Approximation numérique et optimisation. Une introduction à la modélisation mathématique et à la simulation numérique

Grégoire Allaire

Approximation numérique et optimisation

Une introduction à la modélisation mathématique et à la simulation numérique

Grégoire ALLAIRE

École Polytechnique
MAP 411
28 juin 2019

Table des matières

1 INTRODUCTION A LA MODÉLISATION MATHÉMATIQUE ET A LA SIMULATION NUMÉRIQUE 1

1.1 Introduction générale 1
1.2 Un exemple de modélisation 2
1.3 Quelques modèles classiques 7
1.3.1 Équation de la chaleur 7
1.3.2 Équation des ondes 8
1.3.3 Le Laplacien 10
1.3.4 Équation de Schrödinger 10
1.3.5 Système de Lamé 11
1.3.6 Système de Stokes 11
1.3.7 Équations des plaques 12
1.4 Calcul numérique par différences finies 12
1.4.1 Principes de la méthode 12
1.4.2 Résultats numériques pour l'équation de la chaleur 15
1.4.3 Résultats numériques pour l'équation d'advection 19
1.5 Notion de problème bien posé 23

2 MÉTHODE DES DIFFÉRENCES FINIES 25

2.1 Introduction 25
2.2 Différences finies pour l'équation de la chaleur 25
2.2.1 Divers exemples de schémas 25
2.2.2 Consistance et précision 28
2.2.3 Stabilité et analyse de Fourier 29
2.2.4 Convergence des schémas 34
2.2.5 Schémas multiniveaux 36
2.2.6 Le cas multidimensionnel 38
2.3 Autres modèles 41
2.3.1 Équation d'advection 41
2.3.2 Équation des ondes 48

3 FORMULATION VARIATIONNELLE DES PROBLÈMES AUX LIMITES 51

3.1 Approche variationnelle 51
3.1.1 Formules de Green 52

3.1.2 Formulation variationnelle . . . 53
3.1.3 Approximation variationnelle . . . 58
3.2 Éléments finis en dimension $N = 1$ . . . 60
3.2.1 Éléments finis $\mathbb{P}_1$ . . . 61
3.2.2 Convergence et estimation d'erreur . . . 65
3.2.3 Éléments finis $\mathbb{P}_2$ . . . 67
3.3 Problèmes d'évolution . . . 69
3.3.1 Modéles . . . 69
3.3.2 Formulation variationnelle . . . 71
3.3.3 Semi-discrétisation par éléments finis en espace . . . 73
3.3.4 Discrétisation totale en espace-temps . . . 75
3.4 Problèmes aux valeurs propres . . . 78
3.4.1 Modèle et motivation . . . 78
3.4.2 Formulation variationnelle et discrétisation par éléments finis . 81
3.5 Résolution des systèmes linéaires . . . 82
3.5.1 Rappels sur les normes matricielles . . . 83
3.5.2 Conditionnement et stabilité . . . 85
3.5.3 Méthodes directes . . . 86
3.5.4 Méthodes itératives . . . 90
3.5.5 Méthode du gradient conjugué . . . 92
3.5.6 Calcul de valeurs et vecteurs propres . . . 94

4 OPTIMISATION 97
4.1 Motivation et généralités . . . 97
4.1.1 Introduction et exemples . . . 97
4.1.2 Définitions et notations . . . 101
4.1.3 Existence de minima en dimension finie . . . 102
4.1.4 Analyse convexe . . . 104
4.2 Conditions d'optimalité . . . 105
4.2.1 Différentiabilité . . . 106
4.2.2 Inéquations d'Euler et contraintes convexes . . . 109
4.2.3 Multiplicateurs de Lagrange . . . 112
4.3 Point-selle, théorème de Kuhn et Tucker, dualité . . . 121
4.3.1 Point-selle . . . 122
4.3.2 Théorème de Kuhn et Tucker . . . 123
4.3.3 Dualité . . . 125
4.3.4 Vers la programmation linéaire . . . 129
4.4 Algorithmes numériques . . . 136
4.4.1 Introduction . . . 136
4.4.2 Algorithmes de type gradient (cas sans contraintes) . . . 137
4.4.3 Algorithmes de type gradient (cas avec contraintes) . . . 140
4.4.4 Méthode de Newton . . . 147
4.4.5 Méthodes d'approximations successives . . . 149

If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.
John Von Neumann (1903-1957)

Truth is much too complicated to allow anything but approximations.
John Von Neumann (1903-1957)

Introduction

Ce cours est consacré à deux aspects essentiels des mathématiques appliquées : l'approximation numérique et l'optimisation. Avant même de présenter ces deux disciplines, disons tout de suite qu'à travers leur enseignement l'objectif de ce cours est d'introduire le lecteur au monde de la **modélisation mathématique** et de la **simulation numérique** qui ont pris une importance considérable ces dernières décennies dans tous les domaines de la science et des applications industrielles (ou sciences de l'ingénieur). La modélisation mathématique permet de représenter une réalité physique à travers des modèles abstraits, de complexité et de fidélité variable, accessibles à l'analyse et au calcul. La simulation numérique est, bien sûr, le processus qui permet de calculer sur ordinateur les solutions de ces modèles, et donc de simuler la réalité physique.

Plus que pour tout autre discipline l'ordinateur a été une révolution pour les mathématiques : il en a fait une science expérimentale ! On fait des "expériences numériques" comme d'autres font des expériences physiques, et la conception ainsi que l'analyse des méthodes de calcul sur ordinateur sont devenues une nouvelle branche des mathématiques : c'est l'analyse numérique. Par ailleurs, l'optimisation est la branche des mathématiques qui s'intéresse, d'un point de vue théorique et algorithmique, à la minimisation ou à la maximisation de fonctions, représentant par exemple la réponse d'un système dont on cherche à améliorer le fonctionnement. L'optimisation a aussi grandement bénéficié de l'accroissement de la capacité des ordinateurs. Ces progrès ont ainsi permis aux mathématiques de s'attaquer à des problèmes beaucoup plus complexes et concrets, issus de motivations immédiates industrielles, sociétales ou scientifiques, auxquels on peut apporter des réponses à la fois qualitatives mais aussi quantitatives : c'est la modélisation mathématique.

L'approximation numérique et l'optimisation sont deux disciplines, apparemment distinctes mais extrêmement liées en pratique, qui toutes deux concourent à faire de la modélisation mathématique un outil indispensable dans la résolution des grands problèmes de notre époque. Qu'il s'agisse de la conception d'un nouvel avion, d'un réseau de télécommunications, de la production d'énergie, de l'étude du climat ou de la météorologie (pour ne citer que ces exemples parmi tant d'autres), les mathématiques sont concrètement incontournables.

Dans ce cours nous suivons le parti pris de ne considérer que des modèles déterministes et distribués en espace, c'est-à-dire sans aléa et dépendant d'une variable de position dans l'espace physique. Cela nous conduit naturellement à considérer des modèles qui sont des équations aux dérivées partielles, comme c'est souvent le cas en physique ou en mécanique, par exemple.

Les **objectifs de ce cours** sont d'introduire quelques principes d'approximation numérique pour la résolution des équations aux dérivées partielles (linéaires, pour simplifier), et de donner les principes et algorithmes fondamentaux en optimisation. Au delà, l'ambition de ce cours est de familiariser le lecteur avec quelques modèles universels en science et ingénierie et de donner les bases qui permettront aux futurs ingénieurs de bureau d'études ou de recherche et développement de créer

de **nouveaux modèles** et de **nouveaux algorithmes numériques** pour des problèmes plus compliqués non discutés ici. Cependant, même ceux qui ne se destinent pas à une telle carrière ont intérêt à bien comprendre les enjeux de la simulation numérique et de l'optimisation. En effet, de nombreuses décisions industrielles ou politiques se prennent désormais sur la foi de simulations et d'optimisations numériques. Il importe donc que les décideurs aient la capacité de juger de la **qualité** et de la **fiabilité** des calculs qui leur sont présentés. Ce cours leur permettra de connaître les premiers critères qui garantissent la validité et la pertinence des simulations numériques et des algorithmes d'optimisation. Par contre, ce cours n'a pas pour objectif d'établir une théorie d'existence, d'unicité et de propriétés qualitatives des solutions des modèles considérés (ou des problèmes d'optimisation autres qu'en dimension finie). Pour cela nous renvoyons à d'autres cours, plus avancés.

Le plan de ce cours est divisé en quatre parties. Le Chapitre 1 est une introduction à la **modélisation mathématique**. Il présente les principaux modèles "classiques" et explique à travers quelques exemples pourquoi leur résolution numérique fait appel aux mathématiques. Le Chapitre 2 est consacré à l'étude de la méthode numérique des **différences finies**. Le Chapitre 3 est consacré à la résolution numérique par **l'approche variationnelle** de problèmes aux limites. On y présente, en une dimension d'espace, la méthode, dite des **éléments finis**. Etendue à plusieurs dimensions d'espace, la méthode des éléments finis est à la base de nombreux logiciels de calculs industriels ou académiques. Le Chapitre 4 est dédié à **l'optimisation**. Après avoir présenté quelques exemples concrets de problèmes d'optimisation et discuté des généralités, on y étudie les conditions (nécessaires ou suffisantes) d'optimalité des solutions. Ces conditions sont importantes tant du point de vue théorique que numérique. Elles permettent de caractériser les optima et elles sont à la base des algorithmes numériques que nous décrivons.

Les schémas numériques en différences finies, ainsi que la méthode des éléments finis en dimension un, ont été programmés dans le langage du logiciel Scilab développé par INRIA et l'ENPC, disponible gratuitement sur le site web
`http://www.scilab.org`
Ces programmes informatiques, ainsi que d'autres informations sur le cours, sont disponibles sur le site web
`http://www.cmap.polytechnique.fr/~allaire/cours_map411.html`

Ce cours est d'un niveau introductif et n'exige aucun autre prérequis que le niveau de connaissances acquis en classes préparatoires ou en premier cycle universitaire. Il s'inspire fortement d'un cours précédent [1] qui, lui-même, faisait de larges emprunts à ces prédécesseurs (cours de B. Larrouturou, P.-L. Lions, P.-A. Raviart). L'auteur remercie à l'avance tous ceux qui voudront bien lui signaler d'éventuelles erreurs ou imperfections de cette édition, par exemple par courrier électronique à l'adresse `gregoire.allaire@polytechnique.fr` .

G. Allaire
Paris, le 19 Juillet 2014

Chapitre 1

INTRODUCTION A LA MODÉLISATION MATHÉMATIQUE ET A LA SIMULATION NUMÉRIQUE

1.1 Introduction générale

Ce chapitre est une introduction à deux aspects distincts, mais très liés, des mathématiques appliquées : la **modélisation mathématique** et la **simulation numérique**. Un modèle mathématique est une représentation ou une interprétation abstraite de la réalité physique qui est accessible à l'analyse et au calcul. La simulation numérique permet de calculer sur ordinateur les solutions de ces modèles, et donc de simuler la réalité physique. Dans ce cours, les modèles que nous étudierons seront des équations aux dérivées partielles (ou e.d.p. en abrégé), c'est-à-dire des équations différentielles à plusieurs variables (le temps et l'espace, par exemple).

Nous laissons de coté un troisième aspect fondamental des mathématiques appliquées, à savoir l'analyse mathématique des modèles, qui n'est pas dans les objectifs de ce cours et nous renvoyons le lecteur à d'autres ouvrages comme [1], [3], [13]. Néanmoins, il faut bien comprendre que l'analyse mathématique et la simulation numérique sont fortement intriquées et qu'on ne peut faire l'économie de l'une en étudiant l'autre. Nous allons voir, en effet, que le calcul numérique des solutions de ces modèles physiques réserve parfois des surprises (désagréables) qui ne peuvent s'expliquer et s'éviter que par une bonne compréhension de leurs propriétés mathématiques. Rappelons encore une fois le caractère fondamentalement multidisciplinaire des mathématiques appliquées, et donc de la simulation numérique, qui mêlent mathématiques, calcul informatique, et sciences de l'ingénieur.

Bien que la plupart des problèmes et des applications qui motivent les mathématiques appliquées sont souvent **non-linéaires**, nous nous restreignons dans cet ouvrage aux problèmes linéaires par souci de simplicité. De la même façon, nous n'envisageons que des problèmes déterministes, c'est-à-dire sans introduction d'aléatoire ou de stochastique. Enfin, ce chapitre se voulant introductif et attractif, nous res-

terons souvent un peu flou dans l'argumentaire mathématique pour ne pas alourdir inutilement l'exposé.

Le plan de ce chapitre est le suivant. La Section 1.2 est consacrée à un exemple élémentaire de modélisation qui conduit à **l'équation de la chaleur**. La Section 1.3 est une revue rapide des principales équations aux dérivées partielles que l'on rencontre dans les modèles usuels en mécanique, physique, ou sciences de l'ingénieur. La Section 1.4 est une introduction assez informelle au calcul numérique et à la méthode des **différences finies**. Enfin, nous donnons dans la Section 1.5 la définition d'un **problème bien posé** ainsi qu'une classification (sommaire) des équations aux dérivées partielles.

1.2 Un exemple de modélisation

La modélisation représente une part considérable du travail du mathématicien appliqué et nécessite une connaissance approfondie, non seulement des mathématiques appliquées, mais aussi de la discipline scientifique à laquelle elles s'appliquent. En effet, dans de nombreux cas le modèle mathématique n'est pas encore établi, ou bien il faut en sélectionner un pertinent parmi plusieurs disponibles, ou encore il faut simplifier des modèles connus mais trop complexes. Néanmoins, il ne nous est pas possible dans une première présentation de la discipline de rendre compte avec justice de l'importance de cette démarche de modélisation : il faut bien commencer par apprendre les notions de base propres aux mathématiques appliquées ! C'est pourquoi nous nous limitons à décrire un exemple de dérivation d'un modèle physique très classique, et nous renvoyons le lecteur désireux d'en savoir plus à des ouvrages ou cours plus spécialisés.

Le modèle que nous allons décrire est connu sous le nom **d'équation de la chaleur**, ou d'équation de diffusion.

Considérons un domaine Ω de l'espace à N dimensions (noté $\mathbb{R}^N$, avec en général $N = 1, 2$, ou 3) que l'on suppose occupé par un matériau homogène, isotrope, et conducteur de la chaleur. On note x la variable d'espace, c'est-à-dire un point de Ω, et t la variable de temps. Dans Ω les sources de chaleur (éventuellement non uniformes en espace et variables dans le temps) sont représentées par une fonction donnée $f(x,t)$, tandis que la température est une fonction inconnue $\theta(x,t)$. La quantité de chaleur est proportionnelle à la température θ et vaut $c\theta$ où c est une constante physique (qui dépend du type de matériau) appelée chaleur spécifique. Pour déterminer la température θ, nous écrivons la **loi de conservation de l'énergie** ou de la quantité de chaleur. Dans un volume élémentaire V inclus dans Ω, la variation en temps de la quantité de chaleur est le bilan de ce qui est produit par les sources et de ce qui sort ou rentre à travers les parois. Autrement dit,

$$\frac{d}{dt}\left(\int_V c\theta \, dx\right) = \int_V f \, dx - \int_{\partial V} q \cdot n \, ds \tag{1.1}$$

où ∂V est le bord de V (d'élément de surface ds), n est la normale extérieure unité de V, et q est le vecteur flux de chaleur. Si on applique le théorème de Gauss, on

obtient

$$\int_{\partial V} q \cdot n \, ds = \int_V \text{div} q \, dx.$$

Regroupant les différents termes de (1.1) et utilisant le fait que le volume élémentaire V est quelconque, indépendant du temps, on en déduit l'équation de conservation de l'énergie

$$c\frac{\partial \theta}{\partial t} + \text{div} q = f \tag{1.2}$$

qui a lieu en tout point $x \in \Omega$ et à tout temps t. Rappelons que l'opérateur divergence est défini par

$$\text{div} q = \sum_{i=1}^{N} \frac{\partial q_i}{\partial x_i} \text{ avec } q = (q_1, ..., q_N)^t.$$

Il faut maintenant relier le flux de chaleur à la température, et on fait appel à ce qu'on appelle une **loi constitutive**. Dans le cas présent, il s'agit de la loi de Fourier qui relie le flux de chaleur de manière proportionnelle au gradient de température

$$q = -k\nabla\theta, \tag{1.3}$$

où k est une constante positive (qui dépend du type de matériau) appelée conductivité thermique. Rappelons que l'opérateur gradient est défini par

$$\nabla\theta = \left(\frac{\partial \theta}{\partial x_1}, ..., \frac{\partial \theta}{\partial x_N} \right)^t.$$

En combinant la loi de conservation (1.2) et la loi constitutive (1.3), on obtient une équation pour la température θ

$$c\frac{\partial \theta}{\partial t} - k\Delta\theta = f,$$

où $\Delta = \text{div}\nabla$ est l'opérateur Laplacien donné par

$$\Delta\theta = \sum_{i=1}^{N} \frac{\partial^2 \theta}{\partial x_i^2}.$$

Il faut ajouter à cette équation qui est valable dans tout le domaine Ω, une relation, appelée **condition aux limites**, qui indique ce qui se passe à la frontière ou au bord $\partial\Omega$ du domaine, et une autre relation qui indique quel est l'état initial de la température. Par convention, on choisit l'instant $t = 0$ pour être le temps initial, et on impose une **condition initiale**

$$\theta(t = 0, x) = \theta_0(x), \tag{1.4}$$

où θ_0 est la fonction de distribution initiale de température dans le domaine Ω. En ce qui concerne la condition aux limites, cela dépend du contexte physique. Si le domaine est supposé baigner dans un thermostat à température constante, alors,

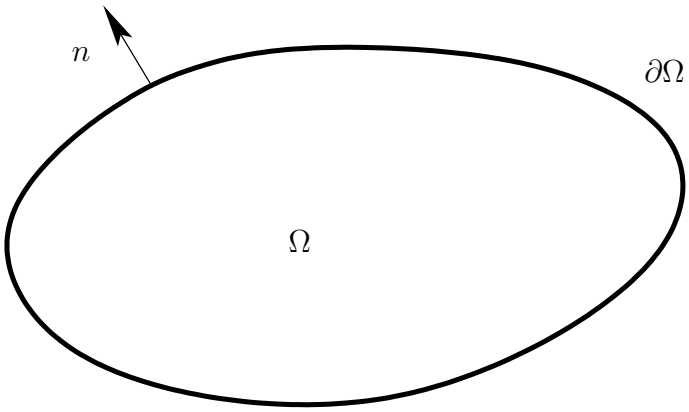

FIGURE 1.1 – Vecteur normal unité orienté vers l'extérieur.

quitte à modifier l'échelle des températures, la température vérifie la condition aux limites de Dirichlet

$$\theta(t,x) = 0 \text{ pour tout } x \in \partial\Omega \text{ et } t > 0. \tag{1.5}$$

Si le domaine est supposé adiabatique ou thermiquement isolé de l'extérieur, alors le flux de chaleur sortant au bord est nul et la température vérifie la condition aux limites de Neumann

$$\frac{\partial\theta}{\partial n}(t,x) \equiv n(x)\cdot\nabla\theta(t,x) = 0 \text{ pour tout } x \in \partial\Omega \text{ et } t > 0, \tag{1.6}$$

où n est la normale extérieure unité de Ω (voir la Figure 1.1). Une situation intermédiaire peut aussi avoir lieu : le flux de chaleur sortant au bord est proportionnel au saut de température entre l'extérieur et l'intérieur, et la température vérifie la condition aux limites de Fourier

$$\frac{\partial\theta}{\partial n}(t,x) + \alpha\theta(t,x) = 0 \text{ pour tout } x \in \partial\Omega, \text{ et } t > 0 \tag{1.7}$$

où α est une constante positive. Puisqu'il faut choisir (c'est une des étapes de la modélisation), nous allons sélectionner la condition aux limites de Dirichlet (1.5). Rassemblant enfin l'équation, la condition initiale, et la condition aux limites satisfaites par la température, on obtient l'équation de la chaleur

$$\begin{cases} c\dfrac{\partial\theta}{\partial t} - k\Delta\theta = f & \text{pour } (x,t) \in \Omega \times \mathbb{R}_*^+ \\ \theta(t,x) = 0 & \text{pour } (x,t) \in \partial\Omega \times \mathbb{R}_*^+ \\ \theta(t=0,x) = \theta_0(x) & \text{pour } x \in \Omega \end{cases} \tag{1.8}$$

Le problème (1.8) est donc constitué d'une équation aux dérivées partielles munie de conditions aux limites et d'une condition initiale. A cause de la présence de conditions aux limites, on dit que (1.8) est un **problème aux limites**, mais on dit aussi que c'est un **problème de Cauchy** à cause de la donnée initiale en temps.

Remarque 1.2.1 Le problème (1.8) n'est pas seulement un modèle de propagation de la chaleur. Il a en fait un caractère universel, et on le retrouve comme modèle de nombreux phénomènes sans aucun rapport entre eux (il faut simplement changer le nom des diverses variables du problème). Par exemple, (1.8) est aussi connue sous le nom **d'équation de diffusion**, et modélise la diffusion ou migration d'une concentration ou densité à travers le domaine Ω (imaginer un polluant diffusant dans l'atmosphère, ou bien une espèce chimique migrant dans un substrat). Dans ce cas, θ est la concentration ou la densité en question, q est le flux de masse, k est la diffusivité, et c est la densité volumique de l'espèce. De même, la loi de conservation (1.2) est un bilan de masse, tandis que la loi constitutive (1.3) est appelée loi de Fick. •

Il existe de nombreuses variantes de l'équation de la chaleur (1.8) dont nous explorons certaines maintenant. Jusqu'ici nous avons supposé que la chaleur se propageait dans un milieu immobile ou au repos. Supposons à présent qu'elle se propage dans un milieu en mouvement comme, par exemple, un fluide animé d'une vitesse $V(x,t)$ (une fonction à valeurs vectorielles dans $\mathbb{R}^N$). Alors, il faut changer la loi constitutive car le flux de chaleur est la somme d'un flux de diffusion (comme précédemment) et d'un flux de convection (proportionnel à la vitesse V), et des considérations similaires à celles qui précèdent nous conduisent à un problème, dit de **convection-diffusion**

$$\begin{cases} c\dfrac{\partial\theta}{\partial t} + cV\cdot\nabla\theta - k\Delta\theta = f & \text{dans } \Omega\times\mathbb{R}_*^+ \\ \theta = 0 & \text{sur } \partial\Omega\times\mathbb{R}_*^+ \\ \theta(t=0,x) = \theta_0(x) & \text{dans } \Omega \end{cases} \tag{1.9}$$

La différence entre (1.8) et (1.9) est l'apparition d'un terme de convection. On mesure la balance entre ce nouveau terme de convection et le terme de diffusion par un nombre sans dimension, appelé **nombre de Péclet**, défini par

$$\mathrm{Pe} = \frac{cVL}{k}, \tag{1.10}$$

où L est une longueur caractéristique du problème (par exemple le diamètre du domaine Ω). Si le nombre de Péclet est très petit, alors les effets diffusifs dominent les effets convectifs, et le modèle (1.8) est suffisant pour décrire le phénomène. Si le nombre de Péclet n'est ni petit, ni grand (on dit qu'il est de l'ordre de l'unité), le modèle (1.9) est plus réaliste que (1.8). Par contre, si le nombre de Péclet est très grand, on peut simplifier (1.9) en supprimant le terme de diffusion. On obtient alors l'équation dite **d'advection**

$$\begin{cases} c\dfrac{\partial\theta}{\partial t} + cV\cdot\nabla\theta = f & \text{dans } \Omega\times\mathbb{R}_*^+ \\ \theta(t,x) = 0 & \text{pour } (x,t)\in\partial\Omega\times\mathbb{R}_*^+ \text{ si } V(x)\cdot n(x) < 0 \\ \theta(t=0,x) = \theta_0(x) & \text{dans } \Omega \end{cases} \tag{1.11}$$

Remarquons la différence dans la condition aux limites de (1.11) par rapport à celle de (1.9) : on n'impose plus à la température θ d'être nulle partout sur le bord $\partial\Omega$ mais seulement en ces points du bord où la vitesse V est rentrante.

Nous venons donc de décrire trois modèles de propagation de la chaleur par convection et diffusion, (1.8), (1.9), (1.11), qui ont des régimes de validité correspondant à des valeurs différentes du nombre de Péclet. Bien sûr, la résolution analytique ou numérique de ces trois modèles est assez différente. Il s'agit là d'une situation courante en modélisation mathématique : plusieurs modèles sont en concurrence et il faut pouvoir choisir le "meilleur".

Afin de mieux comprendre les différences fondamentales qui existent entre ces modèles, nous nous restreignons provisoirement au cas où $\Omega = \mathbb{R}$ est l'espace tout entier en dimension 1 (ce qui évacue la question des conditions aux limites), où le terme source f est nul, et où la vitesse V est constante. On peut alors calculer explicitement des solutions de ces modèles. Par exemple, (1.9) devient

$$\left\{ \begin{array}{ll} \dfrac{\partial \theta}{\partial t} + V \dfrac{\partial \theta}{\partial x} - \nu \dfrac{\partial^2 \theta}{\partial x^2} = 0 & \text{pour } (x,t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t=0,x) = \theta_0(x) & \text{pour } x \in \mathbb{R} \end{array} \right. \tag{1.12}$$

avec $\nu = k/c$, qui admet comme solution

$$\theta(t,x) = \frac{1}{\sqrt{4\pi\nu t}} \int_{-\infty}^{+\infty} \theta_0(y) \exp\left(-\frac{(x - Vt - y)^2}{4\nu t} \right) dy. \tag{1.13}$$

Une solution de (1.8) est facilement obtenue en faisant $V = 0$ dans l'expression (1.13).

Exercice 1.2.1 On suppose que la donnée initiale θ_0 est continue et uniformément bornée sur $\mathbb{R}$. Vérifier que (1.13) est bien une solution de (1.12).

Avec les mêmes hypothèses simplificatrices, l'équation d'advection devient

$$\left\{ \begin{array}{ll} \dfrac{\partial \theta}{\partial t} + V \dfrac{\partial \theta}{\partial x} = 0 & \text{pour } (x,t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t=0,x) = \theta_0(x) & \text{pour } x \in \mathbb{R} \end{array} \right. \tag{1.14}$$

On vérifie que

$$\theta(t,x) = \theta_0(x - Vt) \tag{1.15}$$

est une solution de l'équation (1.14).

Exercice 1.2.2 On suppose que la donnée initiale θ_0 est dérivable et uniformément bornée sur $\mathbb{R}$. Vérifier que (1.15) est bien une solution de (1.14). Montrer que (1.15) est la limite de (1.13) lorsque le paramètre ν tend vers zéro.

Remarque 1.2.2 Le rôle du temps est fondamentalement différent dans les équations (1.8) et (1.11). En effet, supposant que le terme source est nul, $f = 0$, si on change le signe du temps t et celui de la vitesse, l'équation d'advection (1.11) est inchangée (quand on remonte le temps, on remonte le courant). Au contraire, un changement de signe du temps dans l'équation de la chaleur (1.8) ne peut pas être "compensé" par une quelconque variation du signe des données. C'est manifeste dans

la forme des solutions explicites de ces équations : (1.15) est invariant par changement de signe de t et V, alors que (1.13) (avec $V = 0$) décroît en temps ce qui indique la "flèche" du temps. On dit que l'équation d'advection est **réversible** en temps, tandis que l'équation de la chaleur est **irréversible** en temps. Cette observation mathématique est conforme à l'intuition physique : certains phénomènes sont réversibles en temps, d'autres non (comme la diffusion d'une goutte de lait dans une tasse de thé). •

Remarque 1.2.3 Une propriété surprenante (du point de vue de la physique) de l'équation de la chaleur (1.8) est que la solution en (x, t) dépend de toutes les valeurs de la donnée initiale dans $\mathbb{R}$ (voir la formule (1.13)). En particulier, dans le cas de (1.12), si la donnée initiale est positive à support compact, alors pour tout temps $t > 0$ (aussi petit soit-il) la solution est strictement positive sur tout $\mathbb{R}$: autrement dit, l'effet de la chaleur se faire sentir "instantanément" à l'infini. On dit que la chaleur se **propage avec une vitesse infinie** (ce qui est bien sûr une limitation du modèle). Au contraire, dans l'équation d'advection (1.14) la donnée initiale est convectée à la vitesse V (voir la formule (1.15)) : il y a donc **propagation à vitesse finie**. •

Remarque 1.2.4 Grâce aux formules explicites (1.13) et (1.15), on vérifie aisément que les solutions de l'équation de convection-diffusion (1.12) et de l'équation d'advection (1.14) vérifient la propriété

$$\min_{x\in\mathbb{R}} \theta_0(x) \leq \theta(x,t) \leq \max_{x\in\mathbb{R}} \theta_0(x) \text{ pour tout } (x,t) \in \mathbb{R}\times\mathbb{R}^+,$$

appelée **principe du maximum**. Cette propriété (très importante, aussi bien du point de vue mathématique que physique) se généralise aux formes plus générales de l'équation de convection-diffusion (1.9) et de l'équation d'advection (1.11). •

1.3 Quelques modèles classiques

Nous faisons une revue de quelques modèles classiques d'équations aux dérivées partielles. Pour plus de détails sur ces modèles (ou d'autres encore !) nous renvoyons à l'encyclopédie [10].

1.3.1 Équation de la chaleur

Comme nous venons de le voir, l'équation de la chaleur intervient comme modèle dans de nombreux problèmes des sciences de l'ingénieur. Elle s'écrit

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t=0) = u_0 & \text{dans } \Omega \end{cases} \tag{1.16}$$

Il s'agit d'une équation d'ordre 1 en temps et d'ordre 2 en espace (l'ordre est celui des dérivées partielles les plus élevées). On dira que cette équation est parabolique.

Nous avons déjà vu certaines propriétés de cette équation : irréversibilité en temps, propagation à vitesse infinie, et principe du maximum.

Exercice 1.3.1 On se propose de retrouver une propriété de décroissance exponentielle en temps (voir la formule (1.13)) de la solution de l'équation de la chaleur (1.16) dans un domaine borné. En une dimension d'espace, on pose $\Omega = (0,1)$ et on suppose que $f = 0$. Soit $u(t,x)$ une solution régulière de (1.16). En multipliant l'équation par u et en intégrant par rapport à x, établir l'égalité

$$\frac{1}{2}\frac{d}{dt}\left(\int_0^1 u^2(t,x)\,dx\right) = -\int_0^1 \left|\frac{\partial u}{\partial x}(t,x)\right|^2 dx$$

Montrer que toute fonction $v(x)$ continûment dérivable sur $[0,1]$, telle que $v(0) = 0$, vérifie l'inégalité de Poincaré

$$\int_0^1 v^2(x)\,dx \leq \int_0^1 \left|\frac{dv}{dx}(x)\right|^2 dx.$$

En déduire la décroissance exponentielle en temps de $\int_0^1 u^2(t,x)\,dx$.

1.3.2 Équation des ondes

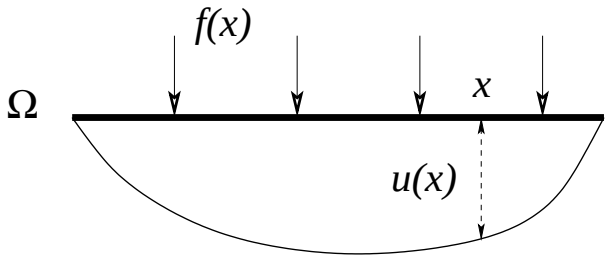

FIGURE 1.2 – Déplacement d'une corde élastique.

L'équation des ondes modélise des phénomènes de propagation d'ondes ou de vibration. Par exemple, en deux dimensions d'espace elle est un modèle pour étudier les vibrations d'une membrane élastique tendue (comme la peau d'un tambour). En une dimension d'espace, elle est aussi appelée équation des cordes vibrantes. Au repos, la membrane occupe un domaine plan Ω. Sous l'action d'une force normale à ce plan d'intensité f, elle se déforme et son déplacement normal est noté u (voir la Figure 1.2). On suppose qu'elle est fixée sur son bord, ce qui donne une condition aux limites de Dirichlet. L'équation des ondes dont u est solution est donnée par

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t=0) = u_0 & \text{dans } \Omega \\ \dfrac{\partial u}{\partial t}(t=0) = u_1 & \text{dans } \Omega \end{cases} \tag{1.17}$$

Remarquons qu'il s'agit d'une équation du deuxième ordre en temps et qu'il faut donc deux conditions initiales pour u.

Exercice 1.3.2 On se place en dimension $N = 1$ d'espace. On suppose que les données initiales u_0 et u_1 sont des fonctions régulières, et que $f = 0$ avec $\Omega = \mathbb{R}$. On note U_1 une primitive de u_1. Vérifier que

$$u(t,x) = \frac{1}{2}\left(u_0(x+t) + u_0(x-t)\right) + \frac{1}{2}\left(U_1(x+t) - U_1(x-t)\right), \tag{1.18}$$

est la solution unique de (1.17) dans la classe des fonctions régulières.

L'équation des ondes partage avec l'équation d'advection (1.11) la propriété importante de **propagation à vitesse finie**. En effet, l'Exercice 1.3.3 montre que sa solution en un point (x,t) ne dépend pas de toutes les valeurs des données initiales mais seulement des valeurs dans un intervalle restreint appelé **domaine de dépendance** (ou cône de lumière ; voir la Figure 1.3). Rappelons que cette propriété n'est pas partagée par l'équation de la chaleur puisqu'il est clair, à travers la formule (1.13), que la solution en (x,t) dépend de toutes les valeurs de la donnée initiale.

Une autre propriété de l'équation des ondes est son invariance par changement du sens du temps. Si on change t en $-t$, la forme de l'équation ne change pas. On peut donc "intégrer" l'équation des ondes vers les temps positifs ou négatifs de la même manière. On dit que l'équation des ondes est **réversible en temps**.

Exercice 1.3.3 Vérifier que la solution (1.18) au point (x,t) ne dépend des données initiales u_0 et u_1 qu'à travers leurs valeurs sur le segment $[x-t, x+t]$. Vérifier aussi que $u(-t,x)$ est solution de (1.17) dans $\Omega \times \mathbb{R}_*^-$ quitte à changer le signe de la vitesse initiale $u_1(x)$.

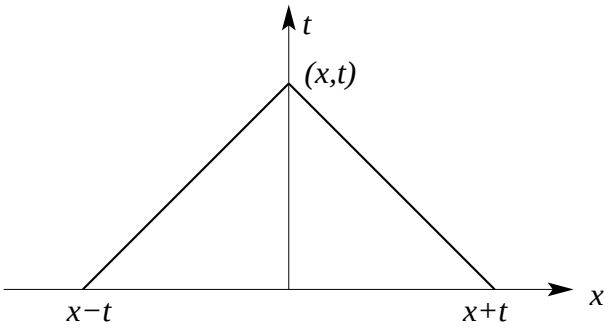

FIGURE 1.3 – Domaine ou cône de dépendance de l'équation des ondes.

Exercice 1.3.4 On se propose de démontrer un principe de conservation de l'énergie pour l'équation des ondes (1.17) sans utiliser la formule explicite (1.18). En une dimension d'espace, on pose $\Omega = (0,1)$ et on suppose $f = 0$. Soit $u(t,x)$ une solution régulière de (1.17). En multipliant l'équation par $\frac{\partial u}{\partial t}$ et en intégrant par rapport à x,

établir l'égalité d'énergie

$$\frac{d}{dt}\left(\int_0^1 \left|\frac{\partial u}{\partial t}(t,x)\right|^2 dx + \int_0^1 \left|\frac{\partial u}{\partial x}(t,x)\right|^2 dx\right) = 0.$$

Conclure et comparer à ce qui se passe pour l'équation de la chaleur.

1.3.3 Le Laplacien

Pour certains choix du terme source f, la solution de l'équation de la chaleur (1.16) atteint un état **stationnaire**, c'est-à-dire que $u(t,x)$ admet une limite $u_\infty(x)$ quand le temps t tend vers l'infini. Souvent, il est intéressant de calculer directement cet état stationnaire. Dans ce cas, pour un terme source $f(x)$ indépendant du temps, on résout une équation du deuxième ordre en espace

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{1.19}$$

que l'on appelle Laplacien ou équation de Laplace. Remarquons que le Laplacien est aussi la version stationnaire de l'équation des ondes (1.18). Le Laplacien intervient aussi dans de très nombreux domaines des sciences de l'ingénieur. Par exemple, (1.19) modélise le déplacement vertical d'une membrane élastique soumise à une force normale f et fixée sur son contour.

1.3.4 Équation de Schrödinger

L'équation de Schrödinger décrit l'évolution de la fonction d'onde u d'une particule soumise à un potentiel V. Rappelons que $u(t,x)$ est une fonction de $\mathbb{R}^+ \times \mathbb{R}^N$ à valeurs dans $\mathbb{C}$ et que son module au carré $|u|^2$ s'interprète comme la densité de probabilité pour détecter que la particule se trouve au point (t,x). Le potentiel $V(x)$ est une fonction à valeurs réelles. La fonction d'onde u est solution de

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \Delta u - Vu = 0 & \text{dans } \mathbb{R}^N \times \mathbb{R}_*^+ \\ u(t=0) = u_0 & \text{dans } \mathbb{R}^N \end{cases} \tag{1.20}$$

Il n'y a pas de condition aux limites apparentes dans (1.20) puisque l'équation a lieu dans tout l'espace (qui n'a pas de bord).

Exercice 1.3.5 On se propose de démontrer des principes de conservation de l'énergie pour l'équation de Schrödinger (1.20). Soit $u(t,x)$ une solution régulière de (1.20) en une dimension d'espace qui décroît vers zéro (ainsi que $\frac{\partial u}{\partial x}$) lorsque $|x| \to +\infty$. Montrer que pour toute fonction dérivable $v(t)$ on a

$$\mathcal{R}\left(\frac{\partial v}{\partial t}\overline{v}\right) = \frac{1}{2}\frac{\partial |v|^2}{\partial t},$$

où $\mathcal{R}$ désigne la partie réelle et $\overline{v}$ le complexe conjugué de v. En multipliant l'équation par $\overline{u}$ et en intégrant par rapport à x, établir l'égalité d'énergie

$$\int_{\mathbb{R}} |u(t,x)|^2 \, dx = \int_{\mathbb{R}} |u_0(x)|^2 \, dx.$$

En multipliant l'équation par $\frac{\partial \overline{u}}{\partial t}$, montrer que

$$\int_{\mathbb{R}} \left(\left| \frac{\partial u}{\partial x}(t,x) \right|^2 + V(x) \, |u(t,x)|^2 \right) dx = \int_{\mathbb{R}} \left(\left| \frac{\partial u_0}{\partial x}(x) \right|^2 + V(x) \, |u_0(x)|^2 \right) dx.$$

1.3.5 Système de Lamé

Le système de Lamé est un cas particulier des équations stationnaires de l'élasticité linéarisée qui modélisent les déformations d'un solide sous l'hypothèse de petites déformations et de petits déplacements. Pour obtenir le système de Lamé, on suppose que le solide est homogène isotrope et qu'il est fixé sur son bord. La principale différence avec les modèles précédents est qu'il s'agit ici d'un **système** d'équations, c'est-à-dire de plusieurs équations couplées entre elles. Le solide au repos occupe un domaine Ω de l'espace $\mathbb{R}^N$. Sous l'action d'une force f il se déforme, et chaque point x se déplace en $x + u(x)$. La force $f(x)$ est une fonction vectorielle de Ω dans $\mathbb{R}^N$, comme le déplacement $u(x)$. Ce dernier est solution de

$$\begin{cases} -\mu \Delta u - (\mu + \lambda) \nabla(\mathrm{div} u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.21}$$

où λ et μ sont deux constantes, dites de Lamé, caractéristiques du matériau homogène isotrope dont est constitué le solide. Pour des raisons mécaniques ces constantes vérifient $\mu > 0$ et $2\mu + N\lambda > 0$. La condition aux limites de Dirichlet pour u traduit le fait que le solide est supposé fixé et immobilisé sur son bord $\partial\Omega$.

Le système (1.21) a été écrit en notation vectorielle. Si on note f_i et u_i, pour $1 \leq i \leq N$, les composantes de f et u dans la base canonique de $\mathbb{R}^N$, (1.21) est équivalent à

$$\begin{cases} -\mu \Delta u_i - (\mu + \lambda) \frac{\partial(\mathrm{div} u)}{\partial x_i} = f_i & \text{dans } \Omega \\ u_i = 0 & \text{sur } \partial\Omega \end{cases}$$

pour $1 \leq i \leq N$. Remarquons que, si $(\mu + \lambda) \neq 0$, alors les équations pour chaque composante u_i sont couplées par le terme de divergence. Évidemment, en dimension $N = 1$, le système de Lamé n'a qu'une seule équation et se réduit au Laplacien.

1.3.6 Système de Stokes

Le système de Stokes modélise l'écoulement d'un fluide visqueux incompressible à petite vitesse. On suppose que le fluide occupe un domaine Ω et qu'il adhère à la paroi de celui-ci, c'est-à-dire que sa vitesse est nulle sur la paroi (ce qui conduit à une condition aux limites de Dirichlet). Sous l'action d'une force $f(x)$ (une fonction

de Ω dans $\mathbb{R}^N$), la vitesse $u(x)$ (un vecteur) et la pression $p(x)$ (un scalaire) sont solutions de

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{dans } \Omega \\ \text{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.22}$$

où $\mu > 0$ est la viscosité du fluide. Remarquons qu'en plus des N équations $\nabla p - \mu \Delta u = f$ (correspondant à la **conservation de la quantité de mouvement**), il y a une autre équation $\text{div} u = 0$ appelée **condition d'incompressibilité** (qui correspond à la **conservation de la masse**). Si la dimension d'espace est $N = 1$, le système de Stokes est sans intérêt car on voit facilement que la vitesse est nulle et que la pression est une primitive de la force. Par contre en dimension $N \geq 2$, le système de Stokes a bien un sens : en particulier, il existe des champs de vitesses incompressibles non triviaux (prendre, par exemple, un rotationnel).

1.3.7 Équations des plaques

On considère la déformation élastique d'une plaque plane d'épaisseur petite (négligeable devant ses autres dimensions). Si on note Ω la surface moyenne de la plaque, et $f(x)$ (une fonction de Ω dans $\mathbb{R}$) la résultante normale des forces, alors la composante normale du déplacement $u(x)$ (un scalaire) est solution de l'équation des plaques (dites en flexion)

$$\begin{cases} \Delta\left(\Delta u\right) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases} \tag{1.23}$$

où on note $\frac{\partial u}{\partial n} = \nabla u \cdot n$ avec n le vecteur normal unité extérieur à $\partial\Omega$. Remarquons qu'il s'agit d'une équation aux dérivées partielles du quatrième ordre en espace (appelée aussi bi-Laplacien). C'est pourquoi il est nécessaire d'avoir deux conditions aux limites. Ces conditions aux limites traduisent l'encastrement de la plaque (pas de déplacement ni de rotation du bord de la plaque).

1.4 Calcul numérique par différences finies

1.4.1 Principes de la méthode

A part dans quelques cas très particuliers, il est impossible de calculer explicitement des solutions des différents modèles présentés ci-dessus. Il est donc nécessaire d'avoir recours au calcul numérique sur ordinateur pour estimer qualitativement et quantitativement ces solutions. Le principe de toutes les méthodes de résolution numérique des équations aux dérivées partielles est d'obtenir des valeurs numériques discrètes (c'est-à-dire en nombre fini) qui **"approchent"** (en un sens convenable à préciser) la solution exacte. Dans ce procédé il faut bien être conscient de deux points fondamentaux : premièrement, on ne calcule pas des solutions exactes mais approchées ; deuxièmement, on **discrétise** le problème en représentant des fonctions par un nombre fini de valeurs, c'est-à-dire que **l'on passe du "continu" au "discret".**

Il existe de nombreuses méthodes d'approximation numérique des solutions d'équations aux dérivées partielles. Nous présentons maintenant une des plus anciennes et des plus simples, appelée méthode des différences finies (nous verrons plus loin une autre méthode, dite des éléments finis). Pour simplifier la présentation, nous nous limitons à la dimension un d'espace (voir la Sous-section 2.2.6 pour les dimensions supérieures). Nous n'abordons pour l'instant que les principes pratiques de cette méthode, c'est-à-dire la construction de ce qu'on appelle des **schémas numériques**. Nous réservons pour le Chapitre 2 la justification théorique de ces schémas, c'est-à-dire l'étude de leur convergence (en quel sens les solutions approchées discrètes sont proches des solutions exactes continues).

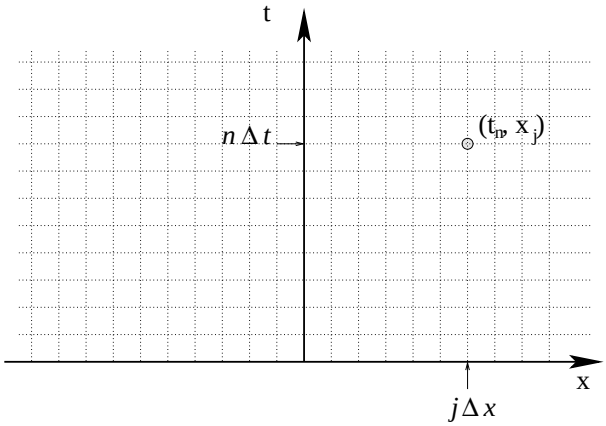

FIGURE 1.4 – Maillage en différences finies.

Pour discrétiser le continuum spatio-temporel, on introduit un **pas d'espace** $\Delta x > 0$ et un **pas de temps** $\Delta t > 0$ qui seront les plus petites échelles représentées par la méthode numérique. On définit un maillage ou des coordonnées discrètes de l'espace et du temps (voir la Figure 1.4)

$$(t_n, x_j) = (n\Delta t, j\Delta x) \text{ pour } n \geq 0, j \in \mathbb{Z}.$$

On note u_j^n la valeur d'une solution discrète approchée au point (t_n, x_j), et $u(t,x)$ la solution exacte (inconnue). Le principe de la méthode des différences finies est de remplacer les dérivées par des différences finies en utilisant des formules de Taylor dans lesquelles on néglige les restes. Par exemple, on approche la dérivée seconde en espace (le Laplacien en dimension un) par

$$-\frac{\partial^2 u}{\partial x^2}(t_n, x_j) \approx \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} \tag{1.24}$$

où l'on reconnaît la formule de Taylor

$$\begin{aligned} -u(t, x-\Delta x) + 2u(t,x) - u(t, x+\Delta x) = \; & -(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}(t,x) \\ & -\frac{(\Delta x)^4}{12}\frac{\partial^4 u}{\partial x^4}(t,x) + \mathcal{O}\Big((\Delta x)^6\Big) \end{aligned} \tag{1.25}$$

Si Δx est "petit", la formule (1.24) est une "bonne" approximation (elle est naturelle mais pas unique). La formule (1.24) est dite **centrée** car elle est symétrique en j.

Pour discrétiser l'équation de convection-diffusion

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0 \tag{1.26}$$

il faut aussi discrétiser le terme de convection. Une formule centrée donne

$$V\frac{\partial u}{\partial x}(t_n, x_j) \approx V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}$$

Il ne reste plus qu'à faire la même chose pour la dérivée en temps. On a encore le choix dans la formule de différences finies : soit centrée, soit décentrée. Examinons trois formules "naturelles".

1. En premier lieu, la différence finie centrée

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$$

conduit au schéma complètement symétrique par rapport à n et j (appelé schéma centré ou **schéma de Richardson**)

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{1.27}$$

Aussi "naturel" et évident soit-il, **ce schéma est incapable de calculer des solutions approchées** de l'équation de convection-diffusion (1.26) (voir l'exemple numérique de la Figure 1.5) ! Nous justifierons cette incapacité du schéma à approcher la solution exacte dans le Lemme 2.2.20. Pour l'instant, indiquons simplement que la difficulté provient du caractère centré de la différence finie qui approche la dérivée en temps.

2. Un deuxième choix est la la différence finie décentrée amont (on remonte le temps ; on parle aussi de **schéma d'Euler rétrograde**)

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^n - u_j^{n-1}}{\Delta t}$$

qui conduit au schéma

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{1.28}$$

3. Le troisième choix est le symétrique du précédent : la différence finie décentrée aval (on avance dans le temps ; on parle aussi de **schéma d'Euler progressif**)

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

conduit au schéma

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu\frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{1.29}$$

La différence principale entre ces deux derniers schémas est que (1.28) est dit **implicite** car il faut résoudre un système d'équations linéaires pour calculer les valeurs $(u_j^n)_{j\in\mathbb{Z}}$ en fonctions des valeurs précédentes $(u_j^{n-1})_{j\in\mathbb{Z}}$, tandis que (1.29) est dit **explicite** puisqu'il donne immédiatement les valeurs $(u_j^{n+1})_{j\in\mathbb{Z}}$ en fonction des $(u_j^n)_{j\in\mathbb{Z}}$. Le décalage de 1 sur l'indice n entre les schémas (1.28) et (1.29) n'est évidemment qu'apparent puisqu'on peut réécrire de manière équivalente (1.29) sous la forme

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + V\frac{u_{j+1}^{n-1} - u_{j-1}^{n-1}}{2\Delta x} + \nu\frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0.$$

Dans les trois schémas que nous venons de définir, il y a bien sûr une donnée initiale pour démarrer les itérations en n : les valeurs initiales $(u_j^0)_{j\in\mathbb{Z}}$ sont définies, par exemple, par $u_j^0 = u_0(j\Delta x)$ où u_0 est la donnée initiale de l'équation de convection-diffusion (1.26). Remarquons que pour le "mauvais" schéma centré (1.27) il y a une difficulté supplémentaire au démarrage : pour $n = 1$ on a aussi besoin de connaître les valeurs $(u_j^1)_{j\in\mathbb{Z}}$ qu'il faut donc calculer autrement (par exemple, par application d'un des deux autres schémas).

1.4.2 Résultats numériques pour l'équation de la chaleur

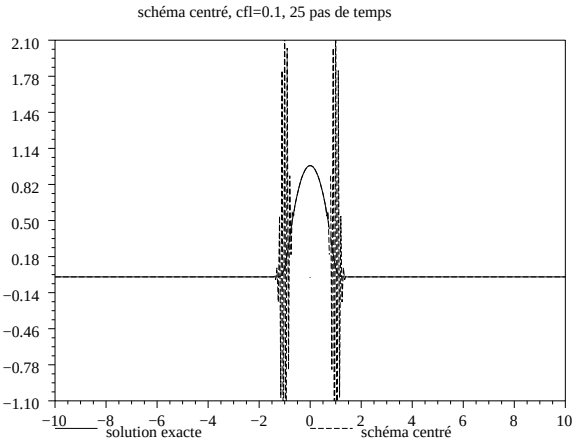

FIGURE 1.5 – Schéma centré instable avec $\nu\Delta t = 0.1(\Delta x)^2$.

Commençons par faire quelques tests numériques très simples dans le cas où $V = 0$ et $\nu = 1$, c'est-à-dire que **l'on résout numériquement l'équation de la chaleur.** On choisit comme condition initiale la fonction

$$u_0(x) = \max(1 - x^2, 0).$$

Pour pouvoir comparer les solutions numériques approchées avec la solution exacte (1.13), nous voudrions travailler sur le domaine infini $\Omega = \mathbb{R}$, c'est-à-dire calculer, pour chaque $n \geq 0$, une infinité de valeurs $(u_j^n)_{j\in\mathbb{Z}}$, mais l'ordinateur ne le permet

pas car sa mémoire est finie ! En première approximation, nous remplaçons donc $\mathbb{R}$ par le "grand" domaine $\Omega = (-10, +10)$ muni de conditions aux limites de Dirichlet. Nous admettrons la validité de cette approximation (qui est confirmée par les comparaisons numériques ci-dessous). Nous fixons le pas d'espace à $\Delta x = 0.05$: il y a donc 401 valeurs $(u_j^n)_{-200 \leq j \leq +200}$ à calculer. Rappelons pour mémoire que les valeurs u_j^n calculées par l'ordinateur sont entachées d'erreurs d'arrondi et ne sont donc pas les valeurs exactes des schémas discrets : néanmoins, dans les calculs présentés ici, ces erreurs d'arrondi sont totalement négligeables et ne sont en aucune manière la cause des différents phénomènes que nous allons observer. Sur toutes les figures nous représentons la solution exacte, calculée avec la formule explicite (1.13), et la solution approchée numérique considérée.

Réglons tout de suite le sort du schéma centré (1.27) : comme nous l'avions annoncé, ce schéma est incapable de calculer des solutions approchées de l'équation de la chaleur. Quel que soit le choix du pas de temps Δt, ce schéma est **instable**, c'est-à-dire que la solution numérique oscille de manière non bornée si l'on diminue les valeurs des pas Δx et Δt. Ce phénomène très caractéristique (et d'apparition très rapide) est illustré par la Figure 1.5. Insistons sur le fait que **quel que soit le choix** des pas Δt et Δx, on observe ces oscillations (non physiques, bien sûr). On dit que le schéma est inconditionnellement instable. Une justification rigoureuse en sera donnée au chapitre suivant (voir le Lemme 2.2.20).

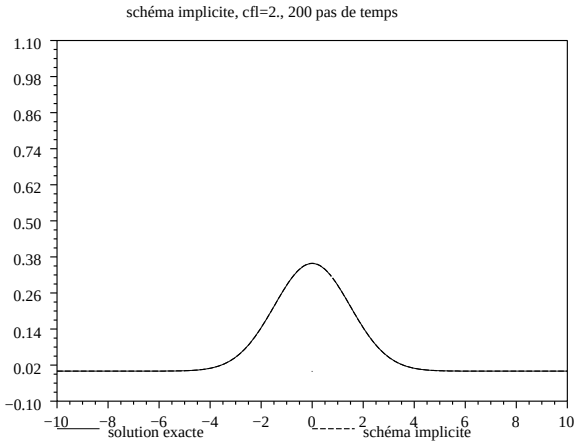

FIGURE 1.6 – Schéma implicite avec $\nu \Delta t = 2(\Delta x)^2$.

