x) = \int_{x_i}^{x} (t - x_i) v''(t) dt + \int_{x}^{x_{i+1}} (t - x_{i+1}) v''(t) dt.$$

Puis en dérivant à nouveau

$$\tilde{g}''(x) = (x - x_i) v''(t) - (x - x_{i+1}) v''(t) = h_i v''(x) = h_i g''(x).$$

Par conséquent la fonction $\tilde{g} - h_i g \in H^2(T_i)$ vaut 0 en x_i et x_{i+1} et sa dérivée seconde est nulle, ce qui entraîne sa nullité.

On a donc

$$|g(x)| = h_i^{-1}|\tilde{g}(x)| \leq h_i^{-1}\int_{x_i}^{x}|(x-x_{i+1})(t-x_i)v''(t)|dt + h_i^{-1}\int_{x}^{x_{i+1}}|(t-x_{i+1})(x-x_i)v''(t)|dt,$$

et pour $x \in T_i$, on peut majorer $|(x-x_{i+1})(t-x_i)|$ et $|(t-x_{i+1})(x-x_i)|$ par h_i^2 ce qui donne

$$|g(x)| \leq h_i \int_{T_o}|v''(t)|dt.$$

En élevant au carré et en appliquant Cauchy-Schwarz, il vient

$$|g(x)|^2 \leq h_i^3 \int_{T_i}|v''(t)|^2 dt.$$

En intégrant sur T_i par rapport à x, on trouve

$$\|g\|^2_{L^2(T_i)} \leq h_i^4 \|v''\|^2_{L^2(T_i)},$$

qui est l'estimation annoncée de $\|(v-\Pi_i v)(x)\|_{L^2(T_i)}$

Pour l'estimation de g' en norme $L^2(T_i)$, on utilise la calcule précèdent en écrivant

$$|g'(x)| = h_i^{-1}|\tilde{g}'(x)| \leq h_i^{-1}\int_{x_i}^{x}|(t-x_i)v''(t)|dt + \int_{x}^{x_{i+1}}|(t-x_{i+1})v''(t)|dt \leq \int_{T_i}|v''(t)|dt.$$

Puis par un calcul similaire au précèdent, on arrive à

$$\|g'\|^2_{L^2(T_i)} \leq h_i^2 \|v''\|^2_{L^2(T_i)},$$

qui est l'estimation annoncée de $\|(v' - (\Pi_i v)')(x)\|_{L^2(T_i)}$. □

On peut déduire de ces estimations locales un résultat d'approximation global pour l'interpolant.

Théorème 6.2.2 *Soit $v \in H^2(\Omega)$ s'annulant en 0 et 1, et soit $h = \max h_i$. On a alors*

$$\|v - \Pi_h v\|_{L^2(\Omega)} \leq h^2 \|v''\|_{L^2(\Omega)},$$

et

$$\|v' - (\Pi_h v)'\|_{L^2(\Omega)} \leq h\|v''\|_{L^2(\Omega)}.$$

On a aussi

$$\|v - \Pi_h v\|_{H^1(\Omega)} \leq Ch\|v''\|_{L^2(\Omega)},$$

avec $C = \sqrt{2}$.

Démonstration. On écrit

$$\|v - \Pi_h v\|^2_{L^2(\Omega)} = \sum_{i=1}^{m} \|v - \Pi_i v\|^2_{L^2(T_i)},$$

par décomposition de l'intégrale sur les sous-domaine T_i. Par application de l'estimation locale du théorème précèdent, il vient

$$\|v-\Pi_h v\|^2_{L^2(\Omega)} \leq \sum_{i=1}^{m} h_i^4 \|v''\|^2_{L^2(T_i)} \leq h^4 \|v''\|^2_{L^2(\Omega)},$$

ce qui donne la première estimation. La deuxième estimation est obtenue de la même manière. La troisième est obtenue en combinant les deux estimations et en observant que $h^2 \leq h$ puisque $h < 1$. □

Remarque 6.2.1 *Dans la preuve du théorème 6.2.1, on a en fait obtenu des majoration ponctuelles sur g et g′ avec $g = v - \Pi_h v$. Un calcul similaire (exercice) nous montre que pour toute fonction $v \in \mathcal{C}^2([0,1])$ s'annulant en* 0 *et* 1 *on a alors les estimations en norme sup*

$$\|v-\Pi_h v\|_{L^\infty(\Omega)} \leq h^2 \|v''\|_{L^\infty(\Omega)},$$

et

$$\|v'-(\Pi_h v)'\|_{L^\infty(\Omega)} \leq h \|v''\|_{L^\infty(\Omega)}.$$

Théorème 6.2.3 *Supposons que la solution u du problème variationnel appartienne à $H^2(\Omega)$, et soit $u_h \in V_h$ la solution obtenue par la méthode de Galerkin. Il existe une constante C indépendante de u et du maillage telle que*

$$\|u-u_h\|_{H^1} \leq Ch \|u''\|_{L^2}, \tag{6.1}$$

Démonstration. La théorie abstraite de l'approximation variationelle nous indique que

$$\|u-u_h\|_{H^1} \leq (C_a/\alpha)^{1/2} \min_{v_h \in V_h} \|u-v_h\|_{H^1},$$

avec dans le cas présent $C_a = \max\{1, \|c\|_{L^\infty}\}$ et $\alpha = \frac{1}{2}$. On a en particulier

$$\|u-u_h\|_{H^1} \leq (C_a/\alpha)^{1/2} \|u-\Pi_h u\|_{H^1},$$

et en appliquant le théorème 6.2.2, on obtient l'estimation (6.1) avec $C = 2\max\{1, \|c\|^{1/2}_{L^\infty}\}$ □

Remarque 6.2.2 *Sous une hypothèse de régularité H^2 sur la solution, on a obtenu une convergence en $O(h)$ en norme H^1, d'où a fortiori, en $O(h)$ en norme de la convergence uniforme par l'injection continue $H^1 \hookrightarrow \mathcal{C}^0$.*

Remarque 6.2.3 *On a vu dans le théorème 4.4.1 que si $f \in L^2(\Omega)$ et $c \in L^\infty(\Omega)$, la solution u appartient en effet à H^2 et sa dérivée seconde faible est donnée par $u'' = cu - f$. La norme L^2 de u'' qui apparait dans l'estimation d'erreur* (6.1) *s'estime par*

$$\|u''\|_{L^2} \leq \|c\|_{L^\infty} \|u\|_{L^2} + \|f\|_{L^2} \leq \|c\|_{L^\infty} \|u\|_V + \|f\|_{L^2} \leq \|f\|_{L^2}(2\|c\|_{L^\infty}+1),$$

d'après le Théorème 4.4.1.

6.3 Eléments finis de degré et régularité plus élevés

Jusqu'ici, les degrés de liberté étaient associés à l'interpolation de Lagrange aux sommets du maillage, c'est-à-dire comme il y a deux sommets par élément, de l'interpolation P_1. Pour définir des méthodes d'éléments finis utilisant des polynômes de degré plus élevé, on va utiliser de façon analogue de l'interpolation utilisant ces polynômes. On part à nouveau d'un maillage de sommets

$$0 = x_0 < x_1 < \cdots < x_m < x_{m+1} = 1.$$

Comme dans la section précèdente, nous noterons parfois

$$T_i =]x_i, x_{i+1}[,$$

l'élément i et

$$h = \max_{i=0,\ldots,m} |x_{i+1} - x_i|$$

la finesse du maillage.

Commençons par des polynômes du second degré. On peut encore faire de l'interpolation de Lagrange, mais on a besoin de trois points, par exemple les sommets x_i, x_{i+1} et leur milieu $x_{i+1/2} = (x_i + x_{i+1})/2$. On parle alors d'éléments finis de Lagrange P_2. On prend comme espace

$$V_h = \{v \in \mathcal{C}^0(\bar{\Omega}) \ : \ v_{|T_i} \in P_2, \ v(0) = v(1) = 0\}.$$

Les degrés de libertés associés à cet espace sont donc les valeurs aux points x_i pour $i = 1, \ldots, m$ et $x_{i+1/2}$ pour $i = 0, \ldots, m$. On peut dans cet espace construire la base nodale pour cet espace en prenant les fonctions w_i et $w_{i+1/2}$ respectivement associés à ces degrés de libertés. Ces fonctions ont des formes différentes. La fonction w_i vaut 1 au point x_i et 0 à tous les autres points x_j et $x_{j+1/2}$, son support est l'intervalle $\overline{T_{i-1} \cup T_i} = [x_{i-1}, x_{i+1}]$. La fonction $w_{i+1/2}$ vaut 1 au point $x_{i+1/2}$ et 0 à tous les autres points x_j et $x_{j+1/2}$, son support est l'intervalle $\overline{T_i}$. On pourra à titre d'exercice écrire l'expression exacte de ces fonctions. Au final on voit que la base de V_h est constituée par les fonctions

$$w_{j/2}, \quad j = 1, \ldots, 2m+1,$$

l'espace étant donc de dimension $2m+1$.

Cette construction se généralise de manière naturelle aux polynômes de degré $k \geq 1$ arbitraire, et on parle d'éléments finis de Lagrange P_k : on définit

$$V_h = \{v \in \mathcal{C}^0(\bar{\Omega}) \ : \ v_{|T_i} \in P_k, \ v(0) = v(1) = 0\};$$

et on introduit les noeuds intermédiaires

$$x_{i+j/k} = x_i + \frac{j}{k}(x_{i+1} - x_i), \quad j = 1, \ldots, k-1, \quad i = 0, \ldots, m,$$

ce qui correspond à subdiviser chaque intervalle T_i en k sous-intervalles de même taille. Les fonctions de V_h sont alors caractérisées par les valeurs en l'ensemble des sommets du maillage et de ces points intermédiaire.

Théorème 6.3.1 *L'application linéaire*

$$v \to (v(x_{j/k})_{j=1,\ldots,km+k-1},$$

définit un isomorphisme entre l'espace V_h des éléments finis de Lagrange P_k et $\mathbb{R}^{km+k-1}$.

Démonstration. Soit $v \in V_h$ telle que $v(j/k) = 0$ pour tout $j = 1,\ldots,km+k-1$. Ceci signifie que pour tout $i = 0,\ldots,m$, la fonction v s'annule aux points $x_{i+j/k}$ pour $j = 0,\ldots,k$ qui sont $k+1$ points distincts dans l'intervalle $\overline{T_i}$. Or v est un polynôme de degré k sur cet intevalle, et par conséquence v est nulle. L'application est donc injective.

Réciproquement soit un vecteur arbitraire $(y_{j/k})_{j=1,\ldots,km+k-1}$ et posons $y_0 = y_{m+1} = 0$. Pour chaque $i = 0,\ldots,m$, il existe un unique polynôme $p_i \in P_k$ tel que

$$p_i(x_{i+j/k}) = y_{i+j/k}, \quad j = 0,\ldots,k,$$

que l'on obtient par interpolation. On définit un fonction polynomiale par morceaux v par

$$v_{|T_i} = p_i, \quad i = 0,\ldots,m.$$

Cette fonction vérifie ainsi les propriétés

$$v(x_{j/k}) = y_{j/k}, \quad j = 1,\ldots,km+k-1.$$

Sur les sommets x_i du maillage on voit que

$$p_{i-1}(x_i) = p_i(x_i) = y_i,$$

ce qui montre la continuité globale de v sur $[0,1]$ On a par ailleurs $v(0) = v(1) = 0$ puisqu'on a pris $y_0 = y_{m+1} = 0$, et par conséquent $v \in V_h$. L'application est donc surjective. □

Le théorème qu'on vient de démontrer nous indique que

$$\dim(V_h) = km+k-1.$$

La base nodale est constituées des fonctions $(w_{j/k})_{j=1,\ldots,km+k-1}$ qui vérifient les conditons

$$w_{j/k}(x_{l/k}) = \delta_{j,l}.$$

Les fonctions w_i associées aux sommets du maillage ont pour support l'intervalle $\overline{T_{i-1} \cup T_i}$, et les fonctions $w_{i+j/k}$ pour $j = 1,\ldots,k-1$ associées aux noeuds intermédiaires sont supportées dans $\overline{T_i}$. Toute fonction $v_h \in V_h$ se décompose suivant

$$v_h = \sum_{j=1}^{km+k-1} v_h(x_{j/k}) w_{j/k},$$

et on peut définir, comme pour les éléments P_1, l'opérateur d'interpolation Π_h qui agit sur toute fonction $v \in \mathcal{C}^0(\overline{\Omega})$ telle que $v(0) = v(1) = 0$ - et donc en particulier sur toute fonction de $H_0^1(\Omega)$ - suivant

$$\Pi_h v = \sum_{j=1}^{km+k-1} v_h(x_{j/k}) w_{j/k}.$$

On note que $\Pi_h v$ coincide avec v en tous les noeuds $x_{j/k}$ et que $\Pi_h v_h = v_h$ pour tout $v_h \in V_h$. On note aussi que sur l'élément T_i on a

$$\Pi_h v = \Pi_i v,$$

où $\Pi_i v$ est l'interpolation polynomiale de degré k de v aux points $(x_{i+j/k})_{j=0,\dots,k}$.

Notons que dans la construction des éléments finis de Lagrange P_k que l'on vient de décrire, on a assuré le caractère globalement continu des fonctions de V_h, mais ces espaces contiennent des fonctions qui ne sont pas de classe $\mathcal{C}^1$. On peut vouloir construire un espace de fonctions $\mathcal{C}^1$, en effet, si l'on veut approcher une solution qui est régulière, il est dommage de le faire avec des fonctions qui ne le sont pas (mais ce n'est pas toujours nécessaire). On voit aisément que dans le cas des éléments de Lagrange P_1, la seule fonction $\mathcal{C}^1$ contenue dans V_h est la fonction identiquement nulle. En revanche, l'espace

$$V_h = \{v \in \mathcal{C}^1(\bar{\Omega}) \ : \ v_{|T_i} \in P_2, \ v(0) = v(1) = 0\},$$

appelé espaces des *splines* quadratiques, contient des fonctions non-triviales et peut sembler un bon candidat pour approcher avec des fonctions régulières. Cependant, pour cet espace il n'existe plus de base qui ait à la fois les propriétés d'interpolation et une forme simple avec des petits supports.