A l'opposé du précédent schéma, le schéma implicite (1.28) calcule de "bonnes" solutions approchées de l'équation de la chaleur **quel que soit** le pas de temps Δt (voir la Figure 1.6). En particulier, on n'observe jamais d'oscillations numériques quel que soit le choix des pas Δt et Δx. On dit que le schéma implicite est inconditionnellement stable.

Considérons maintenant le schéma explicite (1.29) : des expériences numériques montrent facilement que selon les valeurs du pas de temps Δt des oscillations numériques apparaissent ou non (voir la Figure 1.7). La limite de stabilité est facile à trouver expérimentalement : quel que soit le choix des pas Δt et Δx qui **vérifient**

la condition

$$2\nu\Delta t \leq (\Delta x)^2 \tag{1.30}$$

le schéma est stable, tandis que si (1.30) n'est pas vérifiée, alors le schéma est instable. On dit que le schéma explicite est conditionnellement stable. La condition de stabilité (1.30) est **une des remarques les plus simples et les plus profondes de l'analyse numérique**. Elle fut découverte en 1928 (avant l'apparition des premiers ordinateurs !) par Courant, Friedrichs, et Lewy. Elle porte depuis le nom de **condition CFL ou condition de Courant, Friedrichs, et Lewy**.

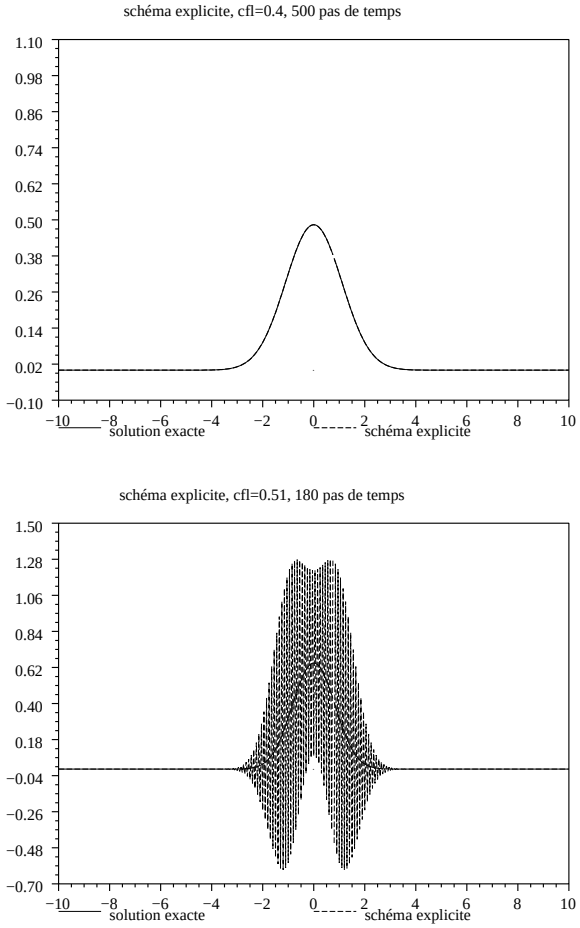

FIGURE 1.7 – Schéma explicite avec $\nu\Delta t = 0.4(\Delta x)^2$ (haut) et $\nu\Delta t = 0.51(\Delta x)^2$ (bas).

Nous allons justifier brièvement cette condition de stabilité (une analyse plus poussée sera effectuée au prochain chapitre). Réécrivons le schéma explicite sous la forme

$$u_j^{n+1} = \frac{\nu\Delta t}{(\Delta x)^2} u_{j-1}^n + \left(1 - 2\frac{\nu\Delta t}{(\Delta x)^2}\right) u_j^n + \frac{\nu\Delta t}{(\Delta x)^2} u_{j+1}^n. \tag{1.31}$$

Si la condition CFL est vérifiée, alors (1.31) montre que u_j^{n+1} est une combinaison convexe des valeurs au temps précédent $u_{j-1}^n, u_j^n, u_{j+1}^n$ (tous les coefficients dans le

membre de droite de (1.31) sont positifs et leur somme vaut 1). En particulier, si la donnée initiale u_0 est bornée par deux constantes m et M telles que

$$m \leq u_j^0 \leq M \text{ pour tout } j \in \mathbb{Z},$$

alors une récurrence facile montre que les mêmes inégalités restent vraies pour tous les temps ultérieurs

$$m \leq u_j^n \leq M \text{ pour tout } j \in \mathbb{Z} \text{ et pour tout } n \geq 0. \tag{1.32}$$

La propriété (1.32) empêche le schéma d'osciller de manière non bornée : il est donc stable sous la condition CFL. La propriété (1.32) est appelée principe du maximum discret : il s'agit de l'équivalent **discret** du principe du maximum **continu** pour les solutions exactes que nous avons vu à la Remarque 1.2.4.

Supposons au contraire que la condition CFL ne soit pas vérifiée, c'est-à-dire que

$$2\nu\Delta t > (\Delta x)^2.$$

Alors, pour certaines données initiales le schéma est instable (il peut être stable pour certaines données initiales "exceptionnelles" : par exemple, si $u_0 \equiv 0$!). Prenons la donnée initiale définie par

$$u_j^0 = (-1)^j$$

qui est bien uniformément bornée. Un calcul simple montre que

$$u_j^n = (-1)^j \left(1 - 4\frac{\nu\Delta t}{(\Delta x)^2}\right)^n$$

qui croît en module vers l'infini lorsque n tend vers l'infini car $1 - 4\frac{\nu\Delta t}{(\Delta x)^2} < -1$. Le schéma explicite est donc instable si la condition CFL n'est pas satisfaite.

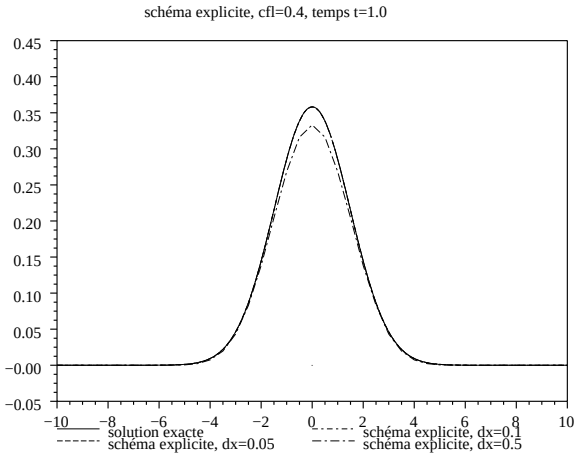

FIGURE 1.8 – Schéma explicite avec $\nu\Delta t = 0.4(\Delta x)^2$ pour diverses valeurs de Δx.

Exercice 1.4.1 Le but de cet exercice est de montrer que le schéma implicite (1.28), avec $V = 0$, vérifie aussi le principe du maximum discret. On impose des conditions aux limites de Dirichlet, c'est-à-dire que la formule (1.28) est valable pour $1 \leq j \leq J$ et on fixe $u_0^n = u_{J+1}^n = 0$ pour tout $n \in \mathbb{N}$. Soit deux constantes $m \leq 0 \leq M$ telles que $m \leq u_j^0 \leq M$ pour $1 \leq j \leq J$. Vérifier que l'on peut bien calculer de manière unique les u_j^{n+1} en fonction des u_j^n. Montrer que pour tous les temps $n \geq 0$ on a encore les inégalités $m \leq u_j^n \leq M$ pour $1 \leq j \leq J$ (et ceci sans condition sur Δt et Δx).

Si nous avons à peu près élucidé la question de la stabilité du schéma explicite, nous n'avons rien dit sur sa convergence, c'est-à-dire sur sa capacité à approcher correctement la solution exacte. Nous répondrons rigoureusement à cette question au prochain chapitre. Remarquons que la stabilité est, bien sûr, une condition nécessaire de convergence, mais pas suffisante. Contentons nous pour l'instant de vérifier expérimentalement la convergence du schéma, c'est-à-dire que lorsque les pas d'espace et de temps deviennent de plus en plus petits les solutions numériques correspondantes convergent et que leur limite est bien la solution exacte (nous pouvons vérifier ce dernier point puisqu'ici la solution exacte est disponible). Sur la Figure 1.8 nous vérifions numériquement que, si l'on raffine le pas d'espace Δx (qui prend les valeurs 0.5, 0.1, et 0.05) ainsi que le pas de temps Δt en gardant constant le rapport $\nu \Delta t/(\Delta x)^2$ (le nombre CFL), alors la solution numérique est de plus en plus proche de la solution exacte. (La comparaison s'effectue au même temps final $t = 1$, donc le nombre de pas de temps augmente lorsque le pas de temps Δt diminue.) Ce procédé de **"vérification numérique de la convergence"** est très simple et on ne doit jamais hésiter à l'utiliser faute de mieux (c'est-à-dire si l'analyse théorique de la convergence est impossible ou trop difficile).

1.4.3 Résultats numériques pour l'équation d'advection

Effectuons une deuxième série d'expériences numériques sur **l'équation de convection-diffusion** (1.26) avec une vitesse $V = 1$ non nulle. Nous reprenons les mêmes données que précédemment et nous choisissons le schéma explicite avec $\nu \Delta t = 0.4(\Delta x)^2$. Nous regardons l'influence de la valeur de la constante de diffusion ν (ou inverse du nombre de Péclet) sur la stabilité du schéma. La Figure 1.9 montre que le schéma est stable pour $\nu = 1$, instable pour $\nu = 0.01$, et que pour la valeur intermédiaire $\nu = 0.1$, le schéma semble stable mais la solution approchée est légèrement différente de la solution exacte. On comprend bien que plus l'inverse du nombre de Péclet ν est petit, plus le terme convectif est prédominant sur le terme diffusif. Par conséquent, la condition CFL (1.30), obtenue lorsque la vitesse V est nulle, est de moins moins en valable au fur et à mesure que ν diminue.

Pour comprendre ce phénomène, examinons **l'équation d'advection** qui s'obtient à la limite $\nu = 0$. Remarquons tout d'abord que la condition CFL (1.30) est automatiquement satisfaite si $\nu = 0$ (quels que soient Δt et Δx), ce qui semble contradictoire avec le résultat expérimental du bas de la Figure 1.9.

Pour l'équation d'advection (c'est-à-dire (1.26) avec $\nu = 0$), le schéma explicite

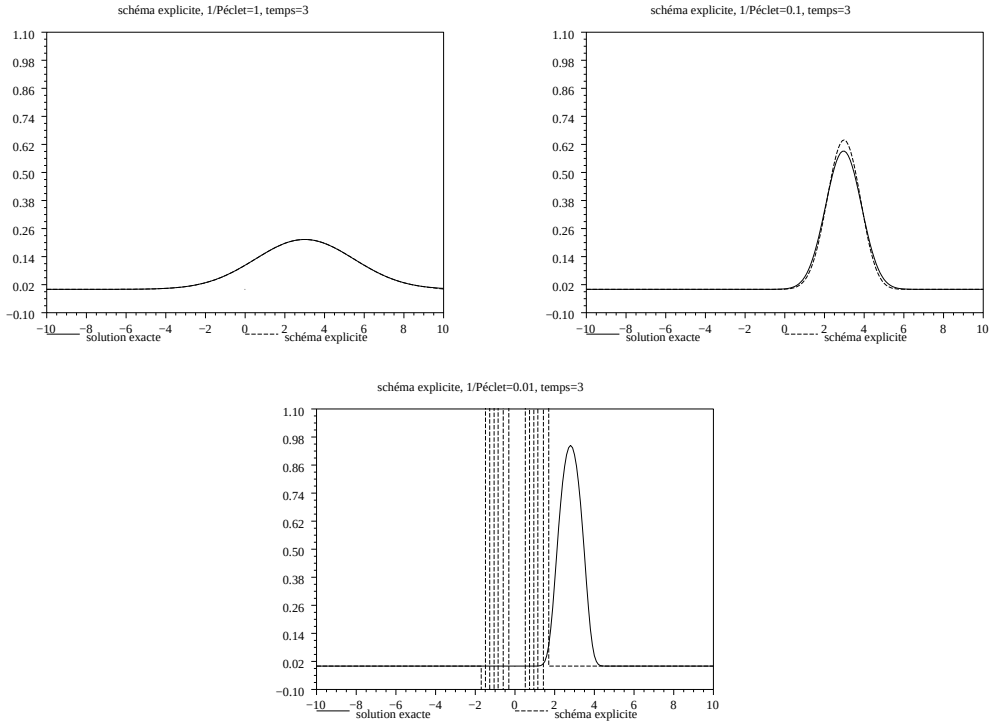

FIGURE 1.9 – Schéma explicite pour l'équation de convection-diffusion avec $\nu\Delta t = 0.4(\Delta x)^2$ et $V = 1$. En haut à gauche $\nu = 1$, en haut à droite $\nu = 0.1$, et en bas $\nu = 0.01$.

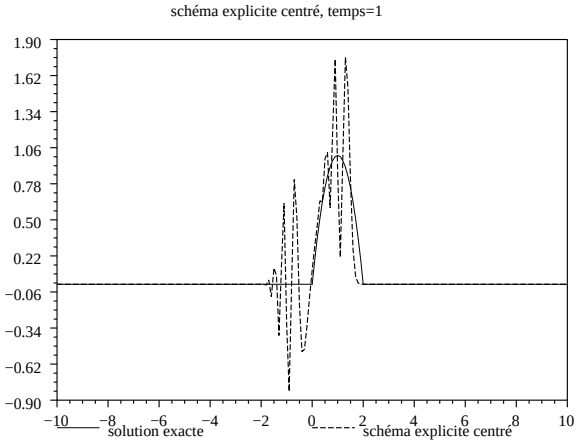

FIGURE 1.10 – Schéma explicite centré pour l'équation d'advection avec $\Delta t = 0.9\Delta x$, $V = 1$, $\nu = 0$.

(1.29) peut se réécrire

$$u_j^{n+1} = \frac{V\Delta t}{2\Delta x} u_{j-1}^n + u_j^n - \frac{V\Delta t}{2\Delta x} u_{j+1}^n. \tag{1.33}$$

Ce schéma conduit aux oscillations de la Figure 1.10 dans les mêmes conditions expérimentales que le bas de la Figure 1.9. On voit bien que u_j^{n+1} n'est jamais (quel que soit Δt) une combinaison convexe de u_{j-1}^n, u_j^n, et u_{j+1}^n. Il ne peut donc y avoir de principe du maximum discret pour ce schéma, ce qui est une indication supplémentaire de son instabilité (une preuve rigoureuse en sera donnée au Lemme 2.3.1). L'origine de cette instabilité est que, dans le schéma explicite (1.33), nous avons choisi de traiter le terme convectif de manière centré. Nous pouvons cependant décentrer ce terme comme nous l'avons fait pour la dérivée en temps. Deux choix sont possibles : décentrer vers la droite ou vers la gauche. Le signe de la vitesse V est bien sûr crucial : ici nous supposons que $V > 0$ (un argument symétrique est valable si $V < 0$). Pour $V > 0$, le décentrement à droite est dit **décentrement aval** : on obtient

$$V\frac{\partial u}{\partial x}(t_n, x_j) \approx V\frac{u_{j+1}^n - u_j^n}{\Delta x}$$

en allant chercher "l'information" en suivant le courant. Ce choix conduit à un schéma décentré aval "désastreux"

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \tag{1.34}$$

qui est tout aussi instable que le schéma centré. Au contraire le **décentrement amont** (c'est-à-dire à gauche si $V > 0$), qui va chercher "l'information" en remontant le courant

$$V\frac{\partial u}{\partial x}(t_n, x_j) \approx V\frac{u_j^n - u_{j-1}^n}{\Delta x}$$

conduit au schéma explicite décentré amont

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \tag{1.35}$$

qui donne les résultats de la Figure 1.11. On vérifie aisément que le schéma (1.35) est stable sous une nouvelle condition CFL (différente de la précédente condition CFL (1.30))

$$|V|\Delta t \leq \Delta x. \tag{1.36}$$

En effet, on peut réécrire (1.35) sous la forme

$$u_j^{n+1} = \frac{V\Delta t}{\Delta x} u_{j-1}^n + \left(1 - \frac{V\Delta t}{\Delta x}\right) u_j^n,$$

qui montre que, si la condition (1.36) est satisfaite, u_j^{n+1} est une combinaison convexe de u_{j-1}^n et u_j^n. Par conséquent, le schéma décentré amont (1.35) vérifie un principe du maximum discret, ce qui entraîne sa stabilité conditionnelle. L'idée du **décentrement amont est une autre idée majeure de l'analyse numérique**. Elle est

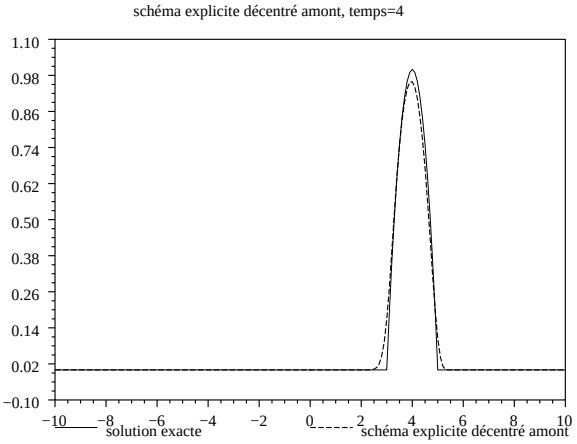

FIGURE 1.11 – Schéma explicite décentré amont pour l'équation d'advection avec $\Delta t = 0.9\Delta x$, $V = 1$.

particulièrement cruciale dans tous les problèmes de mécanique des fluides où elle fut d'abord découverte (en anglais on parle de **upwinding**, c'est-à-dire de remonter le vent ou le courant), mais elle apparaît dans bien d'autres modèles.

La conclusion de cette étude sur l'équation d'advection est que pour le modèle de convection-diffusion avec faible valeur de la constante de diffusion ν, il faut absolument décentrer vers l'amont le terme convectif et suivre la condition CFL (1.36) plutôt que celle (1.30). A ce prix on peut améliorer les résultats de la Figure 1.9.

Exercice 1.4.2 Montrer que, si la condition CFL (1.36) n'est pas satisfaite, le schéma décentré amont (1.35) pour l'équation d'advection est instable pour la donnée initiale $u_j^0 = (-1)^j$.

Exercice 1.4.3 Écrire un schéma explicite centré en espace pour l'équation des ondes (1.17) en une dimension d'espace et sans terme source. Préciser comment démarrer les itérations en temps. Vérifier l'existence d'un cône de dépendance discret analogue à celui continu illustré par la Figure 1.3. En déduire que, si ce schéma converge, les pas de temps et d'espace doivent nécessairement satisfaire la condition (de type CFL) $\Delta t \leq \Delta x$.

Les conclusions de cette section sont nombreuses et vont nourrir les réflexions du prochain chapitre. Tout d'abord, tous les schémas numériques "raisonnables" ne fonctionnent pas, loin s'en faut. On rencontre des problèmes de stabilité (sans parler de convergence) qui nécessitent d'analyser en détails ces schémas : c'est la raison d'être de l'analyse numérique qui concilie objectifs pratiques et études théoriques. Enfin, les "bons" schémas numériques doivent respecter un certain nombre de propriétés (comme par exemple, le principe du maximum discret, ou le décentrement amont) qui ne sont que la traduction (au niveau discret) de propriétés physiques ou mathématiques de l'équation aux dérivées partielles. **On ne peut donc pas faire l'économie d'une bonne compréhension de la modélisation physique et des propriétés mathématiques des modèles si l'on veut réaliser de bonnes**

simulations numériques.

1.5 Notion de problème bien posé

Nous terminons ce chapitre par un certain nombre de définitions qui permettront au lecteur de s'y retrouver dans le vocabulaire employé ici comme dans les ouvrages classiques sur l'analyse numérique.

Définition 1.5.1 *On appelle* **problème aux limites** *une équation aux dérivées partielles munie de conditions aux limites sur la totalité de la frontière du domaine sur lequel elle est posée.*

Par exemple, le Laplacien (1.19) est un problème aux limites. A contrario, l'équation différentielle ordinaire

$$\begin{cases} \frac{dy}{dt} = f(t,y) \text{ pour } 0 < t < T \\ y(t=0) = y_0 \end{cases} \tag{1.37}$$

n'est pas un problème aux limites puisqu'étant posée sur un segment $(0,T)$, avec $0 < T \leq +\infty$, elle n'a de conditions "au bord" qu'en $t = 0$ (et pas en $t = T$).

Définition 1.5.2 *On appelle* **problème de Cauchy** *une équation aux dérivées partielles où, pour au moins une variable (généralement le temps t), les conditions "au bord" sont des conditions initiales (c'est-à-dire ne portent que sur un bord $t = 0$, et pas en $t = T$).*

Par exemple, l'équation différentielle ordinaire (1.37) est un problème de Cauchy, mais pas le Laplacien (1.19) (quel que soit le choix de la composante de la variable d'espace x à qui on ferait jouer le rôle du temps).

De nombreux modèles sont à la fois des problèmes aux limites et des problèmes de Cauchy. Ainsi, l'équation de la chaleur (1.8) est un problème de Cauchy par rapport à la variable de temps t et un problème aux limites par rapport à la variable d'espace x. Tous les modèles que nous allons étudier dans ce cours rentrent dans une de ces deux catégories de problème.

Le fait qu'un modèle mathématique soit un problème de Cauchy ou un problème aux limites n'implique pas automatiquement qu'il s'agisse d'un "bon" modèle. L'expression **bon modèle** n'est pas employée ici au sens de la pertinence physique du modèle et de ses résultats, mais au sens de sa cohérence mathématique. Comme nous allons le voir cette cohérence mathématique est une condition nécessaire avant de pouvoir même envisager des simulations numériques et des interprétations physiques. Le mathématicien Jacques Hadamard a donné une définition de ce qu'est un "bon" modèle, en parlant de **problème bien posé** (un problème mal posé est le contraire d'un problème bien posé). On décide de noter f les données (le second membre, les données initiales, le domaine, etc.), u la solution recherchée, et $\mathcal{A}$ "l'opérateur" qui agit sur u. Il s'agit ici de notations abstraites, $\mathcal{A}$ désignant à la fois

l'équation aux dérivées partielles et le type de conditions initiales ou aux limites. Le problème est donc de trouver u solution de

$$\mathcal{A}(u) = f \tag{1.38}$$

Définition 1.5.3 *On dit que le problème (1.38) est* **bien posé** *si pour toute donnée f il admet une solution unique u, et si cette solution u dépend continûment de la donnée f.*

Examinons en détail cette définition de Hadamard : elle contient en fait trois conditions pour qu'un problème soit bien posé. Premièrement, il faut qu'il existe au moins une solution : c'est bien la moindre des choses à demander à un modèle sensé représenter la réalité ! Deuxièmement, il faut que la solution soit unique : c'est nécessaire si on veut une réponse déterministe et non ambigûe. Troisièmement, et c'est la condition la moins évidente a priori, il faut que la solution dépende continûment des données. Au premier abord, cela semble une fantaisie de mathématicien, mais c'est pourtant crucial dans une perspective **d'approximation numérique**. En effet, faire un calcul numérique d'une solution approchée de (1.38) revient à perturber les données (qui de continues deviennent discrètes) et à résoudre (1.38) pour ces données perturbées. Si de petites perturbations des données conduisent à de grandes perturbations de la solution, il n'y a aucune chance pour que la simulation numérique soit proche de la réalité (ou du moins de la solution exacte). Par conséquent, cette dépendance continue de la solution par rapport aux données est une condition absolument nécessaire pour envisager des simulations numériques précises.

Terminons en avouant qu'à ce niveau de généralité la Définition 1.5.3 est bien floue, et que pour lui donner un sens mathématique précis il faut bien sûr dire dans quels espaces de fonctions on place les données et on cherche la solution, et quelles normes ou topologies on utilise pour la continuité. Il n'est pas rare en effet qu'un changement d'espace (bien anodin en apparence) entraîne des propriétés d'existence ou d'unicité fort différentes !

Exercice 1.5.1 Le but de cet exercice est de montrer que le problème de Cauchy pour le Laplacien est mal posé. Soit le domaine bidimensionnel $\Omega = (0,1) \times (0,2\pi)$. On considère le problème de Cauchy en x et le problème aux limites en y suivant

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^2 u}{\partial y^2} = 0 & \text{dans } \Omega \\ u(x,0) = u(x,2\pi) = 0 & \text{pour } 0 < x < 1 \\ u(0,y) = 0,\ \dfrac{\partial u}{\partial x}(0,y) = -e^{-\sqrt{n}} \sin(ny) & \text{pour } 0 < y < 2\pi \end{cases}$$

Vérifier que $u(x,y) = \frac{e^{-\sqrt{n}}}{n} \sin(ny)\mathrm{sh}(nx)$ est une solution. Montrer que la condition initiale et toutes ses dérivées en $x = 0$ convergent uniformément vers 0, tandis que, pour tout $x > 0$, la solution trouvée $u(x,y)$ et toutes ses dérivées ne sont pas bornés quand n tend vers l'infini. Conclure.

Chapitre 2

MÉTHODE DES DIFFÉRENCES FINIES

2.1 Introduction

Dans ce chapitre nous analysons les schémas numériques de différences finies. Nous définissons la **stabilité** et la **consistance** d'un schéma et nous montrons que, pour les équations aux dérivées partielles linéaires à coefficients constants, la stabilité combinée à la consistance d'un schéma impliquent sa **convergence**.

Le plan de ce chapitre est le suivant. La Section 2.2 traite le cas de l'équation de la chaleur introduite au Chapitre 1. La Section 2.3 généralise les résultats précédents aux cas de l'équation des ondes ou de l'équation d'advection. Un des buts de ce chapitre est de fournir un cadre de conception et d'analyse des schémas de différences finies pour des modèles beaucoup plus généraux. Le lecteur ne devrait pas avoir de mal à étendre les concepts présentés ici à son modèle préféré et à concevoir ainsi des schémas numériques originaux.

2.2 Différences finies pour l'équation de la chaleur

2.2.1 Divers exemples de schémas

Nous nous limitons à la dimension un d'espace et nous renvoyons à la Sous-section 2.2.6 pour le cas de plusieurs dimensions d'espace. Nous considérons l'équation de la chaleur dans le domaine borné $(0,1)$

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = 0 \text{ pour } (x,t) \in (0,1) \times \mathbb{R}_*^+ \\ u(0,x) = u_0(x) \text{ pour } x \in (0,1). \end{cases} \tag{2.1}$$

Pour discrétiser le domaine $(0,1) \times \mathbb{R}^+$, on introduit un pas d'espace $\Delta x = 1/(N+1) > 0$ (avec N un entier positif) et un pas de temps $\Delta t > 0$, et on définit les noeuds d'un maillage régulier

$$(t_n, x_j) = (n\Delta t, j\Delta x) \text{ pour } n \geq 0, j \in \{0, 1, ..., N+1\}.$$

On note u_j^n la valeur d'une solution discrète approchée au point (t_n, x_j), et $u(t,x)$ la solution exacte de (2.1). La donnée initiale est discrétisée par

$$u_j^0 = u_0(x_j) \text{ pour } j \in \{0, 1, ..., N+1\}.$$

Les conditions aux limites de (2.1) peuvent être de plusieurs types, mais leur choix n'intervient pas dans la définition des schémas. Ici, nous utilisons des conditions aux limites de Dirichlet

$$u(t,0) = u(t,1) = 0 \text{ pour tout } t \in \mathbb{R}_*^+$$

qui se traduisent en

$$u_0^n = u_{N+1}^n = 0 \text{ pour tout } n > 0.$$

Par conséquent, à chaque pas de temps nous avons à calculer les valeurs $(u_j^n)_{1\leq j\leq N}$ qui forment un vecteur de $\mathbb{R}^N$. Nous donnons maintenant plusieurs schémas possibles pour l'équation de la chaleur (2.1). Tous ces schémas sont définis par N équations (en chaque point x_j, $1 \leq j \leq N$) qui permettent de calculer les N valeurs u_j^n. Au Chapitre 1 nous avons déjà parlé du **schéma explicite**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0 \tag{2.2}$$

pour $n \geq 0$ et $j \in \{1, ..., N\}$, ainsi que du **schéma implicite**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} = 0. \tag{2.3}$$

Il est facile de vérifier que le schéma implicite (2.3) est effectivement bien défini, c'est-à-dire qu'on peut calculer les valeurs u_j^{n+1} en fonction des u_j^n : en effet, il faut inverser la matrice tridiagonale carrée de taille N

$$\begin{pmatrix} 1+2c & -c & & & 0 \\ -c & 1+2c & -c & & \\ & \ddots & \ddots & \ddots & \\ & & -c & 1+2c & -c \\ 0 & & & -c & 1+2c \end{pmatrix} \text{ avec } c = \frac{\nu \Delta t}{(\Delta x)^2}, \tag{2.4}$$

dont il est aisé de vérifier le caractère défini positif, donc inversible. En faisant une combinaison convexe de (2.2) et (2.3), pour $0 \leq \theta \leq 1$, on obtient le **θ-schéma**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta\nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta)\nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \tag{2.5}$$

Bien sûr, on retrouve le schéma explicite (2.2) si $\theta = 0$, et le schéma implicite (2.3) si $\theta = 1$. Le θ-schéma (2.5) est implicite dès que $\theta \neq 0$. Pour la valeur $\theta = 1/2$, on obtient le **schéma de Crank-Nicolson**.

Tous les schémas qui précèdent sont dits **à deux niveaux** car ils ne font intervenir que deux indices de temps. On peut bien sûr construire des schémas multi-niveaux : les plus populaires sont à trois niveaux. En plus du schéma (instable) de Richardson vu au Chapitre 1, on cite le **schéma de DuFort-Frankel**

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^n + u_j^{n+1} + u_j^{n-1} - u_{j+1}^n}{(\Delta x)^2} = 0, \tag{2.6}$$

le **schéma de Gear**

$$\frac{3u_j^{n+1} - 4u_j^n + u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} = 0. \tag{2.7}$$

Nous voilà en face de beaucoup trop de schémas ! Et la liste ci-dessus n'est pas exhaustive ! Un des buts de l'analyse numérique va être de comparer et de sélectionner les meilleurs schémas suivant des critères de précision, de coût, ou de robustesse.

Remarque 2.2.1 S'il y a un second membre $f(t, x)$ dans l'équation de la chaleur (2.1), alors les schémas se modifient en remplaçant zéro au second membre par une approximation consistante de $f(t, x)$ au point (t_n, x_j). Par exemple, si on choisit l'approximation $f(t_n, x_j)$, le schéma explicite (2.2) devient

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = f(t_n, x_j).$$

•

Remarque 2.2.2 Les schémas ci-dessus ont une écriture plus ou moins compacte, c'est-à-dire qu'il font intervenir un nombre fini, plus ou moins restreint, de valeurs u_j^n. La collection des couples (n', j') qui interviennent dans l'équation discrète au point (n, j) est appelé **stencil** du schéma (terme anglais qu'on peut essayer de traduire par **support**). En général, plus le stencil est large, plus le schéma est coûteux et difficile à programmer (en partie à cause des "effets de bord", c'est-à-dire des cas où certains des couples (n', j') sortent du domaine de calcul). •

Remarque 2.2.3 On peut remplacer les conditions aux limites de Dirichlet dans (2.1) par des conditions aux limites de Neumann, ou bien des conditions aux limites de périodicité (entres autres). Commençons par décrire deux manières différentes de discrétiser les conditions de Neumann

$$\frac{\partial u}{\partial x}(t, 0) = 0 \text{ et } \frac{\partial u}{\partial x}(t, 1) = 0.$$

Tout d'abord, on peut écrire

$$\frac{u_1^n - u_0^n}{\Delta x} = 0 \text{ et } \frac{u_{N+1}^n - u_N^n}{\Delta x} = 0$$

qui permet d'éliminer les valeurs u_0^n et u_{N+1}^n et de ne calculer que les N valeurs $(u_j^n)_{1\leq j\leq N}$. Cette discrétisation de la condition de Neumann n'est que du premier ordre. Si le schéma est du deuxième ordre, cela engendre une perte de précision près du bord. C'est pourquoi on propose une autre discrétisation (du deuxième ordre)

$$\frac{u_1^n - u_{-1}^n}{2\Delta x} = 0 \text{ et } \frac{u_{N+2}^n - u_N^n}{2\Delta x} = 0$$

qui est plus précise, mais nécessite l'ajout de 2 "points fictifs" x_{-1} et x_{N+2}. On élimine les valeurs u_{-1}^n et u_{N+2}^n, correspondant à ces points fictifs, et il reste maintenant $N+2$ valeurs à calculer, à savoir $(u_j^n)_{0\leq j\leq N+1}$.

D'autre part, les conditions aux limites de périodicité s'écrivent

$$u(t, x+1) = u(t, x) \text{ pour tout } x \in [0,1], t \geq 0.$$

Elles se discrétisent par les égalités $u_0^n = u_{N+1}^n$ pour tout $n \geq 0$, et plus généralement $u_j^n = u_{N+1+j}^n$. •

2.2.2 Consistance et précision

Bien sûr, les formules des schémas ci-dessus ne sont pas choisies au hasard : elles résultent d'une approximation de l'équation par développement de Taylor comme nous l'avons expliqué au Chapitre 1. Pour formaliser cette approximation de l'équation aux dérivées partielles par des différences finies, on introduit la notion de **consistance** et de **précision**. Bien que pour l'instant nous ne considérons que l'équation de la chaleur (2.1), nous allons donner une définition de la consistance valable pour n'importe quelle équation aux dérivées partielles que nous notons $F(u) = 0$. Remarquons que $F(u)$ est une notation pour une fonction de u et de ses dérivées partielles en tout point (t, x). De manière générale un schéma aux différences finies est défini, pour tous les indices possibles n, j, par la formule

$$F_{\Delta t,\Delta x}\left(\{u_{j+k}^{n+m}\}_{m^-\leq m\leq m^+,\ k^-\leq k\leq k^+}\right) = 0 \tag{2.8}$$

où les entiers m^-, m^+, k^-, k^+ définissent la largeur du stencil du schéma (voir la Remarque 2.2.2).

Définition 2.2.4 *Le schéma aux différences finies (2.8) est dit consistant avec l'équation aux dérivées partielles $F(u) = 0$, si, pour toute solution $u(t, x)$ suffisamment régulière de cette équation,* **l'erreur de troncature** *du schéma, définie par*

$$F_{\Delta t,\Delta x}\left(\{u(t+m\Delta t, x+k\Delta x)\}_{m^-\leq m\leq m^+,\ k^-\leq k\leq k^+}\right), \tag{2.9}$$

tend vers zéro, uniformément par rapport à (t, x), lorsque Δt et Δx tendent vers zéro indépendemment.

De plus, on dit que le schéma est précis à l'ordre p en espace et à l'ordre q en temps si l'erreur de troncature (2.9) tend vers zéro comme $\mathcal{O}\Big((\Delta x)^p + (\Delta t)^q\Big)$ lorsque Δt et Δx tendent vers zéro.

Remarque 2.2.5 Il faut prendre garde dans la formule (2.8) à une petite ambiguïté quant à la définition du schéma. En effet, on peut toujours multiplier n'importe quelle formule par une puissance suffisamment élevée de Δt et Δx de manière à ce que l'erreur de troncature tende vers zéro. Cela rendrait consistant n'importe quel schéma ! Pour éviter cet inconvénient, on supposera toujours que la formule $F_{\Delta t,\Delta x}(\{u_{j+k}^{n+m}\}) = 0$ a été écrite de telle manière que, pour une fonction régulière $u(t,x)$ qui n'est pas solution de l'équation de la chaleur, la limite de l'erreur de troncature n'est pas nulle. •

Concrètement on calcule l'erreur de troncature d'un schéma en remplaçant u_{j+k}^{n+m} dans la formule (2.8) par $u(t+m\Delta t, x+k\Delta x)$. Comme application de la Définition 2.2.4, nous allons montrer le lemme suivant.

Lemme 2.2.6 *Le schéma explicite (2.2) est consistant, précis à l'ordre 1 en temps et 2 en espace. De plus, si on choisit de garder constant le rapport $\nu\Delta t/(\Delta x)^2 = 1/6$, alors ce schéma est précis à l'ordre 2 en temps et 4 en espace.*

Remarque 2.2.7 Dans la deuxième phrase de l'énoncé du Lemme 2.2.6 on a légèrement modifié la définition de la consistance en spécifiant le rapport entre Δt et Δx lorsqu'ils tendent vers zéro. Ceci permet de tenir compte d'éventuelles compensations entre termes apparaissant dans l'erreur de troncature. En pratique, on observe effectivement de telles améliorations de la précision si on adopte le bon rapport entre les pas Δt et Δx. •

Démonstration. Soit $v(t,x)$ une fonction de classe $\mathcal{C}^6$. Par développement de Taylor autour du point (t,x), on calcule l'erreur de troncature du schéma (2.2)

$$\frac{v(t+\Delta t,x)-v(t,x)}{\Delta t}+\nu\frac{-v(t,x-\Delta x)+2v(t,x)-v(t,x+\Delta x)}{(\Delta x)^2}$$
$$=\Big(v_t-\nu v_{xx}\Big)+\frac{\Delta t}{2}v_{tt}-\frac{\nu(\Delta x)^2}{12}v_{xxxx}+\mathcal{O}\Big((\Delta t)^2+(\Delta x)^4\Big),$$

où v_t, v_x désignent les dérivées partielles de v. Si v est une solution de l'équation de la chaleur (2.1), on obtient ainsi aisément la consistance ainsi que la précision à l'ordre 1 en temps et 2 en espace. Si on suppose en plus que $\nu\Delta t/(\Delta x)^2 = 1/6$, alors les termes en Δt et en $(\Delta x)^2$ se simplifient car $v_{tt} = \nu v_{txx} = \nu^2 v_{xxxx}$. □

Exercice 2.2.1 Pour chacun des schémas de la Sous-section 2.2.1, vérifier que l'erreur de troncature est bien du type annoncé dans le Tableau 2.1. (On remarquera que tous ces schémas sont consistants sauf celui de DuFort-Frankel.)

2.2.3 Stabilité et analyse de Fourier

Dans le Chapitre 1 nous avons évoqué la stabilité des schémas de différences finies sans en donner une définition précise. Tout au plus avons nous expliqué que, numériquement, l'instabilité se manifeste par des oscillations non bornées de la solution numérique. Il est donc temps de donner une définition mathématique de la

Schéma	Erreur de troncature	Stabilité
Explicite (2.2)	$\mathcal{O}\Big(\Delta t + (\Delta x)^2\Big)$	stable L^2 et L^∞ si condition CFL $2\nu\Delta t \leq (\Delta x)^2$
Implicite (2.3)	$\mathcal{O}\Big(\Delta t + (\Delta x)^2\Big)$	stable L^2 et L^∞
Crank-Nicolson (2.5) (avec $\theta = 1/2$)	$\mathcal{O}\Big((\Delta t)^2 + (\Delta x)^2\Big)$	stable L^2
θ-schéma (2.5) (avec $\theta \neq 1/2$)	$\mathcal{O}\Big(\Delta t + (\Delta x)^2\Big)$	stable L^2 si condition CFL $2(1-2\theta)\nu\Delta t \leq (\Delta x)^2$
DuFort-Frankel (2.6)	$\mathcal{O}\Big((\frac{\Delta t}{\Delta x})^2 + (\Delta x)^2\Big)$	stable L^2 si condition CFL $\Delta t/(\Delta x)^2$ borné
Gear (2.7)	$\mathcal{O}\Big((\Delta t)^2 + (\Delta x)^2\Big)$	stable L^2

TABLE 2.1 – Erreurs de troncature et stabilité de divers schémas pour l'équation de la chaleur

stabilité. Pour cela nous avons besoin de définir une norme pour la solution numérique $u^n = (u_j^n)_{1\leq j\leq N}$. Nous reprenons les normes classiques sur $\mathbb{R}^N$ que nous pondérons simplement par le pas d'espace Δx :

$$\|u^n\|_p = \left(\sum_{j=1}^{N} \Delta x |u_j^n|^p\right)^{1/p} \quad \text{pour } 1 \leq p \leq +\infty, \tag{2.10}$$

où le cas limite $p = +\infty$ doit être compris dans le sens $\|u^n\|_\infty = \max_{1\leq j\leq N} |u_j^n|$. Remarquons que la norme ainsi définie dépend de Δx à travers la pondération mais aussi à travers l'entier N car $\Delta x = 1/(N+1)$. Grâce à la pondération par Δx, la norme $\|u^n\|_p$ est identique à la norme $L^p(0,1)$ pour les fonctions constantes par morceaux sur les sous-intervalles $[x_j, x_{j+1}[$ de $[0,1]$. Souvent, on l'appellera donc "norme L^p". En pratique on utilise surtout les normes correspondant aux valeurs $p = 2, +\infty$.

Définition 2.2.8 *Un schéma aux différences finies est dit* **stable** *pour la norme* $\|\ \|$*, définie par (2.10), s'il existe une constante* $K > 0$ *indépendante de* Δt *et* Δx *(lorsque ces valeurs tendent vers zéro) telle que*

$$\|u^n\| \leq K\|u^0\| \text{ pour tout } n \geq 0, \tag{2.11}$$

quelle que soit la donnée initiale u^0*.*

Si (2.11) n'a lieu que pour des pas Δt *et* Δx *astreints à certaines inégalités, on dit que le schéma est* **conditionnellement stable**.

Remarque 2.2.9 Puisque toutes les normes sont équivalentes dans $\mathbb{R}^N$, le lecteur trop rapide pourrait croire que la stabilité par rapport à une norme implique la stabilité par rapport à toutes les normes. Malheureusement il n'en est rien et il existe

des schémas qui sont stables par rapport à une norme mais pas par rapport à une autre (voir plus loin l'exemple du schéma de Lax-Wendroff avec les Exercices 2.3.2 et 2.3.3). En effet, le point crucial dans la Définition 2.2.8 est que la majoration est uniforme par rapport à Δx alors même que les normes définies par (2.10) dépendent de Δx. •

Définition 2.2.10 *Un schéma aux différences finies est dit* **linéaire** *si la formule* $F_{\Delta t,\Delta x}(\{u_{j+k}^{n+m}\}) = 0$ *qui le définit est linéaire par rapport à ses arguments* u_{j+k}^{n+m}.

La stabilité d'un schéma linéaire à deux niveaux est très facile à interpréter. En effet, par linéarité tout schéma linéaire à deux niveaux peut s'écrire sous la forme condensée

$$u^{n+1} = Au^n, \tag{2.12}$$

où A est un opérateur linéaire (une matrice, dite d'itération) de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Par exemple, pour le schéma explicite (2.2) la matrice A vaut

$$\begin{pmatrix} 1-2c & c & & & 0 \\ c & 1-2c & c & & \\ & \ddots & \ddots & \ddots & \\ & & c & 1-2c & c \\ 0 & & & c & 1-2c \end{pmatrix} \text{ avec } c = \frac{\nu\Delta t}{(\Delta x)^2}, \tag{2.13}$$

tandis que pour le schéma implicite (2.3) la matrice A est l'inverse de la matrice (2.4). A l'aide de cette matrice d'itération, on a $u^n = A^n u^0$ (attention, la notation A^n désigne ici la puissance n-ème de A), et par conséquent la stabilité du schéma est équivalente à

$$\|A^n u^0\| \leq K\|u^0\| \quad \forall\, n \geq 0, \forall\, u^0 \in \mathbb{R}^N.$$

Introduisant la norme matricielle subordonnée (voir la Définition 3.5.1)

$$\|M\| = \sup_{u\in\mathbb{R}^N, u\neq 0} \frac{\|Mu\|}{\|u\|},$$

la stabilité du schéma est équivalente à

$$\|A^n\| \leq K \quad \forall\, n \geq 0, \tag{2.14}$$

qui veut dire que la suite des puissances de A est bornée.

Stabilité en norme L^∞.

La stabilité en norme L^∞ est très liée avec le principe du maximum discret que nous avons déjà vu au Chapitre 1. Rappelons la définition de ce principe.

Définition 2.2.11 *Un schéma aux différences finies vérifie le* **principe du maximum discret** *si pour tout* $n \geq 0$ *et tout* $1 \leq j \leq N$ *on a*

$$\min\left(0, \min_{0\leq j\leq N+1} u_j^0\right) \leq u_j^n \leq \max\left(0, \max_{0\leq j\leq N+1} u_j^0\right)$$

quelle que soit la donnée initiale u^0.

Remarque 2.2.12 Dans la Définition 2.2.11 les inégalités tiennent compte non seulement du minimum et du maximum de u^0 mais aussi de zéro qui est la valeur imposée au bord par les conditions aux limites de Dirichlet. Cela est nécessaire si la donnée initiale u^0 ne vérifie pas les conditions aux limites de Dirichlet (ce qui n'est pas exigé), et inutile dans le cas contraire. •

Comme nous l'avons vu au Chapitre 1 (voir (1.32) et l'Exercice 1.4.1), la vérification du principe du maximum discret permet de démontrer le lemme suivant.

Lemme 2.2.13 *Le schéma explicite (2.2) est stable en norme L^∞ si et seulement si la condition CFL $2\nu\Delta t \leq (\Delta x)^2$ est satisfaite. Le schéma implicite (2.3) est stable en norme L^∞ quels que soient les pas de temps Δt et d'espace Δx (on dit qu'il est inconditionnellement stable).*

Exercice 2.2.2 Montrer que le schéma de Crank-Nicolson (2.5) (avec $\theta = 1/2$) est stable en norme L^∞ si $\nu\Delta t \leq (\Delta x)^2$, et que le schéma de DuFort-Frankel (2.6) est stable en norme L^∞ si $2\nu\Delta t \leq (\Delta x)^2$

Stabilité en norme L^2.

De nombreux schémas ne vérifient pas le principe du maximum discret mais sont néanmoins de "bons" schémas. Pour ceux-là, il faut vérifier la stabilité dans une autre norme que la norme L^∞. La norme L^2 se prête très bien à l'étude de la stabilité grâce à l'outil très puissant de l'analyse de Fourier que nous présentons maintenant. Pour ce faire, nous supposons désormais que les conditions aux limites pour l'équation de la chaleur sont des **conditions aux limites de périodicité**, qui s'écrivent $u(t, x+1) = u(t, x)$ pour tout $x \in [0, 1]$ et tout $t \geq 0$. Pour les schémas numériques, elles conduisent aux égalités $u_0^n = u_{N+1}^n$ pour tout $n \geq 0$, et plus généralement $u_j^n = u_{N+1+j}^n$. Il reste donc à calculer $N+1$ valeurs u_j^n.

A chaque vecteur $u^n = (u_j^n)_{0 \leq j \leq N}$ on associe une fonction $u^n(x)$, constante par morceaux, périodique de période 1, définie sur $[0, 1]$ par

$$u^n(x) = u_j^n \text{ si } x_{j-1/2} < x < x_{j+1/2}$$

avec $x_{j+1/2} = (j + 1/2)\Delta x$ pour $0 \leq j \leq N$, $x_{-1/2} = 0$, et $x_{N+1+1/2} = 1$. Ainsi définie, la fonction $u^n(x)$ appartient à $L^2(0, 1)$. Or, d'après l'analyse de Fourier, toute fonction de $L^2(0, 1)$ peut se décomposer en une somme de Fourier (voir [14], [13]). Plus précisément on a

$$u^n(x) = \sum_{k \in \mathbb{Z}} \hat{u}^n(k) \exp(2i\pi kx), \tag{2.15}$$

avec $\hat{u}^n(k) = \int_0^1 u^n(x) \exp(-2i\pi kx)\, dx$ et la formule de Plancherel

$$\int_0^1 |u^n(x)|^2\, dx = \sum_{k \in \mathbb{Z}} |\hat{u}^n(k)|^2. \tag{2.16}$$

Remarquons que même si u^n est une fonction réelle, les coefficients $\hat{u}^n(k)$ de la série de Fourier sont complexes. Une propriété importante pour la suite de la transformée

de Fourier des fonctions périodiques est la suivante : si on note $v^n(x) = u^n(x+\Delta x)$, alors $\hat{v}^n(k) = \hat{u}^n(k)\exp(2i\pi k\Delta x)$.

Expliquons maintenant la méthode sur l'exemple du schéma explicite (2.2). Avec nos notations, on peut réécrire ce schéma, pour $0 \leq x \leq 1$,

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} + \nu \frac{-u^n(x-\Delta x) + 2u^n(x) - u^n(x+\Delta x)}{(\Delta x)^2} = 0.$$

Par application de la transformée de Fourier, il vient

$$\hat{u}^{n+1}(k) = \left(1 - \frac{\nu\Delta t}{(\Delta x)^2}\left(-\exp(-2i\pi k\Delta x) + 2 - \exp(2i\pi k\Delta x)\right)\right)\hat{u}^n(k).$$

Autrement dit

$$\hat{u}^{n+1}(k) = A(k)\hat{u}^n(k) = A(k)^{n+1}\hat{u}^0(k) \text{ avec } A(k) = 1 - \frac{4\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2.$$

Pour $k \in \mathbb{Z}$, le coefficient de Fourier $\hat{u}^n(k)$ est borné lorsque n tend vers l'infini si et seulement si le facteur d'amplification vérifie $|A(k)| \leq 1$, c'est-à-dire

$$2\nu\Delta t(\sin(\pi k\Delta x))^2 \leq (\Delta x)^2. \tag{2.17}$$

Si la condition CFL (1.30), i.e. $2\nu\Delta t \leq (\Delta x)^2$, est satisfaite, alors l'inégalité (2.17) est vraie quelque soit le mode de Fourier $k \in \mathbb{Z}$, et par la formule de Plancherel on en déduit

$$\|u^n\|_2^2 = \int_0^1 |u^n(x)|^2 dx = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 \leq \sum_{k\in\mathbb{Z}} |\hat{u}^0(k)|^2 = \int_0^1 |u^0(x)|^2 dx = \|u^0\|_2^2,$$

ce qui n'est rien d'autre que la stabilité L^2 du schéma explicite. Si la condition CFL n'est pas satisfaite, le schéma est instable. En effet, il suffit de choisir Δx (éventuellement suffisamment petit) et k_0 (suffisamment grand) et une donnée initiale ayant une seule composante de Fourier non nulle $\hat{u}^0(k_0) \neq 0$ avec $\pi k_0\Delta x \approx \pi/2$ (modulo π) de telle manière que $|A(k_0)| > 1$. On a donc démontré le lemme suivant.

Lemme 2.2.14 *Le schéma explicite (2.2) est stable en norme L^2 si et seulement si la condition CFL $2\nu\Delta t \leq (\Delta x)^2$ est satisfaite.*

De la même façon on va démontrer la stabilité du schéma implicite.

Lemme 2.2.15 *Le schéma implicite (2.3) est stable en norme L^2.*

Démonstration. Un raisonnement analogue à celui utilisé pour le schéma explicite conduit, pour $0 \leq x \leq 1$, à

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} + \nu \frac{-u^{n+1}(x-\Delta x) + 2u^{n+1}(x) - u^{n+1}(x+\Delta x)}{(\Delta x)^2} = 0,$$

et par application de la transformée de Fourier

$$\hat{u}^{n+1}(k)\left(1+\frac{\nu\Delta t}{(\Delta x)^2}\left(-\exp(-2i\pi k\Delta x)+2-\exp(2i\pi k\Delta x)\right)\right)=\hat{u}^n(k).$$

Autrement dit

$$\hat{u}^{n+1}(k)=A(k)\hat{u}^n(k)=A(k)^{n+1}\hat{u}^0(k) \text{ avec } A(k)=\left(1+\frac{4\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2\right)^{-1}.$$

Comme $|A(k)|\leq 1$ pour tout mode de Fourier k, la formule de Plancherel permet de conclure à la stabilité L^2 du schéma. □

Remarque 2.2.16 (Essentielle d'un point de vue pratique) Traduisons sous forme de "recette" la méthode de l'analyse de Fourier pour prouver la stabilité L^2 d'un schéma. On injecte dans le schéma un mode de Fourier

$$u_j^n=A(k)^n\exp(2i\pi kx_j) \quad \text{avec} \quad x_j=j\Delta x,$$

et on en déduit la valeur du facteur d'amplification $A(k)$. Rappelons que, pour l'instant, nous nous sommes limités au cas scalaire, c'est-à-dire que $A(k)$ est un nombre complexe dans $\mathbb{C}$. On appelle **condition de stabilité de Von Neumann** l'inégalité

$$|A(k)|\leq 1 \text{ pour tout mode } k\in\mathbb{Z}. \tag{2.18}$$

Si la condition de stabilité de Von Neumann est satisfaite (avec éventuellement des restrictions sur Δt et Δx), alors le schéma est stable pour la norme L^2, si non il est instable.

En général, un schéma stable (et consistant) est convergent (voir la Sous-section 2.2.4). En pratique, un schéma instable est totalement "inutilisable". En effet, même si on part d'une donnée initiale spécialement préparée de manière à ce qu'aucun des modes de Fourier instables ne soit excité par elle, les inévitables erreurs d'arrondi vont créer des composantes non nulles (bien que très petites) de la solution sur ces modes instables. La croissance exponentielle des modes instables entraîne qu'après seulement quelques pas en temps ces "petits" modes deviennent "énormes" et polluent complètement le reste de la solution numérique. •

Exercice 2.2.3 Montrer que le θ-schéma (2.5) est stable en norme L^2 inconditionnellement si $1/2\leq\theta\leq 1$, et sous la condition CFL $2(1-2\theta)\nu\Delta t\leq(\Delta x)^2$ si $0\leq\theta<1/2$.

2.2.4 Convergence des schémas

Nous avons maintenant tous les outils pour démontrer la convergence des schémas de différences finies. Le résultat principal de cette sous-section est le Théorème de Lax qui affirme que, pour un schéma linéaire, **consistance et stabilité impliquent convergence**. La portée de ce résultat dépasse en fait de beaucoup la méthode des différences finies. Pour toute méthode numérique (différences finies, éléments finis, etc.) la convergence se démontre en conjuguant deux arguments :

stabilité et consistance (leurs définitions précises varient d'une méthode à l'autre). D'un point de vue pratique, le Théorème de Lax est très rassurant : si l'on utilise un schéma consistant (ils sont construits pour cela en général) et que l'on n'observe pas d'oscillations numériques (c'est-à-dire qu'il est stable), alors la solution numérique est proche de la solution exacte (le schéma converge).

Théorème 2.2.17 (Lax) *Soit $u(t,x)$ la solution suffisamment régulière de l'équation de la chaleur (2.1) (avec des conditions aux limites appropriées). Soit u_j^n la solution numérique discrète obtenue par un schéma de différences finies avec la donnée initiale $u_j^0 = u_0(x_j)$. On suppose que le schéma est linéaire, à deux niveaux, consistant, et stable pour une norme $\|\ \|$. Alors le schéma est convergent au sens où*

$$\forall\, T > 0, \quad \lim_{\Delta t, \Delta x \to 0} \left(\sup_{t_n \leq T} \|e^n\| \right) = 0, \tag{2.19}$$

avec e^n le vecteur "erreur" défini par ses composantes $e_j^n = u_j^n - u(t_n, x_j)$.

De plus, si le schéma est précis à l'ordre p en espace et à l'ordre q en temps, alors pour tout temps $T > 0$ il existe une constante $C_T > 0$ telle que

$$\sup_{t_n \leq T} \|e^n\| \leq C_T \Big((\Delta x)^p + (\Delta t)^q \Big). \tag{2.20}$$

Remarque 2.2.18 Nous n'avons pas démontré l'existence et l'unicité de la solution de l'équation de la chaleur (2.1) (avec des conditions aux limites de Dirichlet ou périodiques). Ici, nous nous contentons d'admettre ce résultat qui est vrai et dont on peut trouver la démonstration dans [1], [3], [13]. •

Démonstration. Pour simplifier, on suppose que les conditions aux limites sont de Dirichlet. La même démonstration est aussi valable pour des conditions aux limites de périodicité ou des conditions aux limites de Neumann (en supposant ces dernières discrétisées avec le même ordre de précision que le schéma). Un schéma linéaire à deux niveaux peut s'écrire sous la forme condensée (2.12), i.e.

$$u^{n+1} = Au^n,$$

où A est la matrice d'itération (carrée de taille N). Soit u la solution (supposée suffisamment régulière) de l'équation de la chaleur (2.1). On note $\tilde{u}^n = (\tilde{u}_j^n)_{1 \leq j \leq N}$ avec $\tilde{u}_j^n = u(t_n, x_j)$. Comme le schéma est consistant, il existe un vecteur ϵ^n tel que

$$\tilde{u}^{n+1} = A\tilde{u}^n + \Delta t \epsilon^n \text{ avec } \lim_{\Delta t, \Delta x \to 0} \|\epsilon^n\| = 0, \tag{2.21}$$

et la convergence de ϵ^n est uniforme pour tous les temps $0 \leq t_n \leq T$. Si le schéma est précis à l'ordre p en espace et à l'ordre q en temps, alors $\|\epsilon^n\| \leq C((\Delta x)^p + (\Delta t)^q)$. En posant $e_j^n = u_j^n - u(t_n, x_j)$ on obtient par soustraction de (2.21) à (2.12)

$$e^{n+1} = Ae^n - \Delta t \epsilon^n$$

d'où par récurrence

$$e^n = A^n e^0 - \Delta t \sum_{k=1}^{n} A^{n-k} \epsilon^{k-1}. \tag{2.22}$$

Or, la stabilité du schéma veut dire que $\|u^n\| = \|A^n u^0\| \leq K\|u^0\|$ pour toute donnée initiale, c'est-à-dire que $\|A^n\| \leq K$ où la constante K ne dépend pas de n. D'autre part, $e^0 = 0$, donc (2.22) donne

$$\|e^n\| \leq \Delta t \sum_{k=1}^{n} \|A^{n-k}\| \|\epsilon^{k-1}\| \leq \Delta t n K C \Big((\Delta x)^p + (\Delta t)^q \Big),$$

ce qui donne l'inégalité (2.20) avec la constante $C_T = TKC$. La démonstration de (2.19) est similaire. □

Remarque 2.2.19 Le Théorème de Lax 2.2.17 est en fait valable pour toute équation aux dérivées partielles linéaire. Il admet une réciproque au sens où un schéma linéaire consistant à deux niveaux qui converge est nécessairement stable. Remarquer que la vitesse de convergence dans (2.20) est exactement la précision du schéma. Enfin, il est bon de noter que cette estimation (2.20) n'est valable que sur un intervalle borné de temps $[0, T]$ mais qu'elle est indépendante du nombre de points de discrétisation N. •

2.2.5 Schémas multiniveaux

Jusqu'ici nous avons principalement analysé des schémas à deux niveaux, c'est-à-dire des schémas qui relient les valeurs de u^{n+1} à celles de u^n seulement. On peut parfaitement envisager des schémas multiniveaux, et en particulier nous avons déjà introduit des schémas à trois niveaux où u^{n+1} dépend de u^n et u^{n-1} (comme les schémas de Richardson, de DuFort-Frankel, ou de Gear). Examinons comment les résultats précédents se généralisent aux schémas multiniveaux (nous nous limitons par souci de clarté aux schémas à trois niveaux).

La Définition 2.2.8 de la stabilité d'un schéma est indépendante de son nombre de niveaux. Toutefois, l'interprétation de la stabilité en terme de matrice d'itération est un peu plus compliquée pour un schéma linéaire à trois niveaux. En effet, u^{n+1} dépend linéairement de u^n et u^{n-1}, donc on ne peut pas écrire la relation (2.12). Par contre, si on pose

$$U^n = \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}, \tag{2.23}$$

alors il existe deux matrices d'ordre N, A_1 et A_2, telles que

$$U^{n+1} = A\, U^n = \begin{pmatrix} A_1 & A_2 \\ \mathrm{Id} & 0 \end{pmatrix} U^n, \tag{2.24}$$

où la matrice d'itération A est donc de taille $2N$. Comme précédemment, $U^n = A^n U^1$ et la stabilité est équivalente à

$$\|A^n\| = \sup_{U^1 \in \mathbb{R}^{2N}, U^1 \neq 0} \frac{\|A^n U^1\|}{\|U^1\|} \leq K \quad \forall\, n \geq 1.$$

De la même manière la méthode d'analyse de Fourier s'étend aux schémas à trois niveaux grâce à la notation vectorielle (2.23). En guise d'exemple, nous démontrons un résultat pressenti au Chapitre 1.

Lemme 2.2.20 *Le schéma centré (1.27) est instable en norme L^2.*

Démonstration. Avec les notations usuelles le schéma (1.27) s'écrit, pour $x \in [0,1]$,

$$\frac{u^{n+1}(x) - u^{n-1}(x)}{2\Delta t} + \nu \frac{-u^n(x-\Delta x) + 2u^n(x) - u^n(x+\Delta x)}{(\Delta x)^2} = 0,$$

et par application de la transformée de Fourier

$$\hat{u}^{n+1}(k) + \frac{8\nu\Delta t}{(\Delta x)^2}(\sin(\pi k \Delta x))^2 \hat{u}^n(k) - \hat{u}^{n-1}(k) = 0.$$

Autrement dit,

$$\hat{U}^{n+1}(k) = \begin{pmatrix} \hat{u}^{n+1}(k) \\ \hat{u}^n(k) \end{pmatrix} = \begin{pmatrix} -\frac{8\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2 & 1 \\ 1 & 0 \end{pmatrix} \hat{U}^n(k) = A(k)\hat{U}^n(k),$$

et $\hat{U}^{n+1}(k) = A(k)^n \hat{U}^1(k)$. Ici, $A(k)$ est une matrice d'ordre 2 alors que pour les schémas à deux niveaux c'était un scalaire. Pour $k \in \mathbb{Z}$, le vecteur $\hat{U}^n(k)$, et donc le coefficient de Fourier $\hat{u}^n(k)$, est borné lorsque n tend vers l'infini si et seulement si la matrice d'amplification vérifie

$$\|A(k)^n\|_2 = \sup_{U \in \mathbb{R}^2, U \neq 0} \frac{\|A(k)^n U\|_2}{\|U\|_2} \leq K \quad \forall\, n \geq 1, \tag{2.25}$$

où $\|U\|_2$ est la norme euclidienne dans $\mathbb{R}^2$ et K est une constante bornée indépendante de n et k. Par conséquent, si l'inégalité (2.25) est vraie quelque soit le mode de Fourier $k \in \mathbb{Z}$, par la formule de Plancherel on en déduit

$$\|u^n\|_2^2 = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 \leq K \sum_{k\in\mathbb{Z}} \left(|\hat{u}^0(k)|^2 + |\hat{u}^1(k)|^2\right) = \|u^0\|_2^2 + \|u^1\|_2^2,$$

c'est-à-dire la stabilité L^2 du schéma. A l'inverse, s'il existe k_0 tel que $\|A(k_0)^n\|$ n'est pas borné lorsque n tend vers l'infini, alors en choisissant convenablement la donnée initiale avec un seul mode $\hat{u}^0(k_0)$ (ainsi que $\hat{u}^1(k_0)$), on obtient l'instabilité L^2 du schéma.

Comme la matrice d'amplification $A(k)$ est symétrique réelle, on a la propriété $\|A(k)\|_2 = \rho(A(k))$ et $\|A(k)^n\|_2 = \|A(k)\|_2^n$, où $\rho(M)$ désigne le rayon spectral de la matrice M (voir le Lemme 3.5.5). Donc, l'inégalité (2.25) est satisfaite si et seulement si $\rho(A(k)) \leq 1$. Les valeurs propres de $A(k)$ sont les racines du polynôme du deuxième degré

$$\lambda^2 + \frac{8\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2 \lambda - 1 = 0$$

qui admet toujours deux racines réelles distinctes dont le produit vaut -1. Par conséquent, l'une des deux racines est plus grande que 1 (strictement) en valeur absolue, et donc $\rho(A(k)) > 1$. Par conséquent, le schéma centré est inconditionnellement instable en norme L^2. □

Remarque 2.2.21 La méthode de l'analyse de Fourier que nous venons d'utiliser dans la démonstration du Lemme 2.2.20 est un peu plus compliquée dans le cas des schémas multiniveaux que dans le cas à deux niveaux (voir la Remarque 2.2.16). Lorsqu'on injecte dans le schéma un mode de Fourier, on obtient

$$\begin{pmatrix} u_j^{n+1} \\ u_j^n \end{pmatrix} = A(k)^n \begin{pmatrix} u_j^1 \\ u_j^0 \end{pmatrix} \exp(2i\pi k x_j)$$

où $A(k)$ est désormais une **matrice** d'amplification (et non plus un facteur scalaire). On appelle **condition de stabilité de Von Neumann** la condition

$$\rho(A(k)) \leq 1 \text{ pour tout mode } k \in \mathbb{Z}, \tag{2.26}$$

où $\rho(A(k))$ est le rayon spectral de la matrice $A(k)$. Comme pour une matrice quelconque B on a

$$\|B\| \geq \rho(B) \text{ et } \|B^n\| \geq \rho(B)^n,$$

il est clair que la condition de stabilité de Von Neumann est une **condition nécessaire** de stabilité L^2 du schéma (donc de convergence). Lorsque la matrice $A(k)$ est normale, elle vérifie $\|A(k)\|_2 = \rho(A(k))$ et $\|A(k)^n\|_2 = \|A(k)\|_2^n$ (voir le Lemme 3.5.5), donc la condition de Von Neumann (2.26) est nécessaire et suffisante (nous avons eu la "chance" lors de la démonstration du Lemme 2.2.20 de tomber dans ce cas favorable). Cependant, si $A(k)$ n'est pas normale, alors en général la condition de stabilité de Von Neumann n'est **pas suffisante** et il faut faire une analyse beaucoup plus délicate de $A(k)$ (et notamment de sa diagonalisation ou non). •

Exercice 2.2.4 Montrer que le schéma de Gear (2.7) est inconditionnellement stable et donc convergent en norme L^2.

Exercice 2.2.5 Montrer que le schéma de DuFort-Frankel (2.6) est stable en norme L^2, et donc convergent, si le rapport $\Delta t/(\Delta x)^2$ reste borné lorsqu'on fait tendre Δt et Δx vers 0.

2.2.6 Le cas multidimensionnel

La méthode des différences finies s'étend sans difficulté aux problèmes en plusieurs dimensions d'espace. Considérons par exemple l'équation de la chaleur en deux dimensions d'espace (le cas de trois ou plus dimensions d'espace n'est pas plus compliqué, du moins en théorie) dans le domaine rectangulaire $\Omega = (0,1) \times (0,L)$ avec des conditions aux limites de Dirichlet

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} - \nu \dfrac{\partial^2 u}{\partial y^2} = 0 \text{ pour } (x,y,t) \in \Omega \times \mathbb{R}_*^+ \\ u(t=0,x,y) = u_0(x,y) \text{ pour } (x,y) \in \Omega \\ u(t,x,y) = 0 \text{ pour } t \in \mathbb{R}_*^+, (x,y) \in \partial\Omega. \end{cases} \tag{2.27}$$

Pour discrétiser le domaine Ω, on introduit deux pas d'espace $\Delta x = 1/(N_x+1) > 0$ et $\Delta y = L/(N_y+1) > 0$ (avec N_x et N_y deux entiers positifs). Avec le pas de temps $\Delta t > 0$, on définit ainsi les noeuds d'un maillage régulier

$$(t_n, x_j, y_k) = (n\Delta t, j\Delta x, k\Delta y) \text{ pour } n \geq 0, 0 \leq j \leq N_x+1, 0 \leq k \leq N_y+1.$$

On note $u^n_{j,k}$ la valeur d'une solution discrète approchée au point (t_n, x_j, y_k), et $u(t,x,y)$ la solution exacte de (2.27).

Les conditions aux limites de Dirichlet se traduisent, pour $n > 0$, en

$$u^n_{0,k} = u^n_{N_x+1,k} = 0, \ \forall\, k, \quad \text{et} \quad u^n_{j,0} = u^n_{j,N_y+1} = 0, \ \forall\, j.$$

La donnée initiale est discrétisée par

$$u^0_{j,k} = u_0(x_j, y_k) \ \forall\, j, k.$$

La généralisation au cas bidimensionnel du **schéma explicite** est évidente

$$\frac{u^{n+1}_{j,k} - u^n_{j,k}}{\Delta t} + \nu \frac{-u^n_{j-1,k} + 2u^n_{j,k} - u^n_{j+1,k}}{(\Delta x)^2} + \nu \frac{-u^n_{j,k-1} + 2u^n_{j,k} - u^n_{j,k+1}}{(\Delta y)^2} = 0 \qquad (2.28)$$

pour $n \geq 0$, $j \in \{1, ..., N_x\}$ et $k \in \{1, ..., N_y\}$. La seule différence notable avec le cas unidimensionnel est le caractère deux fois plus sévère de la condition CFL.

Exercice 2.2.6 Montrer que le schéma explicite (2.28) est stable en norme L^∞ (et même qu'il vérifie le principe du maximum) sous la condition CFL

$$\frac{\nu \Delta t}{(\Delta x)^2} + \frac{\nu \Delta t}{(\Delta y)^2} \leq \frac{1}{2}.$$

De même, on a le **schéma implicite**

$$\frac{u^{n+1}_{j,k} - u^n_{j,k}}{\Delta t} + \nu \frac{-u^{n+1}_{j-1,k} + 2u^{n+1}_{j,k} - u^{n+1}_{j+1,k}}{(\Delta x)^2} + \nu \frac{-u^{n+1}_{j,k-1} + 2u^{n+1}_{j,k} - u^{n+1}_{j,k+1}}{(\Delta y)^2} = 0. \quad (2.29)$$

Remarquons que le schéma implicite nécessite, pour calculer u^{n+1} en fonction de u^n, la résolution d'un système linéaire sensiblement plus compliqué qu'en une dimension d'espace (la situation serait encore pire en trois dimensions). Rappelons qu'en dimension un, il suffit d'inverser une matrice tridiagonale. Nous allons voir qu'en dimension deux la matrice a une structure moins simple. L'inconnue discrète $u^n_{j,k}$ est indicée par deux entiers j et k, mais en pratique on utilise un seul indice pour stocker u^n sous la forme d'un vecteur dans l'ordinateur. Une manière (simple et efficace) de ranger dans un seul vecteur les inconnues $u^n_{j,k}$ est d'écrire

$$u^n = (u^n_{1,1}, ..., u^n_{1,N_y}, u^n_{2,1}, ..., u^n_{2,N_y}, ..., u^n_{N_x,1}, ..., u^n_{N_x,N_y}).$$

Remarquons qu'on a rangé les inconnues "colonne par colonne", mais qu'on aurait aussi bien pu le faire "ligne par ligne" en "déroulant" d'abord l'indice j plutôt que k (N_x est le nombre de colonnes et N_y celui de lignes). Avec cette convention, le schéma implicite (2.29) requiert l'inversion de la matrice symétrique tridiagonale "par blocs"

$$M = \begin{pmatrix} D_1 & E_1 & & & 0 \\ E_1 & D_2 & E_2 & & \\ & \ddots & \ddots & \ddots & \\ & & E_{N_x-2} & D_{N_x-1} & E_{N_x-1} \\ 0 & & & E_{N_x-1} & D_{N_x} \end{pmatrix}$$

où les blocs diagonaux D_j sont des matrices carrées de taille N_y

$$D_j = \begin{pmatrix} 1+2(c_y+c_x) & -c_y & & & 0 \\ -c_y & 1+2(c_y+c_x) & -c_y & & \\ & \ddots & \ddots & \ddots & \\ & & -c_y & 1+2(c_y+c_x) & -c_y \\ 0 & & & -c_y & 1+2(c_y+c_x) \end{pmatrix}$$

avec $c_x = \frac{\nu\Delta t}{(\Delta x)^2}$ et $c_y = \frac{\nu\Delta t}{(\Delta y)^2}$, et les blocs extra-diagonaux $E_j = (E_j)^t$ sont des matrices carrées de taille N_y

$$E_j = \begin{pmatrix} -c_x & 0 & & & 0 \\ 0 & -c_x & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & -c_x & 0 \\ 0 & & & 0 & -c_x \end{pmatrix}.$$

Au total la matrice M est pentadiagonale et symétrique. Cependant les cinq diagonales ne sont pas contiguës, ce qui entraîne une augmentation considérable du coût de la résolution d'un système linéaire associé à M (voir l'annexe sur l'analyse numérique matricielle et notamment les Remarques 3.5.15 et 3.5.24). La situation serait encore pire en trois dimensions.

Exercice 2.2.7 Montrer que le schéma de Peaceman-Rachford

$$\frac{u_{j,k}^{n+1/2} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^n + 2u_{j,k}^n - u_{j,k+1}^n}{2(\Delta y)^2} = 0$$

$$\frac{u_{j,k}^{n+1} - u_{j,k}^{n+1/2}}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^{n+1} + 2u_{j,k}^{n+1} - u_{j,k+1}^{n+1}}{2(\Delta y)^2} = 0.$$

est précis d'ordre 2 en espace et temps et inconditionnellement stable en norme L^2 (pour des conditions aux limites de périodicité dans chaque direction).

A cause de son coût de calcul élevé, on remplace souvent le schéma implicite par une généralisation à plusieurs dimensions d'espace de schémas unidimensionnels, obtenue par une technique de **directions alternées** (dite aussi de séparation d'opérateurs, ou **splitting** en anglais). L'idée est de résoudre, au lieu de l'équation bidimensionnelle (2.27), alternativement les deux équations unidimensionnelles

$$\frac{\partial u}{\partial t} - 2\nu \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{et} \quad \frac{\partial u}{\partial t} - 2\nu \frac{\partial^2 u}{\partial y^2} = 0$$

dont la "moyenne" redonne (2.27). Par exemple, en utilisant dans chaque direction un schéma de Crank-Nicolson pour un demi pas de temps $\Delta t/2$, on obtient un **schéma**

de directions alternées

$$\frac{u_{j,k}^{n+1/2}-u_{j,k}^{n}}{\Delta t}+\nu\frac{-u_{j-1,k}^{n+1/2}+2u_{j,k}^{n+1/2}-u_{j+1,k}^{n+1/2}}{2(\Delta x)^2}+\nu\frac{-u_{j-1,k}^{n}+2u_{j,k}^{n}-u_{j+1,k}^{n}}{2(\Delta x)^2}=0$$

$$\frac{u_{j,k}^{n+1}-u_{j,k}^{n+1/2}}{\Delta t}+\nu\frac{-u_{j,k-1}^{n+1}+2u_{j,k}^{n+1}-u_{j,k+1}^{n+1}}{2(\Delta y)^2}+\nu\frac{-u_{j,k-1}^{n+1/2}+2u_{j,k}^{n+1/2}-u_{j,k+1}^{n+1/2}}{2(\Delta y)^2}=0 \tag{2.30}$$

L'avantage de ce type de schéma est qu'il suffit, à chaque demi pas de temps, d'inverser une matrice tridiagonale de type unidimensionnel (c'est donc un calcul peu cher). En trois dimensions, il suffit de faire trois tiers-pas de temps et les propriétés du schéma sont inchangées. Ce schéma est non seulement stable mais consistant avec l'équation bidimensionnelle (2.27).

Exercice 2.2.8 Montrer que le schéma de directions alternées (2.30) est précis d'ordre 2 en espace et temps et inconditionnellement stable en norme L^2 (pour des conditions aux limites de périodicité dans chaque direction).

2.3 Autres modèles

2.3.1 Équation d'advection

Nous considérons l'équation d'advection en une dimension d'espace dans le domaine borné $(0,1)$ avec une vitesse constante $V>0$ et des conditions aux limites de périodicité

$$\begin{cases} \dfrac{\partial u}{\partial t}+V\dfrac{\partial u}{\partial x}=0 \text{ pour } (x,t)\in(0,1)\times\mathbb{R}_*^+ \\ u(t,x+1)=u(t,x) \text{ pour } (x,t)\in(0,1)\times\mathbb{R}_*^+ \\ u(0,x)=u_0(x) \text{ pour } x\in(0,1). \end{cases} \tag{2.31}$$

On discrétise toujours l'espace avec un pas $\Delta x=1/(N+1)>0$ (N entier positif) et le temps avec $\Delta t>0$, et on note $(t_n,x_j)=(n\Delta t,j\Delta x)$ pour $n\geq 0, j\in\{0,1,...,N+1\}$, u_j^n la valeur d'une solution discrète approchée au point (t_n,x_j), et $u(t,x)$ la solution exacte de (2.31). Les conditions aux limites de périodicité conduisent aux égalités $u_0^n=u_{N+1}^n$ pour tout $n\geq 0$, et plus généralement $u_j^n=u_{N+1+j}^n$. Par conséquent, l'inconnue discrète à chaque pas de temps est un vecteur $u^n=(u_j^n)_{0\leq j\leq N}\in\mathbb{R}^{N+1}$. Nous donnons quelques schémas possibles pour l'équation d'advection (2.31). Au Chapitre 1 nous avons déjà constaté le mauvais comportement numérique du **schéma explicite centré**

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+V\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}=0 \tag{2.32}$$

pour $n\geq 0$ et $j\in\{0,...,N\}$. Le caractère instable de ce schéma est confirmé par le lemme suivant.

Lemme 2.3.1 *Le schéma explicite centré (2.32) est consistant avec l'équation d'advection (2.31), précis à l'ordre 1 en temps et 2 en espace, mais inconditionnellement instable en norme L^2.*

Démonstration. A l'aide d'un développement de Taylor autour du point (t_n, x_j), on vérifie facilement que le schéma est consistant, précis à l'ordre 1 en temps et 2 en espace. Par analyse de Fourier, on étudie la stabilité L^2. Avec les notations de la Sous-section 2.2.3, les composantes de Fourier $\hat{u}^n(k)$ de u^n vérifient

$$\hat{u}^{n+1}(k) = \left(1 - i\frac{V\Delta t}{\Delta x}\sin(2\pi k\Delta x)\right)\hat{u}^n(k) = A(k)\hat{u}^n(k).$$

On vérifie que le module du facteur d'amplification est toujours plus grand que 1,

$$|A(k)|^2 = 1 + \left(\frac{V\Delta t}{\Delta x}\sin(2\pi k\Delta x)\right)^2 \geq 1,$$

avec inégalité stricte dès que $2k\Delta x$ n'est pas entier. Donc le schéma est instable. □

On peut écrire une version implicite du précédent schéma qui devient stable : c'est le **schéma implicite centré**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} = 0. \tag{2.33}$$

Exercice 2.3.1 Montrer que le schéma implicite centré (2.33) est consistant avec l'équation d'advection (2.31), précis à l'ordre 1 en temps et 2 en espace, inconditionnellement stable en norme L^2, donc convergent.

Si l'on tient absolument à rester centré et explicite, le **schéma de Lax-Friedrichs**

$$\frac{2u_j^{n+1} - u_{j+1}^n - u_{j-1}^n}{2\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \tag{2.34}$$

est un schéma simple, robuste, mais pas très précis.

Lemme 2.3.2 *Le schéma de Lax-Friedrichs (2.34) est stable en norme L^2 sous la condition CFL*

$$|V|\Delta t \leq \Delta x.$$

Si le rapport $\Delta t/\Delta x$ est gardé constant lorsque Δt et Δx tendent vers zéro, il est consistant avec l'équation d'advection (2.31) et précis à l'ordre 1 en espace et temps. Par conséquent, il est conditionnellement convergent.

Démonstration. Par analyse de Fourier on a

$$\hat{u}^{n+1}(k) = \left(\cos(2\pi k\Delta x) - i\frac{V\Delta t}{\Delta x}\sin(2\pi k\Delta x)\right)\hat{u}^n(k) = A(k)\hat{u}^n(k).$$

Le module du facteur d'amplification est donné par

$$|A(k)|^2 = \cos^2(2\pi k\Delta x) + \left(\frac{V\Delta t}{\Delta x}\right)^2 \sin^2(2\pi k\Delta x).$$

On vérifie donc que $|A(k)| \leq 1$ pour tout k si la condition $|V|\Delta t \leq \Delta x$ est satisfaite, tandis qu'il existe des modes instables k tels que $|A(k)| > 1$ si non. Le schéma est donc conditionnellement stable. Pour étudier la consistance, on effectue un développement de Taylor autour de (t_n, x_j) pour la solution u :

$$\frac{2u(t_{n+1},x_j) - u(t_n,x_{j+1}) - u(t_n,x_{j-1})}{2\Delta t} + V\frac{u(t_n,x_{j+1}) - u(t_n,x_{j-1})}{2\Delta x} =$$

$$(u_t + Vu_x)(t_n,x_j) - \frac{(\Delta x)^2}{2\Delta t}\left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right)u_{xx}(t_n,x_j) + \mathcal{O}\Big((\Delta x)^2 + \frac{(\Delta x)^4}{\Delta t}\Big). \quad (2.35)$$

Comme l'erreur de troncature contient un terme en $\mathcal{O}\Big((\Delta x)^2/\Delta t\Big)$, le schéma n'est pas consistant si Δt tend vers zéro plus vite que $(\Delta x)^2$. Par contre, il est consistant et précis d'ordre 1 si le rapport $\Delta t/\Delta x$ est constant. Pour obtenir la convergence on reprend la démonstration du Théorème de Lax 2.2.17. L'erreur e^n est toujours majorée par l'erreur de troncature, et donc ici

$$\|e^n\| \leq \Delta t n K C\Big(\frac{(\Delta x)^2}{\Delta t} + \Delta t\Big).$$

Si on garde fixe le rapport $\Delta x/\Delta t$ l'erreur est donc majoré par une constante fois Δt qui tend bien vers zéro, d'où la convergence. □

Remarque 2.3.3 Le schéma de Lax-Friedrichs n'est pas consistant (stricto sensu suivant la Définition 2.2.4). Néanmoins il est conditionnellement consistant et convergent. Il faut cependant faire attention que si on prend le pas de temps Δt beaucoup plus petit que ce qui est permis par la condition CFL de stabilité, la convergence sera très lente. En pratique le schéma de Lax-Friedrichs n'est pas recommandé. •

Un schéma centré, explicite, plus précis est le **schéma de Lax-Wendroff**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \left(\frac{V^2\Delta t}{2}\right)\frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0. \quad (2.36)$$

Sa dérivation n'est pas immédiate, aussi nous la présentons en détail. On commence par écrire un développement à l'ordre 2 en temps de la solution exacte

$$u(t_{n+1},x_j) = u(t_n,x_j) + (\Delta t)u_t(t_n,x_j) + \frac{(\Delta t)^2}{2}u_{tt}(t_n,x_j) + \mathcal{O}\Big((\Delta t)^3\Big).$$

En utilisant l'équation d'advection on remplace les dérivées en temps par des dérivées en espace

$$u(t_{n+1},x_j) = u(t_n,x_j) - (V\Delta t)u_x(t_n,x_j) + \frac{(V\Delta t)^2}{2}u_{xx}(t_n,x_j) + \mathcal{O}\Big((\Delta t)^3\Big).$$

Enfin, on remplace les dérivées en espace par des formules centrées d'ordre 2

$$\begin{aligned} u(t_{n+1},x_j) &= u(t_n,x_j) - V\Delta t\frac{u(t_n,x_{j+1}) - u(t_n,x_{j-1})}{2\Delta x} \\ &\quad + \frac{(V\Delta t)^2}{2}\frac{u(t_n,x_{j+1}) - 2u(t_n,x_j) + u(t_n,x_{j-1})}{(\Delta x)^2} + \mathcal{O}\Big((\Delta t)^3 + \Delta t(\Delta x)^2\Big). \end{aligned}$$

On retrouve bien le schéma de Lax-Wendroff en négligeant les termes du troisième ordre et en remplaçant $u(t_n, x_j)$ par u_j^n. Remarquer que, par rapport aux précédents schémas, on a discrétisé "simultanément" les dérivées en espace et en temps de l'équation d'advection. Par construction, le schéma de Lax-Wendroff est précis à l'ordre 2 en temps et en espace. On peut montrer qu'il ne vérifie pas le principe du maximum discret (voir l'Exercice 2.3.3). Par contre, il est stable en norme L^2 et donc convergent sous la condition CFL $|V|\Delta t \leq \Delta x$.

Exercice 2.3.2 Montrer que le schéma de Lax-Wendroff est stable et convergent en norme L^2 si $|V|\Delta t \leq \Delta x$.

Exercice 2.3.3 Montrer que le schéma de Lax-Friedrichs préserve le principe du maximum discret si la condition CFL $|V|\Delta t \leq \Delta x$ est satisfaite, tandis que le schéma de Lax-Wendroff ne le préserve pas sauf si $V\Delta t/\Delta x$ vaut $-1, 0$, ou 1.

Exercice 2.3.4 Montrer que le schéma de Lax-Wendroff (2.36) est le seul schéma précis à l'ordre 2 en espace et temps qui soit du type

$$u_j^{n+1} = \alpha u_{j-1}^n + \beta u_j^n + \gamma u_{j+1}^n,$$

où α, β, γ dépendent seulement de $V\Delta t/\Delta x$.

Comme nous l'avons déjà dit au Chapitre 1, une idée fondamentale pour obtenir de "bons" schémas pour l'équation d'advection (2.31) est le **décentrement amont**. Nous donnons la forme générale du **schéma décentré amont**

$$\begin{aligned} &\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad \text{si} \quad V > 0 \\ &\frac{u_j^{n+1} - u_j^n}{\Delta t} + V\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \quad \text{si} \quad V < 0. \end{aligned} \tag{2.37}$$

On a déjà vu au Chapitre 1 que le schéma décentré amont est stable en norme L^∞ si la condition CFL, $|V|\Delta t \leq \Delta x$, est satisfaite. Comme il est consistant et précis d'ordre 1 en espace et temps, il converge en norme L^∞ d'après le théorème de Lax. Le même résultat est vrai en norme L^2 avec la même condition CFL.

Exercice 2.3.5 Montrer que le schéma explicite décentré amont (2.37) est consistant avec l'équation d'advection (2.31), précis à l'ordre 1 en espace et temps, stable et convergent en norme L^2 si la condition CFL $|V|\Delta t \leq \Delta x$ est satisfaite.

Pour comparer ces divers schémas d'un point de vue pratique, un concept pertinent (quoique formel) est celui d'équation équivalente.

Définition 2.3.4 *On appelle* **équation équivalente** *d'un schéma l'équation obtenue en ajoutant au modèle étudié la partie principale (c'est-à-dire le terme d'ordre dominant) de l'erreur de troncature du schéma.*

Schéma	Stabilité	Erreur de troncature
Explicite centré (2.32)	instable	$\mathcal{O}\Big(\Delta t + (\Delta x)^2\Big)$
Implicite centré(2.33)	stable L^2	$\mathcal{O}\Big(\Delta t + (\Delta x)^2\Big)$
Lax-Friedrichs (2.34)	stable L^2 et L^∞ si condition CFL $\lvert V\rvert\Delta t \leq \Delta x$	$\mathcal{O}\Big(\Delta t + \frac{(\Delta x)^2}{\Delta t}\Big)$
Lax-Wendroff (2.36)	stable L^2 si condition CFL $\lvert V\rvert\Delta t \leq \Delta x$	$\mathcal{O}\Big((\Delta t)^2 + (\Delta x)^2\Big)$
Décentré amont (2.37)	stable L^2 et L^∞ si condition CFL $\lvert V\rvert\Delta t \leq \Delta x$	$\mathcal{O}\Big(\Delta t + \Delta x\Big)$

TABLE 2.2 – Résumé des propriétés de divers schémas pour l'équation d'advection

Tous les schémas que nous venons de voir sont consistants. Cependant, si on ajoute à l'équation la partie principale de l'erreur de troncature d'un schéma, alors ce schéma est non seulement encore consistant avec cette nouvelle équation "équivalente", mais est même strictement plus précis pour cette équation équivalente. En d'autres termes, le schéma est "plus consistant" avec l'équation équivalente qu'avec l'équation d'origine. Prenons l'exemple du schéma de Lax-Friedrichs (2.34) pour l'équation d'advection : d'après (2.35), la partie principale de son erreur de troncature est $-\frac{(\Delta x)^2}{2\Delta t}\left(1-\frac{(V\Delta t)^2}{(\Delta x)^2}\right)u_{xx}$. Par conséquent, l'équation équivalente du schéma de Lax-Friedrichs est

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0 \quad \text{avec} \quad \nu = \frac{(\Delta x)^2}{2\Delta t}\left(1-\frac{(V\Delta t)^2}{(\Delta x)^2}\right). \tag{2.38}$$

Cette équation équivalente va nous donner des renseignements précieux sur le comportement numérique du schéma. En effet, le schéma de Lax-Friedrichs est une bonne approximation (à l'ordre 2) de l'équation de convection-diffusion (2.38) où le coefficient de diffusion ν est petit (voire nul si la condition CFL est exactement satisfaite, i.e. $\Delta x = |V|\Delta t$). Remarquons que si le pas de temps est pris trop petit, le coefficient de diffusion ν peut être très grand et le schéma mauvais car trop porté à la diffusion (voir la Figure 2.1). Le coefficient de diffusion ν de l'équation équivalente est appelé **diffusion numérique**. S'il est grand, on dit que le schéma est **diffusif** (ou dissipatif). Le comportement typique d'un schéma diffusif est sa tendance à étaler artificiellement les données initiales au cours du temps. Les schémas trop diffusifs sont donc de "mauvais" schémas.

Exercice 2.3.6 Montrer que l'équation équivalente du schéma décentré amont (2.37) est

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} - \frac{|V|}{2}\left(\Delta x - |V|\Delta t\right)\frac{\partial^2 u}{\partial x^2} = 0.$$

Le schéma décentré amont est aussi diffusif (sauf si la condition CFL est exactement satisfaite, i.e. $\Delta x = |V|\Delta t$). En tout cas, le coefficient de diffusion de l'équation

équivalente ne tend pas vers l'infini si le pas de temps tend vers zéro (à Δx fixé), ce qui est une nette amélioration par rapport au schéma de Lax-Friedrichs (voir la Figure 2.1). Cet effet de diffusion numérique est illustré par la Figure 2.1 où l'on résout l'équation d'advection sur un intervalle de longueur 1 avec des conditions aux limites de périodicité, une donnée initiale sinusoïdale, un pas d'espace $\Delta x = 0.01$, une vitesse $V = 1$ et un temps final $T = 5$. On compare deux valeurs du pas de temps $\Delta t = 0.9\Delta x$ et $\Delta t = 0.45\Delta x$.

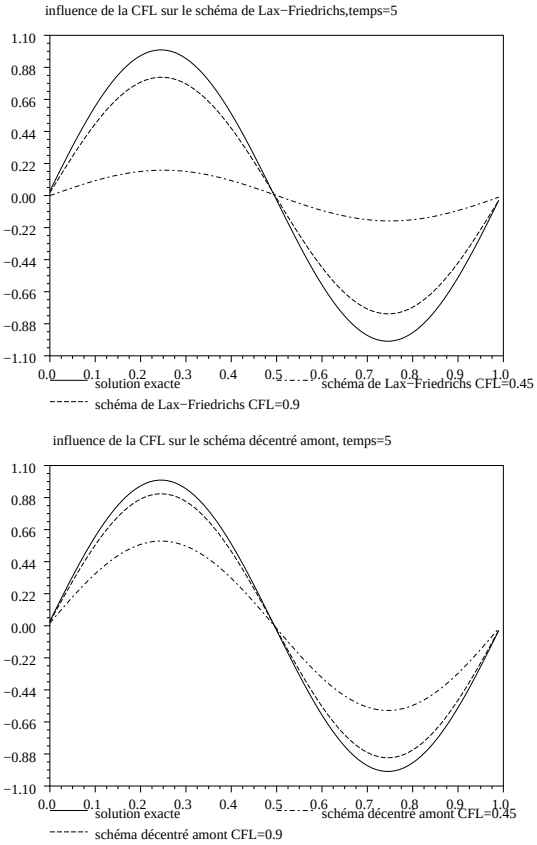

FIGURE 2.1 – Influence de la CFL sur la diffusion numérique du schéma de Lax-Friedrichs (haut) et du schéma décentré amont (bas).

Exercice 2.3.7 Montrer que l'équation équivalente du schéma de Lax-Wendroff (2.36) est

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} + \frac{V(\Delta x)^2}{6}\left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right)\frac{\partial^3 u}{\partial x^3} = 0.$$

Comme le schéma de Lax-Wendroff est précis d'ordre 2, l'équation équivalente ne contient pas de terme de diffusion mais un terme du troisième ordre, dit **dispersif**. Remarquons que le coefficient devant ce terme dispersif est un ordre plus petit que le coefficient de diffusion des équations équivalentes des schémas diffusifs. C'est pourquoi cet effet dispersif ne peut se voir, en général, que sur un schéma non diffusif. Le comportement typique d'un schéma dispersif est qu'il produit des oscillations

lorsque la solution est discontinue (voir la Figure 2.2). En effet, le terme dispersif modifie la vitesse de propagation des ondes planes ou modes de Fourier de la solution (particulièrement des modes de fréquence élevée), alors qu'un terme diffusif ne fait qu'atténuer son amplitude (voir l'Exercice 2.3.8).

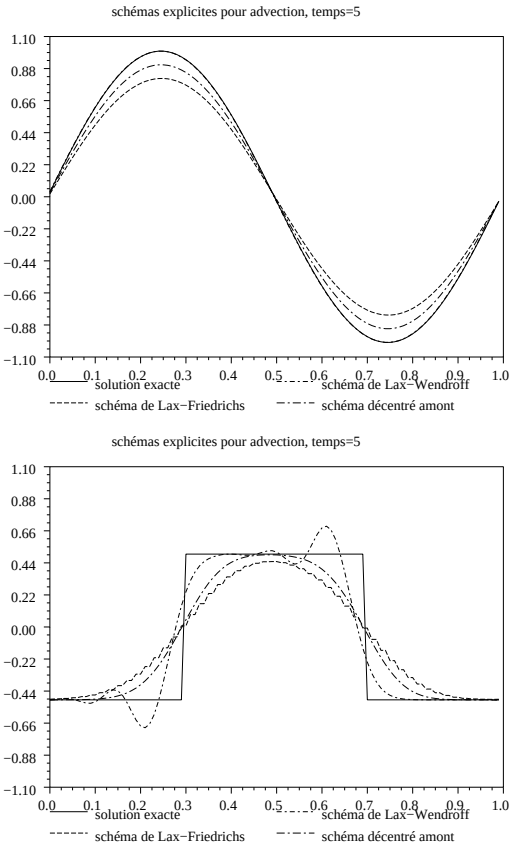

FIGURE 2.2 – Comparaison des schémas de Lax-Friedrichs, de Lax-Wendroff, et décentré amont pour une donnée initiale sinusoïdale (haut) ou en créneau (bas).

Afin d'illustrer notre propos nous présentons des calculs effectués sur un intervalle de longueur 1 avec des conditions aux limites de périodicité, un pas d'espace $\Delta x = 0.01$, un pas de temps $\Delta t = 0.9 * \Delta x$, une vitesse $V = 1$ et un temps final $T = 5$. Deux types de conditions initiales sont testées : tout d'abord une condition initiale très régulière, un sinus, puis une condition initiale discontinue, un créneau (voir la Figure 2.2). Les schémas précis à l'ordre 1 sont clairement diffusifs : ils écrasent la solution. Le schéma de Lax-Wendroff précis à l'ordre 2 est très bon pour une solution régulière mais oscille pour le créneau car il est dispersif. Le concept d'équation équivalente permet de comprendre ces phénomènes numériques.

Exercice 2.3.8 Soit l'équation

$$\begin{cases} \dfrac{\partial u}{\partial t} + V\dfrac{\partial u}{\partial x} - \nu\dfrac{\partial^2 u}{\partial x^2} - \mu\dfrac{\partial^3 u}{\partial x^3} = 0 \text{ pour } (x,t) \in \mathbb{R}\times\mathbb{R}_*^+ \\ u(t=0,x) = \sin(\omega x + \phi) \text{ pour } x \in \mathbb{R}, \end{cases}$$

avec $V, \nu, \mu, \omega, \phi \in \mathbb{R}$. Montrer que sa solution est

$$u(t,x) = \exp(-\nu\omega^2 t)\sin\left(\omega(x-(V+\mu\omega^2)t)+\phi\right)$$

(on admettra son unicité). En déduire que la diffusion atténue l'amplitude de la solution, tandis que la dispersion modifie la vitesse de propagation.

Exercice 2.3.9 On définit le schéma "saute-mouton" (leapfrog, en anglais)

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\Delta t} + V\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0.$$

Étudier la consistance et l'erreur de troncature de ce schéma. Montrer par analyse de Fourier qu'il est stable sous la condition CFL $|V|\Delta t \leq M\Delta x$ avec $M < 1$.

Exercice 2.3.10 On définit le schéma de Crank-Nicolson

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} + V\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{4\Delta x} + V\frac{u_{j+1}^n - u_{j-1}^n}{4\Delta x} = 0.$$

Étudier la consistance et l'erreur de troncature de ce schéma. Montrer par analyse de Fourier qu'il est inconditionnellement stable.

2.3.2 Équation des ondes

Nous considérons l'équation des ondes dans le domaine borné $(0,1)$ avec conditions aux limites de périodicité

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = 0 \text{ pour } (x,t) \in (0,1)\times\mathbb{R}_*^+ \\ u(t,x+1) = u(t,x) \text{ pour } (x,t) \in (0,1)\times\mathbb{R}_*^+ \\ u(t=0,x) = u_0(x) \text{ pour } x \in (0,1) \\ \dfrac{\partial u}{\partial t}(t=0,x) = u_1(x) \text{ pour } x \in (0,1). \end{cases} \tag{2.39}$$

Avec les mêmes notations que précédemment, l'inconnue discrète à chaque pas de temps est un vecteur $u^n = (u_j^n)_{0\leq j\leq N} \in \mathbb{R}^{N+1}$. Les conditions aux limites de périodicité conduisent aux égalités $u_0^n = u_{N+1}^n$ pour tout $n \geq 0$, et plus généralement $u_j^n = u_{N+1+j}^n$. Comme les conditions aux limites ne fixent pas la valeur de u aux extrémités de l'intervalle $(0,1)$ (pensez à l'interprétation en termes de corde vibrante), la solution u peut ne pas rester bornée en temps, ce qui complique l'étude de la stabilité des schémas numériques. Par exemple, si $u_0 \equiv 0$ et $u_1 \equiv C$ dans $(0,1)$, la solution de (2.39) est $u(t,x) = Ct$. Pour éliminer cet effet, on fait donc l'hypothèse que la vitesse initiale est de moyenne nulle

$$\int_0^1 u_1(x)\,dx = 0. \tag{2.40}$$

Pour l'équation des ondes (2.39) le schéma habituel est le **θ-schéma centré** : pour $n \geq 1$ et $j \in \{0, ..., N\}$,

$$\begin{array}{ll} \dfrac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} & +\theta \dfrac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} \\ +(1-2\theta)\dfrac{-u_{j-1}^{n} + 2u_j^{n} - u_{j+1}^{n}}{(\Delta x)^2} & +\theta \dfrac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0 \end{array} \tag{2.41}$$

avec $0 \leq \theta \leq 1/2$. Lorsque $\theta = 0$ on obtient un schéma explicite, tandis que le schéma est implicite si $\theta \neq 0$. Les conditions initiales sont prises en compte par

$$u_j^0 = u_0(x_j) \quad \text{et} \quad \frac{u_j^1 - u_j^0}{\Delta t} = \int_{x_{j-1/2}}^{x_{j+1/2}} u_1(x)\, dx,$$

ce qui garantit que la vitesse initiale discrète vérifie aussi la condition (2.40). Comme chacune des différences finies centrées qui approchent les dérivées secondes dans (2.41) est d'ordre 2, le θ-schéma centré (2.41) est précis à l'ordre 2 en espace et temps. Remarquer que ce schéma est invariant si on change le sens du temps (ce qui est compatible avec la propriété de réversibilité en temps de l'équation des ondes, vue lors de la Sous-section 1.3.2).

Lemme 2.3.5 *Si $1/4 \leq \theta \leq 1/2$, le θ-schéma centré (2.41) est inconditionnellement stable en norme L^2. Si $0 \leq \theta < 1/4$, il est stable sous la condition CFL*

$$\frac{\Delta t}{\Delta x} < \sqrt{\frac{1}{1-4\theta}},$$

et instable si $\Delta t/\Delta x > 1/\sqrt{1-4\theta}$.

Démonstration. Comme précédemment on utilise l'analyse de Fourier pour obtenir

$$\hat{u}^{n+1}(k) - 2\hat{u}^n(k) + \hat{u}^{n-1}(k) + \alpha(k)\left(\theta\hat{u}^{n+1}(k) + (1-2\theta)\hat{u}^n(k) + \theta\hat{u}^{n-1}(k)\right) = 0,$$

avec

$$\alpha(k) = 4\left(\frac{\Delta t}{\Delta x}\right)^2 \sin^2(\pi k \Delta x).$$

Il s'agit d'un schéma à trois niveaux qu'on réécrit

$$\hat{U}^{n+1}(k) = \begin{pmatrix} \hat{u}^{n+1}(k) \\ \hat{u}^n(k) \end{pmatrix} = \begin{pmatrix} \frac{2-(1-2\theta)\alpha(k)}{1+\theta\alpha(k)} & -1 \\ 1 & 0 \end{pmatrix} \hat{U}^n(k) = A(k)\hat{U}^n(k),$$

et $\hat{U}^{n+1}(k) = A(k)^n \hat{U}^1(k)$. Les valeurs propres (λ_1, λ_2) de la matrice $A(k)$ sont les racines du polynôme du deuxième degré

$$\lambda^2 - \frac{2-(1-2\theta)\alpha(k)}{1+\theta\alpha(k)}\lambda + 1 = 0.$$

Le discriminant de cette équation est

$$\Delta = -\frac{\alpha(k)(4-(1-4\theta)\alpha(k))}{(1+\theta\alpha(k))^2}.$$

L'étude de la stabilité du schéma est ici très délicate car $A(k)$ n'est pas une matrice normale et $\|A(k)^n\|_2 \neq \rho(A(k))^n$ où $\rho(A(k)) = \max(|\lambda_1|, |\lambda_2|)$ est le rayon spectral de $A(k)$. On se contente donc de vérifier la condition **nécessaire** de stabilité de Von Neumann, $\rho(A(k)) \leq 1$ (voir la Remarque 2.2.21), et on renvoie à l'Exercice 2.3.11 pour une condition suffisante. Si $\Delta t/\Delta x > 1/\sqrt{1-4\theta}$, un choix judicieux de k (tel que $\sin^2(\pi k \Delta x) \approx 1$) conduit à $\Delta > 0$, et dans ce cas les deux racines λ_1 et λ_2 sont réelles, de produit égal à 1. L'une des deux est forcément strictement plus grande que 1 en valeur absolue, $\rho(A(k)) > 1$, et le schéma est donc instable. Si $\Delta t/\Delta x < 1/\sqrt{1-4\theta}$, alors $\Delta \leq 0$ pour tout k, et les deux racines sont complexes conjuguées de module égal à 1. Par conséquent, $\rho(A(k)) = 1$ et la condition de stabilité de Von Neumann est satisfaite. □

Exercice 2.3.11 Finir la démonstration du Lemme 2.3.5 en calculant $A(k)^n$, et montrer la stabilité du schéma sous condition CFL grâce à (2.40).

Exercice 2.3.12 On considère le cas limite du Lemme 2.3.5, c'est-à-dire $\Delta t/\Delta x = 1/\sqrt{1-4\theta}$ avec $0 \leq \theta < 1/4$. Montrer que le θ-schéma centré (2.41) est instable dans ce cas en vérifiant que $u_j^n = (-1)^{n+j}(2n-1)$ est une solution (remarquez qu'il s'agit d'une instabilité "faible" puisque la croissance de u^n est linéaire et non exponentielle).

Chapitre 3

FORMULATION VARIATIONNELLE DES PROBLÈMES AUX LIMITES

3.1 Approche variationnelle

Dans ce chapitre nous nous intéressons à l'approche variationnelle pour la résolution des problèmes aux limites qui sont, rappelons-le constitué d'équations aux dérivées partielles munies de conditions aux limites et éventuellement de données initiales pour les problèmes dépendant du temps. Ces problèmes aux limites sont des modèles issus de toutes les disciplines scientifiques et des sciences de l'ingénieur. L'approche variationnelle que nous allons suivre consiste à remplacer les équations par une forme intégrale équivalente à l'aide de fonctions tests. Cette nouvelle formulation admet aussi une interprétation physique ou mécanique très naturelle, comme nous le verrons. Un premier intérêt de cette formulation variationnelle est de permettre de démontrer l'existence et l'unicité des solutions des problèmes aux limites. Nous n'en disons rien ici et nous renvoyons le lecteur intéressé à [1], [3]. Un deuxième intérêt, et pas le moindre, est que cette approche variationnelle sera cruciale pour définir et comprendre la méthode numérique des éléments finis.

Nous allons commencer cette présentation en se restreignant à des modèles physiques stationnaires, c'est-à-dire indépendants du temps. Nous verrons par la suite comment elle s'applique aussi à des problèmes d'évolution en temps.

L'exemple prototype d'équation aux dérivées partielles de type elliptique sera le Laplacien pour lequel nous étudierons le problème aux limites suivant

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{3.1}$$

où nous imposons des conditions aux limites de Dirichlet (nous renvoyons à la Sous-section 1.3.3 pour une présentation de ce modèle). Dans (3.1), Ω est un ouvert de l'espace $\mathbb{R}^N$, $\partial\Omega$ est son bord (ou frontière), f est un second membre (une donnée du problème), et u est l'inconnue.

Définition 3.1.1 *Soit Ω un ouvert de $\mathbb{R}^N$, $\overline{\Omega}$ sa fermeture. On note $C(\Omega)$ (respectivement, $C(\overline{\Omega})$) l'espace des fonctions continues dans Ω (respectivement, dans $\overline{\Omega}$). Soit un entier $k \geq 0$. On note $C^k(\Omega)$ (respectivement, $C^k(\overline{\Omega})$) l'espace des fonctions k fois continûment dérivables dans Ω (respectivement, dans $\overline{\Omega}$).*

Le principe de l'approche variationnelle pour la résolution des équations aux dérivées partielles est de remplacer l'équation par une formulation équivalente, dite variationnelle, obtenue en intégrant l'équation multipliée par une fonction quelconque, dite test. Comme il est nécessaire de procéder à des intégrations par parties dans l'établissement de la formulation variationnelle, nous commençons par donner quelques résultats essentiels à ce sujet.

3.1.1 Formules de Green

Dans toute cette sous-section Ω est un ouvert borné de l'espace $\mathbb{R}^N$, dont le bord (ou la frontière) est noté $\partial\Omega$. Nous supposons aussi que Ω est un ouvert **régulier** de classe $\mathcal{C}^1$. Il n'est pas nécessaire de connaître la définition précise d'un ouvert régulier (voir [1]). Il suffit juste de savoir qu'un ouvert régulier de classe $\mathcal{C}^1$ est *grosso modo* un ouvert dont le bord est une hypersurface (une variété de dimension $N-1$) régulière de classe $\mathcal{C}^1$, et que cet ouvert est localement situé d'un seul côté de sa frontière. On définit alors la **normale extérieure** au bord $\partial\Omega$ comme étant le vecteur unité $n = (n_i)_{1\leq i\leq N}$ normal en tout point au plan tangent de Ω et pointant vers l'extérieur de Ω (voir la Figure 1.1). Dans $\Omega \subset \mathbb{R}^N$ on note dx la mesure volumique, ou mesure de Lebesgue de dimension N. Sur $\partial\Omega$, on note ds la mesure surfacique, ou mesure de Lebesgue de dimension $N-1$ sur la variété $\partial\Omega$. Le résultat principal de cette sous-section est le lemme suivant que nous admettrons (voir [13]).