On passe alors à l'étape suivante, à savoir des fonctions P_3 par morceaux. Dans ce cas il est facile d'assurer la classe $\mathcal{C}^1$ globale grâce à l'interpolation d'Hermite. En effet, pour toute fonction v de classe $\mathcal{C}^1$ et chaque indice i, il existe un unique polynôme p_i de P_3 tel que $p_i(x_i) = v(x_i)$, $p_i(x_{i+1}) = v(x_{i+1})$, $p_i'(x_i) = v'(x_i)$, $p_i'(x_{i+1}) = v'(x_{i+1})$. Sur l'espace

$$V_h = \{v \in \mathcal{C}^1(\bar{\Omega}) \ : \ v_{|T_i} \in P_3, \ v(0) = v(1) = 0\},$$

on peut donc définir une bonne base avec des propriétés d'interpolation correspondant à des éléments finis dont les degrés de liberté sont les valeurs de la fonction et de sa dérivée aux sommets. Il suffit de définir $w_i^0, w_i^1 \in V_h$ par

$$\begin{cases} w_i^0(x_j) = \delta_{i,j}, \ (w_i^0)'(x_j) = 0, & 1 \le i \le m, \ 0 \le j \le m+1, \\ w_i^1(x_j) = 0, \ (w_i^1)'(x_j) = \delta_{i,j}, & 0 \le i \le m+1, \ 0 \le j \le m+1. \end{cases}$$

Ces fonctions sont construites en recollant de façon $\mathcal{C}^1$ les polynômes de base de l'interpolation d'Hermite sur chaque intervalle $[x_i, x_{i+1}]$, voir figure 6.3.

On peut généraliser cette construction pour obtenir des éléments de classe C^k avec des polynômes de degré $2k+1$ sur chaque maille T_i : ce sont les éléments finis de Hermite P_{2k+1}, pour lesquels l'espace V_h est défini par

$$V_h = \{v \in \mathcal{C}^k(\bar{\Omega}) \ : \ v_{|T_i} \in P_{2k+1}, \ v(0) = v(1) = 0\},$$

Les degrés de libertés sont alors donnés par les valeurs de la fonction et de ses dérivées jusqu'à l'ordre k aux sommets du maillage comme l'indique le résulat suivant dont la preuve est laissée en exercice car elle est similaire à celle du théorème 6.3.1.

Théorème 6.3.2 *L'application linéaire*

$$v \to (v(x_1), \dots, v(x_m), v'(x_0), \dots, v'(x_{m+1}), \dots, v^{(k)}(x_0), \dots, v^{(k)}(x_{m+1})),$$

définit un isomorphisme entre l'espace V_h des éléments finis de Hermite P_k et $\mathbb{R}^{(k+1)m+2k}$.

Notons l'absence des degrés de liberté $v(x_0)$ et $v(x_m)$ puisqu'on a imposé la nullité en ces points dans la définition de V_h. Cet espace est donc de dimension $(k+1)m+2k$. On définit des fonctions de bases w_i^l associées aux degrés de libertés par les propriétés

$$(w_i^l)^{(p)}(x_j) = \delta_{i,j}\delta_{l,p}, \quad 0 \leq i,j \leq m+1,\ 0 \leq l,p \leq k,$$

en enlevant les fonctions w_0^0 et w_{m+1}^0 qui correspondent aux degrés de libertés $v(x_0)$ et $v(x_m)$ inactifs.

Toute fonction $v_h \in V_h$ se décompose suivant

$$v_h = \sum_{i=1}^{m} v_h(x_i)w_i^0 + \sum_{l=1}^{k}\sum_{i=0}^{m+1} v_h^{(l)}(x_i)w_i^l$$

et on définit l'opérateur d'interpolation Π_h qui agit sur toute fonction $v \in C^k(\overline{\Omega})$ telle que $v(0) = v(1) = 0$ - et donc en particulier sur toute fonction de $H_0^1(\Omega) \cap H^{k+1}(\Omega)$ - suivant

$$\Pi_h v = \sum_{i=1}^{m} v(x_i)w_i^0 + \sum_{l=1}^{k}\sum_{i=0}^{m+1} v^{(l)}(x_i)w_i^l.$$

On note que $\Pi_h v$ coincide avec v en tous les noeuds x_i ainsi que toutes les dérivées jusqu'à l'ordre k, et que $\Pi_h v_h = v_h$ pour tout $v_h \in V_h$. On note aussi que sur l'élément T_i on a

$$\Pi_h v = \Pi_i v,$$

où $\Pi_i v$ est l'interpolation polynomiale de Hermite degré $2k+1$ de v entre les points x_i et x_{i+1}, c'est à dire l'unique polynôme p_i de degré $2k+1$ tel que

$$p^{(l)}(x_i) = v^{(l)}(x_i),\ p^{(l)}(x_{i+1}) = v^{(l)}(x_{i+1}), \quad l = 0,\ldots,k.$$

Remarque 6.3.1 *Lorsqu'on utilise les éléments P_k de Lagrange ou P_{2k+1} de Hermite pour la discrétisation du problème au limite, on aboutit ainsi à un système ou les éléments de la matrice de rigidité sont donnés par l'application de la forme bilinéaire à tous les coubles de fonctions de base. Pour que les éléments non nuls de cete matrice soient situés près de la diagonale, il faut s'assurer que les fonctions de base soient arrangées dans un ordre qui assure que deux fonctions éloignées l'une de l'autre dans la numérotation ont une intersection de supports vide. Dans le cas des éléments P_k de Lagrange, l'ordre naturel $w_{j/k}$ pour $j = 1,2,\ldots$ convient. Pour les éléments P_{2k+1} de Hermite, on peut choisir l'ordre*

$$w_0^1, w_0^2, \ldots, w_0^k, w_1^0, \ldots, w_1^k, \ldots w_m^0, \ldots, w_m^k, w_{m+1}^1, \ldots, w_{m+1}^k.$$

On peut en guise d'exercice calculer la matrice pour un maillage uniforme et une fonction c constante dans le cas des élements de Lagrange P_2 ou des éléments de Hermite P_3.

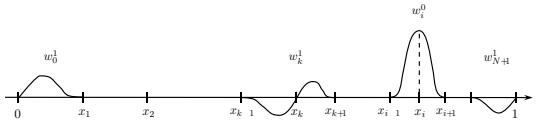


FIGURE 6.3 – Fonctions de base des éléments $\mathcal{C}^1$ Hermite-P_3

6.4 Analyse générale de convergence

Les espaces V_h des éléments de Lagrange P_k et de Hermite P_{2k+1} que l'on vient de définir sont tous des sous-espaces de $V = H_0^1(\Omega)$ et on peut donc les utiliser pour l'approximation variationelles du problème au limites : trouver $u_h \in V_h$ tel que

$$a(u_h, v_h) = \ell(v_h), \quad v_h \in V_h.$$

Le lemme de Cea nous assure que

$$\|u - u_h\|_{H^1} \leq C \min_{v_h \in V_h} \|u - v_h\|_{H^1}, \quad C = (C_a/\alpha)^{1/2}.$$

Comme pour les éléments de Lagrange P_1, l'analyse de convergence de la méthode se ramène donc à la recherche d'une borne de l'erreur de meilleure approximation dans V_h. On a en particulier

$$\min_{v_h \in V_h} \|u - v_h\|_{H^1} \leq \|u - \Pi_h u\|_{H^1},$$

où Π_h est l'opérateur d'interpolation défini précèdement pour l'espace V_h que l'on a choisi d'utiliser.

Nous nous ramenons ainsi à l'étude de l'erreur d'interpolation $u - \Pi_h u$, qui comme dans le cas des éléments de Lagrange P_1 peut se faire élément par élément, en estimant l'erreur $u - \Pi_i u$ sur T_i, où Π_i est l'interpolant de Lagrange de degré k ou l'interpolant de Hermite de degré $2k+1$. Mais contrairement au cas des éléments de Lagrange P_1, cette erreur locale n'admet pas une expression simple permettant de faire des calculs directs. Nous allons l'estimer par une technique plus générale, qui a en particulier l'avantage de bien se généraliser aux éléments finis en plus grande dimension.

L'idée fondamentale est de se ramener au cas où T_i est l'intervalle $]0,1[$ tout entier, que l'on qualifie *d'élément de référence* et que l'on notera ici $\hat{T}$ plutôt que Ω. On note $\widehat{\Pi}$ l'interpolant correspondant à cet élément. Dans le cas des éléments de Lagrange P_k il s'agit donc de l'opérateur d'interpolation de Lagrange de degré k aux noeuds j/k pour $j = 0, \ldots, k$. Dans le cas des éléments de Lagrange P_{2k+1} il s'agit de l'opérateur d'interpolation de Hermite de degré $2k+1$ utilisant les valeurs de la fonctions et de ses dérivées jusqu'à l'ordre k aux points 0 et 1. Nous utiliserons le résultat fondamental suivant.

Théorème 6.4.1 *Soit $k \geq 1$. Pour l'interpolation de Lagrange de degré k sur $\hat{T}$, on a pour tout $1 \leq r \leq k+1$ et pour toute fonction $v \in H^r(\hat{T})$,*

$$\|v - \widehat{\Pi} v\|_{H^r(\hat{T})} \leq C |v|_{H^r(\hat{T})}, \tag{6.2}$$

où la constante C ne dépend que de k et r. Pour l'interpolation de Hermite de degré $2k+1$ sur $\hat{T}$ on a la même estimation pour tout $k+1 \leq r \leq 2k+2$

La preuve de ce théorème utilise des résultats avancés d'analyse fonctionnelle, et pour cette raison nous décalons son exposition à la fin de ce chapitre. Constatons simplement à ce stade que ce résultat est "raisonnable" au sens suivant : si le second membre de l'inégalité (6.2) est nul, cela signifie que $v^{(r)} = 0$, c'est à dire que v est un polynôme de degré $r-1$. Puisqu'on a supposé $r \leq k+1$ dans le cas de l'interpolation de Lagrange, cela signifie que $v \in P_k$ et par conséquent le membre de gauche de (6.2) est nul puisque $v = \widehat{\Pi} v$. De même, dans le cas de l'interpolation de Hermite cela signifie que $v \in P_{2k+1}$ et on arrive à la même conclusion. On remarque aussi qu'il est raisonnable d'imposer $r \geq 1$ pour l'interpolation de Lagrange puisque ceci assure que $v \in \mathcal{C}^0(\hat{T})$ et $r \geq k+1$ pour l'interpolation de Hermite puisque ceci assure que $v \in \mathcal{C}^k(\hat{T})$.

En admettant provisoirement ce théorème, nous allons estimer l'erreur d'interpolation locale $v - \Pi_i v$ sur T_i en fonction de la régularité de v sur T_i et de la longueur de l'élément

$$h_i = |T_i| = x_{i+1} - x_i,$$

par une technique de changement de variable. En effet pour tout $x \in T_i$, on peut écrire

$$x = \phi_i(\hat{x}) := x_i + h_i \hat{x},$$

où $\hat{x}$ appartient à $\hat{T}$. L'application affine ϕ_i est une bijection entre $\hat{T}$ et T_i. Toute fonction v définie sur T_i peut être "transportée" par ce changement de variable vers une fonction $\hat{v} = v \circ \phi_i$ définie sur $\hat{T}$ par

$$\hat{v}(\hat{x}) = v(x).$$

Attention : la notation $\hat{v}$ n'a rien à voir avec la transformée de Fourier de v ! On peut facilement relier les normes L^2 de v et $\hat{v}$ puisqu'on a

$$\int_{T_i} |v(x)|^2 dx = \int_{\hat{T}} |\hat{v}(\hat{x})|^2 h_i d\hat{x},$$

et par conséquent

$$\|v\|_{L^2(T_i)} = h_i^{1/2} \|\hat{v}\|_{L^2(\hat{T})}.$$

On peut en faire autant pour les semi-normes H^l : en dérivant l fois $v(x_i + h_i \hat{x}) = \hat{v}(\hat{x})$, on trouve

$$\hat{v}^{(l)}(\hat{x}) = h_i^l v^{(l)}(x_i + h_i \hat{x}) = h_i^l v^{(l)}(x).$$

On peut ainsi écrire

$$h_i^{2l} \int_{T_i} |v^{(l)}(x)|^2 dx = \int_{\hat{T}} |\hat{v}^{(l)}(\hat{x})|^2 h_i d\hat{x},$$

et par conséquent

$$|v|_{H^l(T_i)} = h_i^{1/2-l} |\hat{v}|_{H^l(\hat{T})}. \tag{6.3}$$

Une propriété essentielle est que ce changement de variable permet aussi de ramener l'interpolation de v à celle de $\hat{v}$. Prenons par exemple l'interpolation de Lagrange de degré k : si $p_i = \Pi_i v \in P_k$ est le polynôme d'interpolation de Lagrange de v aux points $x_{i+j/k}$ pour $j = 0, \ldots, k$, on note $\hat{p}_i = p_i \circ \phi_i$ le transport de p_i. On remarque d'abord que $\hat{p}_i$ est aussi un polynôme de degré k. D'autre part, en remarquant que $x_{i+j/k} = x_i + h_i j/k = a_i(j/k)$, on obtient

$$\hat{p}_i(j/k) = p_i(a_i(j/k)) = p_i(x_{i+j/k}) = v(x_{i+j/k}) = \hat{v}(j/k), \quad j = 0, \ldots, k,$$

ce qui nous montre que $\hat{p}_i$ est aussi le polynôme d'interpolation de Lagrange de $\hat{v}$ aux points j/k pour $j = 0, \ldots, k$, c'est à dire

$$\hat{p}_i = \widehat{\Pi}\hat{v},$$

où $\widehat{\Pi}$ est l'opérateur d'interpolation sur l'élément de référence. Ceci peut aussi s'écrire

$$\Pi_i v \circ \phi_i = \widehat{\Pi}(v \circ \phi_i).$$

En d'autre termes : il est équivalent d'interpoler une fonction puis de la transporter sur l'intervalle de référence, ou de la transporter sur l'intervalle de référence puis de l'interpoler. Cette propriété de commutation est aussi valable pour l'interpolation de Hermite de degré $2k+1$ (exercice).

Munis de ces remarques, nous sommes à présent en mesure d'établir une estimation de l'erreur d'interpolation sur l'intervalle T_i.