Lemme 3.1.2 (Formule de Green) *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^1$. Soit w une fonction de $C^1(\overline{\Omega})$. Alors elle vérifie la formule de Green*

$$\int_\Omega \frac{\partial w}{\partial x_i}(x)\,dx = \int_{\partial\Omega} w(x)n_i(x)\,ds, \tag{3.2}$$

où n_i est la i-ème composante de la normale extérieure unité de Ω.

Le Lemme 3.1.2 a de nombreux corollaires qui sont tous des conséquences immédiates de la formule de Green (3.2). Le lecteur qui voudra économiser sa mémoire ne retiendra donc que la formule de Green (3.2) !

Corollaire 3.1.3 (Formule d'intégration par parties) *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^1$. Soient u et v deux fonctions de $C^1(\overline{\Omega})$. Alors elles vérifient la formule d'intégration par parties*

$$\int_\Omega u(x)\frac{\partial v}{\partial x_i}(x)\,dx = -\int_\Omega v(x)\frac{\partial u}{\partial x_i}(x)\,dx + \int_{\partial\Omega} u(x)v(x)n_i(x)\,ds. \tag{3.3}$$

Démonstration. Il suffit de prendre $w = uv$ dans le Lemme 3.1.2. □

Corollaire 3.1.4 *Soit Ω un ouvert borné régulier de classe $\mathcal{C}^1$. Soit u une fonction de $C^2(\overline{\Omega})$ et v une fonction de $C^1(\overline{\Omega})$. Alors elles vérifient la formule d'intégration par parties*

$$\int_\Omega \Delta u(x)v(x)\,dx = -\int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x)\,ds, \tag{3.4}$$

où $\nabla u = \left(\frac{\partial u}{\partial x_i}\right)_{1\le i\le N}$ est le vecteur gradient de u, et $\frac{\partial u}{\partial n} = \nabla u\cdot n$.

Démonstration. On applique le Corollaire 3.1.3 à v et $\frac{\partial u}{\partial x_i}$ et on somme en i. □

Exercice 3.1.1 Déduire de la formule de Green (3.2) la formule de Stokes

$$\int_\Omega \mathrm{div}\sigma(x)\phi(x)\,dx = -\int_\Omega \sigma(x)\cdot\nabla\phi(x)\,dx + \int_{\partial\Omega} \sigma(x)\cdot n(x)\,\phi(x)\,ds,$$

où ϕ est une fonction scalaire de $C^1(\overline{\Omega})$ et σ une fonction à valeurs vectorielles de $C^1(\overline{\Omega})$.

Exercice 3.1.2 En dimension $N = 3$ on définit le rotationnel d'une fonction de Ω dans $\mathbb{R}^3$, $\phi = (\phi_1, \phi_2, \phi_3)$, comme la fonction de Ω dans $\mathbb{R}^3$ définie par

$$\mathrm{rot}\phi = \left(\frac{\partial\phi_3}{\partial x_2} - \frac{\partial\phi_2}{\partial x_3}, \frac{\partial\phi_1}{\partial x_3} - \frac{\partial\phi_3}{\partial x_1}, \frac{\partial\phi_2}{\partial x_1} - \frac{\partial\phi_1}{\partial x_2}\right).$$

Pour ϕ et ψ, fonctions à valeurs vectorielles de $C^1(\overline{\Omega})$, déduire de la formule de Green (3.2)

$$\int_\Omega \mathrm{rot}\phi\cdot\psi\,dx - \int_\Omega \phi\cdot\mathrm{rot}\psi\,dx = -\int_{\partial\Omega} (\phi\times n)\cdot\psi\,ds.$$

3.1.2 Formulation variationnelle

Nous supposons que l'ouvert Ω est borné et régulier, et que le second membre f de (3.1) est continu sur $\overline{\Omega}$. Le résultat principal de cette sous-section est la proposition suivante.

Proposition 3.1.5 *Soit u une fonction de $C^2(\overline{\Omega})$. Soit V_0 l'espace vectoriel défini par*

$$V_0 = \left\{\phi \in C^1(\overline{\Omega}) \text{ tel que } \phi = 0 \text{ sur } \partial\Omega\right\}.$$

Alors u est une solution du problème aux limites (3.1) si et seulement si u appartient à V_0 et vérifie l'égalité

$$\int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx = \int_\Omega f(x)v(x)\,dx \text{ pour toute fonction } v \in V_0. \tag{3.5}$$

L'égalité (3.5) est appelée la **formulation variationnelle** *du problème aux limites (3.1).*

Remarque 3.1.6 Un intérêt immédiat de la formulation variationnelle (3.5) est qu'elle a un sens si la solution u est seulement une fonction de $C^1(\overline{\Omega})$, contrairement à la formulation "classique" (3.1) qui requiert que u appartienne à $C^2(\overline{\Omega})$. On pressent donc déjà qu'il est plus simple de résoudre (3.5) que (3.1) puisqu'on est moins exigeant sur la régularité de la solution.

Dans la formulation variationnelle (3.5), la fonction v est appelée **fonction test**. La formulation variationnelle est aussi parfois appelée formulation faible du problème aux limites (3.1). En mécanique, la formulation variationnelle est connue sous le nom de "principe des travaux virtuels". En physique, on parle aussi d'équation de bilan ou de formule de réciprocité.

Lorsqu'on prend $v = u$ dans (3.5), on obtient ce qu'il est convenu d'appeler une **égalité d'énergie**, qui exprime généralement l'égalité entre une énergie stockée dans le domaine Ω (le terme de gauche de (3.5)) et une énergie potentielle associée à f (le terme de droite de (3.5)). •

Démonstration. Si u est solution du problème aux limites (3.1), on multiplie l'équation par $v \in V_0$ et on utilise la formule d'intégration par parties du Corollaire 3.1.4

$$\int_\Omega \Delta u(x)v(x)\,dx = -\int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x)\,ds.$$

Or $v = 0$ sur $\partial\Omega$ puisque $v \in V_0$, donc

$$\int_\Omega f(x)v(x)\,dx = \int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx,$$

qui n'est rien d'autre que la formule (3.5). Réciproquement, si $u \in V_0$ vérifie (3.5), en utilisant "à l'envers" la formule d'intégration par parties précédente on obtient

$$\int_\Omega \Big(\Delta u(x) + f(x)\Big)v(x)\,dx = 0 \text{ pour toute fonction } v \in V_0.$$

Comme $(\Delta u + f)$ est une fonction continue, grâce au Lemme 3.1.7 on conclut que $-\Delta u(x) = f(x)$ pour tout $x \in \Omega$. Par ailleurs, comme $u \in V_0$, on retrouve la condition aux limites $u = 0$ sur $\partial\Omega$, c'est-à-dire que u est solution du problème aux limites (3.1). □

Lemme 3.1.7 *Soit Ω un ouvert de $\mathbb{R}^N$. Soit $g(x)$ une fonction continue dans Ω. Si pour toute fonction ϕ de $C^\infty(\Omega)$ à support compact dans Ω, on a*

$$\int_\Omega g(x)\phi(x)\,dx = 0,$$

alors la fonction g est nulle dans Ω.

Démonstration. Supposons qu'il existe un point $x_0 \in \Omega$ tel que $g(x_0) \neq 0$. Sans perte de généralité, on peut supposer que $g(x_0) > 0$ (sinon on prend $-g$). Par continuité, il existe un petit voisinage ouvert $\omega \subset \Omega$ de x_0 tel que $g(x) > 0$ pour

tout $x \in \omega$. Soit alors une fonction test positive, non nulle, ϕ à support inclus dans ω. On a

$$\int_\Omega g(x)\phi(x)\,dx = \int_\omega g(x)\phi(x)\,dx = 0,$$

qui est une contradiction avec l'hypothèse sur g. Donc $g(x) = 0$ pour tout $x \in \Omega$. □

Remarque 3.1.8 Dans (3.1) nous avons considéré une condition aux limites de Dirichlet "homogène", c'est-à-dire nulle, mais nous pouvons aussi bien traiter le cas de conditions non-homogènes. Considérons le problème aux limites

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = u_0 & \text{sur } \partial\Omega, \end{cases} \tag{3.6}$$

où u_0 est la restriction sur $\partial\Omega$ d'une fonction définie dans $C^2(\overline{\Omega})$. Pour obtenir une formulation variationnelle de (3.6) on pose $u = u_0 + \tilde{u}$, et on cherche la solution de

$$\begin{cases} -\Delta \tilde{u} = \tilde{f} = f + \Delta u_0 & \text{dans } \Omega \\ \tilde{u} = 0 & \text{sur } \partial\Omega. \end{cases} \tag{3.7}$$

On applique alors la Proposition 3.1.5 à (3.7) et, par changement d'inconnue, on en déduit une formulation variationnelle pour (3.6). •

Au lieu de considérer le problème (3.1) avec des conditions aux limites de Dirichlet, on étudie maintenant le problème suivant avec des conditions aux limites de Neumann

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = g & \text{sur } \partial\Omega, \end{cases} \tag{3.8}$$

où $f \in C(\overline{\Omega})$ et $g \in C(\partial\Omega)$. On généralise facilement la Proposition 3.1.9 à ce nouveau problème. La seule différence, mais elle est essentielle, est que la condition aux limites de Neumann est incluse dans la formulation variationnelle alors que celle de Dirichlet était inscrite dans le choix de l'espace vectoriel V_0.

Proposition 3.1.9 *Soit u une fonction de $C^2(\overline{\Omega})$. Soit l'espace vectoriel $V = C^1(\overline{\Omega})$. Alors u est une solution du problème aux limites (3.8) si et seulement si u appartient à V et vérifie, pour toute fonction $v \in V$,*

$$\int_\Omega \left(\nabla u(x) \cdot \nabla v(x) + u(x)\, v(x)\right)\, dx = \int_\Omega f(x)v(x)\,dx + \int_{\partial\Omega} g(x)v(x)\,ds. \tag{3.9}$$

Démonstration. Si u est solution du problème aux limites (3.8), on multiplie l'équation par $v \in V$ et on utilise la formule d'intégration par parties du Corollaire 3.1.4

$$\begin{aligned} -\int_\Omega \Delta u(x)v(x)\,dx &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx - \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x)\,ds \\ &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx - \int_{\partial\Omega} g(x)v(x)\,ds, \end{aligned}$$

à cause de la condition aux limites de Neumann. On en déduit donc l'égalité (3.9). Réciproquement, si $u \in V$ vérifie (3.9), en utilisant "à l'envers" la formule d'intégration par parties précédente on obtient, pour toute fonction $v \in V$,

$$\int_\Omega \Big(\Delta u(x) - u(x) + f(x)\Big) v(x)\, dx = \int_{\partial\Omega} \left(\frac{\partial u}{\partial n}(x) - g(x)\right) v(x)\, ds.$$

On choisit d'abord une fonction v à support compact dans Ω, de manière à ce que l'intégrale sur le bord $\partial\Omega$ soit nulle. Comme $(\Delta u - u + f)$ est une fonction continue, le Lemme 3.1.7 nous permet de conclure que $-\Delta u(x) + u(x) = f(x)$ pour tout $x \in \Omega$. En utilisant cette égalité dans la formule précédente on en déduit que, pour toute fonction $v \in V$,

$$\int_{\partial\Omega} \left(\frac{\partial u}{\partial n}(x) - g(x)\right) v(x)\, ds = 0.$$

Comme $\frac{\partial u}{\partial n}(x) - g(x)$ est continu sur $\partial\Omega$, l'analogue du Lemme 3.1.7 pour le bord $\partial\Omega$ indique que $\frac{\partial u}{\partial n}(x) - g(x) = 0$ pour tout $x \in \partial\Omega$. On retrouve donc la condition aux limites de Neumann, c'est-à-dire que u est solution du problème aux limites (3.8). □

Remarque 3.1.10 En notation compacte on peut réécrire les formulations variationnelles (3.5) et (3.9) sous la forme : trouver $u \in V$ tel que

$$a(u, v) = L(v) \text{ pour toute fonction } v \in V,$$

où V est un espace vectoriel réel, $a(\cdot,\cdot)$ est une forme bilinéaire sur V (c'est-à-dire une application de $V \times V$ dans $\mathbb{R}$, linéaire par rapport à chacun de ses deux arguments) et $L(\cdot)$ est une forme linéaire sur V (c'est-à-dire une application linéaire de V dans $\mathbb{R}$). Dans le cas de (3.5) on a $V = V_0$,

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v\, dx \quad \text{et} \quad L(v) = \int_\Omega f\, v\, dx,$$

tandis que pour (3.9) on a $V = C^1(\overline{\Omega})$,

$$a(u, v) = \int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx \quad \text{et} \quad L(v) = \int_\Omega f\, v\, dx + \int_{\partial\Omega} g\, v\, ds.$$

•

On peut munir l'espace vectoriel $V_0 = \{v \in C^1(\overline{\Omega}),\ v = 0 \text{ sur } \partial\Omega\}$ du produit scalaire suivant

$$\langle u, v\rangle_{V_0} = \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx \tag{3.10}$$

et de la norme associée $\|u\|_{V_0} = \sqrt{\langle u, u\rangle_{V_0}}$.

Exercice 3.1.3 Vérifier que la formule (3.10) définit bien un produit scalaire sur V_0. Montrer que la forme bilinéaire $a(u, v)$ de (3.5) est **coercive** sur V_0 au sens où il existe une constante $\nu > 0$ telle que $a(v, v) \geq \nu \|v\|^2_{V_0}$ pour tout $v \in V_0$.

On peut munir l'espace vectoriel $V = C^1(\overline{\Omega})$ du produit scalaire suivant

$$\langle u, v\rangle_V = \int_\Omega \Big(\nabla u(x)\cdot\nabla v(x) + u(x)\,v(x)\Big)\,dx \tag{3.11}$$

et de la norme associée $\|u\|_V = \sqrt{\langle u, u\rangle_V}$.

Exercice 3.1.4 Montrer que (3.10) n'est pas un produit scalaire sur V tandis que (3.11) définit bien un produit scalaire sur V. Montrer que la forme bilinéaire $a(u, v)$ de (3.9) est **coercive** sur V au sens où il existe une constante $\nu > 0$ telle que $a(v, v) \geq \nu\|v\|_V^2$ pour tout $v \in V$.

On peut montrer (mais cela n'est pas le but du cours) que **l'approche variationnelle** permet de démontrer l'existence et l'unicité de la solution de la formulation variationnelle (3.5), ce qui entraînera le même résultat pour l'équation (3.1) à cause de la Proposition 3.1.5. Pour plus de détails nous renvoyons à [1], [3]. Une des difficultés de cette approche variationnelle est qu'elle nécessite que l'espace V soit un espace de Hilbert, ce qui n'est pas le cas pour $V_0 = \{v \in C^1(\overline{\Omega}),\ v = 0 \text{ sur } \partial\Omega\}$ muni du produit scalaire (3.10), ni pour $C^1(\overline{\Omega})$ muni du produit scalaire (3.11). Pour contourner cette difficulté il faut introduire la notion d'espace de Sobolev qui dépasse largement le cadre de ce cours. Néanmoins la formulation variationnelle d'un problème aux limites est aussi intéressante pour des question d'approximation numérique comme nous allons le voir dans les sections suivantes.

Remarque 3.1.11 En fait, il n'est pas nécessaire que les fonctions de V_0 soit continuement dérivables sur $\overline{\Omega}$. Il suffit qu'elles soient continues sur $\overline{\Omega}$ et continuement dérivables "par morceaux". Autrement dit, on peut remplacer $V_0 = \{v \in C^1(\overline{\Omega}),\ v = 0 \text{ sur } \partial\Omega\}$ par un espace vectoriel, sensiblement plus grand, qui est

$$V_0 = \{v \in V,\ v = 0 \text{ sur } \partial\Omega\}\,, \tag{3.12}$$

avec

$$V = \big\{v \in C(\overline{\Omega}),\ \text{il existe une partition } (\omega_i) \text{ de } \Omega \text{ telle que } v \in C^1(\overline{\omega}_i)\big\}\,. \tag{3.13}$$

Cette généralisation est utile pour définir la méthode des éléments finis ci-dessous. Rappelons qu'une partition de Ω est une collection finie d'ouverts $(\omega_i)_{1\leq i\leq I}$ telle que $\omega_i \subset \Omega$, $\omega_i \cap \omega_j = \emptyset$ pour $i \neq j$ et $\overline{\Omega} = \cup_{i=1}^I \overline{\omega_i}$. •

Exercice 3.1.5 On considère le Laplacien avec condition aux limites de Neumann

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega. \end{cases} \tag{3.14}$$

Soit u une fonction de $C^2(\overline{\Omega})$. Montrer que u est une solution du problème aux limites (3.14) si et seulement si u appartient à $C^1(\overline{\Omega})$ et vérifie l'égalité

$$\int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx = \int_\Omega f(x)v(x)\,dx \text{ pour toute fonction } v \in C^1(\overline{\Omega}). \tag{3.15}$$

En déduire qu'une condition nécessaire d'existence d'une solution dans $C^2(\overline{\Omega})$ de (3.14) est que $\int_\Omega f(x)dx = 0$.

Exercice 3.1.6 On considère l'équation des plaques

$$\begin{cases} \Delta\left(\Delta u\right) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases} \tag{3.16}$$

On note V l'espace des fonctions v de $C^2(\overline{\Omega})$ telles que v et $\frac{\partial v}{\partial n}$ s'annulent sur $\partial\Omega$. Soit u une fonction de $C^4(\overline{\Omega})$. Montrer que u est une solution du problème aux limites (3.16) si et seulement si u appartient à V et vérifie l'égalité

$$\int_\Omega \Delta u(x)\Delta v(x)\,dx = \int_\Omega f(x)v(x)\,dx \text{ pour toute fonction } v \in V. \tag{3.17}$$

3.1.3 Approximation variationnelle

Dans cette section nous présentons le processus **d'approximation variationnelle interne** qui nous conduira, par la suite, à la méthode des éléments finis. L'idée centrale de cette méthode d'approximation (appelée parfois méthode de Galerkin) est de remplacer l'espace vectoriel V sur lequel est posée la formulation variationnelle par un sous-espace V_h de dimension finie. Le problème "approché" posé sur V_h se ramène à la simple résolution d'un système linéaire, dont la matrice est appelée **matrice de rigidité**. Par ailleurs, on peut choisir le mode de construction de V_h de manière à ce que le sous-espace V_h soit une bonne approximation de V et que la solution u_h dans V_h de la formulation variationnelle soit **"proche"** de la solution exacte u dans V.

Nous considérons à nouveau les notations générales du formalisme variationnel introduites dans la Remarque 3.1.10. Étant donné un espace vectoriel V, une forme bilinéaire continue et coercive $a(u,v)$, et une forme linéaire continue $L(v)$, on considère la formulation variationnelle :

$$\text{trouver } u \in V \text{ tel que } a(u,v) = L(v) \quad \forall\, v \in V. \tag{3.18}$$

L'approximation variationnelle interne de (3.18) consiste à remplacer l'espace vectoriel V par un sous-espace de dimension finie V_h, c'est-à-dire à chercher la solution de :

$$\text{trouver } u_h \in V_h \text{ tel que } a(u_h,v_h) = L(v_h) \quad \forall\, v_h \in V_h. \tag{3.19}$$

La résolution de l'approximation interne (3.19) est facile comme le montre le lemme suivant.

Lemme 3.1.12 *Soit V un espace vectoriel réel normé, et V_h un sous-espace de dimension finie. Soit $L(v)$ une forme linéaire sur V et $a(u,v)$ une forme bilinéaire, coercive sur V, au sens où il existe une constante $\nu > 0$ telle que $a(v,v) \geq \nu\|v\|^2$ pour tout $v \in V$. Alors l'approximation variationnelle interne (3.19) admet une unique solution. Par ailleurs cette solution peut s'obtenir en résolvant un système linéaire de matrice définie positive (et symétrique si $a(u,v)$ est symétrique).*

Démonstration. Comme V_h est de dimension finie, on introduit une base $(\phi_j)_{1\leq j\leq N_h}$ de V_h. Dans cette base la solution s'écrit $u_h = \sum_{j=1}^{N_h} u_j\phi_j$ et on pose $U_h = (u_1, ..., u_{N_h})$ le vecteur dans $\mathbb{R}^{N_h}$ des coordonnées de u_h. En décomposant aussi la fonction test sur cette base, le problème (3.19) est équivalent à :

$$\text{trouver } U_h \in \mathbb{R}^{N_h} \text{ tel que } a\left(\sum_{j=1}^{N_h} u_j\phi_j, \phi_i\right) = L(\phi_i) \quad \text{pour tout } 1 \leq i \leq N_h,$$

ce qui s'écrit sous la forme d'un système linéaire

$$\mathcal{K}_h U_h = b_h, \tag{3.20}$$

avec, pour $1 \leq i, j \leq N_h$,

$$(\mathcal{K}_h)_{ij} = a(\phi_j, \phi_i), \quad (b_h)_i = L(\phi_i).$$

La coercivité de la forme bilinéaire $a(u, v)$ entraîne le caractère défini positif de la matrice $\mathcal{K}_h$, et donc son inversibilité. En effet, pour tout vecteur $W_h = (w_1, ..., w_{N_h}) \in \mathbb{R}^{N_h}$, en notant $w_h = \sum_{j=1}^{N_h} w_j\phi_j$, on a

$$\mathcal{K}_h W_h \cdot W_h = a(w_h, w_h) \geq \nu \left\| \sum_{j=1}^{N_h} w_j\phi_j \right\|^2 \geq C|W_h|^2 \quad \text{avec} \quad C > 0,$$

car toutes les normes sont équivalentes en dimension finie ($|\cdot|$ désigne la norme euclidienne dans $\mathbb{R}^{N_h}$). Ainsi la matrice $\mathcal{K}_h$ est inversible et il existe bien une unique solution de (3.19) (indépendemment du choix de la base de V_h). De même, la symétrie de $a(u, v)$ implique celle de $\mathcal{K}_h$. Dans les applications mécaniques la matrice $\mathcal{K}_h$ est appelée **matrice de rigidité**. □

Nous allons maintenant comparer l'erreur commise en remplaçant l'espace V par son sous-espace V_h. Plus précisément, nous allons majorer la différence $\|u-u_h\|$ où u est la solution dans V de (3.18) et u_h celle dans V_h de (3.19). Le lemme suivant, dû à Jean Céa, montre que la distance entre la solution exacte u et la solution approchée u_h est majorée **uniformément par rapport au sous-espace** V_h par la distance entre u et V_h.

Lemme 3.1.13 (de Céa) *On se place sous les hypothèses du Lemme 3.1.12 et on suppose aussi que la forme bilinéaire $a(u, v)$ est continue, au sens où il existe une constante $M > 0$ telle que $|a(u, v)| \leq M\|u\| \, \|v\|$ pour tout $u, v \in V$. Soit u la solution de (3.18) et u_h celle de (3.19). On a*

$$\|u - u_h\| \leq \frac{M}{\nu} \inf_{v_h \in V_h} \|u - v_h\|. \tag{3.21}$$

Démonstration. Puisque $V_h \subset V$, on déduit, par soustraction des formulations variationnelles (3.18) et (3.19), que

$$a(u - u_h, w_h) = 0 \quad \forall\, w_h \in V_h.$$

En choisissant $w_h = u_h - v_h$ on obtient

$$\nu\|u-u_h\|^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u-v_h) \leq M\|u-u_h\|\|u-v_h\|,$$

d'où l'on déduit (3.21). □

Exercice 3.1.7 Dans le cadre du Lemme de Céa 3.1.13, démontrer que, si la forme bilinéaire $a(u,v)$ est symétrique, alors on améliore (3.21) en

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

Indication : on utilisera le fait que la solution u_h de (3.19) réalise aussi le minimum d'une énergie.

3.2 Éléments finis en dimension $N = 1$

La méthode des éléments finis est la méthode numérique de référence pour le calcul des solutions de nombreux problèmes aux limites. Le principe de la méthode des éléments finis est de construire des espaces d'approximation interne V_h dont la définition est basée sur la notion géométrique de **maillage** du domaine Ω. Un maillage est un pavage de l'espace en volumes élémentaires très simples : triangles, tétraèdres, parallélépipèdes. Dans ce contexte le paramètre h de V_h correspond à la **taille maximale des mailles** ou cellules qui composent le maillage. Typiquement une base de V_h sera constituée de fonctions dont le support est **localisé** sur une ou quelques mailles. Ceci aura deux conséquences importantes : d'une part, dans la limite $h \to 0$, l'espace V_h sera de plus en plus "gros" et approchera de mieux en mieux l'espace V tout entier, et d'autre part, la matrice de rigidité $\mathcal{K}_h$ du système linéaire (3.20) sera **creuse**, c'est-à-dire que la plupart de ses coefficients seront nuls (ce qui limitera le coût de la résolution numérique).

Dans ce cours nous nous limitons à la définition des éléments finis en une dimension d'espace où, sans trahir les idées générales valables en dimensions supérieures, les aspects techniques sont nettement plus simples. En particulier, la définition géométrique des maillages est particulièrement simple alors que c'est un point délicat en dimension deux et, surtout, trois. On discute des aspects pratiques (assemblage de la matrice de rigidité, formules de quadrature, etc.) autant que théoriques (convergence de la méthode, interpolation et estimation d'erreur). Pour plus de détails sur la méthode des éléments finis nous renvoyons à [3], [7], [9] et aux références citées dans ces ouvrages.

Sans perte de généralité nous choisissons le domaine $\Omega =]0,1[$. En dimension 1 un maillage est simplement constitué d'une collection de points $(x_j)_{0\leq j\leq n+1}$ (comme pour la méthode des différences finies, voir le Chapitre 1) tels que

$$x_0 = 0 < x_1 < ... < x_n < x_{n+1} = 1.$$

Le maillage sera dit **uniforme** si les points x_j sont équidistants, c'est-à-dire que

$$x_j = jh \quad \text{avec} \quad h = \frac{1}{n+1},\ 0 \leq j \leq n+1.$$

Les points x_j sont aussi appelés les **sommets** (ou noeuds) du maillage. Par souci de simplicité nous considérons, pour l'instant, le problème modèle suivant

$$\begin{cases} -u'' = f \ \text{ dans }]0,1[\\ u(0) = u(1) = 0, \end{cases} \tag{3.22}$$

dont il est facile de voir, par intégration, qu'il admet une solution unique dans $C^2([0,1])$ si $f \in C([0,1])$. Dans tout ce qui suit on notera $\mathbb{P}_k$ l'ensemble des polynômes à coefficients réels d'une variable réelle de degré inférieur ou égal à k.

3.2.1 Éléments finis $\mathbb{P}_1$

La méthode des éléments finis $\mathbb{P}_1$ repose sur l'espace discret des fonctions globalement continues et affines sur chaque maille

$$V_h = \left\{ v \in C([0,1]) \text{ tel que } v\,\big|_{[x_j, x_{j+1}]} \in \mathbb{P}_1 \text{ pour tout } 0 \le j \le n \right\}, \tag{3.23}$$

et sur son sous-espace

$$V_{0h} = \{ v \in V_h \text{ tel que } v(0) = v(1) = 0 \}. \tag{3.24}$$

La méthode des éléments finis $\mathbb{P}_1$ est alors simplement la méthode d'approximation variationnelle interne de la Sous-section 3.1.3 appliquée aux espaces V_h ou V_{0h} définis par (3.23) ou (3.24) (qui sont bien des sous-espaces vectoriels de l'espace V défini par (3.13)).

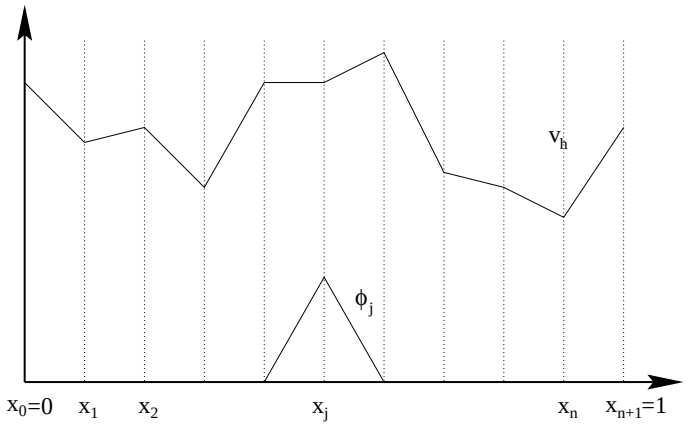

FIGURE 3.1 – Maillage de $\Omega =]0,1[$ et fonction de base en éléments finis $\mathbb{P}_1$.

On peut représenter les fonctions de V_h ou V_{0h}, affines par morceaux, à l'aide de fonctions de base très simples. Introduisons la "fonction chapeau" ϕ définie par

$$\phi(x) = \begin{cases} 1 - |x| & \text{si } |x| \le 1, \\ 0 & \text{si } |x| > 1. \end{cases}$$

Si le maillage est uniforme, pour $0 \le j \le n+1$ on définit les fonctions de base (voir la Figure 3.1)

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h} \right). \tag{3.25}$$

Lemme 3.2.1 *L'espace V_h, défini par (3.23), est un sous-espace vectoriel de V défini par (3.13), qui est de dimension $n+2$, et toute fonction $v_h \in V_h$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{0\le j\le n+1}$*

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j)\phi_j(x) \quad \forall x \in [0,1].$$

De même, V_{0h}, défini par (3.24), est un sous-espace de V_0 défini par (3.12), qui est de dimension n, et toute fonction $v_h \in V_{0h}$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{1\le j\le n}$

$$v_h(x) = \sum_{j=1}^{n} v_h(x_j)\phi_j(x) \quad \forall x \in [0,1].$$

Démonstration. La preuve est immédiate en remarquant que $\phi_j(x_i) = \delta_{ij}$, où δ_{ij} est le symbole de Kronecker qui vaut 1 si $i = j$ et 0 sinon (voir la Figure 3.1). □

Remarque 3.2.2 La base (ϕ_j), définie par (3.25), permet de caractériser une fonction de V_h par ses valeurs aux noeuds du maillage. Dans ce cas on parle **d'éléments finis de Lagrange**. Par ailleurs, comme les fonctions sont localement $\mathbb{P}_1$, on dit que l'espace V_h, défini par (3.23), est l'espace des éléments finis de Lagrange d'ordre 1.

Cet exemple des éléments finis $\mathbb{P}_1$ permet à nouveau de comprendre l'intérêt de la formulation variationnelle. En effet, les fonctions de V_h ne sont pas deux fois dérivables sur le segment $[0,1]$ et cela n'a pas de sens de résoudre, même de manière approchée, l'équation (3.22) (en fait la dérivée seconde d'une fonction de V_h est une somme de masses de Dirac aux noeuds du maillage !). Au contraire, il est parfaitement légitime d'utiliser des fonctions de V_h dans la formulation variationnelle (3.19) qui ne requiert qu'une seule dérivée. •

Décrivons la **résolution pratique** du problème de Dirichlet (3.22) par la méthode des éléments finis $\mathbb{P}_1$. La formulation variationnelle (3.19) de l'approximation interne devient ici :

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_0^1 u_h'(x)v_h'(x)\,dx = \int_0^1 f(x)v_h(x)\,dx \;\forall\, v_h \in V_{0h}. \quad (3.26)$$

On décompose u_h sur la base des $(\phi_j)_{1\le j\le n}$ et on prend $v_h = \phi_i$ ce qui donne

$$\sum_{j=1}^{n} u_h(x_j) \int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \int_0^1 f(x)\phi_i(x)\,dx.$$

En notant $U_h = (u_h(x_j))_{1\le j\le n}$, $b_h = \left(\int_0^1 f(x)\phi_i(x)\,dx\right)_{1\le i\le n}$, et en introduisant la **matrice de rigidité**

$$\mathcal{K}_h = \left(\int_0^1 \phi_j'(x)\phi_i'(x)\,dx\right)_{1\le i,j\le n},$$

la formulation variationnelle dans V_{0h} revient à résoudre dans $\mathbb{R}^n$ le système linéaire

$$\mathcal{K}_h U_h = b_h.$$

Comme les fonctions de base ϕ_j ont un "petit" support, l'intersection des supports de ϕ_j et ϕ_i est souvent vide et la plupart des coefficients de $\mathcal{K}_h$ sont nuls. Un calcul simple montre que

$$\int_0^1 \phi_j'(x)\phi_i'(x)\,dx = \begin{cases} -h^{-1} & \text{si } j = i-1 \\ 2h^{-1} & \text{si } j = i \\ -h^{-1} & \text{si } j = i+1 \\ 0 & \text{sinon} \end{cases}$$

et la matrice $\mathcal{K}_h$ est tridiagonale

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}. \tag{3.27}$$

Pour obtenir le second membre b_h il faut calculer les intégrales

$$(b_h)_i = \int_{x_{i-1}}^{x_{i+1}} f(x)\phi_i(x)\,dx \quad \text{pour tout } 1 \le i \le n.$$

L'évaluation exacte du second membre b_h peut être difficile ou impossible si la fonction f est compliquée. En pratique on a recours à des **formules de quadrature** (ou formules d'intégration numérique) qui donnent une approximation des intégrales définissant b_h. Par exemple, on peut utiliser la formule du "point milieu"

$$\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \psi(x)\,dx \approx \psi\left(\frac{x_{i+1}+x_i}{2}\right),$$

ou la formule des "trapèzes"

$$\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \psi(x)\,dx \approx \frac{1}{2}\left(\psi(x_{i+1})+\psi(x_i)\right). \tag{3.28}$$

Ces deux formules sont exactes pour les fonctions ψ affines. Si la fonction ψ est régulière quelconque, alors ces formules sont simplement approchées avec un reste de l'ordre de $\mathcal{O}(h^2)$.

La résolution du système linéaire $\mathcal{K}_h U_h = b_h$ est la partie la plus coûteuse de la méthode en terme de temps de calcul. C'est pourquoi nous présentons dans la Section 3.5 des méthodes performantes de résolution. Rappelons que la matrice $\mathcal{K}_h$ est nécessairement inversible par application du Lemme 3.1.12.

Remarque 3.2.3 La matrice de rigidité $\mathcal{K}_h$ est très similaire à des matrices déjà rencontrées lors de l'étude des méthodes de différences finies. En fait, $h\mathcal{K}_h$ est la limite de la matrice (2.13) (multipliée par $1/c$) du schéma implicite de résolution de l'équation de la chaleur lorsque le pas de temps tend vers l'infini. Nous verrons à l'Exercice 3.2.3 qu'il ne s'agit pas d'une coïncidence. •

Problème de Neumann. La mise en oeuvre de la méthode des éléments finis $\mathbb{P}_1$ pour le problème de Neumann suivant est très similaire

$$\left\{ \begin{array}{l} -u'' + au = f \text{ dans }]0,1[\\ u'(0) = \alpha, u'(1) = \beta. \end{array} \right. \tag{3.29}$$

On admettra que (3.29) admet une solution unique dans $C^2([0,1])$ si $f \in C([0,1])$, $\alpha, \beta \in \mathbb{R}$, et $a \in C([0,1])$ tel que $a(x) \geq a_0 > 0$ dans $[0,1]$. La formulation variationnelle (3.19) de l'approximation interne devient ici : trouver $u_h \in V_h$ tel que

$$\int_0^1 (u_h'(x)v_h'(x) + a(x)u_h(x)v_h(x))\,dx = \int_0^1 f(x)v_h(x)\,dx - \alpha v_h(0) + \beta v_h(1),$$

pour tout $v_h \in V_h$. En décomposant u_h sur la base des $(\phi_j)_{0\leq j\leq n+1}$, la formulation variationnelle dans V_h revient à résoudre dans $\mathbb{R}^{n+2}$ le système linéaire

$$\mathcal{K}_h U_h = b_h,$$

avec $U_h = (u_h(x_j))_{0\leq j\leq n+1}$, et une nouvelle matrice de rigidité

$$\mathcal{K}_h = \left(\int_0^1 \left(\phi_j'(x)\phi_i'(x) + a(x)\phi_j(x)\phi_i(x)\right)\,dx \right)_{0\leq i,j\leq n+1},$$

et

$$\begin{array}{l} (b_h)_i = \int_0^1 f(x)\phi_i(x)\,dx \quad \text{si } 1 \leq i \leq n, \\ (b_h)_0 = \int_0^1 f(x)\phi_0(x)\,dx - \alpha, \\ (b_h)_{n+1} = \int_0^1 f(x)\phi_{n+1}(x)\,dx + \beta. \end{array}$$

Lorsque $a(x)$ n'est pas une fonction constante, il est aussi nécessaire en pratique d'utiliser des formules de quadrature pour évaluer les coefficients de la matrice $\mathcal{K}_h$ (comme nous l'avons fait dans l'exemple précédent pour le second membre b_h).

Exercice 3.2.1 Appliquer la méthode des éléments finis $\mathbb{P}_1$ au problème

$$\left\{ \begin{array}{l} -u'' = f \text{ dans }]0,1[\\ u(0) = \alpha, u(1) = \beta, \end{array} \right.$$

On suivra la démarche énoncée dans la Remarque 3.1.8. Vérifier que les conditions aux limites de Dirichlet non-homogènes apparaissent dans le second membre du système linéaire qui en découle.

Exercice 3.2.2 On reprend le problème de Neumann (3.29) en supposant que la fonction $a(x) = 0$ dans Ω. Montrer que la matrice du système linéaire issu de la méthode des éléments finis $\mathbb{P}_1$ est singulière. Montrer qu'on peut néanmoins résoudre le système linéaire si les données vérifient la condition de compatibilité

$$\int_0^1 f(x)\,dx = \alpha - \beta.$$

Exercice 3.2.3 Appliquer la méthode des différences finies (voir le Chapitre 2) au problème de Dirichlet (3.22). Vérifier qu'avec un schéma centré d'ordre deux, on obtient un système linéaire à résoudre avec la même matrice $\mathcal{K}_h$ (à un coefficient multiplicatif près) mais avec un second membre b_h différent. Même question pour le problème de Neumann (3.29).

3.2.2 Convergence et estimation d'erreur

Pour démontrer la convergence de la méthode des éléments finis $\mathbb{P}_1$ en une dimension d'espace nous suivons la démarche esquissée à la fin de la Sous-section 3.1.3.

Définition 3.2.4 *On appelle* **opérateur d'interpolation** $\mathbb{P}_1$ *l'application linéaire* r_h *de* $C([0,1])$ *dans* V_h *définie, pour tout* $v \in C([0,1])$*, par*

$$(r_h v)(x) = \sum_{j=0}^{n+1} v(x_j)\phi_j(x).$$

L'interpolée $r_h v$ d'une fonction v est simplement la fonction affine par morceaux qui coïncide avec v sur les sommets du maillage x_j (voir la Figure 3.2).

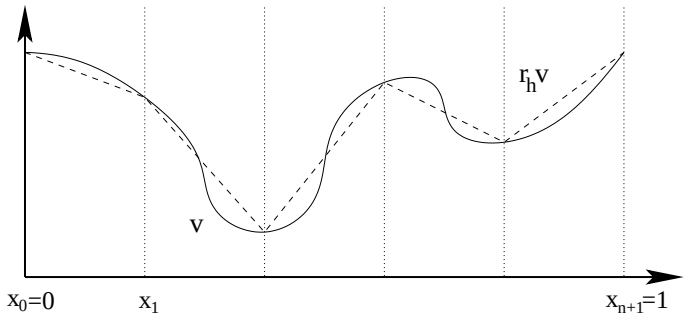

FIGURE 3.2 – Interpolation $\mathbb{P}_1$ d'une fonction de $C^1[0,1]$.

La convergence de la méthode des éléments finis $\mathbb{P}_1$ repose sur le lemme suivant.

Lemme 3.2.5 (d'interpolation) *Soit* r_h *l'opérateur d'interpolation* $\mathbb{P}_1$*. Il existe une constante* C *indépendante de* h *telle que, pour tout* $v \in C^2([0,1])$*,*

$$\|v - r_h v\|_V \leq Ch\|v''\|_{L^2(0,1)},$$

où, suivant (3.11), la norme de V *est définie par* $\|w\|_V^2 = \|w\|_{L^2(0,1)}^2 + \|w'\|_{L^2(0,1)}^2$*.*

Nous repoussons momentanément la démonstration de ce lemme pour énoncer tout de suite le résultat principal de cette sous-section qui établit la convergence de la méthode des éléments finis $\mathbb{P}_1$ pour le problème de Dirichlet (un résultat similaire est valable pour le problème de Neumann).

Théorème 3.2.6 *Soit* $u \in C^2([0,1])$ *la solution de (3.22) et soit* $u_h \in V_{0h}$ *la solution de (3.26). La méthode des éléments finis* $\mathbb{P}_1$ *converge, c'est-à-dire qu'il existe une constante* C *indépendante de* h *et de* f *telle que*

$$\|u - u_h\|_{V_0} \leq Ch\|f\|_{L^2(0,1)}, \tag{3.30}$$

où, suivant (3.10), la norme de V_0 *est définie par* $\|w\|_{V_0} = \|w'\|_{L^2(0,1)}$*.*

Remarque 3.2.7 On peut faire une analogie entre la convergence d'une méthode d'éléments finis et la convergence d'une méthode de différences finies. Rappelons que, d'après le Théorème de Lax 2.2.17, la convergence d'un schéma aux différences finies découle de sa stabilité et de sa consistance. Indiquons quels sont les équivalents (formels) de ces ingrédients dans le contexte des éléments finis. Le rôle de la consistance pour les éléments finis est joué par la propriété d'interpolation du Lemme 3.2.5, tandis que le rôle de la stabilité est tenu par la propriété de coercivité de la forme bilinéaire qui assure la résolution (stable) de toute approximation interne. •

Démonstration. Pour obtenir (3.30), on majore l'estimation du Lemme 3.1.13 de Céa en choisissant $v_h = r_h u$ qui est bien un élément de V_{0h}

$$\|u - u_h\|_{V_0} \leq C \inf_{v_h \in V_{0h}} \|u - v_h\|_{V_0} \leq C\|u - r_h u\|_{V_0},$$

ce qui permet de conclure grâce au Lemme 3.2.5. □

Nous donnons maintenant la démonstration du Lemme 3.2.5 sous la forme d'un autre lemme technique.

Lemme 3.2.8 *Il existe une constante C indépendante de h telle que, pour tout $v \in C^2([0,1])$,*

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch^2 \|v''\|_{L^2(0,1)}, \tag{3.31}$$

et

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq Ch\|v''\|_{L^2(0,1)}. \tag{3.32}$$

Démonstration. Soit $v \in C^2([0,1])$. Par définition, l'interpolée $r_h v$ est une fonction affine et, pour tout $x \in]x_j, x_{j+1}[$, on a

$$\begin{aligned} v(x) - r_h v(x) &= v(x) - \left(v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j}(x - x_j) \right) \\ &= \int_{x_j}^{x} v'(t)\, dt - \frac{x - x_j}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v'(t)\, dt \\ &= (x - x_j)v'(x_j + \theta_x) - (x - x_j)v'(x_j + \theta_j) \\ &= (x - x_j) \int_{x_j + \theta_j}^{x_j + \theta_x} v''(t)\, dt, \end{aligned} \tag{3.33}$$

par application de la formule des accroissements finis avec $0 \leq \theta_x \leq x - x_j$ et $0 \leq \theta_j \leq h$. On en déduit en utilisant l'inégalité de Cauchy-Schwarz

$$|v(x) - r_h v(x)|^2 \leq h^2 \left(\int_{x_j}^{x_{j+1}} |v''(t)|\, dt \right)^2 \leq h^3 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt. \tag{3.34}$$

En intégrant (3.34) par rapport à x sur l'intervalle $[x_j, x_{j+1}]$, on obtient

$$\int_{x_j}^{x_{j+1}} |v(x) - r_h v(x)|^2\, dx \leq h^4 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt,$$

ce qui, par sommation en j, donne exactement (3.31). La démonstration de (3.32) est tout à fait similaire : pour $v \in C^2([0,1])$ et $x \in]x_j, x_{j+1}[$ on écrit

$$\begin{aligned} v'(x) - (r_h v)'(x) \;=\; v'(x) - \frac{v(x_{j+1}) - v(x_j)}{h} \;&=\; \frac{1}{h}\int_{x_j}^{x_{j+1}} \Big(v'(x) - v'(t)\Big)\,dt \\ &=\; \frac{1}{h}\int_{x_j}^{x_{j+1}} \int_t^x v''(y)\,dy. \end{aligned}$$

Élevant au carré cette inégalité, appliquant Cauchy-Schwarz deux fois et sommant en j on obtient (3.32). □

3.2.3 Éléments finis $\mathbb{P}_2$

La méthode des éléments finis $\mathbb{P}_2$ repose sur l'espace discret

$$V_h = \left\{ v \in C([0,1]) \text{ tel que } v\,\big|_{[x_j, x_{j+1}]} \in \mathbb{P}_2 \text{ pour tout } 0 \le j \le n \right\}, \tag{3.35}$$

et sur son sous-espace

$$V_{0h} = \{ v \in V_h \text{ tel que } v(0) = v(1) = 0 \}. \tag{3.36}$$

La méthode des éléments finis $\mathbb{P}_2$ est la méthode d'approximation variationnelle interne de la Sous-section 3.1.3 appliquée à ces espaces V_h ou V_{0h}. Ceux-ci sont composés de fonctions continues, paraboliques par morceaux qu'on peut représenter à l'aide de fonctions de base très simples.

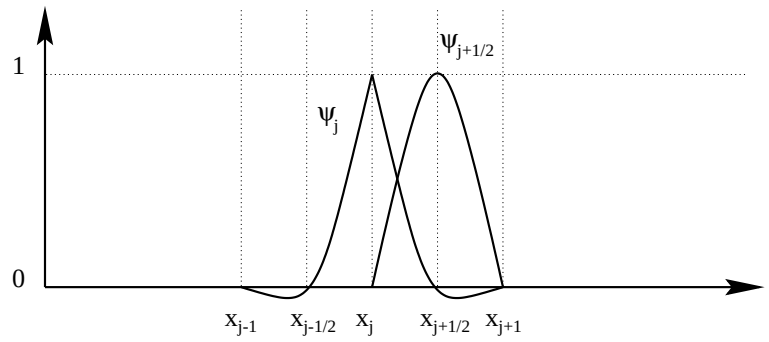

FIGURE 3.3 – Les fonctions de base des éléments finis $\mathbb{P}_2$.

Introduisons tout d'abord les points milieux des segments $[x_j, x_{j+1}]$ définis par $x_{j+1/2} = x_j + h/2$ pour $0 \le j \le n$. On définit aussi deux fonctions "mères"

$$\phi(x) = \begin{cases} (1+x)(1+2x) & \text{si } -1 \le x \le 0, \\ (1-x)(1-2x) & \text{si } 0 \le x \le 1, \\ 0 & \text{si } |x| > 1, \end{cases}$$

et

$$\psi(x) = \begin{cases} 1 - 4x^2 & \text{si } |x| \le 1/2, \\ 0 & \text{si } |x| > 1/2. \end{cases}$$

Si le maillage est uniforme, pour $0 \le j \le n+1$ on définit les fonctions de base (voir la Figure 3.3)

$$\psi_j(x) = \phi\left(\frac{x - x_j}{h}\right),\ 0 \le j \le n+1,\ \text{et}\ \psi_{j+1/2}(x) = \psi\left(\frac{x - x_{j+1/2}}{h}\right),\ 0 \le j \le n.$$

Lemme 3.2.9 *L'espace V_h, défini par (3.35), est un sous-espace de V de dimension $2n+3$, et toute fonction $v_h \in V_h$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{0\le j\le n+1}$ et aux milieux $(x_{j+1/2})_{0\le j\le n}$*

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j)\psi_j(x) + \sum_{j=0}^{n} v_h(x_{j+1/2})\psi_{j+1/2}(x) \quad \forall x \in [0,1].$$

De même, V_{0h}, défini par (3.36), est un sous-espace de V_0 de dimension $2n+1$, et toute fonction $v_h \in V_{0h}$ est définie de manière unique par ses valeurs aux sommets $(x_j)_{1\le j\le n}$ et aux milieux $(x_{j+1/2})_{0\le j\le n}$

$$v_h(x) = \sum_{j=1}^{n} v_h(x_j)\psi_j(x) + \sum_{j=0}^{n} v_h(x_{j+1/2})\psi_{j+1/2}(x) \quad \forall x \in [0,1].$$

Remarque 3.2.10 Ici encore, V_h est un espace d'éléments finis de Lagrange (cf. la Remarque 3.2.2). Comme les fonctions sont localement $\mathbb{P}_2$, on dit que l'espace V_h, défini par (3.35), est l'espace des éléments finis de Lagrange d'ordre 2. •

Démonstration. Les dimensions et les bases de V_h et V_{0h} se trouvent facilement en remarquant que $\psi_j(x_i) = \delta_{ij}$, $\psi_{j+1/2}(x_{i+1/2}) = \delta_{ij}$, $\psi_j(x_{i+1/2}) = 0$, $\psi_{j+1/2}(x_i) = 0$ (voir la Figure 3.3). □

Décrivons la **résolution pratique** du problème de Dirichlet (3.22) par la méthode des éléments finis $\mathbb{P}_2$. La formulation variationnelle (3.19) de l'approximation interne revient à résoudre dans $\mathbb{R}^{2n+1}$ un système linéaire $\mathcal{K}_h U_h = b_h$. On note $(x_{k/2})_{1\le k\le 2n+1}$ les points du maillage et $(\psi_{k/2})_{1\le k\le 2n+1}$ les fonctions de base correspondantes dans V_{0h}. Dans cette base $U_h \in \mathbb{R}^{2n+1}$ est le vecteur des coordonnées de la solution approchée u_h et la matrice de rigidité est ici pentadiagonale

$$\mathcal{K}_h = h^{-1}\begin{pmatrix}
16/3 & -8/3 & 0 & & & & & \\
-8/3 & 14/3 & -8/3 & 1/3 & & & 0 & \\
0 & -8/3 & 16/3 & -8/3 & 0 & & & \\
 & 1/3 & -8/3 & 14/3 & -8/3 & 1/3 & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & 0 & -8/3 & 16/3 & -8/3 & 0 \\
 & 0 & & & 1/3 & -8/3 & 14/3 & -8/3 \\
 & & & & & 0 & -8/3 & 16/3
\end{pmatrix}.$$

Remarquons que cette matrice est plus "pleine" que celle obtenue par la méthode des éléments finis $\mathbb{P}_1$, et donc que la résolution du système linéaire coûtera plus cher en temps de calcul. Pour évaluer le second membre b_h on a recours aux même formules de quadrature que celles présentées dans la méthode $\mathbb{P}_1$.

On admettra le résultat de convergence suivante.

Théorème 3.2.11 *Soit $u \in C^3([0,1])$ la solution exacte de (3.22) et $u_h \in V_{0h}$ la solution approchée par la méthode des éléments finis $\mathbb{P}_2$. Il existe une constante C indépendante de h telle que*

$$\|u - u_h\|_{V_0} \leq Ch^2 \|f'\|_{L^2(0,1)}.$$

Le Théorème 3.2.11 montre l'avantage principal des éléments finis $\mathbb{P}_2$: si la solution est régulière, alors la convergence de la méthode est **quadratique** (la vitesse de convergence est proportionnelle à h^2) alors que la convergence pour les éléments finis $\mathbb{P}_1$ est seulement linéaire (proportionnelle à h). Bien sûr cet avantage a un prix : il y a deux fois plus d'inconnues (exactement $2n+1$ au lieu de n pour les éléments finis $\mathbb{P}_1$) donc la matrice est deux fois plus grande, et en plus la matrice a cinq diagonales non nulles au lieu de trois dans le cas $\mathbb{P}_1$. Remarquons que si la solution u n'est pas régulière il n'y a aucun avantage théorique **mais aussi pratique** à utiliser des éléments finis $\mathbb{P}_2$ plutôt que $\mathbb{P}_1$.

3.3 Problèmes d'évolution

Dans cette section nous étendons l'approche précédente aux problèmes d'évolution en temps (dans les sections précédentes nous avions étudié des problèmes stationnaires sans variable de temps). Nous allons plus particulièrement analyser deux exemples différents d'équations aux dérivées partielles : l'équation de la chaleur et l'équation des ondes. Plus généralement, l'approche développée ici s'étend à beaucoup d'autres problèmes d'évolution en temps, comme, par exemple, les équations de Stokes instationnaires, l'élastodynamique, l'électromagnétisme ou l'équation de Schrödinger de la mécanique quantique.

La résolution numérique de l'équation de la chaleur ou de celle des ondes s'obtient à nouveau grâce au concept de **formulation variationnelle** qui conduit à des méthodes **d'éléments finis**. Plus précisément, conformément à l'usage, nous utiliserons des éléments finis pour la discrétisation spatiale, mais des différences finies pour la discrétisation temporelle.

3.3.1 Modéles

Nous présentons tout d'abord un modèle du premier ordre en temps (et d'ordre deux en espace), dit problème parabolique. L'archétype de ces modèles est **l'équation de la chaleur** dont l'origine physique a déjà été discutée au Chapitre 1. Soit Ω un ouvert borné de $\mathbb{R}^N$ de frontière $\partial\Omega$. Pour des conditions aux limites de Dirichlet ce modèle s'écrit

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(x,0) = u_0(x) & \text{pour } x \in \Omega. \end{cases} \tag{3.37}$$

Le problème aux limites (3.37) modélise l'évolution de la température $u(x,t)$ dans un corps thermiquement conducteur qui occupe le domaine Ω. La distribution de température initiale, à $t = 0$, est donnée par la fonction u_0. Sur le bord $\partial\Omega$ du

corps considéré, la température est maintenue à une valeur constante, utilisée comme valeur de référence (c'est la condition de Dirichlet homogène $u(x,t) = 0$ sur $\partial\Omega\times\mathbb{R}_+$). Les sources de chaleur sont modélisées par la fonction donnée $f = f(x,t)$. Notons que les variables $x \in \Omega$ et $t \in \mathbb{R}_+$ jouent des rôles très différents dans (3.37).

Indiquons qu'il existe d'autres origines physiques du système (3.37). Par exemple, (3.37) modélise aussi la diffusion d'une concentration u dans le domaine Ω, ou bien l'évolution du champ de pression u d'un fluide s'écoulant dans un milieu poreux (système de Darcy), ou encore la loi d'un mouvement brownien dans le domaine Ω.

On peut, bien sûr, associer d'autres conditions aux limites à l'équation de la chaleur (par exemple, une condition de Neumann homogène si la paroi du corps Ω est adiabatique).

Une première généralisation évidente de l'équation de la chaleur s'obtient lorsque l'on remplace le Laplacien par un opérateur elliptique du deuxième ordre plus général. Cette généralisation se rencontre, par exemple, si on étudie la propagation de la chaleur dans un matériau non homogène ou en présence d'un effet convectif. Une deuxième généralisation (moins évidente) concerne le système des équations de Stokes instationnaires qui généralise le cas stationnaire vu dans la Section 1.3. En notant u la vitesse et p la pression d'un fluide visqueux soumis à des forces f, ce système s'écrit

$$\begin{cases} \frac{\partial u}{\partial t} + \nabla p - \mu\Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ \mathrm{div} u = 0 & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(x, t = 0) = u_0(x) & \text{dans } \Omega \end{cases} \tag{3.38}$$

où $\mu > 0$ est la viscosité du fluide. Rappelons que la condition aux limites de Dirichlet homogène modélise l'adhérence du fluide à la paroi de Ω (voir le Chapitre 1), et que le système de Stokes n'est valable que pour des vitesses faibles.

Nous passons maintenant à un modèle du second ordre en temps (et d'ordre deux en espace), dit problème hyperbolique. L'archétype de ces modèles est **l'équation des ondes** dont l'origine physique a déjà été discutée au Chapitre 1. Soit Ω un ouvert borné de $\mathbb{R}^N$ de frontière $\partial\Omega$. Pour des conditions aux limites de Dirichlet ce modèle s'écrit

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0(x) & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{dans } \Omega. \end{cases} \tag{3.39}$$

Le problème aux limites (3.39) modélise, par exemple, la propagation au cours du temps du déplacement vertical d'une membrane élastique, ou bien de l'amplitude d'un champ électrique de direction constante. L'inconnue $u(t,x)$ est ici une fonction scalaire.

Une généralisation est **l'élastodynamique** qui est la version d'évolution en temps des équations de l'élasticité linéarisée (voir le Chapitre 1). Par application du principe fondamental de la dynamique, l'accélération étant la dérivée seconde en temps du déplacement, on obtient un problème d'évolution d'ordre deux en temps comme (3.39). Néanmoins, une différence importante avec (3.39) est que l'inconnue

$u(t,x)$ est désormais une fonction à valeurs vectorielles dans $\mathbb{R}^N$. Plus précisément, si on note $f(t,x)$ la résultante (vectorielle) des forces extérieures, le déplacement $u(t,x)$ est solution de

$$\begin{cases} \rho\frac{\partial^2 u}{\partial t^2} - \operatorname{div}\left(2\mu e(u) + \lambda \operatorname{tr}(e(u))\operatorname{Id}\right) = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t=0) = u_0(x) & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(t=0) = u_1(x) & \text{dans } \Omega, \end{cases} \tag{3.40}$$

où u_0 est le déplacement initial, u_1 la vitesse initiale, et $e(u) = (\nabla u + (\nabla u)^t)/2$ le tenseur des déformations. Supposant homogène isotrope le matériau qui occupe Ω, sa densité est constante $\rho > 0$, de même que ses modules de Lamé qui vérifient $\mu > 0$ et $2\mu + N\lambda > 0$.

3.3.2 Formulation variationnelle

L'idée est d'écrire une formulation variationnelle qui ressemble à une **équation différentielle ordinaire linéaire** en temps (celle-ci est alors facile à résoudre). Autrement dit, on utilise des fonctions test $v(x)$ qui ne dépendent pas du temps t et on n'intègre que par rapport à la variable d'espace.

Nous commençons par le cas de l'équation de la chaleur. On suppose que le terme source et la donnée intiale sont des fonctions continues, $f(x,t) \in C(\overline{\Omega} \times \mathbb{R}^+)$ et $u_0(x) \in C(\overline{\Omega})$.

Proposition 3.3.1 *Soit $u(x,t)$ une fonction de $C^2(\overline{\Omega} \times \mathbb{R}^+)$. Soit V_0 l'espace vectoriel défini par*

$$V_0 = \left\{\phi \in C^1(\overline{\Omega}) \text{ tel que } \phi = 0 \text{ sur } \partial\Omega\right\}.$$

Alors u est une solution l'équation de la chaleur (3.37) si et seulement si, pour tout temps $t \in \mathbb{R}^+$ $u(\cdot,t)$ appartient à V_0 et vérifie, pour toute fonction test $v \in V_0$,

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)v(x)\,dx + \int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,dx = \int_\Omega f(x,t)v(x)\,dx, \tag{3.41}$$

avec la condition initiale $u(x,0) = u_0(x)$. L'égalité (3.41) et la condition initiale constituent la **formulation variationnelle** *de (3.37).*

Remarque 3.3.2 Comme précédemment, un intérêt immédiat de la formulation variationnelle (3.41) est qu'elle a un sens si la solution u est seulement dérivable une fois en espace, contrairement à la formulation "classique" (3.37) qui requiert que u soit deux fois dérivable en x. Par contre, on ne "gagne" pas de régularité en temps. Remarquons aussi que la condition de régularité sur u et le fait que $u(\cdot,t) \in V_0$, pour tout $t \geq 0$, implique que u_0 n'est pas n'importe quelle fonction continue (au moins dans ce cadre). On peut néanmoins diminuer l'hypothèse de régularité $u(x,t) \in C^2(\overline{\Omega} \times \mathbb{R}^+)$. •

Démonstration. Si u est solution de l'équation de la chaleur (3.37), on multiplie l'équation par $v \in V_0$ et on utilise la formule d'intégration par parties du Corollaire 3.1.4, à t fixé,

$$-\int_\Omega \Delta u(x,t)v(x)\,dx = \int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,dx - \int_{\partial\Omega} \frac{\partial u}{\partial n}(x,t)v(x)\,ds.$$

Or $v = 0$ sur $\partial\Omega$ puisque $v \in V_0$, donc

$$\int_\Omega f(x,t)v(x)\,dx = \int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,dx + \int_\Omega \frac{\partial u}{\partial t}(x,t)v(x)\,dx,$$

qui n'est rien d'autre que la formule (3.41). Réciproquement, si u vérifie (3.41), en utilisant "à l'envers" la formule d'intégration par parties précédente on obtient

$$\int_\Omega \Big(\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) - f(x,t)\Big)v(x)\,dx = 0 \text{ pour toute fonction } v \in V_0.$$

Comme $(\frac{\partial u}{\partial t} - \Delta u - f)$ est une fonction continue en x pour tout t, grâce au Lemme 3.1.7 on conclut que cette fonction est nulle pour tout t, x. Par ailleurs, comme $u(\cdot,t) \in V_0$, on retrouve la condition aux limites $u = 0$ sur $\partial\Omega$ pour tout temps t. De plus, la condition initiale $u(\cdot,0) = u_0$ fait partie de la formulation variationnelle. On retrouve donc bien que u est solution de (3.37). □

Un résultat similaire a lieu pour l'équation des ondes et nous en laissons la démonstration au lecteur en exercice. En plus des hypothèses précédentes sur f et u_0, nous supposons aussi que $u_1(x) \in C(\overline{\Omega})$.

Proposition 3.3.3 *Soit $u(x,t)$ une fonction de $C^2(\overline{\Omega}\times\mathbb{R}^+)$. Soit V_0 l'espace vectoriel défini par*

$$V_0 = \left\{\phi \in C^1(\overline{\Omega}) \text{ tel que } \phi = 0 \text{ sur } \partial\Omega\right\}.$$

Alors u est une solution de l'équation des ondes (3.39) si et seulement si, pour tout temps $t \in \mathbb{R}^+$ $u(\cdot,t)$ appartient à V_0 et vérifie, pour toute fonction test $v \in V_0$,

$$\int_\Omega \frac{\partial^2 u}{\partial t^2}(x,t)v(x)\,dx + \int_\Omega \nabla u(x,t)\cdot\nabla v(x)\,dx = \int_\Omega f(x,t)v(x)\,dx, \tag{3.42}$$

avec les conditions initiales $u(x,0) = u_0(x)$ et $\frac{\partial u}{\partial t}(x,0) = u_1(x)$. L'égalité (3.42) et les conditions initiales constituent la **formulation variationnelle** *de (3.39).*

Comme dans la Remarque 3.1.10 on peut introduire des notations abstraites. Pour simplifier nous notons $f(t)$ et $u(t)$ les fonctions $f(\cdot,t)$ et $u(\cdot,t)$. On introduit le produit scalaire de $L^2(\Omega)$ et la forme bilinéaire $a(w,v)$ définis par

$$\langle w,v\rangle_{L^2(\Omega)} = \int_\Omega w(x)v(x)\,dx \quad\text{et}\quad a(w,v) = \int_\Omega \nabla w(x)\cdot\nabla v(x)\,dx.$$

On fait alors la remarque que ni Ω ni $v(x)$ ne varient avec le temps t, ce qui permet de réécrire

$$\int_\Omega \frac{\partial u}{\partial t}(x,t)v(x)\,dx = \frac{d}{dt}\int_\Omega u(x,t)v(x)\,dx.$$

Si l'on se donne un temps final $T > 0$ (éventuellement égal à $+\infty$), la formulation variationnelle de l'équation de la chaleur (3.37) est : trouver $u(t)$ fonction de $[0, T]$ à valeurs dans V_0 telle que

$$\begin{cases} \dfrac{d}{dt}\langle u(t), v\rangle_{L^2(\Omega)} + a\left(u(t), v\right) = \langle f(t), v\rangle_{L^2(\Omega)} \quad \forall\, v \in V_0, \quad 0 < t < T, \\ u(t=0) = u_0. \end{cases} \tag{3.43}$$

De même, la formulation variationnelle de l'équation des ondes (3.39) est : trouver $u(t)$ fonction de $[0, T]$ à valeurs dans V_0 telle que

$$\begin{cases} \dfrac{d^2}{dt^2}\langle u(t), v\rangle_{L^2(\Omega)} + a\left(u(t), v\right) = \langle f(t), v\rangle_{L^2(\Omega)} \quad \forall\, v \in V_0,\ 0 < t < T, \\ u(t=0) = u_0, \quad \dfrac{du}{dt}(t=0) = u_1. \end{cases} \tag{3.44}$$

3.3.3 Semi-discrétisation par éléments finis en espace

Il s'agit de discrétiser **en espace seulement** la formulation variationnelle (3.43) ou (3.44). Pour cela, comme dans le cas des problèmes elliptiques, nous construisons une approximation variationnelle interne en introduisant un sous-espace V_{0h} de V_0, de dimension finie.

Nous commençons par l'équation de la chaleur (3.37). La semi-discrétisation de (3.43) est donc l'approximation variationnelle interne suivante : trouver $u_h(t)$ fonction de $[0, T]$ à valeurs dans V_{0h} telle que

$$\begin{cases} \dfrac{d}{dt}\langle u_h(t), v_h\rangle_{L^2(\Omega)} + a\left(u_h(t), v_h\right) = \langle f(t), v_h\rangle_{L^2(\Omega)} \quad \forall\, v_h \in V_{0h}, 0 < t < T, \\ u_h(t=0) = u_{0,h} \end{cases} \tag{3.45}$$

où $u_{0,h} \in V_{0h}$ est une approximation de la donnée initiale u_0. Cette méthode d'approximation est aussi connue sous le nom de "méthode des lignes".

Nous allons montrer que (3.45) admet une unique solution car il s'agit en fait un système **d'équations différentielles ordinaires** à coefficients constants dont on calcule facilement l'unique solution.

Lemme 3.3.4 *Soit V_{0h} un sous-espace de dimension finie de l'espace vectoriel normé V_0. Soit $a(u, v)$ une forme bilinéaire, coercive sur V_0, au sens où il existe une constante $\nu > 0$ telle que $a(v, v) \geq \nu \|v\|^2_{V_0}$ pour tout $v \in V_0$. Alors l'approximation variationnelle interne (3.45) admet une unique solution.*

Démonstration. Comme V_{0h} est de dimension finie, on introduit une base $(\phi_j)_{1 \leq j \leq N_h}$ de V_{0h}. Dans cette base la solution s'écrit sous la forme

$$u_h(t) = \sum_{i=1}^{n_{dl}} U_i^h(t)\phi_i, \tag{3.46}$$

avec $U^h = (U_i^h)_{1\leq i\leq n_{dl}}$ le vecteur des coordonnées de u_h. Il est important de noter que dans (3.46) les fonctions de base ϕ_i ne dépendent pas du temps et que seules les coordonnées $U_i^h(t)$ sont des fonctions du temps t. De même, on pose

$$u_{0,h} = \sum_{i=1}^{n_{dl}} U_i^{0,h}\phi_i,$$

et (3.45) devient, pour tout $1 \leq j \leq n_{dl}$,

$$\begin{cases} \displaystyle\sum_{i=1}^{n_{dl}} \langle \phi_i, \phi_j\rangle_{L^2(\Omega)} \frac{dU_i^h(t)}{dt} + \sum_{i=1}^{n_{dl}} a(\phi_i,\phi_j) U_i^h(t) = \langle f(t), \phi_j\rangle_{L^2(\Omega)} \\ U_j^h(t=0) = U_j^{0,h} \end{cases}$$

Introduisant la **matrice de masse** $\mathcal{M}_h$ définie par

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j\rangle_{L^2(\Omega)} \quad 1 \leq i,j \leq n_{dl},$$

et la **matrice de rigidité** $\mathcal{K}_h$ définie par

$$(\mathcal{K}_h)_{ij} = a(\phi_i,\phi_j) \quad 1 \leq i,j \leq n_{dl},$$

l'approximation variationnelle (3.45) est équivalente au système linéaire **d'équations différentielles ordinaires** à coefficients constants

$$\begin{cases} \mathcal{M}_h \dfrac{dU^h}{dt}(t) + \mathcal{K}_h U^h(t) = b^h(t), \quad 0 < t < T, \\ U^h(t=0) = U^{0,h} \end{cases} \tag{3.47}$$

avec $b_i^h(t) = \langle f(t), \phi_i\rangle_{L^2(\Omega)}$. Les appellations "matrices de masse et de rigidité" proviennent des applications en mécanique des solides. On a déjà démontré que la matrice de rigidité $\mathcal{K}_h$ est définie positive donc inversible (voir la preuve du Lemme 3.1.12). La même démonstration s'applique à la matrice de masse $\mathcal{M}_h$ qui est aussi définie positive donc inversible. L'existence et l'unicité, ainsi qu'une formule explicite, de la solution de (3.47) s'obtiennent classiquement par simple diagonalisation simultanée de $\mathcal{M}_h$ et $\mathcal{K}_h$ (voir par exemple le théorème 2.3.6 dans [2]). □

Exercice 3.3.1 On se propose de calculer la matrice de masse pour la méthode des éléments finis $\mathbb{P}_1$. On reprend les notations de la Section 3.2. Montrer que la matrice de masse $\mathcal{M}_h$ est donnée par

$$\mathcal{M}_h = h \begin{pmatrix} 2/3 & 1/6 & & & 0 \\ 1/6 & 2/3 & 1/6 & & \\ & \ddots & \ddots & \ddots & \\ & & 1/6 & 2/3 & 1/6 \\ 0 & & & 1/6 & 2/3 \end{pmatrix}.$$

Montrer que, si on utilise la formule de quadrature (3.28), alors on trouve que $\mathcal{M}_h = h\,\mathrm{Id}$ (on appelle cette procédure la condensation de masse ou "mass-lumping").

Nous passons maintenant à l'équation des ondes (3.39). La semi-discrétisation de (3.44) est donc l'approximation variationnelle interne suivante : trouver $u_h(t)$ fonction de $[0,T]$ à valeurs dans V_{0h} telle que

$$\left\{ \begin{array}{l} \dfrac{d^2}{dt^2}\langle u_h(t), v_h\rangle_{L^2(\Omega)} + a\left(u_h(t), v_h\right) = \langle f(t), v_h\rangle_{L^2(\Omega)} \quad \forall\, v_h \in V_{0h},\ 0 < t < T, \\ u_h(t=0) = u_{0,h}, \quad \dfrac{\partial u_h}{\partial t}(t=0) = u_{1,h}, \end{array} \right. \tag{3.48}$$

où $u_{0,h} \in V_{0h}$ et $u_{1,h} \in V_{0h}$ sont des approximations des données initiales u_0 et u_1.

On montre facilement un équivalent du Lemme 3.3.4, c'est-à-dire que (3.48) admet une solution unique car (3.48) est équivalent au système linéaire **d'équations différentielles ordinaires** d'ordre 2 à coefficients constants

$$\left\{ \begin{array}{l} \mathcal{M}_h \dfrac{d^2U^h}{dt^2}(t) + \mathcal{K}_h U^h(t) = b^h(t), \quad 0 < t < T, \\ U^h(t=0) = U^{0,h}, \quad \dfrac{dU^h}{dt}(t=0) = U^{1,h}, \end{array} \right. \tag{3.49}$$

avec les mêmes matrices de masse $\mathcal{M}_h$ et de rigidité $\mathcal{K}_h$ que pour l'équation de la chaleur.

Comme il est difficile et coûteux de diagonaliser (3.47), en pratique on résout numériquement (3.47) par discrétisation et marche en temps. Il existe de nombreuses méthodes classiques de calcul numérique des solutions d'équations différentielles ordinaires. Nous en verrons quelques unes dans la sous-section suivante.

3.3.4 Discrétisation totale en espace-temps

Après avoir discrétisé les équations précédentes en espace par une méthode d'éléments finis, on termine la discrétisation du problème en utilisant une méthode de **différences finies en temps**. Concrètement, on utilise des schémas de différences finies pour résoudre les systèmes d'équations différentielles ordinaires (3.47) et (3.49) issus de la semi-discrétisation en espace. Nous allons donc retrouver de nombreux schémas déjà étudiés au Chapitre 2 ainsi que des notions telle que la stabilité ou l'ordre de précision.

Nous commençons par l'équation de la chaleur. Pour simplifier les notations, nous réécrivons le système (3.47) sans mentionner la dépendance par rapport au paramètre h du maillage spatial

$$\left\{ \begin{array}{l} \mathcal{M}\dfrac{dU}{dt}(t) + \mathcal{K}U(t) = b(t) \\ U(t=0) = U^0 \end{array} \right. \tag{3.50}$$

On découpe l'intervalle de temps $[0,T]$ en n_0 intervalles ou pas de temps $\Delta t = T/n_0$ et on pose

$$t_n = n\Delta t \quad 0 \leq n \leq n_0.$$

On note U^n l'approximation de $U(t_n)$ calculé par un schéma. Pour calculer numériquement des solutions approchées de (3.50) le schéma le plus simple et le plus utilisé est le **θ-schéma** (déjà vu, voir (2.5))

$$\mathcal{M}\frac{U^{n+1}-U^n}{\Delta t}+\mathcal{K}\left(\theta U^{n+1}+(1-\theta)U^n\right)=\theta b(t_{n+1})+(1-\theta)b(t_n). \tag{3.51}$$

Lorsque $\theta=0$, on appelle (3.51) **schéma explicite**, lorsque $\theta=1$, **schéma implicite**, et pour $\theta=1/2$, **schéma de Crank-Nicholson**. On peut réécrire (3.51) sous la forme

$$(\mathcal{M}+\theta\Delta t\mathcal{K})U^{n+1}=(\mathcal{M}-(1-\theta)\Delta t\mathcal{K})U^n+\Delta t(\theta b(t_{n+1})+(1-\theta)b(t_n)). \tag{3.52}$$

Remarquons qu'en général la matrice $\mathcal{M}$ n'est pas diagonale, et donc que, même pour le schéma explicite, il est nécessaire de résoudre un système linéaire pour calculer U^{n+1} en fonction de U^n et du second membre b (sauf si on utilise une formule d'intégration numérique qui rende $\mathcal{M}$ diagonale, voir l'Exercice 3.3.1). Évidemment, on peut construire une foule de schémas en s'inspirant de ceux du Chapitre 2. Ces schémas sont bien sûr consistants (voir la Définition 2.2.4) et on peut facilement analyser leur précision (uniquement par rapport à la variable de temps).

Exercice 3.3.2 Montrer que le schéma de Crank-Nicholson est d'ordre 2 (en temps), tandis que le θ-schéma pour $\theta\neq 1/2$ est d'ordre 1.

Nous donnons maintenant une définition de la stabilité de ces schémas qui est une variante de la Définition 2.2.8.

Définition 3.3.5 *Un schéma aux différence finies pour (3.50) est dit stable si*

$$\mathcal{M}U^n\cdot U^n\leq C \text{ pour tout } 0\leq n\leq n_0=T/\Delta t,$$

où la constante $C>0$ est indépendante de Δt et de la dimension du système n_{dl} (donc du pas du maillage h), mais peut dépendre de la donnée initiale U^0, du second membre b, et de T.

Remarque 3.3.6 Le choix de la norme $\sqrt{\mathcal{M}U\cdot U}$ dans la Définition 3.3.5 s'explique par le fait que $\mathcal{M}U\cdot U=\int_\Omega |u|^2dx$ avec $u\in V_{0h}$ la fonction de coordonnées U dans la base choisie de V_{0h} ($\mathcal{M}$ est bien définie positive). Rappelons que, dans la Définition 2.2.8 de la stabilité au sens des différences finies, on pondérait par Δx la norme euclidienne de U pour retrouver aussi l'analogie avec la norme de u dans $L^2(\Omega)$. •

Lemme 3.3.7 *Si $1/2\leq\theta\leq 1$, le θ-schéma (3.51) est inconditionnellement stable, tandis que, si $0\leq\theta<1/2$, il est stable sous la condition CFL*

$$\max_i \lambda_i\,\Delta t\leq\frac{2}{1-2\theta}, \tag{3.53}$$

où les λ_i sont les valeurs propres de $\mathcal{K}U=\lambda\mathcal{M}U$ (voir (3.71)).

Remarque 3.3.8 On ne reconnaît pas immédiatement dans (3.53) la condition CFL (Courant-Friedrichs-Lewy) usuelle $\Delta t \leq Ch^2$ pour l'équation de la chaleur (voir la Sous-section 2.2.3). En fait, le lecteur peut vérifier en dimension $N = 1$ lors de l'Exercice 3.3.3 ci-dessous qu'on a effectivement $\max_i \lambda_i = \mathcal{O}(h^{-2})$. En pratique on n'utilise pas le θ-schéma pour $\theta < 1/2$ car la condition de stabilité (3.53) est beaucoup trop sévère : elle oblige l'usage de très petits pas de temps qui rendent le calcul beaucoup trop coûteux. •

Démonstration. On réécrit le schéma (3.52) dans la base orthonormale pour $\mathcal{M}$ et diagonale pour $\mathcal{K}$ (voir la démonstration du Lemme 3.4.3)

$$(\,\mathrm{Id} + \theta\Delta t\,\mathrm{diag}(\lambda_i))\tilde{U}^{n+1} = (\,\mathrm{Id} - (1-\theta)\Delta t\,\mathrm{diag}(\lambda_i))\tilde{U}^n + \Delta t\tilde{b}^n, \tag{3.54}$$

avec $\mathcal{M} = PP^*$, $\mathcal{K} = P\,\mathrm{diag}(\lambda_i)P^*$, $\tilde{U}^n = P^*U^n$, et $\tilde{b}^n = P^{-1}(\theta b(t_{n+1}) + (1-\theta)b(t_n))$. On déduit de (3.54) que les composantes $\tilde{U}_i^n$ de $\tilde{U}^n$ vérifient

$$\tilde{U}_i^n = (\rho_i)^n\,\tilde{U}_i^0 + \frac{\Delta t}{1+\theta\Delta t\lambda_i}\sum_{k=1}^{n}(\rho_i)^{k-1}\tilde{b}_i^{n-k}. \tag{3.55}$$

avec

$$\rho_i = \frac{1-(1-\theta)\Delta t\lambda_i}{1+\theta\Delta t\lambda_i}.$$

Dans cette base la condition de stabilité est $\|U^n\|_{\mathcal{M}} = \|\tilde{U}^n\| \leq C$. Par conséquent, une condition nécessaire et suffisante de stabilité est $|\rho_i| \leq 1$ pour tout i, ce qui n'est rien d'autre que la condition (3.53) si $0 \leq \theta < 1/2$, et qui est toujours satisfaite si $\theta \geq 1/2$. □

Remarque 3.3.9 Il est clair dans l'estimation (3.55) que plus θ est grand, plus le coefficient devant le terme $\tilde{b}_i^{n-k}$ est petit. En fait, cette propriété correspond à un amortissement exponentiel par le schéma des contributions passées du terme source. Par conséquent, même si pour toute valeur $1/2 < \theta \leq 1$ le θ-schéma est stable, son maximum de stabilité est atteint pour $\theta = 1$ (les erreurs numériques passées s'amortissent). C'est pourquoi le schéma implicite est plus robuste et souvent utilisé pour des problèmes "raides" bien qu'il soit moins précis que le schéma de Crank-Nicholson. •

Exercice 3.3.3 On résout par éléments finis $\mathbb{P}_1$ et schéma explicite en temps l'équation de la chaleur (3.37) en dimension $N = 1$. On utilise une formule de quadrature qui rend la matrice $\mathcal{M}$ diagonale (voir l'Exercice 3.4.4). On rappelle que la matrice $\mathcal{K}$ est donnée par (3.27) et qu'on a calculé ses valeurs propres lors de l'Exercice 3.4.4. Montrer que dans ce cas la condition CFL (3.53) est bien du type $\Delta t \leq Ch^2$.

Passons maintenant à l'équation des ondes. Pour simplifier les notations, nous réécrivons le système d'équations différentielles ordinaires (3.49) sans mentionner la dépendance spatiale en h

$$\begin{cases} \mathcal{M}\dfrac{d^2U}{dt^2}(t) + \mathcal{K}U(t) = b(t) \\ U(t=0) = U_0, \quad \dfrac{dU}{dt}(t=0) = U_1. \end{cases} \tag{3.56}$$

Avec les mêmes notations, $t_n = n\Delta t$ et U^n l'approximation de $U(t_n)$, pour $0 \leq \theta \leq 1/2$ on propose le **θ-schéma**

$$\begin{aligned} \mathcal{M}\frac{U^{n+1} - 2U^n + U^{n-1}}{(\Delta t)^2} \quad & +\mathcal{K}\left(\theta U^{n+1} + (1-2\theta)U^n + \theta U^{n-1}\right) \\ & = \theta b(t_{n+1}) + (1-2\theta)b(t_n) + \theta b(t_{n-1}). \end{aligned} \tag{3.57}$$

Lorsque $\theta = 0$, on appelle (3.57) **schéma explicite** (il n'est en fait vraiment explicite que si la matrice de masse $\mathcal{M}$ est diagonale). Pour démarrer le schéma il faut connaître U^0 et U^1, ce qu'on obtient grâce aux conditions initiales

$$U^0 = U_0 \quad \text{et} \quad \frac{U^1 - U^0}{\Delta t} = U_1.$$

On peut étudier la stabilité de ce schéma au sens de la Définition 3.3.5. Pour éviter des calculs trop lourds, on se contente d'énoncer la condition nécessaire de stabilité de Von Neumann (voir la Remarque 2.2.21). Le résultat suivant est dans le même esprit que le Lemme 2.3.5.

Lemme 3.3.10 *On considère le θ-schéma. La condition nécessaire de stabilité de Von Neumann est toujours vérifiée si $1/2 \leq 2\theta \leq 1$, tandis que, si $0 \leq 2\theta < 1/2$ elle n'est satisfaite que sous la condition CFL*

$$\max_i \lambda_i \, (\Delta t)^2 < \frac{4}{1-4\theta}, \tag{3.58}$$

où les λ_i sont les valeurs propres de $\mathcal{K}U = \lambda \mathcal{M}U$ (voir (3.71)).

3.4 Problèmes aux valeurs propres

Cette section est consacrée à la théorie spectrale des équations aux dérivées partielles, c'est-à-dire à l'étude des valeurs propres et des fonctions propres de ces équations. La motivation de cette étude est double. D'une part, cela va nous permettre d'étudier des solutions particulières, dites oscillantes en temps (ou vibrantes), des problèmes d'évolution associés à ces équations. D'autre part, il est possible d'en déduire une méthode de résolution générale de ces mêmes problèmes d'évolution.

3.4.1 Modèle et motivation

Donnons tout de suite l'exemple du **problème aux valeurs propres** pour le Laplacien avec condition aux limites de Dirichlet. Si Ω est un ouvert borné de $\mathbb{R}^N$ on cherche les couples $(\lambda, u) \in \mathbb{R} \times C^2(\Omega)$, avec $u \neq 0$, solutions de

$$\begin{cases} -\Delta u = \lambda u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \tag{3.59}$$

Le réel λ est appelé **valeur propre**, et la fonction $u(x)$ **mode propre ou fonction propre**. L'ensemble des valeurs propres est appelé le spectre de (3.59). On peut

faire l'analogie entre (3.59) et le problème plus simple de détermination des valeurs et vecteurs propres d'une matrice A d'ordre n,

$$Au = \lambda u \quad \text{avec} \quad (\lambda, u) \in \mathbb{R} \times \mathbb{R}^n, \tag{3.60}$$

en affirmant que l'opérateur $-\Delta$ est une "généralisation" en dimension infinie d'une matrice A en dimension finie. La résolution de (3.59) sera utile pour résoudre les problèmes d'évolution, de type parabolique ou hyperbolique, associés au Laplacien, c'est-à-dire l'équation de la chaleur (3.63) ou l'équation des ondes (3.65). Néanmoins, les solutions de (3.59) ont aussi une interprétation physique qui leur est propre, par exemple comme modes propres de vibration.

Remarque 3.4.1 Nous ne décrivons pas comment la résolution de (3.59) permet de résoudre des problèmes d'évolution (voir [1] pour des détails). Néanmoins nous en expliquons l'idée centrale en utilisant l'analogie formelle avec (3.60). Supposons que A est une matrice symétrique réelle, définie positive, d'ordre n. On note $\lambda_k > 0$ ses valeurs propres et r_k ses vecteurs propres, $1 \leq k \leq n$, tels que $Ar_k = \lambda_k r_k$. Il est bien connu que le système différentiel du premier ordre

$$\begin{cases} \dfrac{\partial u}{\partial t} + Au = 0 \ \text{ pour } t \geq 0 \\ u(t=0) = u_0, \end{cases} \tag{3.61}$$

admet comme solution unique

$$u(t) = \sum_{k=1}^{n} u_k^0 e^{-\lambda_k t} r_k,$$

avec la décomposition de la donnée initiale sous la forme $u_0 = \sum_{k=1}^n u_k^0 r_k$. De même, la solution unique du système différentiel du deuxième ordre

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + Au = 0 \ \text{ pour } t \geq 0 \\ u(t=0) = u_0, \\ \dfrac{\partial u}{\partial t}(t=0) = u_1, \end{cases} \tag{3.62}$$

est

$$u(t) = \sum_{k=1}^{n} \left(u_k^0 \cos\left(\sqrt{\lambda_k} t\right) + \frac{u_k^1}{\sqrt{\lambda_k}} \sin\left(\sqrt{\lambda_k} t\right) \right) r_k,$$

où $u_1 = \sum_{k=1}^n u_k^1 r_k$. Il est clair sur ces deux exemples que la connaissance du spectre de la matrice A permet de résoudre les problèmes d'évolution (3.61) et (3.62). •

Afin de se convaincre que (3.59) est bien la "bonne" formulation du problème aux valeurs propres pour le Laplacien, on peut passer par un argument de "séparation des variables" dans l'équation de la chaleur ou l'équation des ondes que nous décrivons

formellement. En l'absence de terme source, et en "oubliant" (provisoirement) la condition initiale et les conditions aux limites, nous cherchons une solution $\mathbf{u}$ de ces équations qui s'écrive sous la forme

$$\mathbf{u}(x,t) = \phi(t)u(x),$$

c'est-à-dire que l'on sépare les variables de temps et d'espace. Si $\mathbf{u}$ est solution de l'équation de la chaleur

$$\frac{\partial \mathbf{u}}{\partial t} - \Delta \mathbf{u} = 0, \tag{3.63}$$

on trouve (au moins formellement) que

$$\frac{\phi'(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

où $\lambda \in \mathbb{R}$ est une constante indépendante de t et de x. On en déduit que $\phi(t) = e^{-\lambda t}$ et que u doit être solution du problème aux valeurs propres

$$-\Delta u = \lambda u \tag{3.64}$$

muni de conditions aux limites adéquates.

De la même manière, si $\mathbf{u}$ est solution de l'équation des ondes

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} - \Delta \mathbf{u} = 0, \tag{3.65}$$

on trouve que

$$\frac{\phi''(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

où $\lambda \in \mathbb{R}$ est une constante. Cette fois-ci on en déduit que, si $\lambda > 0$ (ce qui sera effectivement le cas), alors $\phi(t) = a\cos(\sqrt{\lambda}t) + b\sin(\sqrt{\lambda}t)$ et que u doit encore être solution de (3.64). Remarquons que, si le comportement en espace de la solution $\mathbf{u}$ est le même pour l'équation de la chaleur et pour l'équation des ondes, il n'en est pas de même pour son comportement en temps : elle oscille en temps pour les ondes alors qu'elle décroît exponentiellement en temps (car $\lambda > 0$) pour la chaleur.

Exercice 3.4.1 Soit $\Omega = \mathbb{R}^N$. Montrer que $u(x) = \exp(ik \cdot x)$ est une solution de (3.64) si $|k|^2 = \lambda$. Une telle solution est appelée onde plane.

Exercice 3.4.2 Soit un potentiel régulier $V(x)$. Montrer que, si $\mathbf{u}(x,t) = e^{-i\omega t}u(x)$ est solution de

$$i\frac{\partial \mathbf{u}}{\partial t} + \Delta \mathbf{u} - V\mathbf{u} = 0 \ \text{ dans } \mathbb{R}^N \times \mathbb{R}_*^+, \tag{3.66}$$

alors $u(x)$ est solution de

$$-\Delta u + Vu = \omega u \ \text{ dans } \mathbb{R}^N. \tag{3.67}$$

On retrouve le même type de problème spectral que (3.64), à l'addition d'un terme d'ordre zéro près. Pour l'équation de Schrödinger la valeur propre ω s'interprète comme une énergie. La plus petite valeur possible de cette énergie correspond à l'énergie de l'état fondamental du système décrit par (3.66). Les autres valeurs, plus grandes, donnent les énergies des états excités. Sous des conditions "raisonnables" sur le potentiel V, ces niveaux d'énergie sont discrets en nombre infini dénombrable (ce qui est cohérent avec la vision physique des *quanta*).

Exercice 3.4.3 Soit $V(x) = Ax \cdot x$ avec A matrice symétrique réelle définie positive. Montrer que $u(x) = \exp(-A^{1/2}x \cdot x/2)$ est une solution de (3.67) si $\omega = \mathrm{tr}(A^{1/2})$. Une telle solution est appelée état fondamental.

3.4.2 Formulation variationnelle et discrétisation par éléments finis

Nous introduisons la formulation variationnelle du problème aux valeurs propres (3.59).

Proposition 3.4.2 *Soit u une fonction de $C^2(\overline{\Omega})$. Soit V_0 l'espace vectoriel défini par*

$$V_0 = \left\{ \phi \in C^1(\overline{\Omega}) \text{ tel que } \phi = 0 \text{ sur } \partial\Omega \right\}.$$

Alors (λ, u) est une solution du problème aux valeurs propres (3.59) si et seulement si u appartient à V_0 et vérifie l'égalité

$$\int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx = \lambda \int_\Omega u(x)v(x)\, dx \text{ pour toute fonction } v \in V_0. \tag{3.68}$$

Nous laissons au lecteur le soin de démontrer cette proposition, sur le modèle de la Proposition 3.1.5. En reprenant les notations de la Sous-section 3.3.2, la formulation variationnelle (3.68) consiste donc à trouver $\lambda \in \mathbb{R}$ et $u \in V_0 \setminus \{0\}$ tels que

$$a(u, v) = \lambda \langle u, v \rangle_{L^2(\Omega)} \quad \forall\, v \in V_0. \tag{3.69}$$

On considére une approximation interne de cette formulation variationnelle (3.69). Étant donné un sous-espace V_{0h} de l'espace vectoriel V, de dimension finie, on cherche les solutions $(\lambda_h, u_h) \in \mathbb{R} \times V_{0h}$ de

$$a(u_h, v_h) = \lambda_h \langle u_h, v_h \rangle_{L^2(\Omega)} \quad \forall\, v_h \in V_{0h}. \tag{3.70}$$

Typiquement, V_{0h} est un espace d'éléments finis. La résolution de l'approximation interne (3.70) est facile comme le montre le lemme suivant.

Lemme 3.4.3 *Les valeurs propres de (3.70) forment une suite croissante finie*

$$0 < \lambda_1 \leq \cdots \leq \lambda_{n_{dl}} \quad \text{avec } n_{dl} = \dim V_{0h},$$

et il existe une base de V_{0h}, orthonormale dans $L^2(\Omega)$, $(u_{k,h})_{1 \leq k \leq n_{dl}}$ de vecteurs propres associés, c'est-à-dire que

$$u_{k,h} \in V_{0h}, \quad \text{et } a(u_{k,h}, v_h) = \lambda_k \langle u_{k,h}, v_h \rangle_{L^2(\Omega)} \quad \forall\, v_h \in V_{0h}.$$

Démonstration. Soit $(\phi_i)_{1\leq i\leq n_{dl}}$ une base de V_h. On cherche u_h solution de (3.70) sous la forme

$$u_h(x) = \sum_{i=1}^{n_{dl}} U_i^h \phi_i(x).$$

Introduisant la **matrice de masse** $\mathcal{M}_h$ définie par

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_H \quad 1 \leq i, j \leq n_{dl},$$

et la **matrice de rigidité** $\mathcal{K}_h$ définie par

$$(\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl},$$

le problème (3.70) est équivalent à trouver $(\lambda_h, U_h) \in \mathbb{R} \times \mathbb{R}^{n_{dl}}$ solution de

$$\mathcal{K}_h U_h = \lambda_h \mathcal{M}_h U_h. \tag{3.71}$$

On sait que les matrices $\mathcal{M}_h$ et $\mathcal{K}_h$ sont symétriques et définies positives. Le système (3.71) est un problème matriciel aux valeurs propres "généralisé". Le théorème de réduction simultanée (voir par exemple le théorème 2.3.6 dans [2]) affirme qu'il existe une matrice inversible P_h telle que

$$\mathcal{M}_h = P_h P_h^*, \text{ et } \mathcal{K}_h = P_h \operatorname{diag}(\lambda_k) P_h^*.$$

Par conséquent, les solutions de (3.71) sont les valeurs propres (λ_k) et les vecteurs propres $(U_{k,h})_{1\leq k\leq n_{dl}}$ qui sont les vecteurs colonnes de l'inverse de P_h^*. Ces vecteurs colonnes forment donc une base, orthogonale pour $\mathcal{K}_h$ et orthonormale pour $\mathcal{M}_h$. Finalement, les vecteurs $U_{k,h}$ sont simplement les vecteurs des coordonnées dans la base $(\phi_i)_{1\leq i\leq n_{dl}}$ des fonctions $u_{k,h}$ qui forment une base orthonormale de V_h pour le produit scalaire de $L^2(\Omega)$. □

Exercice 3.4.4 On considère le problème aux valeurs propres en dimension $N = 1$

$$\begin{cases} -u_k'' = \lambda_k u_k \quad \text{pour } 0 < x < 1 \\ u_k(0) = u_k(1) = 0. \end{cases}$$

En utilisant la formule pour la matrice de masse $\mathcal{M}_h$ donnée par l'Exercice 3.3.1, montrer que les valeurs propres sont

$$\lambda_k(\mathcal{M}_h) = \frac{h}{3}\left(2 + \cos(k\pi h)\right) \text{ pour } 1 \leq k \leq n.$$

3.5 Résolution des systèmes linéaires

Cette section est principalement consacrée à la résolution des systèmes linéaires avec, dans la dernière sous-section, une introduction au calcul de valeurs et vecteurs propres de matrices. Pour plus de détails nous renvoyons à [2]. On appelle système

linéaire le problème qui consiste à trouver la ou les solutions $x \in \mathbb{R}^n$ (si elle existe) de l'équation algébrique suivante

$$Ax = b, \tag{3.72}$$

où A appartient à l'ensemble $\mathcal{M}_n(\mathbb{R})$ des matrices réelles carrées d'ordre n, et $b \in \mathbb{R}^n$ est un vecteur appelé second membre. Nous allons voir deux types de méthodes de résolution de systèmes linéaires : celles dites directes, c'est-à-dire qui permettent de calculer la solution exacte en un nombre fini d'opérations, et celles dites **itératives**, c'est-à-dire qui calculent une suite de solutions approchées qui converge vers la solution exacte.

3.5.1 Rappels sur les normes matricielles

Nous commençons par rappeler la notion de **norme subordonnée** pour les matrices. Même si l'on considère des matrices réelles, il est nécessaire, pour des raisons techniques qui seront exposées à la Remarque 3.5.3, de les traiter comme des matrices complexes.

Définition 3.5.1 *Soit $\|\cdot\|$ une norme vectorielle sur $\mathbb{C}^n$. On lui associe une norme matricielle, dite subordonnée à cette norme vectorielle, définie par*

$$\|A\| = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Par abus de langage on note de la même façon les normes vectorielle et matricielle subordonnée. On vérifie aisément qu'une norme subordonnée ainsi définie est bien une norme matricielle et qu'elle vérifie le résultat suivant.

Lemme 3.5.2 *Soit $\|\cdot\|$ une norme matricielle subordonnée sur $\mathcal{M}_n(\mathbb{C})$.*

1. *Pour toute matrice A, la norme $\|A\|$ est aussi définie par*

$$\|A\| = \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\| = \sup_{x\in\mathbb{C}^n, \|x\|\leq 1} \|Ax\|.$$

2. *Il existe $x_A \in \mathbb{C}^n, x_A \neq 0$ tel que $\|A\| = \dfrac{\|Ax_A\|}{\|x_A\|}$.*
3. *La matrice identité vérifie $\|\mathrm{Id}\| = 1$.*
4. *Soient A et B deux matrices. On a $\|AB\| \leq \|A\|\,\|B\|$.*

On note $\|A\|_p$ la norme matricielle subordonnée à la norme vectorielle sur $\mathbb{C}^n$ définie pour $p \geq 1$ par $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, et pour $p = +\infty$ par $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$. On peut calculer explicitement certaines de ces normes subordonnées. (Dans tout ce qui suit on note A^* la matrice adjointe de A.)

Exercice 3.5.1 Montrer que

1. $\|A\|_2 = \|A^*\|_2 =$ maximum des valeurs singulières de A,

2. $\|A\|_1 = \max_{1\leq j\leq n} \left(\sum_{i=1}^n |a_{ij}|\right)$,
3. $\|A\|_\infty = \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$.

Remarque 3.5.3 Une matrice réelle peut être considérée soit comme une matrice de $\mathcal{M}_n(\mathbb{R})$, soit comme une matrice de $\mathcal{M}_n(\mathbb{C})$ car $\mathbb{R} \subset \mathbb{C}$. Si $\|\cdot\|_{\mathbb{C}}$ est une norme vectorielle dans $\mathbb{C}^n$, on peut définir sa restriction $\|\cdot\|_{\mathbb{R}}$ à $\mathbb{R}^n$ qui est aussi une norme vectorielle dans $\mathbb{R}^n$. Pour une matrice réelle $A \in \mathcal{M}_n(\mathbb{R})$, on peut donc définir deux normes matricielles subordonnées $\|A\|_{\mathbb{C}}$ et $\|A\|_{\mathbb{R}}$ par

$$\|A\|_{\mathbb{C}} = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Ax\|_{\mathbb{C}}}{\|x\|_{\mathbb{C}}} \text{ et } \|A\|_{\mathbb{R}} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_{\mathbb{R}}}{\|x\|_{\mathbb{R}}}.$$

A priori ces deux définitions peuvent être distinctes. Grâce aux formules explicites de l'Exercice 3.5.1, on sait qu'elles coïncident si $\|x\|_{\mathbb{C}}$ est une des normes $\|x\|_1$, $\|x\|_2$, ou $\|x\|_\infty$. Cependant, pour d'autres normes vectorielles on peut avoir $\|A\|_{\mathbb{C}} > \|A\|_{\mathbb{R}}$. Par ailleurs, dans la preuve de la Proposition 3.5.6 on a besoin de la définition sur $\mathbb{C}$ de la norme subordonnée même si la matrice est réelle. C'est pourquoi on utilise $\mathbb{C}$ dans la Définition 3.5.1 de la norme subordonnée. •

Définition 3.5.4 *Soit A une matrice dans $\mathcal{M}_n(\mathbb{C})$. On appelle rayon spectral de A, et on note $\rho(A)$, le maximum des modules des valeurs propres de A.*

Lemme 3.5.5 *Si U est une matrice unitaire ($U^* = U^{-1}$), on a $\|UA\|_2 = \|AU\|_2 = \|A\|_2$. Par conséquent, si A est une matrice normale ($A^*A = AA^*$), alors $\|A\|_2 = \rho(A)$.*

Démonstration. Comme $U^*U = \mathrm{Id}$, on a

$$\|UA\|_2^2 = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|UAx\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\langle U^*UAx, Ax\rangle}{\langle x, x\rangle} = \|A\|_2^2.$$

D'autre part, le changement de variable $y = Ux$ vérifie $\|x\|_2 = \|y\|_2$, et donc

$$\|AU\|_2^2 = \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|AUx\|_2^2}{\|x\|_2^2} = \sup_{y\in\mathbb{C}^n, y\neq 0} \frac{\|Ay\|_2^2}{\|U^{-1}y\|_2^2} = \sup_{y\in\mathbb{C}^n, y\neq 0} \frac{\|Ay\|_2^2}{\|y\|_2^2} = \|A\|_2^2.$$

Si A est normale, elle est diagonalisable dans une base orthonormée de vecteurs propres et on déduit des résultats précédents que $\|A\|_2 = \|\operatorname{diag}(\lambda_i)\|_2 = \rho(A)$. □

Proposition 3.5.6 *Soit $\|\cdot\|$ une norme subordonnée sur $\mathcal{M}_n(\mathbb{C})$. On a*

$$\rho(A) \leq \|A\|.$$

Réciproquement, pour toute matrice A et pour tout réel $\epsilon > 0$, il existe une norme subordonnée $\|\cdot\|$ (qui dépend de A et ϵ) telle que

$$\|A\| \leq \rho(A) + \epsilon. \tag{3.73}$$

Lemme 3.5.7 *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. Les quatre conditions suivantes sont équivalentes*

1. $\lim_{i\to+\infty} A^i = 0$,
2. $\lim_{i\to+\infty} A^i x = 0$ *pour tout vecteur* $x \in \mathbb{C}^n$,
3. $\rho(A) < 1$,
4. *il existe au moins une norme matricielle subordonnée telle que* $\|A\| < 1$.

3.5.2 Conditionnement et stabilité

Avant de décrire les algorithmes de résolution de systèmes linéaires, il nous faut évoquer les problèmes de précision et de stabilité dus aux erreurs d'arrondi. En effet, dans un ordinateur il n'y a pas de calculs exacts, et la précision est limitée à cause du nombre de bits utilisés pour représenter les nombres réels : d'habitude 32 ou 64 bits (ce qui fait à peu près 8 ou 16 chiffres significatifs). Il faut donc faire très attention aux inévitables erreurs d'arrondi et à leur propagation au cours d'un calcul. Les méthodes numériques de résolution de systèmes linéaires qui n'amplifient pas ces erreurs sont dites stables. En pratique, on utilisera donc des algorithmes qui sont à la fois **efficaces et stables**. Cette amplification des erreurs dépend de la matrice considérée. Pour quantifier ce phénomène, on introduit la notion de conditionnement d'une matrice.

Définition 3.5.8 *Soit une norme matricielle subordonnée $\|A\|$. On appelle conditionnement d'une matrice $A \in \mathcal{M}_n(\mathbb{C})$, relatif à cette norme, la valeur définie par*

$$\mathrm{cond}(A) = \|A\|.\|A^{-1}\|.$$

Cette notion de conditionnement va permettre de mesurer l'amplification des erreurs des données (second membre ou matrice) au résultat.

Proposition 3.5.9 *Soit A une matrice inversible. Soit $b \neq 0$ un vecteur non nul.*

1. *Soit x et $x + \delta x$ les solutions respectives des systèmes*

$$Ax = b \ , \ et \ A(x + \delta x) = b + \delta b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}. \tag{3.74}$$

2. *Soit x et $x + \delta x$ les solutions respectives des systèmes*

$$Ax = b \ , \ et \ (A + \delta A)(x + \delta x) = b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}. \tag{3.75}$$

De plus, ces inégalités sont optimales.

Remarque 3.5.10 On dira qu'une matrice est bien conditionnée si son conditionnement est proche de 1 (sa valeur minimale) et qu'elle est mal conditionnée si son conditionnement est grand. A cause des résultats de la Proposition 3.5.9, en pratique il faudra faire attention aux erreurs d'arrondi si on résout un système linéaire pour une matrice mal conditionnée. •

Démonstration. Pour montrer le premier résultat, on remarque que $A\delta x = \delta b$, et donc $\|\delta x\| \leq \|A^{-1}\|.\|\delta b\|$. Or, on a aussi $\|b\| \leq \|A\|\|x\|$, ce qui donne (3.74). Cette inégalité est optimale au sens suivant : pour toute matrice A, il existe δb et x (qui dépendent de A) tels que (3.74) est en fait une égalité. En effet, d'après une propriété des normes matricielles subordonnées (voir le Lemme 3.5.2) il existe x tel que $\|b\| = \|A\|\|x\|$ et il existe δb tel que $\|\delta x\| = \|A^{-1}\|\|\delta b\|$.

Pour obtenir (3.75) on remarque que $A\delta x + \delta A(x + \delta x) = 0$, et donc $\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x + \delta x\|$, ce qui implique (3.75). Pour en démontrer l'optimalité, on va montrer que pour toute matrice A il existe une perturbation δA et un second membre b pour lesquels il y a égalité. Grâce au Lemme 3.5.2 il existe $y \neq 0$ tel que $\|A^{-1}y\| = \|A^{-1}\|\|y\|$. Soit ϵ un scalaire non nul. On pose $\delta A = \epsilon Id$ et $b = (A+\delta A)y$. On vérifie alors que $y = y + \delta x$ et $\delta x = -\epsilon A^{-1}y$, et comme $\|\delta A\| = |\epsilon|$ on obtient l'égalité dans (3.75). □

Les conditionnements les plus utilisés en pratique sont ceux associés aux normes $\|A\|_p$ avec $p = 1, 2, +\infty$.

Exercice 3.5.2 Soit une matrice $A \in \mathcal{M}_n(\mathbb{C})$. Vérifier que

1. $\mathrm{cond}(A) = \mathrm{cond}(A^{-1}) \geq 1$, $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)\ \forall \alpha \neq 0$,
2. pour une matrice quelconque, $\mathrm{cond}_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$, où $\mu_1(A), \mu_n(A)$ sont respectivement la plus petite et la plus grande valeur singulière de A,
3. pour une matrice normale, $\mathrm{cond}_2(A) = \frac{|\lambda_n(A)|}{|\lambda_1(A)|}$, où $|\lambda_1(A)|, |\lambda_n(A)|$ sont respectivement la plus petite et la plus grande valeur propre en module de A,
4. pour toute matrice unitaire U, $\mathrm{cond}_2(U) = 1$,
5. pour toute matrice unitaire U, $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(A)$.

3.5.3 Méthodes directes

Méthode d'élimination de Gauss

L'idée principale de cette méthode est de se ramener à la résolution d'un système linéaire dont la matrice est triangulaire. En effet, la résolution d'un système linéaire, $Tx = b$, où la matrice T est triangulaire et inversible, est très facile par simple substitution récursive. On appelle ce procédé **remontée** dans le cas d'une matrice triangulaire supérieure et **descente** dans le cas d'une matrice triangulaire inférieure. Remarquons que l'on résout ainsi le système $Tx = b$ sans inverser la matrice T. De la même manière, la méthode d'élimination de Gauss va résoudre le système $Ax = b$ sans calculer l'inverse de la matrice A.

La méthode d'élimination de Gauss se décompose en trois étapes :

(i) élimination : calcul d'une matrice M inversible telle que $MA = T$ soit triangulaire supérieure,
(ii) mise à jour du second membre : calcul simultané de Mb,
(iii) substitution : résolution du système triangulaire $Tx = Mb$ par simple remontée.

L'existence d'une telle matrice M est garantie par le résultat suivant dont on va donner une démonstration constructive qui n'est rien d'autre que la méthode d'élimination de Gauss.