Théorème 6.4.2 *Soit $k \geq 1$. Pour l'interpolation de Lagrange de degré k sur T_i, on a pour tout $1 \leq r \leq k+1$, $l = 0, \ldots, r$, et pour toute fonction $v \in H^r(T_i)$,*

$$|v - \Pi_i v|_{H^l(T_i)} \leq C h_i^{r-l} |v|_{H^r(T_i)}, \tag{6.4}$$

où la constante C ne dépend que de k et r, et est indépendante de l'élément T_i. Pour l'interpolation de Hermite de degré $2k+1$ sur $\hat{T}$ on a la même estimation pour tout $k+1 \leq r \leq 2k+2$ et $l = 0, \ldots, r$.

Démonstration. Notons $p_i = \Pi_i v$ et introduisons la fonction $\hat{v} = v \circ \phi_i$ et son interpolé $\hat{p}_i = \widehat{\Pi}\hat{v}$. Puisque $\hat{v} - \hat{p}_i = (v - p_i) \circ \phi_i$, on peut appliquer (6.3) pour obtenir.

$$|v - \Pi_i v|_{H^l(T_i)} = |v - p_i|_{H^l(T_i)} = h_i^{1/2-l} |\hat{v} - \hat{p}_i|_{H^l(\hat{T})} = h_i^{1/2-l} |\hat{v} - \widehat{\Pi}\hat{v}|_{H^l(\hat{T})}$$

Par application du théorème 6.2, on sait que

$$|\hat{v} - \widehat{\Pi}\hat{v}|_{H^l(\hat{T})} \leq \|\hat{v} - \widehat{\Pi}\hat{v}\|_{H^{r(\hat{T})}} \leq C |\hat{v}|_{H^r(\hat{T})},$$

où C ne dépend que de k et de r. On a ainsi

$$|v - \Pi_i v|_{H^l(T_i)} \leq C h_i^{1/2-l} |\hat{v}|_{H^r(\hat{T})} \leq C h_i^{r-l} |v|_{H^r(T_i)},$$

en appliquant à nouveau (6.3). □

Comme nous l'avons fait dans le cas simple des éléments de Lagrange P_1, il nous faut maintenant aggréger ces estimations afin d'obtenir une estimation globale sur l'erreur d'interpolation $v - \Pi_h v$.

Théorème 6.4.3 *Soit $k \geq 1$. Pour les élément finis de Lagrange P_k, on a pour tout $1 \leq r \leq k+1$, $l = 0, 1$, et pour toute fonction $v \in H_0^1(\Omega) \cap H^r(\Omega)$,*

$$|v - \Pi_h v|_{H^l(\Omega)} \leq C h^{r-l} |v|_{H^r(\Omega)}, \tag{6.5}$$

où la constante C ne dépend que de k et r. Pour les éléments finis de Hermite P_{2k+1}, on a la même estimation pour tout $k+1 \leq r \leq 2k+2$ et $l = 0, 1, \ldots, k+1$.

Démonstration. Dans le cas des éléments de Lagrange, lorsque $v \in H_0^1(\Omega)$, on a aussi $\Pi_h v \in H_0^1(\Omega)$ et on peut écrire

$$\|v - \Pi_h v\|^2_{L^2(\Omega)} = \sum_{i=0}^{m} \|v - \Pi_i v\|^2_{L^2(T_i)},$$

ainsi que

$$\|v' - (\Pi_h v)'\|^2_{L^2(\Omega)} = \sum_{i=0}^{m} \|v' - (\Pi_i v)'\|^2_{L^2(T_i)}.$$

En d'autres terme, pour $l = 0, 1$, on a

$$|v - \Pi_h v|^2_{H^l(\Omega)} = \sum_{i=0}^{m} |v - \Pi_i v|^2_{H^l(T_i)},$$

et par application du théorème 6.4.2 il vient

$$|v - \Pi_h v|^2_{H^l(\Omega)} \leq C \sum_{i=0}^{m} h_i^{2(r-l)} |v|^2_{H^r(T_i)} \leq C h^{2(r-l)} |v|^2_{H^r(\Omega)},$$

qui est l'estimation (6.5). Dans le cas des éléments de Hermite, si de plus $v \in H^{k+1}(\Omega)$, on a aussi $\Pi_h v \in H^{k+1}(\Omega)$ et on peut ainsi écrire, pour $l = 0, \ldots, k+1$,

$$\|v(l) - (\Pi_h v)^{(l)}\|^2_{L^2(\Omega)} = \sum_{i=0}^{m} \|v^{(l)} - (\Pi_i v)^{(l)}\|^2_{L^2(T_i)}.$$

Ceci nous permet d'obtenir l'estimation (6.5) pour les valeurs $l = 0, \ldots, k+1$. □

Nous pouvons à présent revenir à l'approximation variationelle du problème au limite dans les espaces V_h. On a vu que

$$\|u - u_h\|_{H^1} \leq C \|u - \Pi_h u\|_{H^1},$$

avec $C = (C_a/\alpha)^{1/2}$. En appliquant le théorème précèdent, on obtient ainsi immédiatement le résultat suivant.

Corollaire 6.4.4 *Soit $k \geq 1$. Pour les élément finis de Lagrange P_k, pour tout $1 \leq r \leq k+1$, si la solution u appartient à $H^r(\Omega)$, on a*

$$\|u - u_h\|_{H^1} \leq C h^{r-1} |v|_{H^r}, \tag{6.6}$$

où la constante C ne dépend que de C_a, α, k et r. Pour les éléments finis de Hermite P_{2k+1}, on a la même estimation pour tout $k+1 \leq r \leq 2k+2$.

Remarque 6.4.1 *Il est possible de démontrer par une approche plus sophistiquée que la limitation $k+1 \leq r$ pour les éléments de Hermite, nécessaire pour obtenir l'estimation de l'erreur d'interpolation, n'est en fait pas requise pour la validité de* (6.6) *qui reste vérifiée pour $1 \leq r \leq 2k+2$.*

Remarque 6.4.2 *La technique des estimations d'erreur via l'élément de référence se généralise en dimension supérieure. Le fait d'introduire un élément fini de référence est également intéressant pour la construction de la matrice de rigidité. En effet, les fonctions de base dans un élément générique se déduisent aussi des fonctions de base de l'interpolation de Lagrange sur l'élément de référence par le changement de variable.*

En conclusion, plus la solution est régulière, plus on aura une convergence rapide quand h tend vers zéro, à condition d'utiliser des éléments finis de degré suffisament élevé, et donc au prix de plus de calculs. D'après les résultats de régularité indiqués dans la remarque 4.4.4, on obtiendra en particulier l'appartenance de u à $H^r(\Omega)$, et donc la vitesse de convergence $O(h^{r-1})$, si $f \in H^{r-1}(\Omega)$ et $c \in C^{r-1}(\Omega)$. *A contrario*, si la solution n'est pas régulière, il est probablement inutile de dépenser beaucoup d'efforts à utiliser des éléments de haut degré.

Terminons cette section par le schéma général d'étude que nous avons suivi pour établir les estimations d'erreur.

erreur de la méthode	$\leq C$ erreur d'approximation $\leq C$ erreur d'interpolation $\leq Ch^{m-1}$ norme de la dérivée d'ordre m de u

où la constante C change de valeur, mais ne dépend pas de h ni de u.

6.5 Compléments

Le lemme de Cea conduit naturellement une estimation en norme H^1 pour l'erreur $u-u_h$. On a bien entendu

$$\|u-u_h\|_{L^2} \leq \|u-u_h\|_{H^1},$$

et donc la même estimation en $O(h^{r-1})$ en norme L^2 lorsque $u \in H^r(\Omega)$. Il est cependant assez naturel de penser que la fonction doit converger plus rapidement dans cette norme. Le théorème suivant, aussi appelé Lemme de Aubin-Nitsche, donne un contenu concret à cette intuition.

Théorème 6.5.1 *Sous les mêmes hypothèses que celle du corollaire 6.4.4, on a aussi l'estimation*

$$\|u-u_h\|_{L^2} \leq Ch^r|u|_{H^r}, \tag{6.7}$$

où la constante C ne dépend que de $\|c\|_{L^\infty}$, C_a, α, k *et* r.

Démonstration. Notons $g = u-u_h \in H_0^1(\Omega)$ et considèrons le problème auxilliaire

$$-w''(x)+c(x)w(x) = g(x), \quad x \in \Omega, \quad w(0) = w(1) = 0.$$

Comme $g \in L^2(\Omega)$ on sait que $w \in H^2(\Omega)$ avec

$$|w|_{H^2} = \|w''\|_{L^2} \leq \|c\|_{L^\infty}\|w\|_{L^2} + \|g\|_{L^2} \leq C_0\|g\|_{L^2}, \quad C_0 = 1+2\|c\|_{L^\infty}.$$

On peut ensuite écrire en intégrant par partie

$$\|g\|_{L^2}^2 = \int_\Omega g(x)(-w''(x)+c(x)w(x))dx = a(g,w) = a(u-u_h,w).$$

Puisque u_h est la solution de Galerkin, on sait par ailleurs que $a(u-u_h,w_h) = 0$ pour tout $w_h \in V_h$. En prenant $w_h = \Pi_h w$, on peut donc écrire

$$\|g\|_{L^2}^2 = a(g,w-w_h) \leq C_a\|g\|_{H^1}\|w-\Pi_h w\|_{H^1}.$$

D'après le théorème 6.4.3, on a

$$\|w-\Pi_h w\|_{H^1} \leq C_1 h |w|_{H^2} \leq C_0 C_1 h \|g\|_{L^2},$$

où C_1 ne dépend que de k. On a ainsi

$$\|g\|_{L^2}^2 \leq C_0 C_1 C_a h \|g\|_{H^1} \|g\|_{L^2},$$

soit

$$\|u-u_h\|_{L^2} \leq C_0 C_1 C_a h \|u-u_h\|_{H^1}.$$

En appliquant l'estimation en norme H^1 du corollaire 6.4.4, on obtient finalement (6.7). □

Afin d'illustrer l'utilité des éléments finis de Hermite, examinons rapidement un autre exemple de formulation variationnelle construite à partir d'un problème aux limites d'ordre 4 (modélisation d'un fil rigide encastré). Le problème aux limites s'écrit : soit $c \geq 0$ et f données dans $\mathcal{C}^0(\Omega)$, on cherche $u \in \mathcal{C}^1(\overline{\Omega}) \cap \mathcal{C}^4(\Omega)$ telle que

$$\text{(P)} \qquad \begin{cases} u^{(4)}(x) + c(x)u(x) = f(x) \text{ dans } \Omega, \\ u(0) = u(1) = 0,\ u'(0) = u'(1) = 0. \end{cases}$$

En multipliant par une fonction v suffisament régulière et vérifiant les mêmes conditions aux limites que u, puis en effectuant deux intégration par partie, on aboutit à la formuation variationnelle

$$a(u,v) = \ell(v),$$

où la forme linéaire est encore donnée par $\ell(v) = \int_\Omega f(x)v(x)dx$ et la forme bilinéaire est du type

$$a(u,v) = \int_\Omega (u''(x)v''(x) + c(x)u(x)v(x))dx.$$

L'espace de Hilbert naturel pour cette formulation variationnelle est

$$V = H_0^2(\Omega) = \{v \in H^2(\Omega) \ :\ v(0) = v(1) = v'(0) = v'(1) = 0\}.$$

A titre d'exercice, on pourra montrer que toutes les hypothèses de Lax-Milgram sont vérifiées avec ce choix d'espace V dès que $f \in L^2$ et $c \in L^\infty$. Il existe donc une unique solution $u \in V$ et on peut aussi montrer que $u \in H^4(\Omega)$.

Si on veut appliquer la méthode d'approximation variationelle, il nous faut utiliser un espace d'élément fini V_h contenu dans $H^2(\Omega)$, ce qui n'est pas le cas des éléments finis de Lagrange. Un choix naturel est celui des éléments de Hermite, par exemple P_3. Le problème discret est donc : trouver $u_h \in V_h$ tel que

$$a(u_h, v_h) = \ell(v_h), \quad v_h \in V_h.$$

D'après le lemme de Cea, on peut écrire

$$\|u-u_h\|_V \leq C \|u-\Pi_h v\|_V,$$

où Π_h est l'opérateur d'interpolation agissant de V dans V_h et et $C = (C_a/\alpha)^{1/2}$. Finalement, le théorème d'approximation 6.4.3 nous donne l'estimation d'erreur

$$\|u-u_h\|_V \leq C h^2 |u|_{H^4}.$$

Annexe : preuve du théorème 6.4.1

Le point de départ est un résultat de base liant les polynômes et les espaces de Sobolev, appelé théorème de Deny-Lions.

Théorème 6.5.2 *Soit $r > 0$ un entier et I un intervalle ouvert borné. Il existe une constante C qui ne dépend que de r et I telle que*

$$\|v\|_{H^r(I)} \leq C|v|_{H^r(I)},$$

pour toute fonction v telle que $\int_I v(x)p(x)dx = 0$ quelquesoit $p \in P_{r-1}$.

Démonstration. On procède par l'absurde en supposant qu'il n'existe pas de constante C telle que ceci soit vrai. Autrement dit, pour tout entier $n \geq 1$, il existe une fonction v_n orthogonale dans $L^2(I)$ à tous les polynômes de degré $n-1$ et telle que

$$\|v_n\|_{H^r(I)} \geq n|v_n|_{H^r(I)}.$$

En posant $w_n = v_n/\|v_n\|_{H^r(I)}$, on obtient

$$\|w_n\|_{H^r(I)} = 1 \quad \text{et} \quad \lim_{n\to+\infty} |w_n|_{H^r(I)} = 0.$$

et w_n vérifie aussi que $\int_I v_n(x)p(x)dx = 0$ pour tout $p \in P_{r-1}$.

La suite $(w_n)_{n\geq 1}$ est bornée dans $H^r(I)$. Rappelons que le théorème de Rellich nous indique que l'injection de $H^1(I)$ dans $L^2(I)$ est compacte lorsque I est borné. On en déduit aisément par récurrence que l'injection de $H^r(I)$ dans $H^{r-1}(I)$ est compacte. Par conséquent quitte à en extraire une sous-suite, on obtient que la suite $(w_n)_{n\geq 1}$ converge dans $H^{r-1}(I)$ vers une limite $w \in H^{r-1}(I)$.

La suite $(w_n)_{n\geq 1}$ est donc de Cauchy dans $H^{r-1}(I)$ c'est à dire

$$\lim_{m,n\to\infty} \|w_m - w_n\|_{H^{r-1}(I)} = 0.$$

D'autre part, en écrivant $|w_m - w_n|_{H^r(I)} \leq |w_m|_{H^r(I)} + |w_n|_{H^r(I)}$, on voit qu'on a aussi

$$\lim_{m,n\to\infty} |w_m - w_n|_{H^r(I)} = 0.$$

Ceci nous montre que $(w_n)_{n\geq 1}$ est aussi de Cauchy dans $H^r(I)$ et converge donc dans cet espace vers une limite qui est forcément égale (presque partout) à w puisque c'est aussi une limite dans $H^{r-1}(I)$.