Proposition 3.5.11 *Soit A une matrice carrée (inversible ou non). Il existe au moins une matrice inversible M telle que la matrice $T = MA$ soit triangulaire supérieure.*

Démonstration. Le principe est de construire une suite de matrices A^k, $1 \leq k \leq n$, dont les $(k-1)$ premières colonnes sont remplies de zéros sous la diagonale. Par modifications successives, on passe de $A^1 = A$ à $A^n = T$ qui est triangulaire supérieure. On note $\left(a_{ij}^k\right)_{1\leq i,j\leq n}$ les éléments de la matrice A^k, et on appelle pivot de A^k l'élément a_{kk}^k. Pour passer de la matrice A^k à la matrice A^{k+1}, on s'assure tout d'abord que le pivot a_{kk}^k n'est pas nul. S'il l'est, on permute la k-ème ligne avec une autre ligne pour amener en position de pivot un élément non nul. Puis on procède à l'élimination de tous les éléments de la k-ème colonne en dessous de la k-ème ligne en faisant des combinaisons linéaires de la ligne courante avec la k-ème ligne. □

Méthode de la factorisation LU

La méthode LU consiste à factoriser la matrice A en un produit de deux matrices triangulaires $A = LU$, où L est triangulaire inférieure (L pour "lower" en anglais) et U est triangulaire supérieure (U pour "upper" en anglais). Il s'agit en fait du même algorithme que celui de l'élimination de Gauss dans le cas particulier où **on ne pivote jamais**. Une fois établie la factorisation LU de A, la résolution du système linéaire $Ax = b$ est équivalente à la simple résolution de deux systèmes triangulaires $Ly = b$ puis $Ux = y$.

Proposition 3.5.12 *Soit une matrice $A = (a_{ij})_{1\leq i,j\leq n}$ d'ordre n telle que toutes les sous-matrices diagonales d'ordre k, définies par*

$$\Delta^k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$

soient inversibles. Il existe un unique couple de matrices (L, U), avec U triangulaire supérieure, et L triangulaire inférieure ayant une diagonale de 1, tel que

$$A = LU.$$

Calcul pratique de la factorisation LU. On peut calculer la factorisation LU (si elle existe) d'une matrice A par identification de A au produit LU. En posant $A = (a_{ij})_{1\leq i,j\leq n}$, et

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}, \; U = \begin{pmatrix} u_{1,1} & \dots & \dots & u_{1,n} \\ 0 & u_{2,2} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{n,n} \end{pmatrix},$$

comme L est triangulaire inférieure et U triangulaire supérieure, pour $1 \leq i, j \leq n$ il vient

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k}u_{k,j} = \sum_{k=1}^{min(i,j)} l_{i,k}u_{k,j}.$$

En identifiant par ordre croissant les colonnes de A on en déduit les colonnes de L et de U. Ainsi, après avoir calculé les $(j-1)$ premières colonnes de L et de U en fonction des $(j-1)$ premières colonnes de A, on lit la j-ème colonne de A

$$\begin{aligned} a_{i,j} = \sum_{k=1}^{i} l_{i,k}u_{k,j} \quad &\Rightarrow \quad u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}u_{k,j} \qquad \text{pour } 1 \leq i \leq j, \\ a_{i,j} = \sum_{k=1}^{j} l_{i,k}u_{k,j} \quad &\Rightarrow \quad l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k}u_{k,j}}{u_{jj}} \quad \text{pour } j+1 \leq i \leq n. \end{aligned}$$

On calcule donc les j premières composantes de la j-ème colonne de U et les $n-j$ dernières composantes de la j-ème colonne de L en fonction de leurs $(j-1)$ premières colonnes. On divise par le pivot u_{jj} qui doit donc être non nul !

Compte d'opérations. Pour mesurer l'efficacité de l'algorithme de la décomposition LU on compte le nombre d'opérations nécessaires à son accomplissement (qui sera proportionnel à son temps d'exécution sur un ordinateur). On ne calcule pas exactement ce nombre d'opérations, et on se contente du premier terme de son développement asymptotique lorsque la dimension n est grande. De plus, pour simplifier on ne compte que les multiplications et divisions (et pas les additions dont le nombre est en général du même ordre de grandeur).

— Élimination ou décomposition LU : le nombre d'opérations N_{op} est

$$N_{op} = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} (1 + \sum_{k=j+1}^{n} 1),$$

qui, au premier ordre, donne $N_{op} \approx n^3/3$.

— Substitution (ou remontée-descente sur les deux systèmes triangulaires) : le nombre d'opérations N_{op} est donné par la formule

$$N_{op} = 2\sum_{j=1}^{n} j,$$

qui, au premier ordre, donne $N_{op} \approx n^2$.

Au total la résolution d'un système linéaire $Ax = b$ par la méthode de la factorisation LU demande $N_{op} \approx n^3/3$ opérations car n^2 est négligeable devant n^3 quand n est grand.

Méthode de Cholesky

C'est une méthode qui ne s'applique qu'aux matrices symétriques réelles, définies positives. Elle consiste à factoriser une matrice A sous la forme $A = BB^*$ où B est une matrice triangulaire inférieure (et B^* son adjointe ou transposée).

Proposition 3.5.13 *Soit A une matrice symétrique réelle, définie positive. Il existe une unique matrice réelle B triangulaire inférieure, telle que tous ses éléments diagonaux soient positifs, et qui vérifie*

$$A = BB^*.$$

Calcul pratique de la factorisation de Cholesky. En pratique, on calcule le facteur de Cholesky B par identification dans l'égalité $A = BB^*$. Soit $A = (a_{ij})_{1\leq i,j\leq n}$, $B = (b_{ij})_{1\leq i,j\leq n}$ avec $b_{ij} = 0$ si $i < j$. Pour $1 \leq i, j \leq n$, il vient

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{min(i,j)} b_{ik}b_{jk}.$$

En identifiant par ordre croissant les colonnes de A (ou ses lignes, ce qui revient au même puisque A est symétrique) on en déduit les colonnes de B. Ainsi, après avoir calculé les $(j-1)$ premières colonnes de B en fonction des $(j-1)$ premières colonnes de A, on lit la j-ème colonne de A en dessous de la diagonale

$$\begin{aligned} a_{jj} = \sum_{k=1}^{j} (b_{jk})^2 \quad &\Rightarrow \quad b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2} \\ a_{i,j} = \sum_{k=1}^{j} b_{jk}b_{i,k} \quad &\Rightarrow \quad b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{jk}b_{i,k}}{b_{jj}} \text{ pour } j+1 \leq i \leq n. \end{aligned}$$

On calcule donc la j-ème colonne de B en fonction de ses $(j-1)$ premières colonnes. A cause du théorème précédent, on est sûr que, si A est symétrique définie positive, les termes sous les racines carrées sont strictement positifs. Au contraire, si A n'est pas définie positive, on trouvera que $a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2 \leq 0$ pour un certain rang j, ce qui empêche de terminer l'algorithme.

Compte d'opérations. Pour mesurer l'efficacité de la méthode de Cholesky on compte le nombre d'opérations (uniquement les multiplications) nécessaires à son accomplissement. Le nombre de racines carrées est n qui est négligeable dans ce compte d'opérations.

— Factorisation de Cholesky : le nombre d'opérations N_{op} est

$$N_{op} = \sum_{j=1}^{n} \left((j-1) + \sum_{i=j+1}^{n} j \right),$$

qui, au premier ordre, donne $N_{op} \approx n^3/6$.

— Substitution : il faut effectuer une remontée et une descente sur les systèmes triangulaire associés à B et B^*. Le nombre d'opérations est au premier ordre $N_{op} \approx n^2$.

La méthode de Cholesky est donc approximativement **deux fois plus rapide** que celle de Gauss pour une matrice symétrique définie positive.

Matrices bandes et matrices creuses

Lorsqu'une matrice a beaucoup de coefficients nuls, on dit qu'elle est **creuse**. Si les éléments non nuls sont répartis à proximité de la diagonale, on dit que la matrice a une structure **bande**. Pour ces deux types de matrices (qui apparaissent naturellement dans la méthode des éléments finis comme dans la plupart des autres méthodes), on peut améliorer le compte d'opérations et la taille de stockage nécessaire pour résoudre un système linéaire. Ce gain est très important en pratique.

Définition 3.5.14 *Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est dite matrice bande, de demie largeur de bande (hors diagonale) $p \in \mathbb{N}$ si ses éléments vérifient $a_{i,j} = 0$ pour $|i-j| > p$. La largeur de la bande est alors $2p+1$.*

L'intérêt des matrices bandes vient de la propriété suivante.

Exercice 3.5.3 Montrer que les factorisations LU et de Cholesky conservent la structure bande des matrices.

Remarque 3.5.15 Si les factorisations LU et de Cholesky préservent la structure bande des matrices, il n'en est pas de même de leur structure creuse. En général, si A est creuse (même à l'intérieur d'une bande), les facteurs L et U, ou B et B^* sont "pleins" (le contraire de creux) à l'intérieur de la même bande. •

L'exercice suivant permet de quantifier le gain qu'il y a à utiliser des matrices bandes.

Exercice 3.5.4 Montrer que, pour une matrice bande d'ordre n et de demie largeur de bande p, le compte d'opérations de la factorisation LU est $\mathcal{O}(np^2/3)$ et celui de la factorisation de Cholesky est $\mathcal{O}(np^2/6)$.

3.5.4 Méthodes itératives

Les méthodes itératives sont particulièrement intéressantes pour les très grandes matrices ou les matrices creuses. En effet, dans ce cas les méthodes directes peuvent avoir un coût de calcul et de stockage en mémoire prohibitif (se rappeler que la factorisation LU ou de Cholesky demande de l'ordre de n^3 opérations). Commençons par une classe très simple de méthodes itératives.

Définition 3.5.16 *Soit A une matrice inversible. On introduit une décomposition régulière de A (en anglais "splitting"), c'est-à-dire un couple de matrices (M, N) avec M inversible (et facile à inverser dans la pratique) tel que $A = M - N$. La méthode itérative basée sur le splitting (M, N) est définie par*

$$\begin{cases} x_0 \text{ donné dans } \mathbb{R}^n, \\ Mx_{k+1} = Nx_k + b \quad \forall k \geq 1. \end{cases} \tag{3.76}$$

Si la suite de solutions approchées x_k converge vers une limite x quand k tend vers l'infini, alors, par passage à la limite dans la relation de récurrence (3.76), on obtient

$$(M - N)x = Ax = b.$$

Par conséquent, si la suite de solutions approchées converge, sa limite est forcément la solution du système linéaire.

D'un point de vue pratique, il faut savoir quand on peut arrêter les itérations, c'est-à-dire à quel moment x_k est suffisamment proche de la solution inconnue x. Comme on ne connaît pas x, on ne peut pas décider d'arrêter le calcul dès que $\|x - x_k\| \leq \epsilon$ où ϵ est la précision désirée. Par contre on connaît Ax (qui vaut b), et un critère d'arrêt fréquemment utilisé est $\|b - Ax_k\| \leq \epsilon$. Cependant, si la norme de A^{-1} est grande ce critère peut être trompeur car

$$\|x - x_k\| \leq \|A^{-1}\| \, \|b - Ax_k\| \leq \epsilon \|A^{-1}\|$$

qui peut ne pas être petit.

Définition 3.5.17 *On dit qu'une méthode itérative est convergente si,* **quel que soit le choix du vecteur initial** $x_0 \in \mathbb{R}^n$*, la suite de solutions approchées x_k converge vers la solution exacte x.*

On commence par donner une condition nécessaire et suffisante de convergence d'une méthode itérative à l'aide du rayon spectral de la matrice d'itération (voir la Définition 3.5.4 pour la notion de rayon spectral).

Lemme 3.5.18 *La méthode itérative définie par (3.76) converge si et seulement si le rayon spectral de la matrice d'itération $M^{-1}N$ vérifie $\rho(M^{-1}N) < 1$.*

Démonstration. On définit l'erreur $e_k = x_k - x$. On a

$$e_k = (M^{-1}Nx_{k-1} + M^{-1}b) - (M^{-1}Nx + M^{-1}b) = M^{-1}Ne_{k-1} = (M^{-1}N)^k e_0.$$

Par application du Lemme 3.5.7, on en déduit que e_k tend vers 0, quel que soit e_0, si et seulement si $\rho(M^{-1}N) < 1$. □

Définition 3.5.19 (méthode de Jacobi) *Soit $A = (a_{ij})_{1 \leq i,j \leq n}$. On note $D = \mathrm{diag}(a_{ii})$ la diagonale de A. On appelle méthode de Jacobi la méthode itérative associée à la décomposition*

$$M = D, \quad N = D - A.$$

Définition 3.5.20 (méthode de Gauss-Seidel) *Soit $A = (a_{ij})_{1\leq i,j\leq n}$. On décompose A sous la forme $A = D - E - F$ où $D = \mathrm{diag}(a_{ii})$ est la diagonale, $-E$ est la partie triangulaire inférieure (strictement), et $-F$ est la partie triangulaire supérieure (strictement) de A. On appelle méthode de Gauss-Seidel la méthode itérative associée à la décomposition*

$$M = D - E, \quad N = F.$$

Définition 3.5.21 (méthode de relaxation (SOR)) *Soit $\omega \in \mathbb{R}^+$. On appelle méthode de relaxation (SOR en anglais pour "Successive Over Relaxation"), pour le paramètre ω, la méthode itérative associée à la décomposition*

$$M = \frac{D}{\omega} - E, \quad N = \frac{1-\omega}{\omega}D + F$$

Définition 3.5.22 (méthode du gradient) *Soit un paramètre réel $\alpha \neq 0$. On appelle méthode du gradient la méthode itérative associée à la décomposition*

$$M = \frac{1}{\alpha}\,\mathrm{Id} \quad et \quad N = \left(\frac{1}{\alpha}\,\mathrm{Id} - A\right).$$

La méthode du gradient semble encore plus primitive que les méthodes précédentes, mais elle a une interprétation en tant que méthode de minimisation de la fonction $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ qui lui donne une plus grande applicabilité.

3.5.5 Méthode du gradient conjugué

La méthode du gradient conjugué est la méthode itérative de choix pour résoudre des systèmes linéaires dont la matrice est symétrique réelle définie positive.

Proposition 3.5.23 *Soit A une matrice symétrique définie positive, et $x_0 \in \mathbb{R}^n$. Soit (x_k, r_k, p_k) trois suites définies par les relations de récurrence*

$$p_0 = r_0 = b - Ax_0, \ \textit{et pour } 0 \leq k \ \begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases} \tag{3.77}$$

avec

$$\alpha_k = \frac{\|r_k\|^2}{Ap_k \cdot p_k} \ \textit{et}\ \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Alors, la suite (x_k) de la méthode du gradient conjugué converge en moins de n itérations vers la solution du système linéaire $Ax = b$.

On peut montrer que r_k est la suite des résidus, c'est-à-dire $r_k = b - Ax_k$. Par conséquent, dès que $r_k = 0$, l'algorithme a convergé, c'est-à-dire que x_k est la solution du système $Ax = b$. On peut démontrer que la convergence est atteinte en moins de n itérations mais, dans la pratique, on utilise le gradient conjugué comme une méthode itérative et on décide que l'algorithme a convergé dès que $\|r_k\|/\|r_0\| \leq \epsilon$ pour un "petit" paramètre ϵ.

Remarque 3.5.24

1. A chaque itération on n'a besoin de faire qu'un seul produit matrice-vecteur, à savoir Ap_k, car r_k est calculé par la formule de récurrence et non par la relation $r_k = b - Ax_k$.
2. Pour mettre en oeuvre la méthode du gradient conjugué, il n'est pas nécessaire de stocker la matrice A dans un tableau si on sait calculer le produit matrice vecteur Ay pour tout vecteur y.
3. La méthode du gradient conjugué est très efficace et très utilisée. Elle a beaucoup de variantes ou de généralisations, notamment au cas des matrices non symétriques définies positives.

•

La vitesse de convergence de la méthode du gradient conjugué dépend du conditionnement de la matrice A, l'idée du préconditionnement est de pré-multiplier le système linéaire $Ax = b$ par une matrice C^{-1} telle que le conditionnement de $(C^{-1}A)$ soit plus petit que celui de A. En pratique on choisit une matrice C "proche" de A mais plus facile à inverser.

Définition 3.5.25 *Soit à résoudre le système linéaire $Ax = b$. On appelle préconditionnement de A, une matrice C (facile à inverser) telle que* $\mathrm{cond}_2(C^{-1}A)$ *soit plus petit que* $\mathrm{cond}_2(A)$. *On appelle système préconditionné le système équivalent $C^{-1}Ax = C^{-1}b$.*

La technique du préconditionnement est très efficace et essentielle en pratique pour converger rapidement. Nous indiquons trois choix possibles de C du plus simple au plus compliqué. Le préconditionnement le plus simple est le "préconditionnement diagonal" : il consiste à prendre $C = \mathrm{diag}(A)$. Il est malheureusement peu efficace, et on lui préfère souvent le "préconditionnement SSOR" (pour Symmetric SOR). En notant $D = \mathrm{diag}(A)$ la diagonale d'une matrice symétrique A et $-E$ sa partie strictement inférieure telle que $A = D - E - E^*$, pour $\omega \in]0, 2[$, on pose

$$C_\omega = \frac{\omega}{2-\omega}\left(\frac{D}{\omega} - E\right) D^{-1} \left(\frac{D}{\omega} - E^*\right).$$

On vérifie que, si A est définie positive, alors C l'est aussi. Le système $Cz = r$ est facile à résoudre car C est déjà sous une forme factorisée en produit de matrices triangulaires. Le nom de ce préconditionnement vient du fait qu'inverser C revient à effectuer deux itérations successives de la méthode itérative de relaxation (SOR), avec deux matrices d'itérations symétriques l'une de l'autre.

Un dernier exemple est le "préconditionnement de Cholesky incomplet". La matrice C est cherchée sous la forme BB^* où B est le facteur "incomplet" de la factorisation de Cholesky de A (voir la Proposition 3.5.13). Cette matrice triangulaire inférieure B est obtenue en appliquant l'algorithme de factorisation de Cholesky à A en forçant l'égalité $b_{ij} = 0$ si $a_{ij} = 0$. Cette modification de l'algorithme assure, d'une part que le facteur B sera aussi creux que la matrice A, et d'autre part que le calcul de ce facteur incomplet sera beaucoup moins cher (en temps de calcul) que

le calcul du facteur exact si A est creuse (ce qui est le cas pour des matrices de discrétisation par éléments finis).

3.5.6 Calcul de valeurs et vecteurs propres

Dans cette sous-section nous expliquons brièvement comment calculer les valeurs propres et les vecteurs propres d'une matrice symétrique réelle. Pour plus de détails nous renvoyons à [2].

Puisque les valeurs propres d'une matrice A sont les racines de son polynôme caractéristique $\det(A - \lambda\,\mathrm{Id})$, on pourrait penser naïvement que, pour les calculer, il "suffit" de factoriser son polynôme caractéristique. Il n'en est rien : on sait depuis Galois et Abel qu'on ne peut pas calculer par opérations élémentaires (addition, multiplication, extraction de racines) les racines d'un polynôme quelconque de degré supérieur ou égal à 5. Pour s'en convaincre, on peut remarquer que n'importe quel polynôme de degré n,

$$P(\lambda) = (-1)^n \left(\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n\right),$$

est le polynôme caractéristique (le développer par rapport à la dernière colonne) de la matrice

$$A = \begin{pmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

Par conséquent, il ne peut pas exister de méthodes directes (c'est-à-dire qui donnent le résultat en un nombre fini d'opérations) pour le calcul des valeurs propres ! Il n'existe donc que des méthodes itératives pour calculer des valeurs propres (et des vecteurs propres). Il se trouve que le calcul pratique des valeurs et vecteurs propres d'une matrice est une tâche beaucoup plus difficile que la résolution d'un système linéaire. Fort heureusement, le cas des matrices symétriques réelles (auquel nous nous limitons puisqu'il suffit pour nos applications) est bien plus simple que le cas des matrices non auto-adjointes.

Nous nous contentons de présenter la méthode, dite de la puissance, qui permet de calculer très simplement la plus grande ou la plus petite valeur propre (en module) d'une matrice (et un vecteur propre associé). Une limitation de la méthode est que la valeur propre extrême que l'on calcule doit être simple (ou de multiplicité égale à 1, c'est-à-dire que la dimension du sous-espace propre correspondant est 1). Soit A une matrice symétrique réelle d'ordre n, de valeurs propres $(\lambda_1, \cdots, \lambda_n)$ avec $\lambda_n > |\lambda_i|$ pour tout $1 \leq i \leq n-1$. La méthode de la puissance pour calculer la plus grande valeur propre λ_n est définie par l'algorithme ci-dessous.

1. `Initialisation:` $x_0 \in \mathbb{R}^n$ `tel que` $\|x_0\| = 1$.
2. `Itérations: pour` $k \geq 1$
   1. $y_k = Ax_{k-1}$

2. $x_k = y_k/\|y_k\|$

3. `test de convergence: si` $\|x_k - x_{k-1}\| \leq \varepsilon$`, on arrête.`

Dans le test de convergence ε est un petit nombre réel, typiquement égal à 10^{-6}. Si $\delta_k = x_k - x_{k-1}$ est petit, alors x_k est un vecteur propre approché de A de valeur propre approchée $\|y_k\|$ car $Ax_k - \|y_k\|x_k = A\delta_k$.

Proposition 3.5.26 *On suppose que la matrice A est symétrique réelle, de valeurs propres $(\lambda_1, \cdots, \lambda_n)$, associées à une base orthonormée de vecteurs propres $(e_1, \cdots, e_n)$, et que la valeur propre de plus grand module λ_n est simple et positive, c'est-à-dire que $|\lambda_1|, \cdots, |\lambda_{n-1}| < \lambda_n$. On suppose aussi que le vecteur initial x_0 n'est pas orthogonal à e_n. Alors la méthode de la puissance converge, c'est-à-dire que*

$$\lim_{k\to+\infty} \|y_k\| = \lambda_n, \quad \lim_{k\to+\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_n.$$

La vitesse de convergence est proportionnelle au rapport $|\lambda_{n-1}|/|\lambda_n|$

$$|\|y_k\| - \lambda_n| \leq C \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{k}.$$

Démonstration. Soit $x_0 = \sum_{i=1}^n \beta_i e_i$ le vecteur initial, avec $\beta_n \neq 0$. Le vecteur x_k est proportionnel à $A^k x_0 = \sum_{i=1}^n \beta_i \lambda_i^k e_i$, d'où il vient

$$x_k = \frac{\beta_n e_n + \sum_{i=1}^{n-1} \beta_i \left(\frac{\lambda_i}{\lambda_n}\right)^k e_i}{\left(\beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 \left(\frac{\lambda_i}{\lambda_n}\right)^{2k}\right)^{1/2}}.$$

Comme $|\lambda_i| < \lambda_n$ on en déduit que x_k converge vers $\text{sign}(\beta_n)e_n$. De même, on a

$$\|y_{k+1}\| = \lambda_n \frac{\left(\beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 \left(\frac{\lambda_i}{\lambda_n}\right)^{2(k+1)}\right)^{1/2}}{\left(\beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 \left(\frac{\lambda_i}{\lambda_n}\right)^{2k}\right)^{1/2}},$$

qui converge vers λ_n. □

En pratique (et notamment pour le calcul des valeurs propres de la discrétisation d'un problème aux limites elliptiques), on est surtout intéressé par la **plus petite** valeur propre, en module, de A. On peut adapter les idées précédentes, ce qui donne la méthode de la puissance inverse dont l'algorithme est écrit ci-dessous. On considère une matrice symétrique réelle A dont la plus petite valeur propre en module est simple et strictement positive $0 < \lambda_1 < |\lambda_i|$ pour tout $2 \leq i \leq n$.

1. `Initialisation:` $x_0 \in \mathbb{R}^n$ `tel que` $\|x_0\| = 1$.

2. `Itérations: pour` $k \geq 1$

   1. `résoudre` $Ay_k = x_{k-1}$

2. $x_k = y_k/\|y_k\|$

3. `test de convergence: si` $\|x_k - x_{k-1}\| \leq \varepsilon$`, on arrête.`

Si $\delta_k = x_k - x_{k-1}$ est petit, alors x_{k-1} est un vecteur propre approché de valeur propre approchée $1/\|y_k\|$ car $Ax_{k-1} - \frac{x_{k-1}}{\|y_k\|} = -A\delta_k$.

Proposition 3.5.27 *On suppose que la matrice A est symétrique réelle, de valeurs propres $(\lambda_1, \cdots, \lambda_n)$, associées à une base orthonormée de vecteurs propres $(e_1, \cdots, e_n)$, et que la valeur propre de plus petit module λ_1 est simple et strictement positive, c'est-à-dire que $0 < \lambda_1 < |\lambda_2|, \cdots, |\lambda_n|$. On suppose aussi que le vecteur initial x_0 n'est pas orthogonal à e_1. Alors la méthode de la puissance inverse converge, c'est-à-dire que*

$$\lim_{k\to+\infty} \frac{1}{\|y_k\|} = |\lambda_1|, \quad \lim_{k\to+\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_1.$$

La vitesse de convergence est proportionnelle au rapport $\lambda_1/|\lambda_2|$

$$\left|\|y_k\|^{-1} - \lambda_1\right| \leq C\left|\frac{\lambda_1}{\lambda_2}\right|^{2k}, \quad \|x_k - x_\infty\| \leq C\left|\frac{\lambda_1}{\lambda_2}\right|^{k}.$$

La démonstration est similaire à celle de la Proposition 3.5.26 et nous la laissons au lecteur en guise d'exercice.

Remarque 3.5.28 Pour accélérer la convergence, on peut toujours procéder à une translation de la matrice A qu'on remplace par $A - \sigma\,\mathrm{Id}$ avec σ une approximation de λ_1. •

Remarque 3.5.29 Pour calculer les valeurs propres intermédiaires d'une matrice symétrique réelle (c'est-à-dire ni la plus petite, ni la plus grande), on peut utiliser une méthode dite de déflation. Par exemple, après avoir calculé λ_n et un vecteur propre unitaire correspondant e_n tel que $Ae_n = \lambda_n e_n$ et $\|e_n\| = 1$, on applique à nouveau la méthode de la puissance à A en partant d'un vecteur initial x_0 orthogonal à e_n. Cela revient à calculer la plus grande valeur propre de A restreinte au sous-espace orthogonal à e_n (qui est stable par A), c'est-à-dire à calculer λ_{n-1}. En pratique, pour être sûr de rester dans le sous-espace orthogonal à e_n on orthogonalise à chaque itération le vecteur x_k par rapport à e_n. Dans le cadre de la méthode de la puissance inverse, si on a déjà calculé λ_1, de vecteur propre unitaire e_1, pour obtenir λ_2 on démarre d'un vecteur x_0 orthogonal à e_1 et on orthogonalise (à chaque itération) x_k par rapport à e_1. Cette technique ne permet, en pratique, que le calcul de quelques valeurs propres extrêmes de A. Elle n'est pas recommandée si toutes les valeurs propres doivent être calculées. •

Chapitre 4

OPTIMISATION

4.1 Motivation et généralités

4.1.1 Introduction et exemples

L'optimisation est un sujet très ancien qui connaît un nouvel essor depuis l'apparition des ordinateurs et dont les méthodes s'appliquent dans de très nombreux domaines : économie, gestion, planification, logistique, automatique, robotique, conception optimale, sciences de l'ingénieur, traitement du signal, etc. L'optimisation est aussi un sujet très vaste qui touche aussi bien au calcul des variations, qu'à la recherche opérationnelle (domaine de l'optimisation des processus de gestion ou de décision), en passant par le contrôle optimal. Nous ne ferons souvent qu'effleurer ces sujets car il faudrait un polycopié complet pour chacun d'eux si nous voulions les traiter à fond.

D'une certaine manière, l'optimisation peut être vue comme une discipline indépendante de l'analyse numérique des équations aux dérivées partielles que nous avons étudiée dans les chapitres précédents. Cependant, les interactions entre ces deux disciplines sont extrêmement nombreuses et fécondes et il est beaucoup plus naturel de les relier dans un même cours. En effet, après l'étape de **modélisation** d'un phénomène physique ou d'un système industriel (éventuellement à l'aide d'équations aux dérivées partielles), après l'étape de **simulation numérique** sur ce modèle, la démarche du mathématicien appliqué (qu'il soit ingénieur ou chercheur) ne s'arrête pas là : il lui faut souvent **agir** sur le phénomène ou sur le système afin d'en améliorer certaines performances. Cette troisième étape est celle de **l'optimisation**, c'est-à-dire celle de la minimisation (ou de la maximisation) d'une fonction qui dépend de la solution du modèle.

Le plan de ce chapitre est le suivant. Le reste de cette section est consacré à des exemples, à préciser quelques notations, à donner des résultats élémentaires d'existence de solutions optimales et enfin à définir la notion de convexité. La Section 4.2, après avoir rappelé des notions élémentaires de dérivabilité, donne la forme des conditions nécessaires d'optimalité dans deux cas essentiels : lorsque l'ensemble des contraintes est convexe on obtient une **inéquation d'Euler** ; lorsqu'il s'agit de contraintes égalités ou inégalités, on obtient une équation faisant intervenir des **mul-**

tiplicateurs de Lagrange. La Section 4.3 est consacrée au **théorème de Kuhn et Tucker** qui affirme que, sous certaines hypothèses de convexité, les conditions nécessaires d'optimalité sont aussi suffisantes. On y donne aussi un bref aperçu de la théorie de la **dualité**. Finalement, la Section 4.4 traite des **algorithmes numériques d'optimisation**. On étudie principalement les algorithmes de **gradient** qui sont les plus importants en pratique.

Pour plus de détails sur l'optimisation nous renvoyons le lecteur aux ouvrages [4], [6], [8], [12], [15].

Passons en revue quelques problèmes typiques d'optimisation, d'importance pratique ou théorique inégale, mais qui permettent de saisir l'importance et l'ubiquité de l'optimisation.

Exemple 4.1.1 (Problème de transport) Il s'agit d'un exemple de programme linéaire (ou programmation linéaire). Le but est d'optimiser la livraison d'une marchandise (un problème classique en logistique). On dispose de M entrepôts, indicés par $1 \leq i \leq M$, disposant chacun d'un niveau de stocks s_i. Il faut livrer N clients, indicés par $1 \leq j \leq N$, qui ont commandé chacun une quantité r_j. Le coût de transport unitaire entre l'entrepôt i et le client j est donné par c_{ij}. Les variables de décision sont les quantités v_{ij} de marchandise partant de l'entrepôt i vers le client j. On veut minimiser le coût du transport tout en satisfaisant les commandes des clients (on suppose que $\sum_{i=1}^{M} s_i \geq \sum_{j=1}^{N} r_j$). Autrement dit, on veut résoudre

$$\inf_{(v_{ij})} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} v_{ij} \right)$$

sous les contraintes de limites des stocks et de satisfaction des clients

$$v_{ij} \geq 0,\ \sum_{j=1}^{N} v_{ij} \leq s_i,\ \sum_{i=1}^{M} v_{ij} = r_j \ \text{ pour } 1 \leq i \leq M,\ 1 \leq j \leq N.$$

•

Exemple 4.1.2 (Consommation des ménages) Il s'agit d'un modèle classique en économie. On considère un ménage qui peut consommer n types de marchandise dont les prix forment un vecteur $p \in \mathbb{R}^n_+$. Son revenu à dépenser est un réel $b > 0$, et ses choix de consommation sont supposés être modélisés par une fonction d'utilité $u(x)$ de $\mathbb{R}^n_+$ dans $\mathbb{R}$ (croissante et concave), qui mesure le bénéfice que le ménage tire de la consommation de la quantité x des n marchandises. La consommation du ménage sera le vecteur x^* qui réalisera le maximum de

$$\max_{x \in \mathbb{R}^n_+, x \cdot p \leq b} u(x),$$

c'est-à-dire qui maximise l'utilité sous une contrainte de budget maximal (voir la Sous-section 4.3.2 pour la résolution). •

Exemple 4.1.3 (Moindres carrés) En statistique ou bien en analyse de données on rencontre souvent le problème d'estimer les paramètres d'un modèle en fonction de données mesurées ou expérimentales. Lorsque le modèle est linéaire, ce que nous allons supposer, on parle de régression linéaire et cela revient à résoudre le problème suivant, dit "aux moindres carrés",

$$\inf_{x\in\mathbb{R}^n} \|Ax-b\|^2,$$

où $b \in \mathbb{R}^p$ sont les données, $x \in \mathbb{R}^n$ les paramètres inconnus et A, matrice réelle d'ordre $p \times n$, le modèle prédictif linéaire. •

Un exemple algébrique simple est celui du quotient de Rayleigh qui permet de calculer les valeurs et vecteurs propres d'une matrice symétrique.

Exemple 4.1.4 (Première valeur propre) Soit A une matrice carrée d'ordre n, symétrique. On veut caractériser et calculer les solutions de

$$\inf_{x\in\mathbb{R}^n,\|x\|=1} Ax\cdot x,$$

où $\|x\|$ est la norme euclidienne de x. Nous verrons qu'il s'agit bien sûr des vecteurs propres de A associés à sa plus petite valeur propre (cf. la Sous-section 4.2.3). •

Exemple 4.1.5 (Entropie) La notion d'entropie est fondamentale, aussi bien en thermodynamique qu'en physique statistique ou qu'en théorie de l'information. Afin de ne considérer que des problèmes de minimisation, les mathématiciens changent le signe de l'entropie (afin de remplacer sa maximisation par sa minimisation). Ainsi, en théorie de l'information on minimise l'entropie de Shannon

$$\inf_{p\in\mathbb{R}^n_+,\sum_{i=1}^n p_i=1} \sum_{i=1}^n p_i \log p_i,$$

voir l'Exercice 4.2.16 pour la solution. Un autre problème de minimisation d'entropie en théorie cinétique des gaz est proposé à l'Exercice 4.2.17. •

Exemple 4.1.6 (Minimisation d'une énergie mécanique) Il s'agit de minimiser l'énergie mécanique d'une membrane ou bien l'énergie électrostatique d'un conducteur. Soit Ω un ouvert borné de $\mathbb{R}^N$ et f une fonction continue sur $\overline{\Omega}$. Pour résoudre le problème de Dirichlet (3.1) pour le Laplacien, on peut minimiser l'énergie $J(v)$ définie par

$$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega fv\,dx.$$

pour des fonctions $v \in V_0$ avec

$$V_0 = \left\{\phi \in C^1(\overline{\Omega}) \text{ tel que } \phi = 0 \text{ sur } \partial\Omega\right\}.$$

Autrement dit, la résolution du problème aux limites (3.1) est équivalent à la résolution du problème de minimisation

$$\inf_{v\in V_0} J(v).$$

On vérifiera cette équivalence lors de l'Exercice 4.2.9. •

Exemple 4.1.7 (Minimisation de l'énergie complémentaire) En mécanique il est bien connu [16] que pour résoudre le problème aux limites (3.1) on peut soit minimiser l'énergie "en déplacements" $J(v)$ de l'Exemple 4.1.6 ci-dessus, ou bien on peut minimiser une autre énergie, dite **complémentaire** dont la signification physique est tout aussi importante que celle de $J(v)$. Pour un champ de vecteurs $\tau(x)$, de Ω dans $\mathbb{R}^N$, cette énergie complémentaire s'écrit

$$G(\tau) = \frac{1}{2}\int_\Omega |\tau|^2 dx. \tag{4.1}$$

Elle s'accompagne d'une contrainte sur le vecteur τ (qui s'apparente à une contrainte mécanique) qui doit être **statiquement admissible**, c'est-à-dire vérifier $-\mathrm{div}\tau = f$ dans Ω. Autrement dit, la résolution de (3.1) et la minimisation de l'énergie $J(v)$ de l'Exemple 4.1.6 sont équivalentes au problème de minimisation sous contrainte suivant

$$\inf_{-\mathrm{div}\tau = f \text{ dans } \Omega} G(\tau).$$

Nous verrons lors de l'Exercice 4.3.5 comment cette équivalence s'explique par la théorie de la dualité. •

La résolution numérique de la minimisation de l'énergie complémentaire peut se faire par une méthode d'éléments finis. Après discrétisation on obtient le problème suivant de minimisation en dimension finie.

Exemple 4.1.8 (Optimisation quadratique à contraintes linéaires) Soit A une matrice carrée d'ordre n, symétrique définie positive. Soit B une matrice rectangulaire de taille $m \times n$. Soit b un vecteur de $\mathbb{R}^m$. On veut résoudre le problème

$$\inf_{x\in\mathbb{R}^n,\ Bx=b} \left\{ J(x) = \frac{1}{2}Ax\cdot x \right\}.$$

Voir la Sous-section 4.2.3 pour sa résolution. •

Exemple 4.1.9 (Contrôle d'une membrane) On considère une membrane élastique, fixée sur son contour, et se déformant sous l'action d'une force f. Ce problème est modélisé par

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

où u est le déplacement vertical de la membrane et v est une force de contrôle à notre disposition. Ce contrôle est typiquement un actionneur piézo-électrique qui agit sur une partie ω du domaine Ω avec une intensité limitée. On définit donc l'ensemble des contrôles admissibles

$$K = \{v(x) \text{ tel que } v_{min}(x) \le v(x) \le v_{max}(x) \text{ dans } \omega \text{ et } v = 0 \text{ dans } \Omega \setminus \omega\},$$

où v_{min} et v_{max} sont deux fonctions données. On cherche le contrôle qui rend le déplacement u aussi proche que possible d'un déplacement désiré u_0, et qui soit d'un coût modéré. On définit donc un critère

$$J(v) = \frac{1}{2}\int_\Omega \left(|u - u_0|^2 + c|v|^2 \right) dx,$$

avec $c > 0$. Le problème de contrôle s'écrit

$$\inf_{v \in K} J(v).$$

•

Exemple 4.1.10 (Traitement d'images) On représente une image noir et blanc par une fonction définie sur un domaine Ω du plan dont les valeurs indiquent un niveau de gris. Une image acquise, par exemple par un appareil photographique, $f(x)$ est entachée de "bruit" qui correspond à des défauts du capteur. Un problème important est donc de "gommer" ce bruit et une technique idéale serait de lisser ou régulariser cette fonction $f(x)$. Mais les images contiennent des contours (qui sont des lignes de discontinuités du niveau de gris) qu'il faut absolument préserver durant ce lissage. C'est pourquoi une bonne régularisation $u(x)$ de l'image originale $f(x)$ peut être obtenue par la méthode de minimisation de la variation totale

$$\inf_{u(x):\Omega \to \mathbb{R}} \left\{ J(u) = \int_\Omega |\nabla u| \, dx + \ell \int_\Omega |f - u|^2 \, dx \right\},$$

où $\ell > 0$ est un paramètre qui permet de moduler le lissage. •

L'optimisation a de très nombreux liens avec la théorie de la commande optimale, le calcul des variations, les problèmes inverses, etc. On pourrait multiplier les exemples à l'infini...

4.1.2 Définitions et notations

L'optimisation a un vocabulaire particulier : introduisons quelques notations et définitions classiques. Nous considérons principalement des problèmes de minimisation (sachant qu'il suffit d'en changer le signe pour obtenir un problème de maximisation).

Tout d'abord, l'espace dans lequel est posé le problème, noté V, est supposé être un espace vectoriel normé, c'est-à-dire muni d'une norme notée $\|v\|$. On se donne également un sous-ensemble $K \subset V$ où l'on va chercher la solution : on dit que K est l'ensemble des éléments **admissibles** du problème, ou bien que K définit les **contraintes** s'exerçant sur le problème considéré. Enfin, le **critère**, ou la **fonction coût**, ou la **fonction objectif**, à minimiser, noté J, est une fonction définie sur K à valeurs dans $\mathbb{R}$. Le problème étudié sera donc noté

$$\inf_{v \in K \subset V} J(v). \tag{4.2}$$

Lorsque l'on utilise la notation inf pour un problème de minimisation, cela indique que l'on ne sait pas, a priori, si la valeur du minimum est atteinte, c'est-à-dire s'il existe $\overline{v} \in K$ tel que

$$J(\overline{v}) = \inf_{v \in K \subset V} J(v).$$

Si l'on veut indiquer que la valeur du minimum est atteinte, on utilise de préférence la notation

$$\min_{v \in K \subset V} J(v),$$

mais il ne s'agit pas d'une convention universelle (quoique fort répandue). Pour les problèmes de maximisation, les notations sup et max remplacent inf et min, respectivement. Précisons quelques définitions de base.

Définition 4.1.1 *On dit que u est un minimum (ou un point de minimum) local de J sur K si et seulement si*

$$u \in K \quad et \quad \exists \delta > 0 \ , \ \forall v \in K \ , \ \|v - u\| < \delta \Longrightarrow J(v) \geq J(u) \ .$$

On dit que u est un minimum (ou un point de minimum) global de J sur K si et seulement si

$$u \in K \quad et \quad J(v) \geq J(u) \quad \forall v \in K \ .$$

Définition 4.1.2 *On appelle infimum de J sur K (ou, plus couramment, valeur minimum), que l'on désigne par la notation (4.2), la borne supérieure dans $\mathbb{R}$ des constantes qui minorent J sur K. Si J n'est pas minorée sur K, alors l'infimum vaut $-\infty$. Si K est vide, par convention l'infimum est $+\infty$.*

Une suite minimisante de J dans K est une suite $(u^n)_{n\in\mathbb{N}}$ telle que

$$u^n \in K \ \forall \, n \quad et \quad \lim_{n\to+\infty} J(u^n) = \inf_{v\in K} J(v).$$

Par la définition même de l'infimum de J sur K il existe toujours des suites minimisantes. Pour s'en convaincre il suffit d'utiliser un argument de contradiction avec la définition de l'infimum.

4.1.3 Existence de minima en dimension finie

Intéressons nous maintenant à la question de l'existence de minima pour des problèmes d'optimisation posés en dimension finie. Nous supposerons dans cette sous-section que $V = \mathbb{R}^N$ que l'on munit du produit scalaire usuel $u \cdot v = \sum_{i=1}^N u_i v_i$ et de la norme euclidienne $\|u\| = \sqrt{u \cdot u}$.

Un résultat assez général garantissant l'existence d'un minimum est le suivant.

Théorème 4.1.3 (Existence d'un minimum en dimension finie) *Soit K un ensemble fermé non vide de $\mathbb{R}^N$, et J une fonction continue sur K à valeurs dans $\mathbb{R}$ vérifiant la propriété, dite "infinie à l'infini",*

$$\forall (u^n)_{n\geq 0} \text{ suite dans } K \ , \ \lim_{n\to+\infty} \|u^n\| = +\infty \Longrightarrow \lim_{n\to+\infty} J(u^n) = +\infty \ . \tag{4.3}$$

Alors il existe au moins un point de minimum de J sur K. De plus, on peut extraire de toute suite minimisante de J sur K une sous-suite convergeant vers un point de minimum sur K.

Démonstration. Soit (u^n) une suite minimisante de J sur K. La condition (4.3) entraîne que u^n est bornée puisque $J(u^n)$ est une suite de réels majorée. Donc, il existe une sous-suite (u^{n_k}) qui converge vers un point u de $\mathbb{R}^N$. Mais $u \in K$ puisque K est fermé, et $J(u^{n_k})$ converge vers $J(u)$ par continuité, d'où $J(u) = \inf_{v\in K} J(v)$ d'après la Définition 4.1.2. □

Remarque 4.1.4 Notons que la propriété (4.3), qui assure que toute suite minimisante de J sur K est bornée, est automatiquement vérifiée si K est borné. Lorsque l'ensemble K n'est pas borné, cette condition exprime que, dans K, J est **infinie à l'infini**. •

Exercice 4.1.1 Montrer par des exemples que le fait que K est fermé ou que J est continue est en général nécessaire pour l'existence d'un minimum. Donner un exemple de fonction continue et minorée de $\mathbb{R}$ dans $\mathbb{R}$ n'admettant pas de minimum sur $\mathbb{R}$.

Exercice 4.1.2 Montrer qu'il existe un minimum pour les Exemples 4.1.1 et 4.1.4.

L'existence d'un minimum en dimension infinie n'est **absolument pas garantie** par des conditions du type de celles utilisées dans l'énoncé du Théorème 4.1.3. Cette difficulté est intimement liée au fait qu'en dimension infinie les fermés bornés ne sont pas compacts ! Nous donnons un exemple abstrait qui explique bien le mécanisme de "fuite à l'infini" qui empêche l'existence d'un minimum.

Exemple 4.1.11 Soit l'espace de Hilbert (de dimension infinie) des suites de carré sommable dans $\mathbb{R}$

$$\ell_2(\mathbb{R}) = \left\{ x = (x_i)_{i\geq 1} \text{ tel que } \sum_{i=1}^{+\infty} x_i^2 < +\infty \right\},$$

muni du produit scalaire $\langle x, y\rangle = \sum_{i=1}^{+\infty} x_i y_i$ et de la norme associée $\|x\| = \sqrt{\langle x, x\rangle}$. On considère la fonction J définie sur $\ell_2(\mathbb{R})$ par

$$J(x) = \left(\|x\|^2 - 1\right)^2 + \sum_{i=1}^{+\infty} \frac{x_i^2}{i}.$$

Prenant $K = \ell_2(\mathbb{R})$, on considère le problème

$$\inf_{x\in\ell_2(\mathbb{R})} J(x) \ , \tag{4.4}$$

pour lequel nous allons montrer qu'il n'existe pas de point de minimum. Vérifions tout d'abord que

$$\left(\inf_{x\in\ell_2(\mathbb{R})} J(x)\right) = 0.$$

Introduisons la suite x^n dans $\ell_2(\mathbb{R})$ définie par $x_i^n = \delta_{in}$ (symbole de Kronecker) pour tout $i \geq 1$. On vérifie aisément que $\|x^n\| = 1$ et

$$J(x^n) = \frac{1}{n} \to 0 \ \text{ quand } \ n \to +\infty.$$

Comme J est positive, on en déduit que x^n est une suite minimisante et que la valeur du minimum est nulle. Cependant, il est évident qu'il n'existe aucun $\overline{x} \in \ell_2(\mathbb{R})$ tel que $J(\overline{x}) = 0$. Par conséquent, il n'existe pas de point de minimum pour (4.4). On voit dans cet exemple que la suite minimisante x^n "part à l'infini" (ce qui est impossible dans un espace de dimension finie) et n'est pas compacte dans $\ell_2(\mathbb{R})$ (bien qu'elle soit bornée). •

Pour obtenir l'existence de minima à des problèmes d'optimisation en dimension infinie il faut avoir recours à des hypothèses supplémentaires comme la compacité ou la convexité. Cela dépasse largement le niveau de ce cours et nous renvoyons le lecteur aux ouvrages [1], [4], [11] pour plus de détails.

4.1.4 Analyse convexe

La convexité est une propriété très importante en optimisation car elle permet un grand nombre de simplifications. En particulier, nous verrons qu'il n'y a pas de différence entre minima locaux et globaux pour les fonctions convexes et, plus loin dans la Section 4.2.2, que les conditions d'optimalité sont "meilleures" pour les problèmes convexes.

Dans tout ce qui suit, nous supposerons que V est un espace vectoriel muni d'un produit scalaire $\langle u, v\rangle$ et d'une norme associée $\|v\|$. Rappelons qu'un ensemble K est convexe s'il contient tous les segments reliant deux quelconques de ses points. Donnons quelques propriétés des fonctions convexes.

Définition 4.1.5 *On dit qu'une fonction J définie sur un ensemble convexe non vide $K \in V$ et à valeurs dans $\mathbb{R}$ est convexe sur K si et seulement si*

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) \quad \forall\, u, v \in K\ ,\ \forall\, \theta \in [0,1]\ . \tag{4.5}$$

De plus, J est dite strictement convexe si l'inégalité (4.5) est stricte lorsque $u \neq v$ et $\theta \in]0,1[$.

Exercice 4.1.3 Soient J_1 et J_2 deux fonctions convexes sur V, $\lambda > 0$, et φ une fonction convexe croissante sur un intervalle de $\mathbb{R}$ contenant l'ensemble $J_1(V)$. Montrer que $J_1 + J_2$, $\max(J_1, J_2)$, λJ_1 et $\varphi \circ J_1$ sont convexes.

Exercice 4.1.4 Soit $(L_i)_{i\in I}$ une famille (éventuellement infinie) de fonctions affines sur V. Montrer que $\sup_{i\in I} L_i$ est convexe sur V. Réciproquement, soit J une fonction convexe continue sur V. Montrer que J est égale au $\sup_{L_i \leq J} L_i$ où les fonctions L_i sont affines.

Pour les fonctions convexes il n'y a pas de différence entre minima locaux et globaux comme le montre le résultat élémentaire suivant.

Proposition 4.1.6 *Si J est une fonction convexe sur un ensemble convexe K, tout point de minimum local de J sur K est un minimum global et l'ensemble des points de minimum est un ensemble convexe (éventuellement vide).*

Si de plus J est strictement convexe, alors il existe au plus un point de minimum.

Démonstration. Soit u un minimum local de J sur K. D'après la Définition 4.1.1, nous pouvons écrire

$$\exists\, \delta > 0\ ,\ \forall\, w \in K\ ,\ \|w - u\| < \delta \implies J(w) \geq J(u)\ . \tag{4.6}$$

Soit $v \in K$. Pour $\theta \in]0,1[$ suffisamment petit, $w_\theta = \theta v + (1-\theta)u$ vérifie $\|w_\theta - u\| < \delta$ et $w_\theta \in K$ puisque K est convexe. Donc, $J(w_\theta) \geq J(u)$ d'après (4.6), et la convexité

de J implique que $J(u) \leq J(w_\theta) \leq \theta J(v) + (1-\theta)J(u)$, ce qui montre bien que $J(u) \leq J(v)$, c'est-à-dire que u est un minimum global sur K.

D'autre part, si u_1 et u_2 sont deux minima et si $\theta \in [0,1]$, alors $w = \theta u_1 + (1-\theta)u_2$ est un minimum puisque $w \in K$ et que

$$\inf_{v\in K} J(v) \leq J(w) \leq \theta J(u_1) + (1-\theta)J(u_2) = \inf_{v\in K} J(v) \ .$$

Le même raisonnement avec $\theta \in]0,1[$ montre que, si J est strictement convexe, alors nécessairement $u_1 = u_2$. □

Une propriété agréable des fonctions convexes "propres" (c'est-à-dire qui ne prennent pas la valeur $+\infty$) est d'être continues.

Exercice 4.1.5 Soit $v_0 \in V$ et J une fonction convexe majorée sur une boule de centre v_0. Montrer que J est minorée et continue sur cette boule.

4.2 Conditions d'optimalité

Dans cette section, nous allons chercher à obtenir des conditions nécessaires et parfois suffisantes de minimalité. L'objectif est d'une certaine manière beaucoup plus pratique que la question de l'existence d'un minimum, puisque ces conditions d'optimalité seront souvent utiles pour caractériser un minimum (sans avoir même su démontrer son existence !). Les conditions d'optimalité sont aussi à la base de la plupart des méthodes numériques d'optimisation. En l'absence de contraintes, l'idée générale des conditions d'optimalité est la même que celle qui, lorsque l'on calcule l'extremum d'une fonction sur $\mathbb{R}$ tout entier, consiste à écrire que sa dérivée doit s'annuler. Ces conditions vont donc s'exprimer à l'aide de la dérivée première (conditions d'ordre 1) ou seconde (conditions d'ordre 2). Nous obtiendrons surtout des conditions **nécessaires** d'optimalité, mais l'utilisation de la dérivée seconde ou l'introduction d'hypothèses de convexité permettront aussi d'obtenir des conditions **suffisantes**, et de distinguer entre minima et maxima.

En présence de contraintes, les conditions d'optimalité généralisent la remarque élémentaire suivante : si J est une fonction dérivable de l'intervalle $[a,b] \subset \mathbb{R}$ dans $\mathbb{R}$ et que x_0 est un point de minimum local de J sur $[a,b]$, alors on a

$$J'(x_0) \geq 0 \text{ si } x_0 = a \ , \ J'(x_0) = 0 \text{ si } x_0 \in]a,b[\ , \ J'(x_0) \leq 0 \text{ si } x_0 = b \ .$$

Même si elle est bien connue du lecteur, rappelons la démonstration de cette remarque : si $x_0 \in [a,b[$, on peut choisir $x = x_0 + h$ avec $h > 0$ petit et écrire $J(x) \geq J(x_0)$, d'où $J(x_0) + hJ'(x_0) + o(h) \geq J(x_0)$, ce qui donne $J'(x_0) \geq 0$ en divisant par h et en faisant tendre h vers 0. De même obtient-on $J'(x_0) \leq 0$ si $x_0 \in]a,b]$ en considérant $x = x_0 - h$. Remarquons également (c'est la condition d'ordre 2) que si $x_0 \in]a,b[$ et si J' est dérivable en x_0, on a alors $J''(x_0) \geq 0$ (en effet, on a $J(x_0) + \dfrac{h^2}{2}J''(x_0) + o(h^2) \geq J(x_0)$ pour h assez petit).

La stratégie d'obtention et de démonstration des conditions de minimalité est donc claire : on tient compte des contraintes ($x \in [a,b]$ dans l'exemple ci-dessus)

pour tester la minimalité de x_0 dans des directions particulières qui respectent les contraintes ($x_0 + h$ avec $h > 0$ si $x_0 \in [a, b[$, $x_0 - h$ avec $h > 0$ si $x_0 \in]a, b]$) : on parlera de **directions admissibles**. On utilise ensuite la définition de la dérivée (et les formules de Taylor à l'ordre 2) pour conclure. C'est exactement ce que nous allons faire dans ce qui suit !