Par continuité de la norme au passage à la limite, on voit que

$$\|w\|_{H^r(I)} = 1 \quad \text{et} \quad |w|_{H^r(I)} = 0.$$

La deuxième propriété signifie que w est un polynôme de degré inférieur ou égal à $r-1$. Ce fait n'est pas complètement évident car il s'agit de l'annulation de la dérivée faible d'ordre r. On peut le montrer par récurrence sur r. On a déjà démontré au lemme 5 (chapitre 3) que cette

propriété est vraie pour $r = 0$ (si $v' = 0$ alors v est constante). Supposons la vraie pour $r-1$. Et soit $v \in H^r(I)$ tel que $|v|_{H^{r-1}} = 0$, donc tel que $v^{(r)} = 0$. Clairement, $v' \in H^{r-1}(I)$ est tel que $(v')^{(r-1)} = 0$. Par l'hypothèse de récurrence, il existe $q \in P_{r-2}$ tel que $v' = q$. Soit $Q \in P_{r-1}$ une primitive de q (au sens classique). Alors $v - Q \in H^1(I)$ est tel que $(v-Q)' = 0$. Il existe donc une constante c telle que $v - Q = c$, soit encore $v = Q + c \in P_{r-1}$.

On remarque finalement que puisque

$$\Big| \int_I p(x)w_n(x)dx - \int_I p(x)w(x)dx \Big| \leq \|p\|_{L^2} \|w_n - w\|_{L^2},$$

on a aussi par passage à la limite la propriété $\int_I w(x)p(x)dx = 0$ pour tout $p \in P_{r-1}$ et en particulier pour $p = w$. Ceci entraîne que w est le polynôme nul et est en contradiction avec le fait que $\|w\|_{H^r(I)} = 1$. □

Pour démontrer le théorème 6.4.1, on se place dans le cas $I = \hat{T}$. Pour toute fonction $v \in H^r(\hat{T})$ on note q sa projection $L^2(\hat{T})$-orthogonale sur P_{r-1}. On a ainsi

$$\int_{\hat{T}} (v(x) - q(x))p(x)dx = 0,$$

et on peut donc appliquer le théorème de Deny-Lions à la fonction $v - q$, ce qui nous donne

$$\|v - q\|_{H^r(\hat{T})} \leq C|v - q|_{H^r(\hat{T})},$$

En remarquant que $q^{(r)} = 0$, on obtient ainsi le résultat d'approximation pour toute fonction $v \in H^r(\hat{T})$

$$\min_{p \in P_{r-1}} \|v - p\|_{H^r(\hat{T})} \leq C|v|_{H^r(\hat{T})},$$

où C ne dépend que de r. On note ensuite que pour tout l tel que $r \leq l+1$ on a $P_{r-1} \subset P_r$ et par conséquent

$$\min_{p \in P_l} \|v - p\|_{H^r(\hat{T})} \leq C|v|_{H^r(\hat{T})}.$$

Dans le cas de l'interpolation de Lagrange de degré k, puisqu'on a supposé $r \leq k+1$, on a ainsi

$$\min_{p \in P_k} \|v - p\|_{H^r(\hat{T})} \leq C|v|_{H^r(\hat{T})}, \tag{6.8}$$

D'autre part puisque $r \geq 1$ l'espace $H^r(\hat{T})$ s'injecte continuement dans $\mathcal{C}^0(\overline{\hat{T}})$. En écrivant

$$\widehat{\Pi} v = \sum_{j=0}^{k} v(j/k)\hat{w}_{j/k},$$

on voit que $\widehat{\Pi}$ est un opérateur continu de $\mathcal{C}^0(\overline{\hat{T}})$ dans P_k et donc de $H^r(\hat{T})$ dans P_k muni de n'importe quelle norme puisque c'est un espace de dimension finie. On note ainsi

$$M = \|\widehat{\Pi}\|_{\mathcal{L}(H^r)} = \max_{\|v\|_{H^r(\hat{T})}=1} \|\widehat{\Pi} v\|_{H^r(\hat{T})} < \infty.$$

La constante M ne dépend que de k et r. Pour toute fonction $p \in P_k$ on peut écrire

$$\|v - \widehat{\Pi} v\|_{H^r(\hat{T})} \leq \|v - p\|_{H^r(\hat{T})} + \|p - \widehat{\Pi} v\|_{H^r(\hat{T})} = \|v - p\|_{H^r(\hat{T})} + \|\widehat{\Pi} p - \widehat{\Pi} v\|_{H^r(\hat{T})},$$

et par conséquent

$$\|v-\widehat{\Pi}v\|_{H^r(\hat{T})} \leq (1+M)\|v-p\|_{H^r(\hat{T})}.$$

Puisque le choix de p est arbitraire, on obtien

$$\|v-\widehat{\Pi}v\|_{H^r(\hat{T})} \leq (1+M)\min_{p\in P_k}\|v-p\|_{H^r(\hat{T})},$$

qui combiné à (6.8) nous conduit à l'estimation (6.2) annoncée.

Dans le cas de l'interpolation de Hermite de degré $2k+1$, puisqu'on a supposé $r \leq 2k+2$, on a cette fois

$$\min_{p\in P_{2k+1}}\|v-p\|_{H^r(\hat{T})} \leq C|v|_{H^r(\hat{T})}. \tag{6.9}$$

Puisque $r \geq 1$ l'espace $H^r(\hat{T})$ s'injecte continuement dans $\mathcal{C}^k(\overline{\hat{T}})$, et l'expression de l'interpolation de Hermite permet de montrer que $\widehat{\Pi}$ est continu de $H^r(\hat{T})$ dans P_{2k+1}. On obtient de la même façon que

$$\|v-\widehat{\Pi}v\|_{H^r(\hat{T})} \leq (1+M)\min_{p\in P_{2k+1}}\|v-p\|_{H^r(\hat{T})},$$

et on conclut en combinant avec (6.9).

Chapitre 7

Méthodes de bases hilbertiennes

Dans ce chapitre nous abordons plusieurs méthodes d'approximations variationelles dont le point commun est le suivant : les espaces V_n sont engendrés par les n premiers éléments $e_1, \ldots, e_n$ d'une base orthonormée de fonction $(e_n)_{n\geq 1}$ pour un espace de Hilbert H. De telles familles sont appelées bases hilbertiennes. On note que dans ce cas les espaces V_n résultants vérifient la propriété d'emboitement

$$V_n \subset V_{n+1},$$

qui n'était pas nécessairement vérifiée dans le cas des espaces d'éléments finis. L'étude de convergence de la méthode d'approximation variatonelle dans les espaces V_n se ramène alors celle des coefficients de la solution u dans cette base.

7.1 Bases hilbertiennes

Voici tout d'abord quelques notions générales qu'il convient de connaître sur les bases hilbertiennes.

Définition 7.1.1 *Soit H un espace de Hilbert de dimension infinie. Une suite $(e_n)_{n\geq 1}$ est appellée base hilbertienne de H si et seulement si les deux conditions suivantes sont satisfaites :*

(i) La suite est orthonormée, c'est à dire $(e_n, e_m)_H = \delta_{m,n}$ pour tout $m, n \geq 1$.

(ii) La suite est totale, c'est à dire pour tout $v \in H$, il existe une suite $(v_n)_{n\geq 1}$ telle que $v_n \in V_n = \text{vect}(e_1, \ldots, e_n)$ et $\|v - v_n\|_H \to 0$ quand $n \to \infty$.

La notion de base hilbertienne remplace ainsi la notion classique de base orthonormée en dimension finie, la propriété de famille génératrice étant remplacée par la propriété (ii) de totalité. Cette propriété peut aussi s'exprimer par le fait que l'ensemble des combinaisons linéaires finies des e_n est dense dans H, ou que l'union de tous les espaces V_n est dense dans H.

Remarque 7.1.1 *On dit rappelle qu'un espace de Banach X est séparable si et seulement si il existe une partie dénombrable dense dans X. Il est facile de montrer que l'existence d'une base Hilbertienne implique une telle propriété sur H (considérer les combinaisons à coefficients rationnels des $(e_1, \ldots, e_n)$ pour $n \geq 1$). Réciproquement dans tout espace de Hilbert séparable, on peut construite une base Hilbertienne en partant d'une partie dénombrable dense et en appliquant un procédé d'orthonormalisaiton de type Gramm-Schmidt. En pratique, pour des*

espaces de Hilbert telles que $L^2(\Omega)$ ou les espaces de Sobolev, on s'intéressera à des bases hibertiennes plus spécifique dont les éléments ont des expressions simples.

Quelques propriétés importantes découlent facilement de la définition : puisque pour tout $v \in H$ on peut trouver $v_n \in V_n$ qui converge vers V quand $n \to \infty$, en notant ici P_{V_n} le projecteur orthogonal sur V_n, on obtient

$$\|v - P_{V_n} v\|_H \leq \|v - v_n\|_H \to 0.$$

On vérifie aisément que

$$P_{V_n} v = \sum_{j=1}^{n} c_j(v) e_j, \quad c_j(v) := (v, e_j)_H$$

ce qui entraîne la convergence vers v dans H de la série ci-dessus quand $n \to \infty$:

$$v = \sum_{j=1}^{\infty} c_j(v) e_j,$$

les coefficients $c_j(v) = (v, e_j)_H$ sont les coordonnées de v dans la base hilbertienne. Elles sont uniquement déterminées : si $v = \sum_{j=1}^{\infty} c_j e_j$ au sens de la convergence dans H, alors en prenant le produit scalaire avec e_j on voit que l'on a nécessairement $c_j = (v, e_j)_H = c_j(v)$. Puisque

$$\|P_{V_n} v\|_H^2 = \sum_{j=1}^{n} |c_j(v)|^2,$$

on obtient par passage à la limite l'égalité de Parseval

$$\|v\|_H^2 = \sum_{j=1}^{\infty} |c_j(v)|^2,$$

qui montre que la série des coordonnées appartient à $\ell^2(\mathbb{N})$. Réciproquement, si $(c_j)_{j \geq 1}$ est une suite quelconque de $\ell^2(\mathbb{N})$, la série $\sum_j c_j e_j$ converge vers une fonction $v \in H$ telle que $c_j = (v, e_j)_H$. La base hilbertienne définit ainsi une isométrie entre H et ℓ^2.

Nous allons utiliser certaines bases hilbertiennes afin de résoudre numériquement les problèmes aux limites par la méthode de Galerkin appliquées dans les espaces V_n. Comme on l'a observé, dans l'estimation fondamentale (5.2), l'analyse de cette méthode se ramène à étudier l'erreur de meilleure approximation de la solution u dans V_n, c'est à dire

$$\min_{v_n \in V_n} \|u - v_n\|_V,$$

où V est la norme de l'espace de Hilbert dans lequel la formulation variationelle est posée et vérifie les hypothèse de Lax-Milgram.

Si on a utilisé une base orthonormée pour le produit scalaire de V, cette erreur d'approximation est donnée par

$$\min_{v_n \in V_n} \|u - v_n\|_V = \|u - P_{V_n} u\|_V = \Big(\sum_{j>n} |c_j(u)|^2\Big)^{1/2}, \quad c_j(u) = (u, e_j)_V,$$

et son évaluation se ramène ainsi à comprendre les propriétés des coefficients de u.

Nous allons faire cette étude pour trois type de bases hilbertiennes $(e_n)_{n\geq 1}$. Dans la suite, on considérera plus généralement des espaces V_n de la forme

$$V_n = \text{vect}\{e_1, \ldots, e_{d_n}\},$$

où $(d_n)_{n\geq 1}$ est une suite croissante d'entiers positifs, et on aura ainsi $\dim(V_n) = d_n$.

7.2 Bases de Fourier

Les séries de Fourier sont adaptées à la représentation des fonctions périodiques. Pour les fonctions de période 1, elles peuvent s'écrire de façon équivalente sous la forme réelle

$$v(x) = \sum_{k\geq 0} a_k \cos(2\pi kx) + \sum_{k\geq 1} b_k \sin(2\pi kx), \tag{7.1}$$

ou la forme complexe

$$v(x) = \sum_{k\in\mathbb{Z}} c_k e^{i2\pi kx}. \tag{7.2}$$

On adoptera ici la forme complexe, plus simple, tout en remarquant que $c_{-k} = \overline{c}_k$ dans le cas où v est réelle. Si la série converge uniformément, ou simplement dans $L^1(]0,1[)$, il est facile de vérifier par intégration que le coefficient c_k est donné par

$$c_k = c_k(v) = \int_{-\pi}^{\pi} v(x) e^{-i2\pi kx} dx.$$

Dans toute cette section, on posera à nouveau

$$\Omega =]0,1[$$

La série de Fourier (7.2) s'interprete comme une décomposition de v dans une base hilbertienne de $L^2(\Omega)$ en écrivant

$$v = \sum_{k\in\mathbb{Z}} (v, e_k)_{L^2} e_k, \quad e_k(x) := e^{i2\pi kx}.$$

La seule différence avec le cadre général présenté dans la section précèdente est que les fonctions e_k sont naturellement indexées par $\mathbb{Z}$ au lieu de $\mathbb{N}$ (on pourrait les re-indexer au prix de notations moins naturelles), et que ce sont des fonctions à valeur complexe. Le produit scalaire est alors défini par

$$(v,w)_{L^2} = \int_\Omega v(x)\overline{w(x)} dx.$$

L'orthonormalité de la suite $(e_k)_{k\in\mathbb{Z}}$ est immédiate en calculant les produits scalaires.