4.2.1 Différentiabilité

Désormais (et nous ne le rappellerons plus systématiquement), nous supposerons que V est un espace vectoriel, muni d'un produit scalaire noté $\langle u, v\rangle$ et d'une norme associée $\|u\|$. On suppose aussi que le critère J est une fonction continue à valeurs dans $\mathbb{R}$.

Commençons par introduire la notion de dérivée première de J puisque nous en aurons besoin pour écrire des conditions d'optimalité. Lorsqu'il y a plusieurs variables (c'est-à-dire si l'espace V n'est pas $\mathbb{R}$), la "bonne" notion théorique de dérivabilité, appelée différentiabilité au sens de Fréchet, est donnée par la définition suivante.

Définition 4.2.1 *On dit que la fonction J, définie sur un voisinage de $u \in V$ à valeurs dans $\mathbb{R}$, est dérivable (ou différentiable) au sens de Fréchet en u s'il existe une forme linéaire continue sur V, notée L, telle que*

$$J(u + w) = J(u) + L(w) + o(w) \quad , \quad \textit{avec} \quad \lim_{w\to 0} \frac{|o(w)|}{\|w\|} = 0 \, . \tag{4.7}$$

On appelle L la dérivée (ou la différentielle, ou le gradient) de J en u et on note $L = J'(u)$.

Remarque 4.2.2 Rappelons qu'une forme linéaire L sur V est une application linéaire de V dans $\mathbb{R}$. Elle est continue s'il existe une constante $C > 0$ telle que $|L(w)| \leq C\|w\|$ pour tout $w \in V$. En dimension finie (par exemple, si $V = \mathbb{R}^N$) toute forme linéaire est continue, donc il n'y a rien à vérifier en ce qui concerne la continuité ! Toujours en dimension finie (et même dans un espace de Hilbert), on sait, grâce au théorème de représentation de Riesz (théorème 12.1.18 dans [1], ou bien [14]), que pour toute forme linéaire continue L il existe un unique $p \in V$ tel que $\langle p, w\rangle = L(w)$. Autrement dit, la relation (4.7) peut s'écrire de manière équivalente et parfois plus simple

$$J(u + w) = J(u) + \langle p, w\rangle + o(w) \quad , \quad \text{avec} \quad \lim_{w\to 0} \frac{|o(w)|}{\|w\|} = 0 \, . \tag{4.8}$$

On note aussi parfois $p = J'(u)$ qui est cette fois un élément de V (et non plus une forme linéaire sur V). •

Dans la plupart des applications, il suffit souvent de déterminer la forme linéaire continue $L = J'(u)$ car on n'a pas besoin de l'expression explicite de $p = J'(u) \in V$. En pratique, il est parfois plus facile de trouver l'expression explicite de L que celle de p, comme le montre l'exercice suivant.

Exercice 4.2.1 (essentiel !) Soit $a(\cdot,\cdot)$ une forme bilinéaire symétrique sur $V \times V$ (cf. la définition de la Remarque 3.1.10), continue par rapport à chacun de ses arguments. Soit L une forme linéaire continue sur V. On pose $J(u) = \frac{1}{2}a(u,u) - L(u)$. Montrer que J est dérivable sur V et que $J'(u)(w) = a(u,w) - L(w)$ pour tout $u, w \in V$.

Lorsque $V = \mathbb{R}^N$ on peut plus facilement trouver l'expression explicite de $J'(u)$ comme le montre l'exercice suivant.

Exercice 4.2.2 Soit A une matrice symétrique $N \times N$ et $b \in \mathbb{R}^N$. Pour $x \in \mathbb{R}^N$, on pose $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Montrer que J est dérivable et que $J'(x) = Ax - b$ pour tout $x \in \mathbb{R}^N$.

Exercice 4.2.3 Montrer que (4.7) implique la continuité de J en u. Montrer aussi que, si deux formes linéaires continues L_1, L_2 vérifient

$$\begin{cases} J(u+w) \geq J(u) + L_1(w) + o(w)\ , \\ J(u+w) \leq J(u) + L_2(w) + o(w)\ , \end{cases} \tag{4.9}$$

alors J est dérivable et $L_1 = L_2 = J'(u)$.

Remarque 4.2.3 Il existe d'autres notions de différentiabilité, plus faible que celle au sens de Fréchet. Par exemple, on rencontre souvent la définition suivante. On dit que la fonction J, définie sur un voisinage de $u \in V$ à valeurs dans $\mathbb{R}$, est différentiable au sens de Gâteaux en u s'il existe une forme linéaire continue sur V, notée L, telle que

$$\forall\, w \in V \quad , \quad \lim_{\delta \to 0,\, \delta > 0} \frac{J(u+\delta w) - J(u)}{\delta} = L(w)\ . \tag{4.10}$$

L'intérêt de cette notion est que la vérification de (4.10) est plus aisée que celle de (4.7). Cependant, si une fonction dérivable au sens de Fréchet l'est aussi au sens de Gâteaux, la réciproque est fausse, même en dimension finie, comme le montre l'exemple suivant dans $\mathbb{R}^2$

$$J(x,y) = \frac{x^6}{(y-x^2)^2 + x^8} \quad \text{pour} \quad (x,y) \neq (0,0) \quad , \quad J(0,0) = 0\ .$$

Convenons que, dans ce qui suit, nous dirons qu'une fonction est dérivable lorsqu'elle l'est au sens de Fréchet, sauf mention explicite du contraire. •

Examinons maintenant les propriétés de base des fonctions convexes dérivables.

Proposition 4.2.4 *Soit J une application différentiable de V dans $\mathbb{R}$. Les assertions suivantes sont équivalentes*

$$J \quad \textit{est convexe sur} \quad V\ , \tag{4.11}$$

$$J(v) \geq J(u) + \langle J'(u), v-u\rangle \quad \forall\, u, v \in V\ , \tag{4.12}$$

$$\langle J'(u) - J'(v), u - v\rangle \geq 0 \quad \forall\, u, v \in V\ . \tag{4.13}$$

Remarque 4.2.5 La condition (4.12) a une interprétation géométrique simple. Elle signifie que la fonction convexe $J(v)$ est toujours au dessus de son plan tangent en u (considéré comme une fonction affine de v). La condition (4.13) est une hypothèse de monotonie ou de croissance de J'. •

Démonstration. Montrons que (4.11) implique (4.12). La condition de convexité (4.5) s'écrit, pour $0 < \theta < 1$,

$$J\big((1-\theta)u + \theta v\big) \leq (1-\theta)J(u) + \theta J(v) \ ,$$

d'où

$$\frac{J\big(u + \theta(v-u)\big) - J(u)}{\theta} \leq J(v) - J(u) \ .$$

En faisant tendre θ vers 0, on trouve (4.12). Pour obtenir (4.13) il suffit d'additionner (4.12) avec lui-même en échangeant u et v.

Montrons que (4.13) implique (4.11). Pour $u, v \in V$ et $t \in \mathbb{R}$, on pose $\varphi(t) = J(u + t(v-u))$. Alors φ est dérivable sur $\mathbb{R}$ et $\varphi'(t) = \langle J'(u + t(v-u)), v-u\rangle$. En remplaçant u par $u + t(v-u)$ et v par $u + s(v-u)$, l'inégalité (4.13) conduit à

$$\varphi'(t) - \varphi'(s) \geq 0 \quad \text{si} \quad t \geq s \ . \tag{4.14}$$

Soit $\theta \in]0,1[$. En intégrant l'inégalité (4.14) de $t = \theta$ à $t = 1$ et de $s = 0$ à $s = \theta$, on obtient

$$\theta\varphi(1) + (1-\theta)\varphi(0) - \varphi(\theta) \geq 0 \ ,$$

c'est-à-dire (4.11). □

Corollaire 4.2.6 *Soit J une fonction convexe différentiable de V dans $\mathbb{R}$. Si $u \in V$ vérifie $J'(u) = 0$, alors u est un point de minimum global de J.*

Démonstration. De la propriété (4.12) on déduit immédiatement que $J(v) \geq J(u)$ pour tout $v \in V$. □

Exercice 4.2.4 Montrer qu'une fonction J dérivable sur V est strictement convexe si et seulement si

$$J(v) > J(u) + \langle J'(u), v-u\rangle \quad \forall\, u, v \in V \quad \text{avec} \quad u \neq v \ ,$$

ou encore

$$\langle J'(u) - J'(v), u-v\rangle > 0 \quad \forall\, u, v \in V \quad \text{avec} \quad u \neq v \ .$$

Terminons cette sous-section en définissant la **dérivée seconde** de J.

Définition 4.2.7 *Soit J une fonction de V dans $\mathbb{R}$, dérivable en $u \in V$ de dérivée $J'(u)$. On dit que J est deux fois dérivable en u si elle vérifie pour tout $w \in V$*

$$J(u+w) = J(u) + J'(u)w + \frac{1}{2}J''(u)(w,w) + o(\|w\|^2), \ \textit{avec} \ \lim_{w\to 0} \frac{o(\|w\|^2)}{\|w\|^2} = 0 \ , \tag{4.15}$$

où $J''(u)$ est identifiée à une forme bilinéaire continue sur $V \times V$.

En pratique, on calcule donc la dérivée seconde $J''(u)(w,w)$ en faisant un développement de Taylor à l'ordre 2 suivant la formule (4.15). On peut vérifier que cette dérivée seconde est elle-même la dérivée de $J'(u)$. Pour cela il faut d'abord généraliser la Définition 4.2.1 de différentiabilité au cas d'une fonction f définie sur V à valeurs dans un autre espace vectoriel W (et non pas seulement dans $\mathbb{R}$), ce qui est facile mais que nous ne faisons pas par souci de simplicité (voir [1], [4] si nécessaire).

Lorsque J est deux fois dérivable on retrouve la condition usuelle de convexité : si la dérivée seconde est positive, alors la fonction est convexe.

Exercice 4.2.5 Montrer que si J est deux fois dérivable sur V les conditions de la Proposition 4.2.4 sont équivalentes à

$$J''(u)(w,w) \geq 0 \qquad \forall\, u, w \in V\ . \tag{4.16}$$

4.2.2 Inéquations d'Euler et contraintes convexes

Nous commençons par formuler les conditions de minimalité lorsque l'ensemble des contraintes K est convexe, cas où les choses sont plus simples (nous supposerons toujours que K est fermé non vide et que J est continue sur un ouvert contenant K). L'idée essentielle du résultat qui suit est que, pour tout $v \in K$, on peut tester l'optimalité de u dans la "direction admissible" $(v-u)$ car $u + h(v-u) \in K$ si $h \in [0,1]$.

Théorème 4.2.8 (Inéquation d'Euler, cas convexe) *Soit $u \in K$ convexe. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, alors*

$$\langle J'(u), v-u\rangle \geq 0 \quad \forall\, v \in K\ . \tag{4.17}$$

Si $u \in K$ vérifie (4.17) et si J est convexe, alors u est un minimum global de J sur K.

Remarque 4.2.9 On appelle (4.17), "inéquation d'Euler". Il s'agit d'une condition **nécessaire** d'optimalité qui devient **nécessaire et suffisante** si J est convexe. La condition (4.17) exprime que la dérivée directionnelle de J au point u dans toutes les directions $(v-u)$, qui sont **rentrantes** dans K, est positive, c'est-à-dire que la fonction J ne peut que croître localement à l'intérieur de K (voir la Figure 4.1). Il faut aussi remarquer que, dans deux cas importants, (4.17) **se réduit simplement à l'équation d'Euler** $J'(u) = 0$. En premier lieu, si $K = V$, $v - u$ décrit tout V lorsque v décrit V, et donc (4.17) entraîne $J'(u) = 0$. D'autre part, si u est intérieur à K, la même conclusion s'impose. •

Démonstration. Pour $v \in K$ et $h \in]0,1]$, $u + h(v-u) \in K$, et donc

$$\frac{J(u+h(v-u)) - J(u)}{h} \geq 0\ . \tag{4.18}$$

On en déduit (4.17) en faisant tendre h vers 0. Le caractère suffisant de (4.17) pour une fonction convexe découle immédiatement de la propriété de convexité (4.12). □

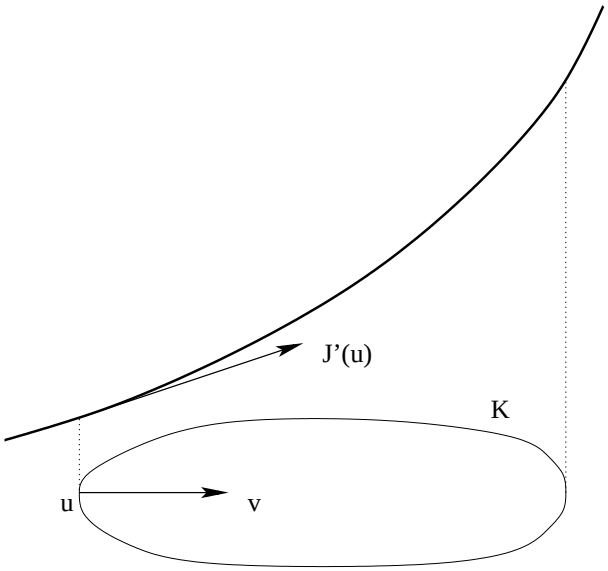

FIGURE 4.1 – Inéquation d'Euler : l'angle entre la dérivée $J'(u)$ et la direction rentrante $(v-u)$ est aigu.

Exercice 4.2.6 Soit K un convexe fermé non vide de V. Pour $x \in V$, on cherche la projection orthogonale $x_K \in K$ de x sur K

$$\|x - x_K\| = \min_{y \in K} \|x - y\|.$$

Montrer que la condition nécessaire et suffisante (4.17) est exactement

$$x_K \in K, \langle x_K - x, x_K - y \rangle \leq 0 \quad \forall y \in K. \tag{12.1}$$

Exercice 4.2.7 Montrer que le problème de minimisation "aux moindres carrés" de l'Exemple 4.1.3 admet toujours une solution et écrire l'équation d'Euler correspondante.

Exercice 4.2.8 Soit A une matrice carrée d'ordre n, symétrique positive. Soit B une matrice rectangulaire de taille $m \times n$ avec $m \leq n$. Soit b un vecteur de $\mathbb{R}^n$. On veut résoudre le problème

$$\inf_{x \in \mathrm{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}.$$

Montrer qu'il existe une solution si A est positive et $b \in (\mathrm{Ker} A \cap \mathrm{Ker} B)^{\perp}$, et qu'elle est unique si A est définie positive. Montrer que tout point de minimum $\overline{x} \in \mathbb{R}^n$ vérifie

$$A\overline{x} - b = B^* p \text{ avec } p \in \mathbb{R}^m.$$

Exercice 4.2.9 On reprend l'Exemple 4.1.6 et ses notations. Montrer que l'énergie

$$J(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 dx - \int_\Omega f v \, dx$$

est strictement convexe. Montrer que l'équation d'Euler vérifiée par un possible point de minimum $u \in V_0$ de

$$\inf_{v \in V_0} J(v)$$

est précisément la formulation variationnelle

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall\, v \in V_0$$

du problème aux limites (3.1).

Exercice 4.2.10 Soit K un convexe fermé non vide de V, soit a une forme bilinéaire symétrique continue coercive sur V, et soit L une forme linéaire continue sur V. Si V est de dimension finie, montrer que $J(v) = \frac{1}{2}a(v,v) - L(v)$ admet un unique point de minimum dans K, noté u. Montrer qu'un point de minimum u est aussi solution du problème (appelé inéquation variationnelle)

$$u \in K \quad \text{et} \quad a(u, v-u) \geq L(v-u) \quad \forall\, v \in K\ ,$$

et réciproquement.

Exercice 4.2.11 Soit J_1 et J_2 deux fonctions convexes continues sur une partie convexe fermée non vide $K \subset V$. On suppose que J_1 seulement est dérivable. Montrer que $u \in K$ est un minimum de $J_1 + J_2$ si et seulement si

$$\langle J_1'(u), v-u \rangle + J_2(v) - J_2(u) \geq 0 \quad \forall\, v \in K\ .$$

Terminons cette sous-section en donnant une **condition d'optimalité du deuxième ordre**.

Proposition 4.2.10 *On suppose que $K = V$ et que J est deux fois dérivable en u. Si u est un point de minimum local de J, alors*

$$J'(u) = 0 \quad et \quad J''(u)(w,w) \geq 0 \quad \forall\, w \in V\ . \tag{4.19}$$

Réciproquement, si, pour tout v dans un voisinage de u,

$$J'(u) = 0 \quad et \quad J''(v)(w,w) \geq 0 \quad \forall\, w \in V\ , \tag{4.20}$$

alors u est un minimum local de J.

Démonstration. Si u est un point de minimum local, on sait déjà que $J'(u) = 0$ et la formule (4.15) nous donne (4.19). Réciproquement, si u vérifie (4.20), on écrit un développement de Taylor à l'ordre deux (au voisinage de zéro) avec reste exact pour la fonction $\phi(t) = J(u + tw)$ avec $t \in \mathbb{R}$ et on en déduit aisément que u est un minimum local de J (voir la Définition 4.1.1). □

4.2.3 Multiplicateurs de Lagrange

Cherchons maintenant à écrire des conditions de minimalité lorsque l'ensemble K n'est pas convexe. Plus précisément, nous étudierons des ensembles K définis par des **contraintes d'égalité** ou des **contraintes d'inégalité** (ou les deux à la fois). Nous commençons par une remarque générale sur les **directions admissibles**.

Définition 4.2.11 *En tout point $v \in K$, l'ensemble*

$$K(v) = \left\{ \begin{array}{l} w \in V \text{ tel que } \exists\, (v^n)_{n\geq 0} \in K\,,\, \exists\, (\varepsilon^n)_{n\geq 0} \in \mathbb{R}_+^* \text{ vérifiant} \\ \lim_{n\to+\infty} v^n = v\,,\ \lim_{n\to+\infty} \varepsilon^n = 0\,,\ \lim_{n\to+\infty} \frac{v^n - v}{\varepsilon^n} = w \end{array} \right\}$$

est appelé le cône des directions admissibles au point v.

En termes géométriques, $K(v)$ est l'ensemble de tous les vecteurs w qui sont tangents en v à une courbe contenue dans K et passant par v (si K est une variété régulière, $K(v)$ n'est rien d'autre que l'espace tangent à K en v). Autrement dit, $K(v)$ est l'ensemble de toutes les directions possibles de variations à partir de v qui restent infinitésimalement dans K.

Il est facile de vérifier que $0 \in K(v)$ (prendre la suite constante $v^n = v$) et que l'ensemble $K(v)$ est un cône, c'est-à-dire que $\lambda w \in K(v)$ pour tout $w \in K(v)$ et tout $\lambda \geq 0$.

Exercice 4.2.12 Montrer que $K(v)$ est un cône fermé et que $K(v) = V$ si v est intérieur à K. Donner un exemple où $K(v)$ est réduit à $\{0\}$.

L'intérêt du cône des directions admissibles réside dans le résultat suivant, qui donne une condition **nécessaire** d'optimalité.

Proposition 4.2.12 (Inéquation d'Euler, cas général) *Soit u un minimum local de J sur K. Si J est différentiable en u, on a*

$$\langle J'(u), w\rangle \geq 0 \quad \forall\, w \in K(u)\,.$$

Démonstration. Soit $w \in K(u)$ et v^n une des suites correspondantes dans la Définition 4.2.11 de $K(u)$. Pour n suffisament grand, puisque $v^n \in K$ converge vers u et que u est un minimum local de J sur K, on a $J(v^n) \geq J(u)$. On pose alors $w^n = (v^n - u)/\varepsilon^n$ qui, par définition, converge vers w et on effectue un développement de Taylor

$$J(v^n) = J(u + \varepsilon^n w^n) = J(u) + \varepsilon^n \langle J'(u), w^n\rangle + o(\varepsilon^n) \geq J(u).$$

On en déduit

$$\langle J'(u), w^n\rangle + o(1) \geq 0,$$

qui donne le résultat voulu lorsque n tend vers l'infini. □

Nous allons maintenant préciser la condition nécessaire de la Proposition 4.2.12 dans le cas où K est donné par des **contraintes d'égalité** ou **d'inégalité**.

Contraintes d'égalité

Dans ce premier cas on suppose que K est donné par

$$K = \{v \in V \ , \quad F(v) = 0\} \ , \tag{4.21}$$

où $F(v) = (F_1(v), ..., F_M(v))$ est une application de V dans $\mathbb{R}^M$, avec $M \geq 1$. La condition **nécessaire** d'optimalité prend alors la forme suivante.

Théorème 4.2.13 *Soit $u \in K$ où K est donné par (4.21). On suppose que J est dérivable en $u \in K$ et que les fonctions $(F_i)_{1\leq i\leq M}$ sont continûment dérivables dans un voisinage de u. On suppose de plus que les vecteurs $\left(F_i'(u)\right)_{1\leq i\leq M}$ sont linéairement indépendants. Alors, si u est un minimum local de J sur K, il existe $\lambda_1, \ldots, \lambda_M \in \mathbb{R}$, appelés* **multiplicateurs de Lagrange**, *tels que*

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0 \ . \tag{4.22}$$

Démonstration. Soit $w \in K(u)$ et v^n une des suites correspondantes dans la Définition 4.2.11 de $K(u)$. Puisque $v^n \in K$ on a $F_i(v^n) = 0$ pour tout indice $1 \leq i \leq M$. On pose alors $w^n = (v^n - u)/\varepsilon^n$ qui, par définition, converge vers w et on effectue un développement de Taylor

$$0 = F_i(v^n) = F_i(u + \varepsilon^n w^n) = F_i(u) + \varepsilon^n \langle F_i'(u), w^n\rangle + o(\varepsilon^n).$$

Comme $F_i(u) = 0$, on en déduit

$$\langle F_i'(u), w^n\rangle + o(1) = 0,$$

qui, lorsque n tend vers l'infini, conduit à

$$\langle F_i'(u), w\rangle = 0 \quad \text{ pour tout } w \in K(u).$$

Par conséquent

$$K(u) \subset \{w \in V \ , \quad \langle F_i'(u), w\rangle = 0 \quad \text{pour} \quad i = 1, \ldots, M\} \, . \tag{4.23}$$

L'hypothèse que les vecteurs $\left(F_i'(u)\right)_{1\leq i\leq M}$ sont linéairement indépendants permet d'appliquer le théorème des fonctions implicites à la définition (4.21) de K qui est ainsi une variété régulière dont l'espace tangent en u est $K(u)$ qui est égal à (et non pas seulement inclus dans) l'espace vectoriel de droite dans (4.23). Autrement dit, le théorème des fonctions implicites permet de démontrer l'inclusion inverse et donc l'égalité dans (4.23) : nous ne détaillons pas la preuve qui est technique mais classique (voir par exemple [4]) et qui requiert que chaque fonction F_i soit continûment différentiable (c'est-à-dire que la fonction $u \to F_i'(u)$ soit continue). De façon équivalente, on obtient donc que

$$K(u) = \bigcap_{i=1}^{M} [F_i'(u)]^{\perp} \, . \tag{4.24}$$

Comme $K(u)$ est un espace vectoriel en vertu de (4.24), on peut prendre successivement w et $-w$ dans la Proposition 4.2.12, ce qui conduit à

$$\langle J'(u), w\rangle = 0 \quad \forall\, w \in \bigcap_{i=1}^{M} [F_i'(u)]^{\perp},$$

c'est-à-dire

$$J'(u) \in \left(\bigcap_{i=1}^{M} [F_i'(u)]^{\perp}\right)^{\perp} = [F_1'(u), ..., F_M'(u)]\,.$$

En d'autres termes, $J'(u)$ est engendré par les $\left(F_i'(u)\right)_{1\leq i\leq M}$ (notons que les multiplicateurs de Lagrange sont définis de manière unique). □

Remarque 4.2.14 Lorsque les vecteurs $\left(F_i'(u)\right)_{1\leq i\leq M}$ sont linéairement indépendants (ou libres), on dit que l'on est dans un **cas régulier**. Dans le cas contraire, on parle de **cas non régulier** et la conclusion du Théorème 4.2.13 est fausse comme le montre l'exemple suivant.

Prenons $V = \mathbb{R}$, $M = 1$, $F(v) = v^2$, $J(v) = v$, d'où $K = \{0\}$, $u = 0$, $F'(u) = 0$: il s'agit donc d'un cas non régulier. Comme $J'(u) = 1$, (4.22) n'a pas lieu. •

Pour bien comprendre la portée du Théorème 4.2.13, nous l'appliquons sur le problème de l'Exercice 4.2.8

$$\min_{x\in \mathrm{Ker}B} \left\{ J(x) = \frac{1}{2}Ax\cdot x - b\cdot x \right\},$$

où A est symétrique définie positive d'ordre n, et B de taille $m \times n$ avec $m \leq n$. On note $(b_i)_{1\leq i\leq m}$ les m lignes de B et on a donc m contraintes $b_i \cdot x = 0$. Pour simplifier on suppose que le rang de B est m, c'est-à-dire que les vecteurs (b_i) sont libres. Si le rang de B est $m' < m$, alors $(m - m')$ lignes de B sont engendrées par m' autres lignes libres de B. Il y a donc $(m - m')$ contraintes redondantes que l'on peut éliminer et on se ramène au cas d'une matrice B' de taille $m' \times n$ et de rang maximal m'. Comme le rang de B est m, les (b_i) sont libres et on peut appliquer la conclusion (4.22). Il existe donc un multiplicateur de Lagrange $p \in \mathbb{R}^m$ tel que un point de minimum $\overline{x}$ vérifie

$$A\overline{x} - b = \sum_{i=1}^{m} p_i b_i = B^* p.$$

Comme A est inversible, on en déduit la valeur $\overline{x} = A^{-1}(b + B^*p)$. Par ailleurs $B\overline{x} = 0$ et, comme B est de rang maximal, la matrice $BA^{-1}B^*$ est inversible, ce qui conduit à

$$p = -\left(BA^{-1}B^*\right)^{-1} BA^{-1}b \quad \text{et} \quad \overline{x} = A^{-1}\left(\mathrm{Id} - B^*\left(BA^{-1}B^*\right)^{-1} BA^{-1}\right) b.$$

Notons que le multiplicateur de Lagrange p est unique. Si B n'est pas de rang m, l'Exercice 4.2.8 montre qu'il existe quand même p solution de $BA^{-1}B^*p = -BA^{-1}b$ mais qui n'est unique qu'à l'addition d'un vecteur du noyau de B^* près.

Exercice 4.2.13 Généraliser les résultats ci-dessus pour cette variante de l'Exercice 4.2.8

$$\min_{Bx=c} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

où $c \in \mathbb{R}^m$ est un vecteur donné, A est symétrique définie positive et le rang de B est m.

Exercice 4.2.14 Appliquer le Théorème 4.2.13 à l'Exemple 4.1.4 et en déduire que les points de minimum de J sur la sphère unité sont des vecteurs propres de A associés à la plus petite valeur propre.

Exercice 4.2.15 Soit A une matrice $n \times n$ symétrique définie positive et $b \in \mathbb{R}^n$ non nul.

1. Montrer que les problèmes

$$\sup_{Ax \cdot x \leq 1} b \cdot x \quad \text{et} \quad \sup_{Ax \cdot x = 1} b \cdot x$$

   sont équivalents et qu'ils ont une solution. Utiliser le Théorème 4.2.13 pour calculer cette solution et montrer qu'elle est unique.

2. On introduit un ordre partiel dans l'ensemble des matrices symétriques définies positives d'ordre n en disant que $A \geq B$ si et seulement si $Ax \cdot x \geq Bx \cdot x$ pour tout $x \in \mathbb{R}^n$. Déduire de la question précédente que, si $A \geq B$, alors $B^{-1} \geq A^{-1}$.

Exercice 4.2.16 Montrer que l'entropie de Shannon de l'Exemple 4.1.5 admet un unique point de minimum que l'on calculera. Montrer aussi que, pour tout $p \in \mathbb{R}^n_+$ tel que $\sum_{i=1}^n p_i = 1$,

$$-\sum_{i=1}^n p_i \log p_i = \inf_{q \in \mathbb{R}^n_+, \sum_{i=1}^n q_i = 1} -\sum_{i=1}^n p_i \log q_i.$$

Exercice 4.2.17 En théorie cinétique des gaz les molécules de gaz sont représentées en tout point de l'espace par une fonction de répartition $f(v)$ dépendant de la vitesse microscopique $v \in \mathbb{R}^N$. Les quantités macroscopiques, comme la densité du gaz ρ, sa vitesse u, et sa température T, se retrouvent grâce aux moments de la fonction $f(v)$

$$\rho = \int_{\mathbb{R}^N} f(v)\, dv\ , \quad \rho u = \int_{\mathbb{R}^N} v\, f(v)\, dv\ , \quad \frac{1}{2}\rho u^2 + \frac{N}{2}\rho T = \frac{1}{2}\int_{\mathbb{R}^N} |v|^2 f(v)\, dv\ . \tag{4.25}$$

Boltzmann a introduit l'entropie cinétique $H(f)$ définie par

$$H(f) = \int_{\mathbb{R}^N} f(v) \log\big(f(v)\big)\, dv\ .$$

Montrer que H est strictement convexe sur l'espace des fonctions $f(v) > 0$ mesurables telle que $H(f) < +\infty$. On minimise H sur cet espace sous les contraintes de moment

(4.25), et on admettra qu'il existe un unique point de minimum $M(v)$. Montrer que ce point de minimum est une Maxwellienne définie par

$$M(v) = \frac{\rho}{(2\pi T)^{N/2}} \exp\left(-\frac{|v-u|^2}{2T}\right).$$

Exercice 4.2.18 On se propose de trouver les dimensions optimales, hauteur h et rayon r de la base, d'une casserole cylindrique de volume fixé. En supposant que la quantité de chaleur fournie à la base est la même quel que soit le rayon et que les pertes thermiques, que l'on souhaite minimiser, sont proportionnelles à la surface non chauffée, montrer qu'à l'optimum $h = r$.

Remarque 4.2.15 Pour obtenir un nouvel éclairage sur le Théorème 4.2.13 on introduit la fonction $\mathcal{L}$, appelée **Lagrangien** du problème de minimisation de J sur K, définie sur $V \times \mathbb{R}^M$ par

$$\mathcal{L}(v,\mu) = J(v) + \sum_{i=1}^{M} \mu_i F_i(v) = J(v) + \mu \cdot F(v).$$

Si $u \in K$ est un minimum local de J sur K, le Théorème 4.2.13 nous dit alors que, dans le cas régulier, il existe $\lambda \in \mathbb{R}^M$ tel que

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = 0 \quad , \quad \frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = 0 \ ,$$

puisque $\dfrac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = F(u) = 0$ si $u \in K$ et $\dfrac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda F'(u) = 0$ d'après (4.22). On peut ainsi écrire la contrainte et la condition d'optimalité comme l'annulation du gradient (la stationnarité) du Lagrangien. Remarquons que le Lagrangien permet en quelque sorte d'éliminer les contraintes d'égalité, au prix du rajout de la variable $\mu \in \mathbb{R}^M$, car

$$\sup_{\mu \in \mathbb{R}^M} \mathcal{L}(v,\mu) = \begin{cases} J(v) & \text{si } F(v) = 0, \\ +\infty & \text{si } F(v) \neq 0, \end{cases}$$

et donc

$$\inf_{v \in V,\ F(v)=0} J(v) \ = \ \inf_{v \in V} \sup_{\mu \in \mathbb{R}^M} \mathcal{L}(v,\mu).$$

•

Contraintes d'inégalité

Dans ce deuxième cas on suppose que K est donné par

$$K = \{v \in V \ , \quad F_i(v) \leq 0 \quad \text{pour} \quad 1 \leq i \leq M\} \ , \tag{4.26}$$

où $F_1, \ldots, F_M$ sont des fonctions dérivables de V dans $\mathbb{R}$. Lorsque l'on veut déterminer le cône des directions admissibles $K(v)$, la situation est un peu plus compliquée

que précédemment car toutes les contraintes dans (4.26) ne jouent pas le même rôle selon le point v où l'on calcule $K(v)$. En effet, si $F_i(v) < 0$, il est clair que, pour toute direction $w \in V$ et pour ε suffisamment petit, on aura aussi $F_i(v+\varepsilon w) \leq 0$ (on dit que la contrainte i est inactive en v). Par contre, si $F_i(v) = 0$, il faudra imposer des conditions sur le vecteur $w \in V$ pour que, pour tout $\varepsilon > 0$ suffisamment petit, $F_i(v + \varepsilon w) \leq 0$. Afin que toutes les contraintes dans (4.26) soient satisfaites pour $(v + \varepsilon w)$ il va donc falloir imposer des conditions sur w, appelées **conditions de qualification**. Grosso modo, ces conditions vont garantir que l'on peut faire des "variations" autour d'un point v afin de tester son optimalité. Il existe différents types de conditions de qualification plus ou moins sophistiquées et générales. On s'inspire ici du cas des contraintes d'égalité pour donner une définition simple (mais pas optimale) de qualification des contraintes d'inégalité.

Définition 4.2.16 *Soit $u \in K$. L'ensemble $I(u) = \{i \in \{1, \ldots, M\}\ ,\ F_i(u) = 0\}$ est appelé l'ensemble des contraintes* **actives** *en u.*

Définition 4.2.17 *On dit que les contraintes (4.26) sont* **qualifiées** *en $u \in K$ si la famille*

$$(F_i'(u))_{i \in I(u)} \quad \textit{est libre.} \tag{4.27}$$

Remarque 4.2.18 On peut donner des conditions de qualification plus générales (i.e. plus souvent vérifiées), mais plus compliquées. Par exemple, la condition suivante utilise le problème **linéarisé**. On dit que les contraintes (4.26) sont qualifiées en $u \in K$ s'il existe une direction $\overline{w} \in V$ telle que l'on ait pour tout $i \in I(u)$,

$$\begin{aligned} &\text{ou bien} \quad \langle F_i'(u), \overline{w} \rangle < 0\ , \\ &\text{ou bien} \quad \langle F_i'(u), \overline{w} \rangle = 0 \quad \text{et} \quad F_i \quad \text{est affine}\ . \end{aligned} \tag{4.28}$$

On vérifie que (4.27) entraîne (4.28). La direction $\overline{w}$ est en quelque sorte une "direction rentrante" puisque on déduit de (4.28) que $F(u + \epsilon\overline{w}) \leq 0$ pour tout $\epsilon \geq 0$ suffisamment petit. Un cas particulier très important en pratique est le cas où toutes les contraintes F_i sont **affines** : on peut prendre $\overline{w} = 0$ et les contraintes sont **automatiquement qualifiées**. Le Théorème 4.2.19 ci-dessous reste valide si on remplace la condition de qualification (4.27) par celle, moins restrictive, (4.28) mais, bien sûr, sa démonstration change (voir [1]). •

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (4.26).

Théorème 4.2.19 *On suppose que K est donné par (4.26), que les fonctions J et $F_1, \ldots, F_M$ sont dérivables en u et que les contraintes sont qualifiées en u. Alors, si u est un minimum local de J sur K, il existe $\lambda_1, \ldots, \lambda_M \geq 0$, appelés multiplicateurs de Lagrange, tels que*

$$\begin{aligned} &J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0, \quad \lambda_i \geq 0, \quad F_i(u) \leq 0, \\ &\lambda_i = 0 \quad si \quad F_i(u) < 0 \quad \forall\, i \in \{1, \ldots, M\}. \end{aligned} \tag{4.29}$$

Remarque 4.2.20 On peut réécrire la condition (4.29) sous la forme suivante

$$J'(u) + \lambda \cdot F'(u) = 0 \ , \quad \lambda \geq 0 \ , \quad F(u) \leq 0 \ , \quad \lambda \cdot F(u) = 0 \ ,$$

où $\lambda \geq 0$ signifie que chacune des composantes du vecteur $\lambda = (\lambda_1, ..., \lambda_M)$ est positive (même notation pour $F(u) \leq 0$). En effet, pour tout indice $i \in \{1, \ldots, M\}$, on a soit $F_i(u) = 0$ (la contrainte est active), soit $\lambda_i = 0$ (la contrainte est inactive). Comme chacun des produits $\lambda_i F_i(u)$ est négatif, la condition $\lambda \cdot F(u) = 0$ est bien équivalente à ce que chaque produit $\lambda_i F_i(u) = 0$. La condition $\lambda \cdot F(u) = 0$ est appelée condition des **écarts complémentaires**. Notons que cette condition est de type **combinatoire**, c'est-à-dire qu'il faut, a priori, essayer toutes les combinaisons de contraintes actives pour espérer résoudre (4.29) et trouver le couple (u, λ) optimal (au contraire du cas des contraintes d'égalité qui toutes jouent le même rôle). •

Démonstration. Commençons par caractériser le cône des directions admissibles. Soit $w \in K(u)$ et v^n une des suites correspondantes dans la Définition 4.2.11 de $K(u)$ (rappelons que v^n converge vers u). Puisque $v^n \in K$ on a $F_i(v^n) \leq 0$ pour tout indice $1 \leq i \leq M$. Si la contrainte i est inactive en u, c'est-à-dire $F_i(u) < 0$, alors il n'y a rien à en tirer comme information pour w. Par contre si la contrainte i est active, c'est-à-dire $i \in I(u)$ et $F_i(u) = 0$, alors on effectue un développement de Taylor

$$0 \geq F_i(v^n) = F_i(u + \varepsilon^n w^n) = F_i(u) + \varepsilon^n \langle F_i'(u), w^n \rangle + o(\varepsilon^n)$$

avec la notation $w^n = (v^n - u)/\varepsilon^n$ qui, par définition, converge vers w. Comme $F_i(u) = 0$, on en déduit

$$\langle F_i'(u), w^n \rangle + o(1) \leq 0,$$

qui, lorsque n tend vers l'infini, conduit à

$$\langle F_i'(u), w \rangle \leq 0 \quad \text{pour tout } w \in K(u) \text{ et pour tout } i \in I(u).$$

Par conséquent

$$K(u) \subset K^e(u) = \{w \in V \ , \quad \langle F_i'(u), w \rangle \leq 0 \quad \forall\, i \in I(u)\} \ . \tag{4.30}$$

Réciproquement, soit l'ensemble

$$K^i(u) = \{w \in V \ , \quad \langle F_i'(u), w \rangle < 0 \quad \forall\, i \in I(u)\} \ .$$

Pour tout $w \in K^i(u)$ et toute suite ε^n tendant vers 0, il est facile de voir que la suite $v^n = u + \varepsilon^n w$ vérifie pour n suffisamment grand

$$F_i(v^n) = F_i(u + \varepsilon^n w) = F_i(u) + \varepsilon^n \langle F_i'(u), w \rangle + o(\varepsilon^n) \leq 0,$$

et donc que $w \in K(u)$, suivant la Définition 4.2.11. Au total, on a donc les inclusions

$$K^i(u) \subset K(u) \subset K^e(u),$$

mais comme la fermeture de $K^i(u)$ est $K^e(u)$ et que $K(u)$ est fermé (cf. Exercice 4.2.12), on en déduit l'égalité $K(u) = K^e(u)$ dans (4.30).

On peut maintenant appliquer la Proposition 4.2.12 qui nous dit que

$$\langle J'(u), w\rangle \geq 0 \quad \forall\, w \in K(u).$$

On termine la démonstration grâce au Lemme de Farkas 4.2.21 ci-dessous (avec $p = J'(u)$, $a_i = F_i'(u)$ et $\mathcal{K} = K(u)$) qui permet de conclure que

$$J'(u) + \sum_{i\in I(u)} \lambda_i F_i'(u) = 0, \qquad \text{avec } \lambda_i \geq 0.$$

□

Lemme 4.2.21 (de Farkas) *Soient $a_1, \ldots, a_M$ une famille libre de V. On considère les ensembles*

$$\mathcal{K} = \left\{ w \in V\,, \quad \langle a_i, w\rangle \leq 0 \ \ \textit{pour} \ \ 1 \leq i \leq M \right\},$$

et

$$\hat{\mathcal{K}} = \left\{ q \in V\,, \quad \exists\, \lambda_1, \ldots, \lambda_M \geq 0\,,\ q = -\sum_{i=1}^{M} \lambda_i a_i \right\}.$$

Alors pour tout $p \in V$, on a l'implication

$$\langle p, w\rangle \geq 0\ \forall\, w \in \mathcal{K} \implies p \in \hat{\mathcal{K}}\,. \tag{4.31}$$

(La réciproque étant évidente, il s'agit en fait d'une équivalence.)

Remarque 4.2.22 L'énoncé ci-dessus est un cas particulier (et facile) du Lemme de Farkas qui, en toute généralité, est valable pour **toute** famille de vecteurs $a_1, \ldots, a_M$ dans V, pas forcément libre. Comme nous avons fait l'hypothèse de qualification des contraintes au sens de la Définition 4.2.17, il n'est pas nécessaire d'utiliser la version générale du Lemme de Farkas, dont la démonstration est plus subtile (voir [1], [4]). •

Remarque 4.2.23 L'ensemble $\{p \in V$ tel que $\langle p, w\rangle \geq 0\ \forall\, w \in \mathcal{K}\}$ est appelé le cône dual de $\mathcal{K}$. Le Lemme de Farkas affirme donc que le cône dual de $\mathcal{K}$ est $\hat{\mathcal{K}}$. Si les inégalités dans la définition de $\mathcal{K}$ et dans celle du cône dual étaient remplacés par des égalités et si le signe des coefficients λ_i dans la définition de $\hat{\mathcal{K}}$ est quelconque, alors le Lemme de Farkas se réduit à dire que l'orthogonal de $\mathcal{K}$ est simplement $\hat{\mathcal{K}}$, le sous-espace vectoriel engendré par les vecteurs $a_1, \ldots, a_M$. Evidemment le cas avec inégalités est plus technique. •

Démonstration. Caractérisons d'abord le cône $\mathcal{K}$. Soit A le sous-espace vectoriel de V engendré par la famille libre $a_1, \ldots, a_M$ et soit $A^\perp$ son orthogonal dans V. Tout vecteur $w \in V$ peut se décomposer de manière unique en

$$w = w_1 + w_2 \quad \text{avec } w_1 \in A, w_2 \in A^\perp.$$

Par conséquent, $w \in \mathcal{K}$ si et seulement si $w_1 \in \mathcal{K}$ et il n'y a aucune restriction au choix de w_2. Comme $w_1 \in A$, il existe des réels $\mu_1, \dots, \mu_M$ tels que

$$w_1 = -\sum_{i=1}^{M} \mu_i a_i.$$

La condition $w_1 \in \mathcal{K}$ est équivalente à

$$\sum_{i=1}^{M} \mu_i \langle a_i, a_j \rangle \geq 0 \quad \text{pour tout } 1 \leq j \leq M,$$

c'est-à-dire, en introduisant la matrice symétrique $M \times M$ définie par ses éléments $D_{ij} = \langle a_i, a_j \rangle$, et le vecteur $\mu = (\mu_1, \dots, \mu_M)$, $D\mu \geq 0$. Or la matrice D est inversible car définie positive puisque la famille $a_1, \dots, a_M$ est libre. On en déduit qu'il suffit de prendre $\mu = D^{-1}q$ avec $q \in (\mathbb{R}_+)^M$. Autrement dit, on a caractérisé

$$\mathcal{K} = \left\{ w \in V \ \text{ tel que } w_1 = -\sum_{i=1}^{M} \mu_i a_i \text{ et } \mu = D^{-1}q \,,\, q \in (\mathbb{R}_+)^M \right\}.$$

Etudions maintenant l'implication (4.31). Lorsque l'on prend $w_1 = 0$ et successivement w_2 et $-w_2$ quelconque dans la condition $\langle p, w \rangle \geq 0$, $\forall\, w \in \mathcal{K}$, on obtient que $p \in A$. Donc, il existe des réels $\lambda_1, \dots, \lambda_M$ (sans signe prescrit pour l'instant) tels que

$$p = -\sum_{i=1}^{M} \lambda_i a_i.$$

Enfin, pour $w_1 \in \mathcal{K}$ et $w_2 = 0$ la condition $\langle p, w \rangle \geq 0$, $\forall\, w \in \mathcal{K}$ conduit à

$$\sum_{i,j=1}^{M} \lambda_i \mu_j \langle a_i, a_j \rangle = (D\mu) \cdot \lambda \geq 0 \text{ avec } \mu = D^{-1}q \,,\, q \in (\mathbb{R}_+)^M.$$

Autrement dit, $q \cdot \lambda \geq 0$ pour tout $q \in (\mathbb{R}_+)^M$. On en déduit que nécessairement $\lambda \in (\mathbb{R}_+)^M$, c'est-à-dire que $p \in \hat{\mathcal{K}}$. □

Remarque 4.2.24 Pour obtenir un nouvel éclairage sur le Théorème 4.2.19 on introduit la fonction $\mathcal{L}$, appelée **Lagrangien** du problème de minimisation de J sur K, définie sur $V \times (\mathbb{R}^+)^M$ par

$$\mathcal{L}(v, \mu) = J(v) + \sum_{i=1}^{M} \mu_i F_i(v) = J(v) + \mu \cdot F(v) \qquad \forall (v, \mu) \in V \times (\mathbb{R}^+)^M.$$

La nouvelle variable **positive** $\mu \in (\mathbb{R}^+)^M$ est appelée **multiplicateur de Lagrange** pour la contrainte $F(v) \leq 0$. La maximisation du Lagrangien en $\mu \in (\mathbb{R}^+)^M$ permet de faire "disparaître" la contrainte. En effet, on vérifie facilement que

$$\inf_{v \in V,\ F(v) \leq 0} J(v) = \inf_{v \in V} \sup_{\mu \in (\mathbb{R}^+)^M} \mathcal{L}(v, \mu), \tag{4.32}$$

car

$$\sup_{\mu\in(\mathbb{R}^+)^M} \mathcal{L}(v,\mu) = \begin{cases} J(v) & \text{si } F(v)\le 0, \\ +\infty & \text{sinon.} \end{cases}$$

La condition nécessaire d'optimalité (4.29) est alors équivalente à la stationnarité du Lagrangien puisque

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda \cdot F'(u) = 0,$$

et que la condition $\lambda \ge 0$, $F(u) \le 0$, $\lambda \cdot F(u) = 0$ est équivalente à l'inéquation d'Euler (4.17) pour la maximisation par rapport à μ dans le convexe fermé $(\mathbb{R}^+)^M$

$$\frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda)\cdot(\mu-\lambda) = F(u)\cdot(\mu-\lambda) \le 0 \quad \forall \mu \in (\mathbb{R}^+)^M.$$

•

Exercice 4.2.19 Soit A une matrice symétrique définie positive d'ordre n, et B une matrice de taille $m \times n$ avec $m \le n$ et de rang m. On considère le problème de minimisation

$$\min_{x\in\mathbb{R}^n,\ Bx\le c} \left\{ J(x) = \frac{1}{2}Ax\cdot x - b\cdot x \right\},$$

Appliquer le Théorème 4.2.19 pour obtenir l'existence d'un multiplicateur de Lagrange $p \in \mathbb{R}^m$ tel qu'un point de minimum $\overline{x}$ vérifie

$$A\overline{x} - b + B^*p = 0\ , \quad p \ge 0\ , \quad p\cdot(B\overline{x} - c) = 0.$$

Exercice 4.2.20 Soit f une fonction définie sur un ouvert borné Ω. Pour $\epsilon > 0$ on considère le problème de régularisation suivant

$$\inf_{u\in V_0\ ,\ \int_\Omega |u-f|^2dx \le \epsilon^2} \int_\Omega |\nabla u|^2 dx,$$

où $V_0 = \{v \in C^1(\overline{\Omega}),\ v = 0 \text{ sur } \partial\Omega\}$. Montrer qu'un point de minimum u_ϵ vérifie, soit $u_\epsilon = f$, soit il existe $\lambda \ge 0$ tel que u_ϵ est solution de

$$\begin{cases} -\Delta u_\epsilon + \lambda(u_\epsilon - f) = 0 & \text{dans } \Omega, \\ u_\epsilon = 0 & \text{sur } \partial\Omega. \end{cases}$$

4.3 Point-selle, théorème de Kuhn et Tucker, dualité

Nous avons vu aux Remarques 4.2.15 et 4.2.24 comment il est possible d'interpréter le couple (u,λ) (point de minimum, multiplicateur de Lagrange) comme **point stationnaire d'un Lagrangien** $\mathcal{L}$. Nous allons dans cette section préciser la nature de ce point stationnaire comme **point-selle** et montrer comment cette

formulation permet de caractériser un minimum (ce qui veut dire que, sous certaines hypothèses, nous verrons que les conditions **nécessaires** de stationnarité du Lagrangien sont aussi **suffisantes**). Nous explorerons brièvement la **théorie de la dualité** qui en découle. Outre l'intérêt théorique de cette caractérisation, son intérêt pratique du point de vue des algorithmes numériques sera illustré dans la Section 4.4.

4.3.1 Point-selle

De manière abstraite, V et Q étant deux espaces vectoriels munis de produit scalaire, un Lagrangien $\mathcal{L}$ est une application de $V \times Q$ (ou d'une partie $U \times P$ de $V \times Q$) dans $\mathbb{R}$. Dans le cadre des contraintes d'égalité (Remarque 4.2.15) nous avions $U = V$, $P = Q = \mathbb{R}^M$ et $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$. En ce qui concerne les contraintes d'inégalité (Remarque 4.2.24) le Lagrangien était le même $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ mais il était essentiel de se limiter aux valeurs positives des multiplicateurs de Lagrange, $U = V$, $Q = \mathbb{R}^M$ et $P = (\mathbb{R}_+)^M$.

Donnons maintenant la définition d'un point-selle, appelé également min-max ou col.

Définition 4.3.1 *On dit que $(u, p) \in U \times P$ est un point-selle de $\mathcal{L}$ sur $U \times P$ si*

$$\forall\, q \in P \qquad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in U\ . \tag{4.33}$$

Autrement dit, (4.33) dit que $v \to \mathcal{L}(v, p)$ est minimum en u à p fixé, tandis que $q \to \mathcal{L}(u, q)$ est maximum en q à u fixé (d'où le nom de min-max).

Le résultat suivant montre le lien entre cette notion de point-selle et les problèmes de minimisation avec contraintes d'égalité (4.21) ou contraintes d'inégalité (4.26) étudiés dans la section précédente. Pour simplifier, nous utiliserons de nouveau des inégalités entre vecteurs, notant parfois $q \geq 0$ au lieu de $q \in (\mathbb{R}_+)^M$.

Proposition 4.3.2 *On suppose que les fonctions $J, F_1, \ldots, F_M$ sont continues sur V, et que l'ensemble K est défini par (4.21) ou (4.26). On note $P = \mathbb{R}^M$ dans le cas de contraintes d'égalité (4.21) et $P = (\mathbb{R}_+)^M$ dans le cas de contraintes d'inégalité (4.26). Soit U un ouvert de V contenant K. Pour $(v, q) \in U \times P$, on considère*

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v).$$

Supposons que (u, p) soit un point-selle de $\mathcal{L}$ sur $U \times P$. Alors $u \in K$ et u est un minimum global de J sur K. De plus, si J et $F_1, \ldots, F_M$ sont dérivables en u, on a

$$J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0\ . \tag{4.34}$$

Démonstration. Écrivons la condition de point-selle

$$\forall\, q \in P \quad J(u) + q \cdot F(u) \leq J(u) + p \cdot F(u) \leq J(v) + p \cdot F(v) \quad \forall\, v \in U\ . \tag{4.35}$$

Examinons d'abord le cas de contraintes d'égalité. Puisque $P = \mathbb{R}^M$, la première inégalité dans (4.35) montre que $F(u) = 0$, i.e. $u \in K$. Il reste alors $J(u) \leq J(v) + p \cdot F(v) \quad \forall\, v \in U$, qui montre bien (en prenant $v \in K$) que u est un minimum global de J sur K.

Dans le cas de contraintes d'inégalité, on a $P = (\mathbb{R}_+)^M$ et la première inégalité de (4.35) montre maintenant que $F(u) \leq 0$ et que $p \cdot F(u) = 0$. Ceci prouve encore que $u \in K$, et permet de déduire facilement de la deuxième inégalité que u est un minimum global de J sur K.

Enfin, si J et $F_1, \ldots, F_M$ sont dérivables en u, la deuxième inégalité de (4.35) montre que u est un point de minimum sans contrainte de $J + p \cdot F$ dans l'ouvert U, ce qui implique que la dérivée s'annule en u, $J'(u) + p \cdot F'(u) = 0$ (cf. la Remarque 4.2.9). □

4.3.2 Théorème de Kuhn et Tucker

Nous revenons au problème de minimisation sous contraintes d'inégalité pour lequel l'ensemble K est donné par (4.26), c'est-à-dire

$$K = \{v \in V \ , \quad F_i(v) \leq 0 \quad \text{pour} \quad 1 \leq i \leq m\} \ . \tag{4.36}$$

Le Théorème 4.2.19 a donné une condition nécessaire d'optimalité. Dans cette sous-section nous allons voir que cette condition est aussi **suffisante** si les contraintes et la fonction coût sont **convexes**. En effet, la Proposition 4.3.2 affirme que, si (u, p) est un point-selle du Lagrangien, alors u réalise le minimum de J sur K. Pour un problème de minimisation convexe avec des contraintes d'inégalités convexes, nous allons établir une réciproque de ce résultat, c'est-à-dire que, si u réalise le minimum de J sur K, alors il existe $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit point-selle du Lagrangien. On suppose désormais que $J, F_1, \ldots, F_M$ sont convexes continues sur V.

Le théorème de Kuhn et Tucker (appelé aussi parfois théorème de Karush, Kuhn et Tucker) affirme que, dans le cas convexe, la condition nécessaire d'optimalité du Théorème 4.2.19 est en fait une condition **nécessaire et suffisante**.