La propriété de totalité de la suite peut se démontrer de plusieurs manières. On renvoie en particulier les étudiants aux cours introductifs sur les séries de Fourier et les conditions assurant la convergence ponctuelle ou uniforme des sommes partielles

$$S_n v(x) := \sum_{|k|\leq n} (v, e_k)_{L^2} e_k(x) = \sum_{|k|\leq n} c_k(v) e^{i2\pi kx},$$

vers v lorsque $n \to +\infty$. Notons que S_n est le projecteur orthogonal sur l'espace

$$V_n := \text{vect}\{e_{-n}, \ldots, e_n\} \tag{7.3}$$

des *polynômes trigonométriques* de degré n et de période 1. On a clairement

$$d_n = \dim(V_n) = 2n+1.$$

L'étude générale de la convergence des série de Fourier, en particulier au sens ponctuel, L^p ou de la norme de la convergence uniforme, est à l'origine de nombreux développements mathématiques depuis le début du XIXème siècle. On rappelle en particulier qu'il ne suffit pas qu'une fonction v soit continue et 2π-périodique pour avoir la convergence uniforme ou même ponctuelle, mais que celle-ci peut-être obtenue si on fait des hypothèse supplémentaire de régularité sur v : elle est assurée par exemple lorsque v est 1-périodique et de classe $\mathcal{C}^1$ sur $\mathbb{R}$ (ou simplement s-Hölderienne pour un $s > 0$). On note $\mathcal{C}^m_{per}(\Omega)$ l'ensemble des fonctions 1-périodique qui sont de classe $\mathcal{C}^m$ sur $\mathbb{R}$.

L'union sur $n \geq 0$ des espaces V_n est donc dense dans $\mathcal{C}^1_{per}$ au sens de la norme L^∞, et donc de la norme de $L^2(\Omega)$. Or $\mathcal{C}^1_{per}$ est dense dans $L^2(\Omega)$ (on peut par exemple prolonger par périodicité les fonctions de $\mathcal{D}(\Omega)$ qui sont dense dans $L^2(\Omega)$). On en déduit la densité de l'union des espaces V_n dans $L^2(\Omega)$, et on conclut que la suite $(e_k)_{k\in\mathbb{Z}}$ est totale dans cet espace. L'égalité de Parseval

$$\int_\Omega |v(x)|^2 dx = \sum_{k\in\mathbb{Z}} |c_k(v)|^2,$$

exprime une isométrie ente $L^2(\Omega)$ et $\ell^2(\mathbb{Z})$.

Remarque 7.2.1 *On peut facilement adapter les séries de Fourier à des fonctions de période quelconque en chaque variable : pour des fonctions de période F, il suffit de modifier la définition de e_k suivant $e_k := \frac{1}{\sqrt{F}} e^{i\frac{2\pi k\cdot x}{F}}$. L'analyse étant tout à fait similaire, on considére ici uniquement le cas $F = 1$ qui a l'avantage de simplifier les notations. Un autre cas souvent considéré est $F = 2\pi$ et $\Omega =]-\pi, \pi[$ qui permet d'enlever le facteur 2π dans l'exponentielle mais le fait apparaître dans la normalisation.*

Afin d'étudier l'erreur d'approximation par les espace V_n et la relier à la régularité de la fonction v, on introduit la version périodique des espaces de Sobolev $H^m(\Omega)$ dont la définition est très similaire.

Définition 7.2.1 *Soit m un entier positif. Une fonction $v \in L^2(\Omega)$ appartient à $H^m_{per}(\Omega)$ si et seulement si pour tout $l \leq m$, il existe $w_l \in L^2(\Omega)$ telle que*

$$\int_\Omega w_l \varphi = (-1)^l \int_\Omega v \varphi^{(l)},$$

pour tout $\varphi \in \mathcal{C}^\infty_{per}(\Omega)$ où $\mathcal{C}^\infty_{per}(\Omega)$ est l'ensemble des fonctions $\mathcal{C}^\infty$ sur $\mathbb{R}$ et de période 1.

Comme pour les espaces de Sobolev non-périodiques on note alors $w_l = v^{(l)}$, la dérivée faible d'ordre l. On a en particulier la récurrence

$$H^m_{per}(\Omega) = \{v \in H^1_{per}(\Omega) \ : \ v' \in H^{m-1}_{per}(\Omega)\}.$$

On définit les normes et semi-normes H^m_{per} de la même manière que celle des espaces H^m. Les propriétés des espaces de Sobolev périodiques sont très similaire à celles des espaces non-périodiques et se démontrent de la même façon : complétude, densité des fonctions $\mathcal{C}^\infty_{per}(\Omega)$, injection continue de $H^1_{per}(\Omega)$ dans $\mathcal{C}^0_{per}(\Omega)$ et plus généralement de $H^m_{per}(\Omega)$ dans $\mathcal{C}^{m-1}_{per}(\Omega)$. Notons que $H^m_{per}(\Omega) \subset H^m(\Omega)$ et que cette inclusion est stricte lorsque $m \geq 1$ à cause de la continuité des dérivées jusqu'à l'ordre $m-1$. On a cependant $H^0_{per}(\Omega) = H^0(\Omega) = L^2(\Omega)$.

En appliquant la définition ci-dessus à la fonction $\varphi(x) = e^{-i2\pi kx}$, on obtient la relation importante

$$c_k(v^{(l)}) = (i2\pi k)^l c_k(v),$$

qui permet de relier la régularité de v aux propriétés de décroissance de ses coefficients de Fourier. En particulier, l'égalité de Parseval nous indique qu'une fonction v appartient à l'espace de Sobolev $H^m_{per}(\Omega)$ si et seulement si

$$\sum_{k\in\mathbb{Z}} |2\pi k|^{2l} |c_k|^2 < \infty, \quad l \leq m, \quad c_k = c_k(v).$$

La norme H^m_{per} de v s'écrit

$$\|v\|^2_{H^m_{per}} = \sum_{k\in\mathbb{Z}} \Big(\sum_{l\leq m} |2\pi k|^{2l} \Big) |c_k|^2, \tag{7.4}$$

et la semi-norme

$$|v|^2_{H^m_{per}} = \sum_{k\in\mathbb{Z}} |2\pi k|^{2l} |c_k|^2.$$

On remarque au passage que les fonctions e_k sont orthogonales pour le produit scalaire de H^m_{per} mais pas orthonormales. On obtient une base orthonormée si on les renormalise en définissant

$$f_k = \Big(\sum_{l\leq m} |2\pi k|^{2l} \Big)^{-1/2} e_k.$$

On note que

$$S_n f = \sum_{|k|\leq n} (f, e_k)_{L^2} e_k = \sum_{|k|\leq n} (f, f_k)_{H^m_{per}} f_k,$$

et par conséquent S_n est aussi le projecteur orthogonal sur V_n pour la norme H^m_{per}.

La caractérisation par les séries de Fourier des espaces de Sobolev $H^m_{per}(T)$ nous permet d'établir un résultat d'approximation élémentaire par les sommes partielles de Fourier en norme L^2 et H^l_{per}.

Théorème 7.2.1 *Soit $m > 0$ un entier. Pour tout $v \in H^m_{per}(\Omega)$, on a*

$$\|v - S_n v\|_{L^2} \leq C n^{-m} |v|_{H^m_{per}}, \quad n > 0, \quad C = (2\pi)^{-m} \tag{7.5}$$

Si l est un entier tel que $0 \leq l \leq m$, on a

$$|v - S_n v|_{H^l_{per}} \leq C n^{-(m-l)} |v|_{H^m_{per}}, \quad n > 0, \quad C = (2\pi)^{-(m-l)}, \tag{7.6}$$

ainsi que

$$\|v - S_n v\|_{H^l_{per}} \leq C n^{-(m-l)} |v|_{H^m_{per}}, \quad n > 0, \quad C = \sqrt{l+1}(2\pi)^{-(m-l)}. \tag{7.7}$$

Démonstration. Comme $f - S_n f = \sum_{|k|>n} c_k e_k$ avec $c_k = c_k(f)$, on peut écrire

$$\begin{aligned} \|f - S_n f\|_{L^2}^2 &= \sum_{|n|>n} |c_k|^2 \\ &\leq (2\pi n)^{-2m} \sum_{k\in\mathbb{Z}^d} (2\pi kl)^{2m} |c_k|^2 \\ &= (2\pi n)^{-2m} |f|_{H^m_{per}}^2, \end{aligned}$$

et on obtient ainsi (7.5). De la même manière on a

$$\begin{aligned} |f - S_n f|_{H^l_{per}}^2 &= \sum_{|n|>n} (2\pi k)^{2l} |c_k|^2 \\ &\leq (2\pi n)^{-2(m-l)} \sum_{k\in\mathbb{Z}^d} (2\pi k)^{2m} |c_k|^2 \\ &= (2\pi n)^{-2(m-l)} |f|_{H^m_{per}}^2, \end{aligned}$$

et on obtient ainsi (7.6). En sommant les estimations en normes L^2 et semi-normes H^j_{per} pour $j = 1, \ldots, l$ élevées au carré, on obtient (7.7). □

Les bases de Fourier sont donc bien adaptées à l'approximation des fonctions appartenant aux espaces de Sobolev H^m_{per}. Dans le cas des solutions de problèmes aux limites ou d'évolution, l'appartenance à de tels espaces sera assurée de manière naturelle si l'on remplace les conditions de Dirichlet par des conditions aux limites périodiques. Nous considérons ici l'exemple suivant

$$-u''(x) + c(x)u(x) = f(x), \quad x \in]0,1[, \quad u(0) = u(1), \quad u'(0) = u'(1). \tag{7.8}$$

où c est une fonction de $\mathcal{C}^0(\Omega)$ telle que $c(x) > \eta$ pour tout $x \in \Omega$, avec $\eta > 0$, et où $f \in L^2(\Omega)$. En multipliant cette équation par une fonction $v \in H^1_{per}(\Omega)$ et en intégrant par partie, on trouve

$$\int_\Omega (u'(x)v'(x) + c(x)u(x)v(x))dx = \int_\Omega f(x)v(x) + [u'v]_0^1,$$

et le dernier terme $[u'v]_0^1$ est nul du fait des périodicités de u' et de v. Cela nous conduit à la formulation variationelle : trouver $u \in V := H^1_{per}(\Omega)$ tel que

$$a(u,v) = \ell(v), \quad v \in V,$$

où a et ℓ ont la même forme que pour le problème avec conditions homogènes de Dirichlet. Seul l'espace V a change : on a remplacé $H^1_0(\Omega)$ par $H^1_{per}(\Omega)$. Notons que cet espace contient la condition de périodicité $u(0) = u(1)$ puisqu'il s'injecte dans $\mathcal{C}^0_{per}(\Omega)$.

Les hypothèses de Lax-Milgram sont clairement vérifiées avec

$$C_\ell = \|f\|_{L^2}, \quad C_a = \max\{1, \|c\|_{L^\infty}\}, \quad \alpha = \min\{1, \eta\}.$$

Ceci nous assure l'existence et l'unicité de la solution $u \in V$. On remarque aussi que cette solution est en fait dans $H^2_{per}(\Omega)$: pour toute fonction $\varphi \in \mathcal{C}^\infty_{per}(\Omega)$ on a, puisque $\varphi \in V$,

$$-\int_\Omega u'(x)\varphi'(x)dx = \int_\Omega (c(x)u(x) - f(x))\varphi(x)dx,$$

qui nous montre que la dérivée faible de u' au sens de H^1_{per} est égale à $cu - f \in L^2(\Omega)$. Ceci nous montre aussi que l'équation $-u'' + cu = f$ est vérifiée dans L^2. On en déduit en particulier que $u \in \mathcal{C}^1_{per}(\Omega)$ et l'on retrouve ainsi la condition de périodicité pour u'. Tout comme dans le cas des conditions de Dirichlet - voir remarque 4.4.4 - on peut monter en régularité et démontrer par récurrence le résultat plus général suivant.

Proposition 7.2.1 *Si $f \in H^m_{per}(\Omega)$ et $c \in \mathcal{C}^m_{per}(\overline{\Omega})$ alors la solution $u \in V = H^1_{per}(\Omega)$ du problème variationel appartient $H^{m+2}_{per}(\Omega)$.*

Appliquons à présent la méthode de Galerkin en utilisant les espaces V_n des polynômes trigonométriques de degré n. Ces espaces sont contenus dans $\mathcal{C}^\infty_{per}(\Omega)$ et donc a-fortiori dans V. La solution discrete est donc $u_n \in V_n$ telle que

$$a(u_n, v_n) = \ell(v_n), \quad v_n \in V_n.$$

D'après le lemme de Cea, on sait que

$$\|u - u_n\|_V \leq C \min_{v_n \in V_n} \|u - v_n\|_V = C\|u - S_n u\|_V,$$

où $C = (C_a/\alpha)^{1/2} = \left(\frac{\max\{1, \|c\|_{L^\infty}\}}{\min\{1,\eta\}}\right)^{1/2}$. Puisque $V = H^1_{per}$, on déduit du théorème 7.2.1, que si la solution u appartient à $H^{m+1}_{per}(\Omega)$ pour un entier $m \geq 1$, on a l'estimation d'erreur

$$\|u - u_n\|_V \leq C|u|_{H^m_{per}} n^{-(m-1)},$$

où $C = \sqrt{2}(2\pi)^{-(m-1)} \left(\frac{\max\{1, \|c\|_{L^\infty}\}}{\min\{1,\eta\}}\right)^{1/2}$.

On voit ici apparaître une caractéristique de la méthode d'approximation variationelle fondée sur les séries de Fourier : la vitesse de convergence $O(n^{-(m-1)})$ augmente avec la régularité de la solution, sans limitation liée à l'ordre de la méthode. D'après les résultats de régularité de la proposition 7.2.1, on obtiendra en particulier cette vitesse si $f \in H^{m-1}_{per}$ et $c \in \mathcal{C}^{m-1}_{per}(\Omega)$. On note la différence avec les méthodes d'éléments finis P_k pour lesquelles on a vu que la vitesse de convergence est limitée par $O(h^k)$ même si la solution est très régulière. Les méthodes numériques sans limitation d'ordre sont appelées méthodes *spectrales*.

Remarque 7.2.2 *Dans le calcul de la solution u_n, il est intéressant d'utiliser la base de Fourier*

$$f_k = \left(1 + |2\pi k|^2\right)^{-1/2} e_k,$$

qui est orthonormée pour le produit scalaire H^1_{per}. En effet, si $v_n \in V_n$ et $\overline{v}_n$ est son vecteur de coordonnée dans cette base, on a

$$\|v_n\|^2_{H^1_{per}} = \|\overline{v}_n\|^2_{\ell^2}.$$

D'autre part, la matrice de rigidité A_n dont les éléments sont les $a(f_k, f_l)$, pour $|k|, |l| \leq n$, vérifie

$$(A_n \overline{v}_n, \overline{v}_n)_{\ell^2} = a(v_n, v_n)$$

Or on sait que

$$\alpha\|v\|^2_{H^1_{per}} \leq a(v,v) \leq C_a\|v\|^2_{H^1_{per}}, \quad v \in H^1_{per}(\Omega),$$

et par conséquence pour tout vecteur $\overline{v}_n$, on a

$$\alpha\|\overline{v}_n\|^2_{\ell^2} \leq (A_n\overline{v}_n, \overline{v}_n)_{\ell^2} \leq C_a\|\overline{v}_n\|^2_{\ell^2},$$

ce qui montre que $\lambda_{\min}(A_n) \geq \alpha$ et $\lambda_{\max}(A_n) \leq C_a$. On voit ainsi que le nombre de conditionnement de A_n est borné suivant

$$\kappa(A_n) \leq \frac{C_a}{\alpha},$$

indépendamment de n. Contrairement ce qui a été observé pour la méthode des éléments finis, on peut donc ici éviter la dégradation du nombre de conditionnement lorsque la précision augmente par une simple renormalisation de la base de V_n.

Le défaut principal des approximations fondées sur les séries de Fourier reste néammoins qu'elles sont adaptées aux conditions aux limites périodiques qui ne sont pas souvent celles qu'on rencontre en pratique. Nous allons voir que des méthodes spectrales peuvent être définies avec les conditions de Dirichlet en utilisant les espaces polynomiaux classiques.

7.3 Bases polynomiales

Les bases orthogonales de polynômes peuvent être définies de manières classiques en appliquant le procédé d'orthogonalisation de Gramm-Schmidt aux fonctions $\{1, x, x^2, \ldots, x^n, \ldots\}$, par rapport un produit scalaire Hilbertien donné. De manière classique, on choisit de travailler avec le produit scalaire L^2 sur l'intervalle symmétrique $]-1,1[$. On posera ainsi dans cette section

$$\Omega =]-1,1[,$$

tout en remarquant qu'il est possible de traiter le cas d'un intervalle Ω quelconque. Le procédé de Gramm-Schmidt conduit ainsi à la suite $(\tilde{L}_n)_{n\geq 1}$ des polynômes de Legendre, qui est caractérisée par les propriétés suivantes

(i) $\tilde{L}_0 = 1$
(ii) $\tilde{L}_n(x) = x^n + q_{n-1}$ avec $q_{n-1} \in P_{n-1}$.
(iii) $\int_\Omega \tilde{L}_n\tilde{L}_m = 0$ si $m \neq n$.

Le polynôme $\tilde{L}_n$ est donc de degré exactement égal à n et est défini en soustrayant la fonction $x \mapsto x^n$ sa projection orthogonale sur P_{n-1}. Par construction $\{\tilde{L}_0, \ldots, \tilde{L}_n\}$ est une base orthogonale de P_n. Notons que la suite ainsi définie n'est pas orthonormale et peut être orthonormalisée en posant

$$L^*_n = \|\tilde{L}_n\|^{-1}_{L^2(\Omega)}\tilde{L}_n.$$

Théorème 7.3.1 *La suite $(L^*_n)_{n\geq 0}$ des polynômes de Legendre renormalisés est une base Hilbertienne de $L^2(\Omega)$.*

Démonstration. On sait déjà qu'il s'agit d'une suite orthonormée. Il suffit donc de démontrer que l'union des espaces P_n est dense dans $L^2(\Omega)$. Or le théorème de Stone-Weierstrass nous indique que l'on peut approcher toute fonction continue sur $[-1,1]$ uniformément par une suite de polynômes, et donc a-fortiori en norme L^2. Comme $\mathcal{C}^0(\overline{\Omega})$ est lui-même dense dans $L^2(\Omega)$, on en déduit la densité de l'union des espaces P_n dans $L^2(\Omega)$. □

Ainsi toute fonction $v \in L^2(\Omega)$ peut se décomposer de manière unique suivant

$$v = \sum_{n\geq 0} c_n(v) L_n^*, \quad c_n(v) = \int_\Omega v(x) L_n^*(x) dx, \tag{7.9}$$

et on a $\|v\|_{L^2}^2 = \sum_{n\geq 0} |c_n(v)|^2$.

Remarque 7.3.1 *Il est évidemment possible d'obtenir des suites orthonormée de polynômes dans $L^2(\Omega)$ pour d'autres intervalles Ω en appliquant des changement de variables, comme on l'a déjà remarqué pour les séries de Fourier. On peut aussi fabriquer par le même procédé des suites de polynômes orthogonaux dans des espaces $L^2(\Omega, d\mu)$ où μ n'est pas la mesure de Lebesgue. On obtient en particulier les exemples classiques suivants :*

- Polynômes de Chebychev : $\Omega =]-1,1[$ et $d\mu = (1-x^2)^{-1/2}dx$.

- Polynômes de Hermite : $\Omega = \mathbb{R}$ et $d\mu = \exp(-x^2)dx$.

- Polynômes de Laguerre : $\Omega = \mathbb{R}_+$ et $d\mu = \exp(-x)dx$.

Notons qu'il est nécessaire d'avoir un poids décroissant rapidement à l'infini dans la mesure μ lorsque Ω est non-borné puisque l'existence d'une base orthonormée polynomiale implique en particulier que $\int_\Omega |p|^2 d\mu < \infty$ pour tout polynôme p.

La normalisation la plus souvent appliquée des polynômes de Legendre consiste cependant à imposer la valeur 1 au point $x = 1$, en posant

$$L_n = (\tilde{L}_n(1))^{-1} \tilde{L}_n.$$

On note en particulier que $L_0 = 1$ et $L_1(x) = x$. La propriété $\tilde{L}_n(1) \neq 0$ découle de la première propriété dans le résultat suivant qui compile quelques propriétés importantes.

Théorème 7.3.2 *Les polynômes de Legendre vérifient les propriétés suivantes :*

(i) Leurs n racines sont réelles, simples, et contenues dans $]-1,1[$.

(ii) Si n est pair, L_n est pair, si n est impair, L_n est impair.

(iii) Formule de Rodrigues : on a $L_n = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}\left((1-x^2)^n\right)$

(iv) Coefficient dominant : on a $L_n(x) = k_n x^n + q_{n-1}(x)$ où $q_{n-1} \in P_{n-1}$ et $k_n = \frac{(2n)!}{2^n (n!)^2}$.

(v) Equation différentielle : $((1-x^2)L_n')' + n(n+1)L_n = 0$.

(vi) Norme L^2 et récurrence : $\|L_n\|_{L^2}^2 = (n+1/2)^{-1}$ et $L_{n+1} = \frac{2n+1}{n+1} x L_n - \frac{n}{n+1} L_{n-1}$.

Démonstration. (i) Soient $\zeta_1, \zeta_2, \ldots, \zeta_k$ les racines de $\tilde{L}_n$ qui appartiennent à $]-1,1[$, de multiplicité impaire $2\alpha_j + 1$ et $\eta_1, \eta_2, \ldots, \eta_l$ les racines de $\tilde{L}_n$ qui appartiennent à $]-1,1[$, de multiplicité paire $2\beta_i$. On peut donc écrire que

$$\tilde{L}_n(x) = q(x) \prod_{i=1}^{l} (x-\eta_i)^{2\beta_i} \prod_{j=1}^{k} (x-\zeta_j)^{2\alpha_j+1},$$

(avec la convention que si $l=0$ ou $k=0$, le produit correspondant vaut 1) où q n'a pas de racine dans $]-1,1[$. Par conséquent, q garde un signe constant sur cet intervalle. Soit $p(x)=\prod_{j=1}^{k}(x-\zeta_j)$. Si L_n a une racine qui n'est pas dans $]-1,1[$ ou si toutes ses racines sont bien dans $]-1,1[$, mais l'une d'entre elle est de multiplicité strictement supérieure à 1, alors $\deg p<n$ (en effet, dans ce cas $k=\deg p<\sum_{i=1}^{l}(2\beta_i)+\sum_{i=1}^{k}(2\alpha_j+1)\leq \deg L_n$). Par conséquent, par orthogonalité

$$0=\int_\Omega \tilde{L}_n p=\int_\Omega q(x)\Big(\prod_{i=1}^{l}(x-\eta_i)^{\beta_i}\prod_{j=1}^{k}(x-\zeta_j)^{\alpha_j+1}\Big)^2\,dx$$

mais l'intégrande du membre de droite ne change pas de signe dans $]-1,1[$, ce qui est une contradiction. Donc toutes les racines de $\tilde{L}_n$ sont dans $]-1,1[$ et elles sont simples.

(ii) Considérons les polynômes $\tilde{L}_n$, et soit $p_n(x)=(-1)^n\tilde{L}_n(-x)$. Alors p_n satisfait visiblement la propriété de suite orthogonale telle que $p_n(x)=x^n+q_{n-1}(x)$ avec $q_{n-1}\in P_{n-1}$. Par unicité on en déduite que $p_n(x)=\tilde{L}_n(x)$ ce qui entraîne les propriétés de parité et d'imparité annoncées.

(iii) Le polynôme $(1-x^2)^n$ est de degré $2n$, donc sa dérivée $n$ième est de degré n. Il s'annule ainsi que ses $n-1$ premières dérivées aux points -1 et 1. Par conséquent, pour toute fonction q, n fois dérivable, on obtient en intégrant n fois par parties

$$\int_\Omega \frac{d^n}{dx^n}\big((1-x^2)^n\big)q(x)\,dx=(-1)^n\int_\Omega (1-x^2)^n\frac{d^n q}{dx^n}(x)\,dx,$$

car tous les termes tout intégrés disparaissent. En particulier, on a

$$\int_\Omega \frac{d^n}{dx^n}\big((1-x^2)^n\big)q(x)\,dx=0,\quad q\in P_{n-1}$$

Donc, $\frac{d^n}{dx^n}\big((1-x^2)^n\big)$ est un polynôme de P_n, orthogonal à P_{n-1}. Il existe donc un scalaire λ_n tel que

$$\frac{d^n}{dx^n}\big((1-x^2)^n\big)=\lambda_n L_n(x).$$

Pour identifier ce scalaire, on considère ici la valeur prise en $x=1$ par les deux membres de cette égalité. D'un côté, $L_n(1)=1$, et de l'autre, comme $(1-x^2)^n=(1-x)^n(1+x)^n$, on voit que sa dérivée n-ème au point $x=1$ et la même que celle de $(-x)^n(2+x)^n$ au point $x=0$ soit $\lambda_n=(-1)^n 2^n n!$. On obtient ainsi la formule de Rodrigues.

(iv) En utilisant la formule de Rodrigues et en remarquant que $(1-x^2)^n=(-1)^n x^{2n}+\ldots$, on en déduit par dérivation d'ordre n la valeur annoncée du coefficient dominant k_n.

(v) Comme L_n est de degré n, on voit que $((1-x^2)L_n')'$ est de degré n. Pour tout polynôme $q\in P_{n-1}$, on a

$$\begin{aligned}\textstyle\int_\Omega\big((1-x^2)L_n'\big)'q &= [(1-x^2)L_n'q]_{-1}^{1}-\textstyle\int_\Omega(1-x^2)L_n'q'\\ &= -\textstyle\int_\Omega L_n\big((1-x^2)q'\big)'=0,\end{aligned}$$

puisque $\big((1-x^2)q'\big)'$ est de degré inférieur ou égal à $n-1$. Par conséquent, il existe un scalaire λ_n tel que

$$\big((1-x^2)L_n'\big)'=\lambda_n L_n.$$

en identifiant les termes de plus haut degré de part et d'autre, on trouve $\lambda_n=-n(n+1)$ et on obtient ainsi l'équation différentielle.

(vi) A cause de la normalisation $L_n(x) = 1$, on sait que $L_0 = 1$ et $L_1(x) = x$. Par récurrence, supposons que $\|L_k\|_{L^2}^2 = (k+1/2)^{-1}$ pour $k = 0, \ldots, n$. On regarde alors la fonction définie par la formule de récurrence

$$g_{n+1}(x) = \frac{2n+1}{n+1} x L_n(x) - \frac{n}{n+1} L_{n-1}(x).$$

C'est un polynôme de degré $n+1$, et puisque

$$\int_\Omega x L_n p = \int_\Omega L_{n-1} p = 0, \quad p \in P_{n-2},$$

ce polynôme est orthogonal à $L_0, \ldots, L_{n-2}$. Son produit scalaire avec L_n est nul car $\int_\Omega x L_n^2 = 0$ par parité de L_n^2. Son produit scalaire avec L_{n-1} est donné par

$$\int_\Omega g_{n+1} L_{n-1} = \frac{2n+1}{n+1} \int_\Omega x L_n L_{n-1} dx - \frac{n}{n+1} \|L_{n-1}\|_{L^2}^2$$

L'intégrale $\int_\Omega x L_n(x) L_{n-1}(x)$ se calcule en remarquant que d'après l'orthogonalité de L_n à P_{n-1} on a $\|L_n\|_{L^2}^2 = \int_\Omega k_n x^n L_n dx$ et $\int_\Omega x L_n L_{n-1} = \int_\Omega k_{n-1} x^n L_n$. Par conséquent

$$\int_\Omega x L_n L_{n-1} = \frac{k_{n-1}}{k_n} \|L_n\|_{L^2}^2 = \frac{(2n-2)!}{2^{n-1}((n-1)!)^2} \frac{2^n (n!)^2}{(2n)!} \frac{2}{2n+1} = \frac{2n}{(2n-1)(2n+1)},$$

où on a utilisé la valeur de k_n et l'hypothèse de récurrence. Par conséquence

$$\int_\Omega g_{n+1} L_{n-1} = \frac{2n}{(2n-1)(n+1)} - \frac{n}{n+1} \|L_{n-1}\|_{L^2}^2 = 0,$$

à nouveau par l'hypothèse de récurrence. On a ainsi montré que g_{n+1} est orthogonal à P_n et sa définition nous montre aussi que $g_{n+1}(1) = 1$. On a donc $g_{n+1} = L_{n+1}$ et la formule de récurrence est démontrée. D'autre part, cette formule montre que la norme L^2 de L_{n+1} est donnée par

$$\|L_{n+1}\|_{L^2}^2 = \frac{2n+1}{n+1} \int_\Omega x L_{n+1} L_n = \frac{2n}{(2n-1)(2n+1)} \frac{2(n+1)}{(2n+1)(2n+3)} = (n+3/2)^{-1},$$

ce qui conclut la récurrence. □

Remarque 7.3.2 *Toutes ces propriétés se traduisent immédiatement sur les polynômes de Legendre L_n^* normalisés dans L^2 et qui sont définit par*

$$L_n^* = \sqrt{n+1/2} L_n,$$

et donc $L_n^*(1) = \sqrt{n+1/2}$. On a en particulier la récurrence

$$L_{n+1}^* = \frac{(2n+1)\sqrt{n+3/2}}{(n+1)\sqrt{n+1/2}} x L_n^* - \frac{n\sqrt{n+3/2}}{(n+1)\sqrt{n-1/2}} L_{n-1}^*$$

L'équation différentielle reste inchangée.

Remarque 7.3.3 *Pour un intervalle $\Omega =]a,b[$ quelconque, on obtient une base Hilbertienne de polynômes pour $L^2(\Omega)$ en posant*

$$L_n^{[a,b]}(x) = \sqrt{\frac{2}{b-a}} L_n^*\Big(\frac{2}{b-a}(x - \frac{a+b}{2})\Big),$$

et on peut effectuer une analyse similaire à celle qui va suivre, mais par simplicité on conservera le choix $\Omega =]-1,1[$

Nous allons à présent établir des résultats d'approximation d'une fonction u dans les espaces polynomiaux P_n. On suit ici une démarche analogue à celle employée pour les séries de Fourier : on considère une fonction $v \in L^2(\Omega)$ et sa décomposition (7.9) et on montre que ses propriétés de régularité se traduisent par des propriétés de sommabilité de ses coefficients $c_n(v)$ avec des poids croissants, comme dans (7.4) pour les coefficients de Fourier.

Théorème 7.3.3 *Pour tout entier $m \geq 1$, il existe une constante C qui ne dépend que de m telle que*

$$\sum_{n\geq 0} |n|^{2m} |c_n(v)|^2 \leq C \|v\|_{H^m}^2, \tag{7.10}$$

pour tout $v \in H^m(\Omega)$.

Démonstration. Nous allons utiliser l'équation différentielle vérifiée par les polynômes de Legendre, qui peut s'écrire

$$DL_n^* + n(n+1)L_n^* = 0$$

où D est l'opérateur différentiel $Dv = ((1-x^2)v')'$. On remarque que si v et w sont suffisament régulières, par exemple dans $\mathcal{C}^2(\overline{\Omega})$, on a

$$\int_\Omega Dv(x)w(x)dx = -\int_\Omega (1-x^2)v'(x)w'(x)dx = \int_\Omega v(x)Dw(x)dx,$$

ce qui nous montre que l'opérateur D est en ce sens *auto-adjoint*. En appliquant cela à $w = L_n^*$ et en utilisant l'équation différentielle, on obtient

$$c_n(Dv) = \int_\Omega (Dv)L_n^* = \int_\Omega v(DL_n^*) = -n(n+1)c_n(v).$$

Notons que si $v \in \mathcal{C}^2(\Omega)$ on a $Dv(x) = (1-x^2)v''(x) - 2xv'(x)$ mais ce calcul garde un sens si $v \in H^2(\Omega)$. On a dans ce cas $Dv \in L^2(\Omega)$ avec

$$\|Dv\|_{L^2} \leq \|v''\|_{L^2} + 2\|v'\|_{L^2} \leq \sqrt{3}(\|v''\|_{L^2}^2 + \|v'\|_{L^2}^2)^{1/2} \leq \sqrt{3}\|v\|_{H^2}.$$

On peut ainsi écrire

$$\|Dv\|_{L^2}^2 = \sum_{n\geq 0} |c_n(Dv)|^2 = \sum_{n\geq 0} |n(n+1)|^2 |c_n(v)|^2.$$

ce qui conduit l'estimation, pour tout $v \in H^2(\Omega)$,

$$\sum_{n\geq 0} |n|^4 |c_n(v)|^2 \leq 3\|v\|_{H^2}^2.$$

On peut itérer ce raisonnement en remarquant que pour tout $k \geq 1$ et $v \in H^{2k}$ on a

$$c_n(D^k v) = -n(n+1)c_n(D^{k-1}v) = \cdots = (-n(n+1))^k c_n(v),$$

et que

$$\|D^k v\|_{L^2}^2 \leq C\|v\|_{H^{2k}}^2,$$

où la constante C ne dépend que de k. En remarquant que $n^2 \leq n(n+1)$, on aboutit ainsi à l'estimation pour tout $v \in H^{2k}(\Omega)$.

$$\sum_{n\geq 0} |n|^{4k}|c_n(v)|^2 \leq C\|v\|_{H^{2k}}^2.$$

On a ainsi obtenu (7.10) pour les valeurs paires de m. On peut obtenir une estimation similaire pour les espaces de régularité impaires $H^{2k+1}(\Omega)$: pour cela, on écrit pour v suffisament régulière, par exemple dans $\mathcal{C}^{4k+2}(\Omega)$,

$$-\int_\Omega (D^{2k+1}v)v = -\int_\Omega D^{k+1}vD^k v = \int_\Omega (1-x)^2 (D^k v)'(D^k v)' \leq \|(D^k v)'\|_{L^2}^2.$$

Puisque $D^{2k+1}v = -\sum_{n\geq 0}(n(n+1))^{2k+1}c_n(v)L_n^*$, on obtient ainsi

$$\sum_{n\geq 0} |n|^{4k+2}|c_n(v)|^2 = -\int_\Omega (D^{2k+1}v)v \leq \|(D^k v)'\|_{L^2}^2 \leq C\|v\|_{H^{2k+1}}^2,$$

où la constante C ne dépend que de k. Par densité, cette estimation est valable pour tout $v \in H^{2k+1}(\Omega)$. □

Le résultat ci-dessus nous conduit naturellement vers un résultat d'approximation.

Théorème 7.3.4 *Pour tout entier $m \geq 0$, il existe une constante C qui ne dépend que de m telle que,*

$$\min_{v_n \in P_n} \|v - v_n\|_{L^2} \leq C(n+1)^{-m}\|v\|_{H^m}, \tag{7.11}$$

pour tout $v \in H^m(\Omega)$.

Démonstration. En notant Π_n le projecteur orthogonal pour $L^2(\Omega)$ sur P_n, on peut écrire

$$\|v - \Pi_n v\|_{L^2}^2 = \sum_{k>n} |c_k(v)|^2 \leq (n+1)^{-2m} \sum_{k\geq 0} |k|^{2m}|c_k(v)|^2 \leq C(n+1)^{-2m}\|v\|_{H^m}^2,$$

où la dernière inégalité découle de (7.10), et on obtient ainsi l'estimation (7.11). □

Nous souhaitons appliquer la méthode d'approximation variationnelle dans les espaces de polynômes, au problème aux limite

$$-u''(x) + c(x)u(x) = f(x), \quad x \in \Omega =]-1,1[,$$

avec les conditions aux limites homogène de Dirichlet $u(-1) = u(1) = 0$. Il nous faut tout d'abord modifier l'espace P_n afin de tenir compte des conditions aux limites. Nous définissons pour cela l'espace

$$V_n := \{v \in P_n \ : v(-1) = v(1) = 0\}.$$

Il est immédiat de vérifier que cet est espace est constitué des fonctions de la forme

$$v(x) = (1-x^2)w(x), \quad w \in P_{n-2},$$

et que sa dimension est celle de P_{n-2}, c'est à dire

$$d_n = \dim(V_n) = n-1.$$

Cet espace est contenu dans l'espace $V = H_0^1(\Omega)$ qui est l'espace naturel de la formulation variationelle : trouver $u \in V$ telle que

$$a(u,v) = \ell(v), \quad v \in V,$$

avec $a(u,v) := \int_\Omega (c(x)u(x)v(x) + u'(x)v'(x))dx$ et $\ell(v) := \int_\Omega f(x)v(x)dx$. La solution discrète est donc définie par : trouver $u_n \in V_n$ telle que

$$a(u_n, v_n) = \ell(v_n), \quad v_n \in V_n.$$

Pour l'estimation d'erreur on va utiliser la norme $\|v\|_V = \|v\|_{H_0^1} = \|v'\|_{L^2}$ dont on rappelle qu'elle est équivalente à la norme H^1 sur V.

On a en particulier pour la forme bilinaire a la propriété de continuité avec $C_a = 1 + \|c\|_{L^\infty} C_P^2$, où C_P est la constante dans l'inégalité de Poincaré

$$\|v\|_{L^2} \leq C_P \|v'\|_{L^2}, \quad v \in H_0^1(\Omega).$$

ainsi que la propriété de coercivité avec $\alpha = 1$. D'après le lemme de Cea, on a l'estimation d'erreur

$$\|u - u_n\|_V \leq C \min_{v_n \in V_n} \|u - v_n\|_V, \quad C = (C_a/\alpha)^{1/2} = (1 + \|c\|_{L^\infty} C_P^2)^{1/2}.$$

Il nous reste à estimer $\min_{v_n \in V_n} \|u - v_n\|_V$, et pour cela il nous faut un resultat d'approximation similaire au théorème 7.3.4 mais pour l'erreur d'approximation mesurée en norme H_0^1 au lieu de L^2. Remarquons tout d'abord qu'on peut exhiber simplement une base de V_n qui sera orthonormée pour le produit scalaire de V. Il suffit de considérer les fonctions primitives des polynômes de Legendre normalisées dans L^2,

$$Q_k(x) = \int_{-1}^{x} L_k^*(t)dt,$$

pour $k = 1, \ldots, n-1$. Au vu de l'équation différentielle, ces fonctions sont aussi données par

$$Q_k(x) = -\frac{1}{k(k+1)}(1-x^2)(L_k^*)'(x).$$

Théorème 7.3.5 *Les fonctions $(Q_1, \ldots, Q_{n-1})$ forment une base orthonormée de V_n et la famille $(Q_k)_{k\geq 1}$ est une base orthonormée de V.*

Démonstration. On remarque tout d'abord que pour tout $k \geq 1$ la fonction Q_k est un polynôme de degré $k+1$. On a d'autre part

$$\int_\Omega L_k^*(t)dt = \int_\Omega L_0^*(t)L_k^*(t)dt = 0,$$

ce qui entraîne immédiatement que $Q_k(-1) = Q_k(1) = 0$. On voit ainsi que $Q_k \in V$. L'orthonormalité dans V est immédiate puisque

$$(Q_j, Q_k)_V = \int_\Omega Q_j' Q_k' = \int_\Omega L_j^* L_k^* = \delta_{j,k}.$$

On en déduite que $(Q_1, \ldots, Q_{n-1})$ est une base orthonormée de V_n puisqu'on a exactement $n-1$ fonctions orthonormées et que $\dim(V_n) = n-1$. D'autre part notons $\overline{\Pi}_n$ l'opérateur de projection V-orthogonale sur V_n. Celui-ci peut s'écrire

$$\overline{\Pi}_n v = \sum_{k=1}^{n-1} (v, Q_k)_V Q_k = \sum_{k=1}^{n-1} (\int_\Omega v' L_k^*) Q_k.$$

Ainsi

$$(\overline{\Pi}_n v)' = \sum_{k=1}^{n-1} c_k(v') L_k^* = \Pi_{n-1} v',$$

où Π_{n-1} est le projecteur L^2-orthogonal sur P_{n-1}, puisque

$$c_0(v') = (v', L_0^*)_{L^2} = \frac{1}{\sqrt{2}} \int_\Omega v' = \frac{1}{\sqrt{2}} (v(1) - v(-1)) = 0.$$

On voit ainsi que

$$\|\overline{\Pi}_n v - v\|_V = \|(\overline{\Pi}_n v)' - v'\|_{L^2} = \|\Pi_{n-1} v' - v'\|_{L^2} \to 0,$$

quand $n \to \infty$, et on conclut ainsi que $(Q_k)_{k \geq 1}$ est une base orthonormée de V. □

En utilisant cette nouvelle base orthonormée, on obtient aisément un résultat d'approximation dans V en utilisant celui qu'on a auparavant démontré dans L^2 pour les espaces P_n.

Théorème 7.3.6 *Pour tout entier $m \geq 1$, il existe une constante C qui ne dépend que de m telle que,*

$$\min_{v_n \in V_n} \|v - v_n\|_V \leq C n^{-(m-1)} \|v\|_{H^m}, \tag{7.12}$$

pour tout $v \in H_0^1(\Omega) \cap H^m(\Omega)$.

Démonstration. En utilisant les propriétés du projecteur $\overline{\Pi}_n$ utilisées dans le théorème 7.3.5, on peut écrire

$$\|v - \overline{\Pi}_n v\|_V = \|v' - \Pi_{n-1} v'\|_{L^2}.$$

D'après le théorème 7.3.4, on a l'estimation

$$\|v' - \Pi_{n-1} v'\|_{L^2} \leq C n^{-(m-1)} \|v'\|_{H^{m-1}},$$

où la constante C ne dépend que de m. On en déduit (7.12) □

Revenons à l'approximation variationelle du problème aux limites dans l'espace V_n. En combinant l'estimation du Lemme de Cea et le théorème 7.3.6, on en déduit que si que si la solution u appartient à $H^{m+1}(\Omega)$ pour un entier $m \geq 1$, on a alors l'estimation d'erreur

$$\|u - u_n\|_V \leq C\|u\|_{H^m} n^{-(m-1)},$$

où C depend de m et des constantes C_P et $\|c\|_{L^\infty}$. Tout comme pour les méthodes utilisant les bases de Fourier, la vitesse de convergence $O(n^{-(m-1)})$ augmente avec la régularité de la solution, sans limitation liée à l'ordre de la méthode. D'après les résultats de régularité indiqués dans la remarque 4.4.4, on obtiendra en particulier cette vitesse si $f \in H^{m-1}$ et $c \in C^{m-1}(\Omega)$. On parle ici de méthodes *spectrales* polynomiales.

Remarque 7.3.4 *La méthode spectrale polynomiale demande le calcul exact d'un certain nombre d'intégrales. Comme ce calcul n'est en général pas possible, on doit dans la pratique utiliser des méthodes d'intégration numérique. Parmi celles-ci, les méthodes de quadratures de Gauss ont l'avantage d'être exactes sur les polynômes de degré le plus élevé possible. On peut analyser l'erreur induite sur l'approximation à l'aide des techniques générales décrites dans la section 5.3*

Remarque 7.3.5 *On peut utiliser la base V-orthonormée $(Q_1, \ldots, Q_{n-1})$ pour le calcul de la solution discrète u_n. Par un raisonnement analogue à celui effectué dans la remarque 7.2.2 pour les bases de Fourier, on obtient que la matrice de rigidité résultante A_n a un nombre de conditionnement borné suivant*

$$\kappa(A_n) \leq \frac{C_a}{\alpha} = 1 + \|c\|_{L^\infty} C_P^2,$$

indépendamment de n.

7.4 Bases d'ondelettes

La théorie des bases d'ondelettes a été mise en place au cours des années 1980. Ces bases jouent aujourd'hui un rôle important en traitement du signal, en particulier pour la compression et la restauration des images. Elles ont été aussi utilisées pour la discrétisation de certaines équations différentielles et aux dérivées partielles, même si, comme les méthodes spectrales, elles sont moins souvent employées que les méthodes d'éléments finis et différences finies. Dans cette section, on se limite à la présentation de deux exemples élémentaires de telles bases, bien adaptés à la représentation de fonctions définies sur $\Omega = [0,1]$.

Commençons par le *système de Haar*. On note $\varphi = \chi_{[0,1]}$ la fonction constante égale à 1 sur Ω, et on définit

$$\psi = \chi_{[0,1/2[} - \chi_{[1/2,1[},$$

la fonction constante par morceaux qui vaut 1 sur $[0,1/2[$ et -1 sur $[1/2,1[$. On remarque que $\int_\Omega \varphi\psi = 0$ c'est à dire que ψ est orthogonale φ dans $L^2(\Omega)$.

Notons V_1 l'espace des fonctions constantes par morceaux sur les intervalles $[0,1/2[$ et $[1/2,1[$ auquel appartiennent φ et ψ. Cet espace est de dimension 2, et on voit ainsi que $\{\varphi,\psi\}$ forment une base orthogonale de V_1. Si l'on considère ensuite les deux fonctions

$$x \mapsto \psi(2x), \quad x \mapsto \psi(2x-1),$$

on voit qu'elles sont orthogonales dans L^2 aux fonctions de V_1. En notant V_2 l'espace des fonctions constantes par morceaux sur les intervalles deux fois plus petits $[k/4,(k+1)/4[$ pour $k=0,1,2,3$, on voit ainsi que les 4 fonctions $\{\varphi,\psi,\psi(2\cdot),\psi(2\cdot-1)\}$ constituent une base orthogonale de V_2.

On itére ce processus en définissant les espaces de fonctions constantes par morceaux sur les intervalles de longueur 2^{-j} obtenus par raffinements successifs

$$I_{j,k} := [2^{-j}k, 2^{-j}(k+1)[, \quad k=0,\dots 2^j-1.$$

On définit ainsi

$$V_j := \{v \,:\, v_{|I_{j,k}} \in P_0,\ k=0,\dots 2^j-1\},$$

pour tout entier $j \geq 0$. L'espace V_j est de dimension 2^j et on a la propriété d'emboitement

$$V_j \subset V_{j+1}.$$

Les fonctions

$$x \mapsto \psi(2^j(x-2^{-j}k)) = \psi(2^jx-k), \quad k=0,\dots,2^{j-1},$$

appartiennent à V_{j+1} et sont orthogonales aux fonctions de V_j. Elles forment ainsi une base orthogonale du supplémentaire orthogonal W_j de V_j dans V_{j+1} qui vérifie

$$\dim(W_j) = \dim(V_{j+1}) - \dim(V_i) = 2^j.$$

On peut normaliser ces fonctions dans L^2 en posant

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^jx-k).$$

et on obtient ainsi que la famille constituée de la fonction φ et des fonctions

$$\psi_{j,k}, \quad k=0,\dots,2^{j-1},\ j=0,\dots,n-1, \tag{7.13}$$

forme une base orthonormée de l'espace

$$V_n = V_0 \oplus^\perp W_0 \oplus^\perp W_1 \oplus^\perp \cdots \oplus^\perp W_{n-1}.$$

Notons la structure multiéchelle de cette base où, à l'exception de φ, toutes les fonctions sont obtenues $\psi_{j,k}$ à partir de la fonctions ψ par mise aux échelle 2^{-j} et translations de $k2^{-j}$.

Notons qu'une base orthonormée plus simple de V_n est donnée par les fonctions indicatrices des intervalles $I_{n,k}$ convenablement normalisées, c'est à dire

$$\varphi_{n,k}(x) = 2^{n/2}\varphi(2^nx-k), \quad k=0,\dots,2^n-1. \tag{7.14}$$

Cette base change completement à chaque fois qu'on passe de n à $n+1$, alors que la base d'ondelette définie par (7.13) s'enrichit d'un niveau à l'autre. La projection orthogonale $\Pi_n v$ d'une fonction v est définie par sa moyenne sur chaque intervalle

$$\Pi_n v_{|I_{n,k}} = 2^n \int_{I_{n,k}} v(x)dx.$$

Il est facile de voir que $\Pi_n v \to v$ dans L^2 pour tout $v \in L^2(\Omega)$, par exemple en montrant d'abord la convergence uniforme donc L^2 pour tout $v \in \mathcal{C}^0(\overline{\Omega})$ puis en raisonnant par densité (exercice).

Ceci nous montre que la famille infinie constituée de la fonction φ et des fonctions

$$\psi_{j,k}, \quad k = 0, \ldots, 2^{j-1},\ j \geq 0, \tag{7.15}$$

forme une base Hilbertienne de $L^2(\Omega)$. Cette base appelée système de Haar est connue depuis le début du XXème siècle. Notons qu'on peut la réindexer en une base orthonormée $(e_n)_{n\geq 0}$ en posant $e_0 = \varphi$ et $e_n = \psi_{j,k}$ pour $n = 2^j + k$, mais on garde le plus souvent la notation $\psi_{j,k}$.

La théorie récente des bases d'ondelettes systématise cette construction à partir d'espaces emboités de la forme V_j qui généralisent les espaces de fonctions constantes par morceaux : on peut par exemple considérer des fonctions polynomiales par morceaux sur les intervalles $I_{j,k}$ de degré et régularité globale prescrits. La difficulté est alors de savoir construire une fonction ψ de forme simple telle que ses versions mise à l'echelle 2^{-j} et translatées de $k2^{-j}$ engendrent un complémentaire (orthogonal ou plus général) W_j de V_j dans V_{j+1}. Nous ne rentrons pas plus loin dans cette théorie générale mais présentons un deuxième exemple élémentaire, lui aussi connu depuis plus longtemps.

Avant cela, expliquons l'un des intérêts fondamentaux de la base d'ondelette (7.13) par rapport à la base plus classique (7.14) pour décrire les fonctions du même espace V_n. Les coefficients d'une fonction v projetée sur V_n dans cette base sont donnés par

$$d_{j,k} = d_{j,k}(v) = \int_{I_{j,k}} v\psi_{j,k}.$$

Or la fonction $\psi_{j,k}$ est supportée dans l'intervalle $I_{j,k}$ on a $\int_\Omega \psi_{j,k} = 0$. Ceci montre que si la fonction v est constante dans cet intervalle alors $d_{j,k} = 0$. Ainsi, si la fonction v est constante par morceaux, avec des sauts de discontinuités en m points $x_1, \ldots, x_m$, on voit que les seuls coefficients $d_{j,k}$ non-nuls seront ceux tels que les intervalles $I_{j,k}$ contiennent l'un de ces m points. Pour un niveau d'échelle j fixé, il n'y a qu'un intervalle $I_{j,k}$ qui contient un point x_i donc en tout au plus $m \times n$ coeffcients non-nuls pour la description de $\Pi_n v$ au lieu de 2^n ce qui est une grande économie de données.

Plus généralement considérons une fonction v bornée par une constante $M_0 = \|v\|_{L^\infty}$, avec des sauts de discontinuités en m points $x_1, \ldots, x_m$, et de classe $\mathcal{C}^1$ avec une dérivée uniformément bornée par M_1 entre ces sauts. Pour tous les coefficients, on peut utiliser la borne uniforme sur v pour obtenir l'estimation

$$|d_{j,k}| = \Big| \int_{I_{j,k}} v\psi_{j,k} \Big| \leq M_0 \|\psi_{j,k}\|_{L^1} = M_0 2^{-j/2}.$$

En utilisant à nouveau $\int_\Omega \psi_{j,k} = 0$, on peut aussi écrire

$$|d_{j,k}| = \Big| \int_{I_{j,k}} v\psi_{j,k} \Big| = \Big| \int_{I_{j,k}} (v - v(2^{-j}k))\psi_{j,k} \Big|,$$

Lorque $I_{j,k}$ ne contient pas de point x_i de discontinuité, on a par le théorème des accroissements finis

$$|v(x) - v(2^{-j}k)| \leq M_1 2^{-j}, \quad x \in I_{j,k},$$

et on en déduit l'estimation

$$|d_{j,k}| \leq M_1 2^{-j} \|\psi_{j,k}\|_{L^1} = M_1 2^{-3j/2}.$$

On voit ainsi que les coefficients des ondelettes dont les supports ne contiennent pas de points x_i ne sont pas nuls mais décroissent beaucoup plus vite avec le niveau d'échelle j que ceux des ondelettes dont les supports contiennent l'un de ces points (pour lequels on a seulement la première estimation en $O(2^{-j/2})$). On pourra ainsi *comprimer* significativement la représentation de la fonction v en oubliant les coefficients qui sont en dessous d'un seuil que l'on se fixe. L'algorithme JPEG 2000, qui constitue l'état de l'art actuel pour le codage efficace des images, tire ses fondements de ce principe, mais utilise des ondelettes plus sophistiquées que celles du système de Haar, ainsi que des techniques de quantification permettant de coder les valeurs des coefficients retenus avec un nombre fini de bits 0 ou 1.

Le système de Haar présente le désavantage d'être constitué de fonctions discontinues et qui par conséquent n'appartiennent pas à l'espace $H_0^1(\Omega)$. On ne peut donc pas l'utiliser pour l'approximation variationelle du problème aux limites. Nous présentons à présent un deuxième exemple de base d'ondelette mieux adapté à cette tâche. On l'obtient tout simplement en considérant les primitives des fonctions $\psi_{j,k}$ du système de Haar, c'est à dire

$$w_{j,k}(x) = \int_0^x \psi_{j,k}(t)dt.$$

Les fonctions $w_{j,k}$ ainsi obtenues appartiennent à H_0^1 puisque $\int_0^1 \psi_{j,k} = 0$ Elles peuvent se déduire de la fonction chapeau centrée en $1/2$

$$w(x) = 1/2 - |x - 1/2|,$$

dont la dérivée est $w' = \psi$. On a en effet par changement d'échelle

$$w_{j,k}(x) = 2^{-j/2} w(2^j x - k).$$

La structure multiéchelle de cette base, appellée *base hiérarchique* ou *base de Schauder* ressemble à celle de la base de Haar, mais l'orthonormalité est à présent vérifiée pour la norme H_0^1 puisqu'on a

$$\int_\Omega w'_{j,k} w'_{i,l} = \int_\Omega h_{j,k} h_{i,l} = \delta_{j,i}\delta_{k,l}.$$

Notons V_n l'espace engendré par les fonctions

$$\{w_{j,k} : k = 0 =, \ldots, 2^{j-1},\ j = 0, \ldots, n-1\}.$$

Il s'agit exactement de l'espace des l'espace d'éléments fini P_1 pour la partition

$$0 = x_0^n < x_1^n < \cdots < x_{2^n}^n = 1, \quad \text{avec} \quad x_k^n := k2^{-n},$$

c'est à dire $V_n = V_{h_n}$ avec $h_n = 2^{-n}$. En effet, toutes les fonctions $w_{j,k}$ appartiennent à V_{h_n}, elles constituent une base (orthonormée dans H_0^1), et leur nombre est exactement

$$1 + 2 + \cdots + 2^{n-1} = 2^n - 1 = \dim(V_{h_n}).$$

En ce sens, l'espace V_n n'est pas nouveau, il a été étudié en détail dans le chapitre 6. La base hiérarchique est simplement un choix différent de celui de la base nodale classique pour représenter les fonctions de cet espace. Ce nouveau choix présente plusieurs avantages.

Considérons l'approximation variationnelle dans V_n du problème aux limites

$$-u''(x) + c(x)u(x) = f(x), \quad x \in \Omega =]0,1[,$$

avec les conditions aux limites homogène de Dirichlet $u(-1) = u(1) = 0$. Pour un élément $v_n = \sum_{j=0}^{n-1} d_{j,k} w_{j,k}$, on note $\overline{d} = (d_{j,k})$ le vecteur des coordonnées (après avoir choisi un ordre d'énumération de la base hiérarchique, par exemple l'ordre lexicographique croissant de (j,k)). On a ainsi

$$\|v_n\|_{H_0^1}^2 = \|\overline{d}\|_{\ell^2}^2.$$

D'autre part, la matrice de rigidité A_n vérifie

$$(A_n \overline{d}, \overline{d})_{\ell^2} = a(v_n, v_n),$$

Or on sait que

$$\alpha \|v\|_{H_0^1}^2 \leq a(v,v) \leq C_a \|v\|_{H_0^1}^2, \quad v \in H_0^1(\Omega).$$

En raisonnant comme pour les bases de Fourier et les bases polynomiales, on obtient ainsi $\lambda_{\min}(A_n) \geq \alpha$ et $\lambda_{\max}(A_n) \leq C_a$, et la borne

$$\kappa(A_n) \leq \frac{C_a}{\alpha},$$

indépendante de n, par contraste avec l'utilisation de la base nodale pour laquelle on a vu que le nombre de conditionnement se comporte en $O(h_n^{-2}) = O(2^2 n)$. L'utilisation de la base hiérarchique peut ainsi être vu comme une opération de *précondionnement*, qui s'incrit dans un ensemble de techniques puissantes appelées *méthodes multigrilles*.

D'autre part, comme pour la base de Haar, on peut décider de réduite la dimension de l'espace V_n en ommettant les fonctions de bases dont les supports se situent dans les régions où l'on sait que la solution va peu varier ou être très régulière. Si par exemple on ne retiens que les fonctions de bases dont les supports contiennent certains points $x_1, \ldots, x_m$ autours desquels la fonctions est peu régulière ou varie beaucoup, l'espace obtenu sera de petite dimension nm par rapport à $\dim(V_n) = 2^n$. On pourra vérifier que ceci correspond à remplacer le maillage uniforme de l'espace V_n par un maillage adaptatif qui se concentre au voisinage de ces points de singularité, ce qui est intuitivement raisonnable si l'on souhaite allouer plus efficacement les degrés de libertés qu'avec un maillage uniforme.