Théorème 4.3.3 (de Kuhn et Tucker) *On suppose que les fonctions J, $F_1, \ldots, F_M$ sont convexes continues sur V et dérivables sur l'ensemble K (4.36). On introduit le Lagrangien $\mathcal{L}$ associé*

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \quad \forall\, (v, q) \in V \times (\mathbb{R}_+)^M \ .$$

Soit $u \in K$ un point de K où les contraintes sont qualifiées au sens de la Définition 4.2.17. Alors u est un minimum global de J sur K si et seulement si il existe $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit un point-selle du Lagrangien $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$ ou, de manière équivalente, tel que

$$F(u) \leq 0 \ , \ p \geq 0 \ , \ p \cdot F(u) = 0 \ , \ J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0 \ . \tag{4.37}$$

Démonstration. Si (u,p) est point-selle du Lagrangien, on a déjà montré à la Proposition 4.3.2 que u est un minimum global de J sur K. Réciproquement, si u est un minimum de J sur K, alors on peut appliquer le Théorème 4.2.19, qui donne exactement la condition (nécessaire) d'optimalité (4.37). Des trois premières conditions de (4.37) on déduit que, pour tout $q \geq 0$,

$$J(u) + q \cdot F(u) \leq J(u) + p \cdot F(u).$$

D'autre part, la fonction $v \to J(v) + p \cdot F(v)$ est convexe comme somme de fonctions convexes (avec poids p positif). La dernière condition de (4.37) dit que la dérivée de cette fonction convexe s'annule en u. Par application du Corollaire 4.2.6, on en déduit que u est un point de minimum

$$J(u) + p \cdot F(u) \leq J(v) + p \cdot F(v) \quad \forall\, v \in V,$$

autrement dit, (u,p) est point-selle de $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$. □

Remarque 4.3.4 Le Théorème 4.3.3 de Kuhn et Tucker ne s'applique qu'aux contraintes d'inégalité, et pas aux contraintes d'égalité, en général. Cependant, il est bon de remarquer que des contraintes **d'égalité affines** $Av = b$ peuvent s'écrire sous la forme de contraintes d'inégalité (affines donc convexes) $Av - b \leq 0$ et $b - Av \leq 0$. C'est une évidence qui permet cependant d'appliquer le Théorème 4.3.3 de Kuhn et Tucker à un problème de minimisation avec contraintes d'égalité affines. Dans ce cas il faut utiliser la notion de qualification des contraintes, un peu plus compliquée, donnée dans la Remarque 4.2.18. Notons aussi que les multiplicateurs de Lagrange pour ces contraintes d'égalité affines n'ont alors pas de signe imposé. •

L'exercice suivant permet d'interpréter les multiplicateurs de Lagrange p_i comme la sensibilité de la valeur minimale de J aux variations des contraintes F_i : en économie, ces coefficients mesurent des prix ou des coûts marginaux , en mécanique des forces de liaison correspondant à des contraintes cinématiques, etc...

Exercice 4.3.1 On considère le problème d'optimisation, dit perturbé

$$\inf_{F_i(v) \leq u_i,\ 1 \leq i \leq m} J(v), \tag{4.38}$$

avec $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$. On se place sous les hypothèses du Théorème 4.3.3 de Kuhn et Tucker. On note $m^*(u)$ la valeur minimale du problème perturbé (4.38).

1. Montrer que si $p \in \mathbb{R}^m$ est le multiplicateur de Lagrange pour le problème non perturbé (c'est-à-dire (4.38) avec $u = 0$), alors

$$m^*(u) \geq m^*(0) - p \cdot u \ . \tag{4.39}$$

2. Déduire de (4.39) que si $u \mapsto m^*(u)$ est dérivable, alors

$$p_i = -\frac{\partial m^*}{\partial u_i}(0).$$

Interpréter ce résultat (cf. l'Exemple 4.1.2 en économie).

4.3.3 Dualité

Donnons un bref aperçu de la théorie de la dualité pour les problèmes d'optimisation. Nous l'appliquerons au problème de minimisation convexe avec contraintes d'inégalité de la sous-section précédente. Nous avons associé à ce problème de minimisation un problème de recherche d'un point-selle (u, p) pour le Lagrangien $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$. Mais nous allons voir que, à l'existence d'un point-selle (u, p) du Lagrangien, on peut associer inversement non pas un mais **deux** problèmes d'optimisation (plus précisément, un problème de minimisation et un problème de maximisation), qui seront dits **duaux** l'un de l'autre. Nous expliquerons ensuite sur deux exemples simples en quoi l'introduction du **problème dual** peut être utile pour la résolution du problème d'origine, dit **problème primal** (par opposition au dual).

Revenons un instant au cadre général de la Définition 4.3.1 où V et Q sont deux espaces vectoriels munis d'un produit scalaire.

Définition 4.3.5 *Soit $\mathcal{L}$ un Lagrangien défini sur une partie $U \times P$ de $V \times Q$. On suppose qu'il existe un point-selle (u, p) de $\mathcal{L}$ sur $U \times P$*

$$\forall\, q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in U\ . \tag{4.40}$$

Pour $v \in U$ et $q \in P$, posons

$$\mathcal{J}(v) = \sup_{q \in P} \mathcal{L}(v, q) \qquad \mathcal{G}(q) = \inf_{v \in U} \mathcal{L}(v, q)\ . \tag{4.41}$$

On appelle problème primal le problème de minimisation

$$\inf_{v \in U} \mathcal{J}(v)\ , \tag{4.42}$$

et problème dual le problème de maximisation

$$\sup_{q \in P} \mathcal{G}(q)\ . \tag{4.43}$$

Remarque 4.3.6 Bien sûr, sans hypothèses supplémentaires, il peut arriver que $\mathcal{J}(v) = +\infty$ pour certaines valeurs de v ou que $\mathcal{G}(q) = -\infty$ pour certaines valeurs de q. Mais l'existence supposée du point-selle (u, p) dans la Définition 4.3.5 nous assure que les **domaines** de $\mathcal{J}$ et $\mathcal{G}$ (i.e. les ensembles $\{v \in U\ ,\ \mathcal{J}(v) < +\infty\}$ et $\{q \in P\ ,\ \mathcal{G}(q) > -\infty\}$ sur lesquels ces fonctions sont à valeurs finies) ne sont pas vides, puisque (4.40) montre que $\mathcal{J}(u) = \mathcal{G}(p) = \mathcal{L}(u, p)$. Les problèmes primal et dual ont donc bien un sens. Le résultat suivant montre que ces deux problèmes sont étroitement liés au point-selle (u, p). •

Théorème 4.3.7 (de dualité) *Le couple (u, p) est un point-selle de $\mathcal{L}$ sur $U \times P$ si et seulement si*

$$\mathcal{J}(u) = \min_{v \in U} \mathcal{J}(v) = \max_{q \in P} \mathcal{G}(q) = \mathcal{G}(p)\ . \tag{4.44}$$

Remarque 4.3.8 Par la Définition (4.41) de $\mathcal{J}$ et $\mathcal{G}$, (4.44) est équivalent à

$$\mathcal{J}(u) = \min_{v\in U}\left(\sup_{q\in P}\mathcal{L}(v,q)\right) = \max_{q\in P}\left(\inf_{v\in U}\mathcal{L}(v,q)\right) = \mathcal{G}(p)\ . \tag{4.45}$$

Si le sup et l'inf sont atteints dans (4.45) (c'est-à-dire qu'on peut les écrire max et min, respectivement), on voit alors que (4.45) traduit la possibilité d'échanger l'ordre du min et du max appliqués au Lagrangien $\mathcal{L}$. Ce fait (qui est faux si $\mathcal{L}$ n'admet pas de point selle) explique le nom de min-max qui est souvent donné à un point-selle. •

Démonstration. Soit (u,p) un point-selle de $\mathcal{L}$ sur $U\times P$. Notons $\mathcal{L}^* = \mathcal{L}(u,p)$. Pour $v\in U$, il est clair d'après (4.41) que $\mathcal{J}(v)\geq\mathcal{L}(v,p)$, d'où $\mathcal{J}(v)\geq\mathcal{L}^*$ d'après (4.40). Comme $\mathcal{J}(u)=\mathcal{L}^*$, ceci montre que $\mathcal{J}(u)=\inf_{v\in U}\mathcal{J}(v)=\mathcal{L}^*$. On montre de la même façon que $\mathcal{G}(p)=\sup_{q\in P}\mathcal{G}(q)=\mathcal{L}^*$.

Réciproquement, supposons que (4.44) a lieu et posons $\mathcal{L}^*=\mathcal{J}(u)$. La définition (4.41) de $\mathcal{J}$ montre que

$$\mathcal{L}(u,q)\leq\mathcal{J}(u)=\mathcal{L}^*\quad\forall\, q\in P\ . \tag{4.46}$$

De même, on a aussi :

$$\mathcal{L}(v,p)\geq\mathcal{G}(p)=\mathcal{L}^*\quad\forall\, v\in U\ , \tag{4.47}$$

et on déduit facilement de (4.46)-(4.47) que $\mathcal{L}(u,p)=\mathcal{L}^*$, ce qui montre que (u,p) est point-selle. □

Remarque 4.3.9 Même si le Lagrangien $\mathcal{L}$ n'admet pas de point selle sur $U\times P$, on a tout de même l'inégalité élémentaire suivante, dite de **dualité faible**

$$\inf_{v\in U}\left(\sup_{q\in P}\mathcal{L}(v,q)\right)\geq\sup_{q\in P}\left(\inf_{v\in U}\mathcal{L}(v,q)\right). \tag{4.48}$$

En effet, pour tout $v\in U$ et $q\in P$, $\mathcal{L}(v,q)\geq\inf_{v'\in U}\mathcal{L}(v',q)$, donc $\sup_{q\in P}\mathcal{L}(v,q)\geq\sup_{q\in P}\inf_{v'\in U}\mathcal{L}(v',q)$, et puisque ceci est vrai pour tout $v\in V$, $\inf_{v\in V}\sup_{q\in P}\mathcal{L}(v,q)\geq\sup_{q\in P}\inf_{v'\in U}\mathcal{L}(v',q)$, ce qui donne (4.48). La différence (positive) entre les deux membres de l'inégalité (4.48) est appelée **saut de dualité**. •

Exercice 4.3.2 Soit le Lagrangien $\mathcal{L}$ défini pour $(v,q)\in\mathbb{R}^N\times\mathbb{R}^N$ par

$$\mathcal{L}(v,q)=\frac{1}{2}Av\cdot v - v\cdot q,$$

où A est une matrice symétrique définie positive. Calculer la fonction duale $\mathcal{G}(q)$.

Exercice 4.3.3 Donner un exemple de Lagrangien pour lequel l'inégalité (4.48) est stricte avec ses deux membres finis.

Exercice 4.3.4 Soit U (respectivement P) un convexe compact non vide de V (respectivement Q). On suppose que le Lagrangien est tel que $v\to\mathcal{L}(v,q)$ est convexe sur U pour tout $q\in P$, et $q\to\mathcal{L}(v,q)$ est concave sur P pour tout $v\in U$. Montrer alors l'existence d'un point selle de $\mathcal{L}$ sur $U\times P$.

Application

Nous appliquons ce résultat de dualité au problème précédent de minimisation convexe avec contraintes d'inégalité convexes

$$\inf_{v\in V\ ,\ F(v)\leq 0} J(v) \tag{4.49}$$

avec J et $F=(F_1,\ldots,F_M)$ convexes sur V. On introduit le Lagrangien

$$\mathcal{L}(v,q)=J(v)+q\cdot F(v)\quad \forall\,(v,q)\in V\times(\mathbb{R}_+)^M\ .$$

Dans ce cadre, on voit facilement que, pour tout $v\in V$,

$$\mathcal{J}(v)=\sup_{q\in(\mathbb{R}_+)^M}\mathcal{L}(v,q)=\begin{cases} J(v) & \text{si } F(v)\leq 0\\ +\infty & \text{sinon ,}\end{cases} \tag{4.50}$$

ce qui montre que le problème primal $\inf_{v\in V}\mathcal{J}(v)$ est exactement le problème d'origine (4.49) ! D'autre part, la fonction $\mathcal{G}(q)$ du problème dual est bien définie par (4.41), car (4.41) est ici un problème de minimisation convexe. De plus, $\mathcal{G}(q)$ est une fonction concave car elle est l'infimum de fonctions affines (voir l'Exercice 4.1.4). Par conséquent, le problème dual

$$\sup_{q\in(\mathbb{R}_+)^M}\mathcal{G}(q)\ ,$$

est un problème de maximisation concave **plus simple** que le problème primal (4.49) car les contraintes sont linéaires ! Cette particularité est notamment exploitée dans des algorithmes numériques (cf. l'algorithme d'Uzawa). Une simple combinaison des Théorèmes de Kuhn et Tucker 4.3.3 et de dualité 4.3.7 nous donne le résultat suivant.

Corollaire 4.3.10 *On suppose que les fonctions $J,F_1,\ldots,F_M$ sont convexes et dérivables sur V. Soit $u\in V$ tel que $F(u)\leq 0$ et les contraintes sont qualifiées en u au sens de la Définition 4.2.17. Alors, si u est un minimum global de $\mathcal{J}$ sur V, il existe $p\in(\mathbb{R}_+)^M$ tel que*

1. *p est un maximum global de $\mathcal{G}$ sur $(\mathbb{R}_+)^M$,*
2. *(u,p) est un point-selle du Lagrangien $\mathcal{L}$ sur $V\times(\mathbb{R}_+)^M$,*
3. *$(u,p)\in V\times(\mathbb{R}_+)^M$ vérifie la condition d'optimalité nécessaire et suffisante*

$$F(u)\leq 0\ ,\ p\geq 0\ ,\ p\cdot F(u)=0\ ,\ J'(u)+p\cdot F'(u)=0\ . \tag{4.51}$$

L'application la plus courante du Corollaire 4.3.10 est la suivante. Supposons que le problème dual de maximisation est plus facile à résoudre que le problème primal (c'est le cas en général car ses contraintes sont plus simples). Alors pour calculer la solution u du problème primal on procède en deux étapes. Premièrement, on calcule la solution p du problème dual. Deuxièmement, on dit que (u,p) est un

point selle du Lagrangien, c'est-à-dire que l'on calcule u, solution du problème de minimisation **sans contrainte**

$$\min_{v\in V} \mathcal{L}(v,p) \ .$$

Précisons qu'avec les hypothèses faites il n'y a pas a priori d'existence, ni d'unicité, des solutions pour tous ces problèmes.

Remarque 4.3.11 Pour illustrer le Corollaire 4.3.10 et l'intérêt de la dualité, nous considérons un problème de minimisation quadratique dans $\mathbb{R}^N$ avec contraintes d'inégalité affines

$$\min_{v\in\mathbb{R}^N,\ F(v)=Bv-c\leq 0} \left\{ J(v) = \frac{1}{2}Av\cdot v - b\cdot v \right\} \ , \tag{4.52}$$

où A est une matrice $N\times N$ symétrique définie positive, $b\in\mathbb{R}^N$, B une matrice $M\times N$ et $c\in\mathbb{R}^M$. Le Lagrangien est donné par

$$\mathcal{L}(v,q) = \frac{1}{2}Av\cdot v - b\cdot v + q\cdot(Bv-c) \quad \forall\,(v,q)\in\mathbb{R}^N\times(\mathbb{R}_+)^M \ . \tag{4.53}$$

Nous avons déjà fait dans (4.50) le calcul de $\mathcal{J}$, et dit que le problème primal est exactement (4.52). Examinons maintenant le problème dual. Pour $q\in(\mathbb{R}_+)^M$, le problème

$$\min_{v\in\mathbb{R}^N} \mathcal{L}(v,q)$$

a une solution unique puisque $v\to\mathcal{L}(v,q)$ est une fonction continue, strictement convexe et infinie à l'infini. Cette solution vérifie $\frac{\partial\mathcal{L}}{\partial v}(v,q) = Av - b + B^*q = 0$, soit $v = A^{-1}(b-B^*q)$. On obtient donc

$$\mathcal{G}(q) = \mathcal{L}\big(A^{-1}(b-B^*q),q\big) \ ,$$

et le problème dual s'écrit finalement

$$\sup_{q\geq 0}\left(-\frac{1}{2}q\cdot BA^{-1}B^*q + (BA^{-1}b-c)\cdot q - \frac{1}{2}A^{-1}b\cdot b\right) \ . \tag{4.54}$$

Certes, la fonctionnelle à maximiser dans (4.54) n'a pas une allure particulièrement sympathique. Il s'agit encore d'un problème avec fonctionnelle quadratique et contraintes affines. Cependant, le Corollaire 4.3.10 nous assure qu'il a une solution. On peut voir d'ailleurs que cette solution n'est pas forcément unique (sauf si la matrice B est de rang M car la matrice $BA^{-1}B^*$ est alors définie positive). Mais l'avantage important du problème dual (4.54) vient du fait que les contraintes $(q\geq 0)$ s'expriment sous une forme particulièrement simple, bien plus simple que pour le problème primal ; et nous verrons à la Sous-section 4.4.3 que cet avantage peut être utilisé pour mettre au point un algorithme de calcul de la solution du problème primal. •

Exercice 4.3.5 Nous reprenons l'Exemple 4.1.7 de minimisation de l'énergie complémentaire. On considère le problème de minimisation sous contrainte

$$\inf_{-\mathrm{div}\tau=f \text{ dans } \Omega} \left\{ G(\tau) = \frac{1}{2}\int_\Omega |\tau|^2 dx \right\}. \tag{4.55}$$

Pour $\tau : \Omega \to \mathbb{R}^N$ et $v : \Omega \to \mathbb{R}$, on introduit le Lagrangien correspondant

$$\mathcal{L}(\tau, v) = \frac{1}{2}\int_\Omega |\tau|^2 dx + \int_\Omega v \cdot (\mathrm{div}\tau + f) dx.$$

Montrer que la fonction duale $\mathcal{D}(v)$ correspondante n'est rien d'autre que (l'opposée de) l'énergie $-J(v)$ de l'Exemple 4.1.6. En admettant que $-J(v)$ admette un point de maximum u dans V_0 et que (4.55) admette un point de minimum σ, montrer que (σ, u) est un point selle du Lagrangien et que $\sigma = \nabla u$.

4.3.4 Vers la programmation linéaire

Au regard de leur importance en pratique (cf. l'exemple 4.1.1) nous consacrons cette sous-section au cas particulier où la fonction objectif et les contraintes sont affines. Dans ce cas les problèmes d'optimisation s'appellent des **programmes linéaires**. Comme les fonctions affines sont convexes, l'ensemble des résultats précédents s'appliquent mais des résultats et des méthodes spécifiques permettent d'étudier ces programmes linéaires. Nous ne donnons ici qu'un très bref aperçu de ce domaine qu'est la programmation linéaire et nous renvoyons aux très nombreux ouvrages qui lui sont consacrés pour plus de détails (voir [6], [8], [12]).

Considérons le problème suivant, dit **programme linéaire sous forme standard**,

$$\inf_{x\in\mathbb{R}^n \text{ tel que } Ax=b,\ x\geq 0} c \cdot x, \tag{4.56}$$

où A est une matrice de taille $m \times n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, et la contrainte $x \geq 0$ signifie que toutes les composantes de x sont positives ou nulles. Dans tout ce qui suit on supposera que $m \leq n$ et que le rang de A est exactement m. En effet, si $\mathrm{rg}(A) < m$, certaines lignes de A sont liées et deux possibilités se présentent : soit les contraintes (correspondantes à ces lignes) sont incompatibles, soit elles sont redondantes et on peut donc éliminer les lignes inutiles.

Le problème (4.56) semble être un cas particulier de programme linéaire puisque les contraintes d'inégalités sont seulement du type $x \geq 0$. Il n'en est rien, et tout programme linéaire du type

$$\inf_{x\in\mathbb{R}^n \text{ tel que } Ax\geq b,\ A'x=b'} c \cdot x.$$

peut se mettre sous la forme standard (4.56) quitte à changer la taille des données. En effet, remarquons tout d'abord que les contraintes d'égalité $A'x = b'$ sont évidemment équivalentes aux contraintes d'inégalité $A'x \leq b'$ et $A'x \geq b'$. On peut donc se restreindre au cas suivant (qui ne contient que des contraintes d'inégalité)

$$\inf_{x\in\mathbb{R}^n \text{ tel que } Ax\geq b} c \cdot x. \tag{4.57}$$

Dans (4.57) on peut remplacer la contrainte d'inégalité en introduisant de nouvelles variables, dites **d'écarts**, $\lambda \in \mathbb{R}^m$. La contrainte d'inégalité $Ax \geq b$ est alors équivalente à $Ax = b + \lambda$ avec $\lambda \geq 0$. Ainsi (4.57) est équivalent à

$$\inf_{(x,\lambda)\in\mathbb{R}^{(n+m)} \text{ tel que } Ax=b+\lambda,\ \lambda\geq 0} c \cdot x. \tag{4.58}$$

Finalement, si on décompose chaque composante de x en partie positive et négative, c'est-à-dire si on pose $x = x^+ - x^-$ avec $x^+ = \max(0, x)$ et $x^- = -\min(0, x)$, on obtient que (4.57) est équivalent à

$$\inf_{(x^+,x^-,\lambda)\in\mathbb{R}^{(2n+m)} \text{ tel que } Ax^+-Ax^-=b+\lambda,\ x^+\geq 0, x^-\geq 0, \lambda\geq 0} c \cdot (x^+ - x^-). \tag{4.59}$$

qui est bien sous forme standard (mais avec plus de variables). Il n'y a donc aucune perte de généralité à étudier le programme linéaire standard (4.56).

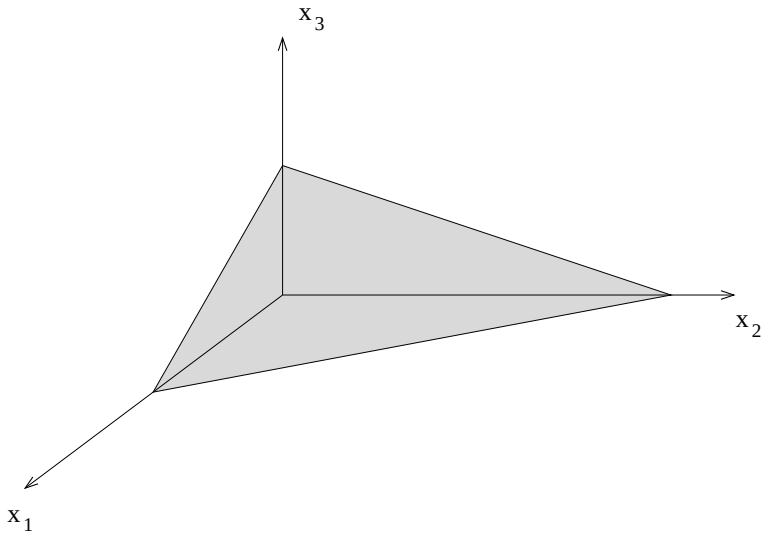

FIGURE 4.2 – Ensemble admissible pour l'exemple (4.60).

Considérons pour l'instant un exemple simple qui va nous permettre de comprendre quelques aspects essentiels d'un programme linéaire

$$\min_{\substack{x_1\geq 0, x_2\geq 0, x_3\geq 0 \\ 2x_1+x_2+3x_3=6}} x_1 + 4x_2 + 2x_3 \ . \tag{4.60}$$

Sur la Figure 4.2 nous avons tracé l'ensemble des (x_1, x_2, x_3) qui vérifient les contraintes : un triangle plan T. C'est un fermé compact de $\mathbb{R}^3$, donc la fonction continue $x_1 + 4x_2 + 2x_3$ y atteint son minimum que l'on note M. Pour déterminer ce minimum on peut considérer la famille de plans parallèles $x_1 + 4x_2 + 2x_3 = c$ paramétrée par c. En augmentant la valeur de c à partir de $-\infty$, on "balaie" l'espace $\mathbb{R}^3$ jusqu'à atteindre le triangle T, et le minimum M est obtenu lorsque le plan "touche" ce triangle. Autrement dit, tout point de minimum de (4.60) est sur le bord du triangle T. Une autre façon de le voir est de dire que la fonction $x_1 + 4x_2 + 2x_3$ a un gradient

non nul dans T donc ses extréma se trouvent sur le bord de T. Pour l'exemple (4.60) le point de minimum (unique) est le sommet $(0, 3, 0)$ de T. Nous verrons qu'il s'agit d'un fait général : un point de minimum (s'il existe) peut toujours se trouver en un des sommets de l'ensemble géométrique des vecteurs x qui vérifient les contraintes. Il "suffit" alors d'énumérer tous les sommets afin de trouver le minimum.

Pour établir cette propriété en toute généralité pour le programme linéaire standard (4.56), nous avons besoin de quelques définitions qui permettent de préciser le vocabulaire.

Définition 4.3.12 *L'ensemble X_{ad} des vecteurs de $\mathbb{R}^n$ qui satisfont les contraintes de (4.56), c'est-à-dire*

$$X_{ad} = \{x \in \mathbb{R}^n \text{ tel que } Ax = b,\ x \geq 0\},$$

est appelé ensemble des **solutions admissibles**. *On appelle sommet ou point extrémal de X_{ad} tout point $\overline{x} \in X_{ad}$ qui ne peut pas se décomposer en une combinaison convexe (non triviale) de deux autres points de X_{ad}, c'est-à-dire que, s'il existe $y, z \in X_{ad}$ et $\theta \in]0, 1[$ tels que $\overline{x} = \theta y + (1-\theta)z$, alors $y = z = \overline{x}$.*

Remarque 4.3.13 Le vocabulaire de l'optimisation est trompeur pour les néophytes. On appelle solution (admissible) un vecteur qui satisfait les contraintes. Par contre, un vecteur qui atteint le minimum de (4.56) est appelé **solution optimale** (ou point de minimum). •

On vérifie facilement que l'ensemble X_{ad} est un **polyèdre** (éventuellement vide). (Rappelons qu'un polyèdre est une intersection finie de demi-espaces de $\mathbb{R}^n$.) Ses points extrémaux sont donc les sommets de ce polyèdre. Lorsque X_{ad} est vide, par convention on note que

$$\inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x = +\infty.$$

Lemme 4.3.14 *Il existe au moins une solution optimale (ou point de minimum) du programme linéaire standard (4.56) si et seulement si la valeur du minimum est finie*

$$-\infty < \inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x < +\infty.$$

Démonstration. Soit $(x^k)_{k \geq 1}$ une suite minimisante de (4.56). Si cette suite est bornée (par exemple, parce que l'ensemble X_{ad} est borné), alors une sous-suite converge vers une limite $\overline{x}$ qui est solution de (4.56). Si aucune sous-suite de $(x^k)_{k \geq 1}$ ne converge (ce qui peut arriver si X_{ad} n'est pas borné), alors il faut utiliser un argument plus subtil. On introduit la matrice $\mathcal{A}$ et le cône C définis par

$$\mathcal{A} = \begin{pmatrix} c^* \\ A \end{pmatrix} \quad \text{et} \quad C = \left\{ \sum_{i=1}^{n} x_i \mathcal{A}_i \text{ avec } x_i \geq 0 \right\},$$

où les $\mathcal{A}_i$ sont les colonnes de la matrice $\mathcal{A}$. La suite $\mathcal{A}x^k$ appartient à C qui, d'après le Lemme de Farkas (voir sa version générale dans [1]) est fermé, ce qui implique que

$$\lim_{k\to+\infty} \mathcal{A}x^k = \begin{pmatrix} z_0 \\ b \end{pmatrix} \in C,$$

donc il existe $\overline{x} \geq 0$ tel que

$$\begin{pmatrix} z_0 \\ b \end{pmatrix} = \begin{pmatrix} c \cdot \overline{x} \\ A\overline{x} \end{pmatrix},$$

et le minimum est atteint en $\overline{x}$. □

Définition 4.3.15 *On appelle* **base** *associée à (4.56) une base de $\mathbb{R}^m$ formée de m colonnes de A. On note B cette base qui est une sous-matrice de A, carrée d'ordre m inversible. Après permutation de ses colonnes on peut écrire A sous la forme (B, N) où N est une matrice de taille $m \times (n-m)$. De la même façon on peut décomposer x en $(x_B, x_N) \in \mathbb{R}^m \times \mathbb{R}^{n-m}$ de sorte qu'on a*

$$Ax = Bx_B + Nx_N.$$

Une **solution basique** *(ou de base) est un vecteur $x \in X_{ad}$ tel que $x_N = 0$.*

La notion de solution basique correspond à celle de sommet de X_{ad}.

Lemme 4.3.16 *Les sommets du polyèdre X_{ad} sont exactement les solutions basiques.*

Démonstration. Si $x \in X_{ad}$ est une solution basique, dans une certaine base de $\mathbb{R}^n$ on a $x = (x_1, ..., x_m, 0, ..., 0)$, $A = (B, N)$ avec $B = (b_1, ..., b_m)$, une base de $\mathbb{R}^m$ telle que $\sum_{i=1}^m x_i b_i = b$. Supposons qu'il existe $0 < \theta < 1$ et $y, z \in X_{ad}$ tels que $x = \theta y + (1-\theta)z$. Nécessairement, les $n-m$ dernières composantes de y et z sont nulles et, comme y et z appartiennent à X_{ad}, on a $\sum_{i=1}^m y_i b_i = b$ et $\sum_{i=1}^m z_i b_i = b$. Par unicité de la décomposition dans une base, on en déduit que $x = y = z$, et donc x est un sommet de X_{ad}.

Réciproquement, si x est un sommet de X_{ad}, on note k le nombre de ses composantes non nulles, et après un éventuel réarrangement on a $b = \sum_{i=1}^k x_i a_i$ où les (a_i) sont les colonnes de A. Pour montrer que x est une solution basique il suffit de prouver que la famille $(a_1, ..., a_k)$ est libre dans $\mathbb{R}^m$ (on obtient une base B en complétant cette famille). Supposons que ce ne soit pas le cas : il existe alors $y \neq 0$ tel que $\sum_{i=1}^k y_i a_i = 0$ et $(y_{k+1}, ..., y_n) = 0$. Comme les composantes $(x_1, ..., x_k)$ sont strictement positives, il existe $\epsilon > 0$ (petit) tel que $(x + \epsilon y) \in X_{ad}$ et $(x - \epsilon y) \in X_{ad}$. Le fait que $x = (x+\epsilon y)/2 + (x-\epsilon y)/2$ contredit le caractère extrémal de x, donc x est une solution basique. □

Le résultat fondamental suivant nous dit qu'il est suffisant de chercher une solution optimale parmi les sommets du polyèdre X_{ad}.

Proposition 4.3.17 *S'il existe une solution optimale du programme linéaire standard (4.56), alors il existe une solution optimale basique.*

Démonstration. La démonstration est très similaire à celle du Lemme 4.3.16. Soit $x \in X_{ad}$ une solution optimale de (4.56). On note k le nombre de ses composantes non nulles, et après un éventuel réarrangement on a

$$b = \sum_{i=1}^{k} x_i a_i,$$

où les (a_i) sont les colonnes de A. Si la famille $(a_1, ..., a_k)$ est libre dans $\mathbb{R}^m$, alors x est une solution optimale basique. Si $(a_1, ..., a_k)$ est lié, alors il existe $y \neq 0$ tel que

$$\sum_{i=1}^{k} y_i a_i = 0 \text{ et } (y_{k+1}, ..., y_n) = 0.$$

Comme les composantes $(x_1, ..., x_k)$ sont strictement positives, il existe $\epsilon > 0$ tel que $(x \pm \epsilon y) \in X_{ad}$. Comme x est un point de minimum, on a nécessairement

$$c \cdot x \leq c \cdot (x \pm \epsilon y),$$

c'est-à-dire $c \cdot y = 0$. On définit alors une famille de points $z_\epsilon = x + \epsilon y$ paramétrée par ϵ. En partant de la valeur $\epsilon = 0$, si on augmente ou on diminue ϵ on reste dans l'ensemble X_{ad} jusqu'à une valeur ϵ_0 au delà de laquelle la contrainte $z_\epsilon \geq 0$ est violée. Autrement dit, $z_{\epsilon_0} \in X_{ad}$ possède au plus $(k-1)$ composantes non nulles et est encore solution optimale. On répète alors l'argument précédent avec $x = z_{\epsilon_0}$ et une famille de $(k-1)$ colonnes (a_i). A force de diminuer la taille de cette famille, on obtiendra finalement une famille libre et une solution optimale basique. □

Grâce à la Proposition 4.3.17 on a donc une méthode pour résoudre le programme linéaire (4.56). Il suffit d'énumérer les sommets du polyèdre X_{ad} (qui sont en nombre fini) et de sélectionner celui qui donne le plus petit coût. Malheureusement, en pratique, le nombre de sommets du polyèdre X_{ad} est gigantesque car il peut être exponentiel par rapport au nombre de variables. Il faut donc parcourir les sommets de manière intelligente : c'est précisément ce que fait l'algorithme du simplexe, dû à G. Dantzig à la fin des années 1940, qui est toujours extraordinairement efficace en pratique (même si d'autres méthodes plus récentes, comme les méthodes de points intérieurs, sont parfois plus performantes). Cet algorithme parcourt les sommets en faisant décroître la fonction coût à chaque nouveau sommet. Passer d'un sommet à un autre revient à échanger une des colonnes de la base B avec une colonne de la matrice N. Cet échange est dicté par la réduction du coût et les contraintes sont toujours satisfaites. Nous ne pouvons en dire plus ici et nous renvoyons à [1], [6], [8], [12] pour une description complète.

Exercice 4.3.6 Résoudre le programme linéaire suivant

$$\max_{x_1 \geq 0, x_2 \geq 0} x_1 + 2x_2$$

sous les contraintes

$$\begin{cases} -3x_1 + 2x_2 & \leq & 2, \\ -x_1 + 2x_2 & \leq & 4, \\ x_1 + x_2 & \leq & 5. \end{cases}$$

La théorie de la dualité (déjà évoquée lors de la Sous-section 4.3.3) est très utile en programmation linéaire. Considérons à nouveau le programme linéaire standard que nous appellerons primal (par opposition au dual)

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x, \tag{4.61}$$

où A est une matrice de taille $m\times n$, $b\in\mathbb{R}^m$, et $c\in\mathbb{R}^n$. Pour $p\in\mathbb{R}^m$, on introduit le Lagrangien de (4.61)

$$L(x,p)=c\cdot x+p\cdot(b-Ax), \tag{4.62}$$

où l'on a seulement "dualisé" les contraintes d'égalité. On introduit la fonction duale associée

$$G(p)=\min_{x\geq 0} L(x,p),$$

qui, après calcul, vaut

$$G(p)=\begin{cases} p\cdot b & \text{si } A^*p-c\leq 0\\ -\infty & \text{sinon.}\end{cases} \tag{4.63}$$

Le problème dual de (4.61) est donc

$$\sup_{p\in\mathbb{R}^m \text{tel que } A^*p-c\leq 0} p\cdot b. \tag{4.64}$$

L'espace de solutions admissibles du problème dual (4.64) est noté

$$P_{ad}=\left\{p\in\mathbb{R}^m \text{ tel que } A^*p-c\leq 0\right\}.$$

Rappelons que l'espace de solutions admissibles de (4.61) est

$$X_{ad}=\left\{x\in\mathbb{R}^n \text{ tel que } Ax=b,\ x\geq 0\right\}.$$

Les programmes linéaires (4.61) et (4.64) sont dits en **dualité**. L'intérêt de cette notion vient du résultat suivant qui est un cas particulier du Théorème de dualité 4.3.10.

Théorème 4.3.18 *Si (4.61) ou (4.64) a une valeur optimale finie, alors il existe $\overline{x}\in X_{ad}$ solution optimale de (4.61) et $\overline{p}\in P_{ad}$ solution optimale de (4.64) qui vérifient*

$$\left(\min_{x\in\mathbb{R}^n \text{ tel que } Ax=b,\ x\geq 0} c\cdot x\right)=c\cdot\overline{x}=\overline{p}\cdot b=\left(\max_{p\in\mathbb{R}^m \text{ tel que } A^*p-c\leq 0} p\cdot b\right) \tag{4.65}$$

De plus, $\overline{x}$ et $\overline{p}$ sont solutions optimales de (4.61) et (4.64) si et seulement si elles vérifient les conditions d'optimalité de Kuhn et Tucker

$$A\overline{x}=b,\ \overline{x}\geq 0,\ A^*\overline{p}-c\leq 0,\ \overline{x}\cdot(c-A^*\overline{p})=0. \tag{4.66}$$

Si (4.61) ou (4.64) a une valeur optimale infinie, alors l'ensemble des solutions admissibles de l'autre problème est vide.

Remarque 4.3.19 Une conséquence immédiate du Théorème 4.3.18 de dualité est que, si $x \in X_{ad}$ et $p \in P_{ad}$ sont deux solutions admissibles de (4.61) et (4.64), respectivement, elles vérifient

$$c \cdot x \geq b \cdot p.$$

De même, si $\overline{x} \in X_{ad}$ et $\overline{p} \in P_{ad}$ vérifient

$$c \cdot \overline{x} = b \cdot \overline{p}$$

alors $\overline{x}$ est solution optimale de (4.61) et $\overline{p}$ de (4.64). Ces deux propriétés permettent de trouver facilement des bornes pour les valeurs optimales de (4.61) et (4.64), et de tester si un couple $(\overline{x}, \overline{p})$ est optimal. •

Démonstration. Pour simplifier la démonstration nous supposons que X_{ad} et P_{ad} sont non vides (si l'un des deux est vide nous renvoyons à [1] pour une preuve complète). Soit $x \in X_{ad}$ et $p \in P_{ad}$. Comme $x \geq 0$ et $A^*p \leq c$, on a

$$c \cdot x \geq A^*p \cdot x = p \cdot Ax = p \cdot b,$$

puisque $Ax = b$. En particulier, cette inégalité implique que les valeurs optimales des deux problèmes, primal et dual, sont finies, donc qu'ils admettent des solutions optimales en vertu du Lemme 4.3.14. L'égalité (4.65) et la condition d'optimalité (4.66) sont alors une conséquence du Théorème de dualité 4.3.10. □

L'intérêt de la dualité pour résoudre le programme linéaire (4.61) est multiple. D'une part, selon l'algorithme choisi, il peut être plus facile de résoudre le problème dual (4.64) (qui a m variables et n contraintes d'inégalités) que le problème primal (4.61) (qui a n variables, m contraintes d'égalités et n contraintes d'inégalités). D'autre part, on peut construire des algorithmes numériques très efficaces pour la résolution de (4.61) qui utilisent les deux formes primale et duale du programme linéaire.

Exercice 4.3.7 Utiliser la dualité pour résoudre "à la main" (et sans calculs !) le programme linéaire

$$\min_{x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0} 8x_1 + 9x_2 + 4x_3 + 6x_4$$

sous les contraintes

$$\begin{cases} 4x_1 + x_2 + x_3 + 2x_4 \geq 1 \\ x_1 + 3x_2 + 2x_3 + x_4 \geq 1 \end{cases}$$

Exercice 4.3.8 Trouver le problème dual de (4.61) lorsqu'on dualise aussi la contrainte $x \geq 0$, c'est-à-dire qu'on introduit le Lagrangien

$$L(x, p, q) = c \cdot x + p \cdot (b - Ax) - q \cdot x$$

avec $q \in \mathbb{R}^n$ tel que $q \geq 0$. Comparer avec (4.64) et interpréter la nouvelle variable duale q. En déduire qu'il n'y a pas d'intérêt à "dualiser" aussi la contrainte $x \geq 0$.

Exercice 4.3.9 Vérifier que le problème dual de (4.64) est à nouveau (4.61).

4.4 Algorithmes numériques

4.4.1 Introduction

L'objet de cette section est de présenter et analyser quelques algorithmes permettant de calculer, ou plus exactement d'**approcher** la solution des problèmes d'optimisation étudiés précédemment. Tous les algorithmes étudiés ici sont effectivement utilisés en pratique pour résoudre sur ordinateur des problèmes concrets d'optimisation. Ils utilisent aussi tous la connaissance de la dérivée première (voire seconde) de la fonction à optimiser. Il existe des algorithmes qui n'utilisent pas de dérivées mais ils sont très nettement moins performants !

Ces algorithmes sont aussi tous de nature itérative : à partir d'une donnée initiale u^0, chaque méthode construit une suite $(u^n)_{n\in\mathbb{N}}$ dont nous montrerons qu'elle converge, sous certaines hypothèses, vers la solution u du problème d'optimisation considéré. Après avoir montré la **convergence de ces algorithmes** (c'est-à-dire, la convergence de la suite (u^n) vers u quel que soit le choix de la donnée initiale u^0), nous dirons aussi un mot de leur vitesse de convergence.

Dans toute cette section nous supposerons qu'en plus d'être dérivable, la fonction objectif à minimiser J est, non seulement convexe, mais α-convexe (on dit aussi fortement convexe). Cette hypothèse d'α-convexité (voir la Définition 4.4.2 ci-dessous) est assez forte, mais nous verrons qu'elle est cruciale pour les démonstrations de convergence des algorithmes. L'application des algorithmes présentés ici à la minimisation de fonctions convexes qui ne sont pas fortement convexes peut soulever quelques petites difficultés (il peut y avoir plusieurs points de minimum entre lesquels l'algorithme oscille sans converger), sans parler des **grosses** difficultés qui apparaissent lorsque l'on cherche à approcher le minimum d'une fonction non convexe ! Typiquement dans ce dernier cas, ils peuvent converger vers un minimum local, très loin d'un minimum global, et le "choix" de ce minimum local dépend de manière instable des paramètres du calcul.

Enfin, pour simplifier la présentation nous nous limiterons à des problèmes de minimisation en dimension finie, en prenant $V = \mathbb{R}^N$. Néanmoins, les algorithmes présentés s'étendent au cas des espaces de Hilbert en dimension infinie (voir [1]).

Remarque 4.4.1 Nous nous limitons aux seuls algorithmes déterministes et nous ne disons rien des algorithmes de type stochastique (recuit simulé, algorithmes génétiques, etc.). Outre le fait que leur analyse fait appel à la théorie des probabilités (que nous n'abordons pas dans ce cours), leur utilisation est très différente. Pour schématiser simplement, disons que les algorithmes déterministes sont les plus efficaces pour la minimisation de fonctions convexes, tandis que les algorithmes stochastiques permettent d'approcher des minima **globaux** (et pas seulement locaux) de fonctions non convexes (à un prix toutefois assez élevé en pratique). •

Définition 4.4.2 *On dit qu'une fonction J définie sur un ensemble convexe K (inclus dans un espace vectoriel normé V) est fortement convexe ou α-convexe s'il existe*

$\alpha > 0$ tel que, pour tout $u, v \in K$,

$$J\left(\frac{u+v}{2}\right) \leq \frac{J(u)+J(v)}{2} - \frac{\alpha}{8}\|u-v\|^2 \ . \tag{4.67}$$

Dans la Définition 4.4.2, la forte convexité de J n'est testée que pour des combinaisons convexes de poids $\theta = 1/2$. Cela n'est pas une restriction pour les fonctions continues comme le montre l'exercice suivant. Par ailleurs, d'après l'Exercice 4.1.5 les fonctions convexes à valeurs finies sont automatiquement continues.

Exercice 4.4.1 Si J est continue et α-convexe, montrer que, pour tout $\theta \in [0,1]$,

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha\theta(1-\theta)}{2}\|u-v\|^2 \ . \tag{4.68}$$

La notion de forte convexité est donc **plus restrictive** que la stricte convexité et la constante α mesure en quelque sorte la convexité minimale de la fonction, comme l'indique l'exercice suivant.

Exercice 4.4.2 Soit A une matrice symétrique d'ordre N et $b \in \mathbb{R}^N$. Pour $x \in \mathbb{R}^N$, on pose $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Montrer que J est convexe si et seulement si A est semi-définie positive, et que J est strictement convexe si et seulement si A est définie positive. Dans ce dernier cas, montrer que J est aussi fortement ou α-convexe avec $\alpha = \lambda_1 > 0$, la plus petite valeur propre de A.

En dimension finie les fonctions α-convexes admettent un unique point de minimum.

Exercice 4.4.3 Soit J une fonction α-convexe sur $V = \mathbb{R}^N$. Montrer que J est infinie à l'infini et que, par conséquent, il existe un unique point de minimum pour J dans V.

Exercice 4.4.4 Montrer qu'une fonction α-convexe J vérifie

$$J(v) \geq J(u) + \langle J'(u), v-u\rangle + \frac{\alpha}{2}\|v-u\|^2 \quad \forall\, u, v \in V \ , \tag{4.69}$$

Indication : on s'inspirera de la démonstration de la Proposition 4.2.4.

4.4.2 Algorithmes de type gradient (cas sans contraintes)

Commençons par étudier la résolution pratique de problèmes d'optimisation en l'absence de contraintes. Soit J une fonction α-convexe différentiable définie sur $V = \mathbb{R}^N$. On considère le problème sans contrainte

$$\min_{v \in V} J(v) \ , \tag{4.70}$$

et on note u son unique point de minimum (cf. l'Exercice 4.4.3), caractérisé d'après la Remarque 4.2.9 par l'équation d'Euler

$$J'(u) = 0 \ .$$

Algorithme de gradient à pas optimal

L'algorithme de gradient consiste à "se déplacer" d'une itérée u^n en suivant la ligne de plus grande pente associée à la fonction coût $J(v)$. La direction de descente correspondant à cette ligne de plus grande pente issue de u^n est donnée par le gradient $J'(u^n)$. En effet, si l'on cherche u^{n+1} sous la forme

$$u^{n+1} = u^n - \mu^n w^n \,, \tag{4.71}$$

avec $\mu^n > 0$ petit et w^n unitaire dans V, un développement de Taylor à l'ordre 1 donne

$$J(u^{n+1}) = J(u^n) - \mu^n \langle J'(u^n), w^n \rangle + o(\mu^n).$$

Si on néglige le reste $o(\mu^n)$ (en l'absence d'autres informations sur les dérivées supérieures ou les itérées antérieures), il est clair que la diminution maximale de la fonction J est obtenue avec le choix de la direction $w_n = \dfrac{J'(u^n)}{\|J'(u^n)\|}$.

Cette remarque simple nous conduit, parmi les méthodes du type (4.71) qui sont appelées "méthodes de descente", à l'algorithme de **gradient à pas optimal**, dans lequel on résout une succession de problèmes de minimisation à une seule variable réelle. A partir de u^0 quelconque dans V, on construit la suite (u^n) définie par

$$u^{n+1} = u^n - \mu^n \, J'(u^n) \,, \tag{4.72}$$

où $\mu^n \in \mathbb{R}$ est choisi à chaque étape tel que

$$J(u^{n+1}) = \inf_{\mu \in \mathbb{R}} J\big(u^n - \mu \, J'(u^n)\big) \,. \tag{4.73}$$

Cet algorithme converge comme l'indique le résultat suivant.

Théorème 4.4.3 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur tout borné de V, c'est-à-dire que, pour tout $M > 0$, il existe $C_M > 0$ tel que*

$$\|v\| + \|w\| \leq M \;\Rightarrow\; \|J'(v) - J'(w)\| \leq C_M \|v - w\| \,. \tag{4.74}$$

Alors l'algorithme de gradient à pas optimal converge : quel que soit u^0, la suite (u^n) définie par (4.72) et (4.73) converge vers la solution u de (4.70).

Démonstration. La fonction $f(\mu) = J\big(u^n - \mu \, J'(u^n)\big)$ est fortement convexe et dérivable sur $\mathbb{R}$ (si $J'(u^n) \neq 0$; sinon, on a déjà convergé, $u^n = u$!). Le problème de minimisation (4.73) a donc bien une solution unique, caractérisée par la condition $f'(\mu) = 0$, ce qui s'écrit aussi

$$\langle J'(u^{n+1}),\, J'(u^n) \rangle = 0 \,. \tag{4.75}$$

Ceci montre que deux "directions de descente" consécutives sont orthogonales.

Puisque (4.75) implique que $\langle J'(u^{n+1}), u^{n+1} - u^n \rangle = 0$, on déduit de l'$\alpha$-convexité de J (voir l'Exercice 4.4.4) que

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2} \|u^n - u^{n+1}\|^2, \tag{4.76}$$

ce qui prouve que la suite $J(u^n)$ est décroissante. Comme elle est minorée par $J(u)$, elle converge et (4.76) montre que $u^{n+1}-u^n$ tend vers 0. D'autre part, l'α-convexité de J et le fait que la suite $J(u^n)$ est bornée montrent que la suite (u^n) est bornée : il existe une constante M telle que

$$\|u^n\| \leq M \ .$$

Écrivant (4.74) pour $v=u^n$ et $w=u^{n+1}$ et utilisant (4.75), on obtient

$$\|J'(u^n)\|^2 \leq \|J'(u^n)\|^2 + \|J'(u^{n+1})\|^2 = \|J'(u^n)-J'(u^{n+1})\|^2 \leq C_M^2\|u^{n+1}-u^n\|^2,$$

ce qui prouve que $J'(u^n)$ tend vers 0. L'α-convexité de J donne alors

$$\alpha\|u^n-u\|^2 \leq \langle J'(u^n)-J'(u), u^n-u\rangle = \langle J'(u^n), u^n-u\rangle \leq \|J'(u^n)\| \ \|u^n-u\| \ ,$$

qui implique $\alpha\|u^n-u\| \leq \|J'(u^n)\|$, d'où l'on déduit la convergence. □

Algorithme de gradient à pas fixe

L'algorithme de gradient à pas fixe consiste simplement en la construction d'une suite u^n définie par

$$u^{n+1} = u^n - \mu\, J'(u^n) \ , \tag{4.77}$$

où μ est un paramètre positif fixé. Cette méthode est donc plus simple que l'algorithme de gradient à pas optimal, puisqu'on fait à chaque étape l'économie de la résolution de (4.73). Le résultat suivant montre sous quelles hypothèses on peut choisir le paramètre μ pour assurer la convergence.

Théorème 4.4.4 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V, c'est-à-dire qu'il existe une constante $C>0$ telle que*

$$\|J'(v)-J'(w)\| \leq C\|v-w\| \quad \forall\, v,w \in V \ . \tag{4.78}$$

Alors, si $0<\mu<2\alpha/C^2$, l'algorithme de gradient à pas fixe converge : quel que soit u^0, la suite (u^n) définie par (4.73) converge vers la solution u de (4.70).

Démonstration. Posons $v^n = u^n - u$. Comme $J'(u)=0$, on a $v^{n+1} = v^n - \mu\big(J'(u^n)-J'(u)\big)$, d'où il vient

$$\begin{aligned} \|v^{n+1}\|^2 &= \|v^n\|^2 - 2\mu\langle J'(u^n)-J'(u), u^n-u\rangle + \mu^2\big\|J'(u^n)-J'(u)\big\|^2 \\ &\leq \big(1-2\alpha\mu+C^2\mu^2\big)\ \|v^n\|^2 \ , \end{aligned} \tag{4.79}$$

d'après (4.78) et l'α-convexité. Si $0<\mu<2\alpha/C^2$, il est facile de voir que $1-2\alpha\mu+C^2\mu^2 \in]0,1[$, et la convergence se déduit de (4.79). De manière équivalente, la même démonstration montre que l'application $v \mapsto v-\mu J'(v)$ est strictement contractante lorsque $0<\mu<2\alpha/C^2$, donc elle admet un unique point fixe (qui n'est autre que u) vers lequel converge la suite u^n. □

Remarque 4.4.5 Il existe de nombreux autres algorithmes de descente du type (4.71) que nous ne décrirons pas ici. On rencontre notamment dans cette classe d'algorithmes la méthode du gradient conjugué dans laquelle la direction de descente w^n dépend non seulement du gradient $J'(u^n)$ mais aussi des directions de descente utilisées aux itérations précédentes. Nous présentons cette méthode dans la Sous-section 3.5.5, pour le cas particulier d'une fonctionnelle quadratique du type $\frac{1}{2}Ax \cdot x - b \cdot x$. •

4.4.3 Algorithmes de type gradient (cas avec contraintes)

On étudie maintenant la résolution de problèmes d'optimisation avec contraintes

$$\min_{v \in K} J(v) \ , \tag{4.80}$$

où J est une fonction α-convexe différentiable définie sur K, sous-ensemble convexe fermé non vide de $V = \mathbb{R}^N$. Le Théorème 4.1.3 et la Proposition 4.1.6 assurent alors l'existence et l'unicité de la solution u de (4.80), caractérisée d'après le Théorème 4.2.8 par la condition

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall\, v \in K \ . \tag{4.81}$$

Selon les algorithmes étudiés ci-dessous, nous serons parfois amenés à préciser des hypothèses supplémentaires sur l'ensemble K.

Algorithme de gradient à pas fixe avec projection

L'algorithme de gradient à pas fixe s'adapte au cas du problème (4.80) avec contraintes à partir de la remarque suivante. Pour tout réel $\mu > 0$, (4.81) s'écrit

$$\langle u - \big(u - \mu J'(u)\big), v - u \rangle \geq 0 \quad \forall\, v \in K \ . \tag{4.82}$$

Notons P_K l'opérateur de projection orthogonal sur l'ensemble convexe K. Alors, (4.82) n'est rien d'autre que la caractérisation de u comme la projection orthogonale de $u - \mu J'(u)$ sur K. Autrement dit,

$$u = P_K\big(u - \mu J'(u)\big) \quad \forall\, \mu > 0 \ . \tag{4.83}$$

Il est facile de voir que (4.83) est en fait équivalent à (4.81), et caractérise donc la solution u de (4.80). L'algorithme de **gradient à pas fixe avec projection** (ou plus simplement de gradient projeté) est alors défini par l'itération

$$u^{n+1} = P_K\big(u^n - \mu J'(u^n)\big) \ , \tag{4.84}$$

où μ est un paramètre positif fixé.

Théorème 4.4.6 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V (de constante C, voir (4.78)). Alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme de gradient à pas fixe avec projection converge : quel que soit $u^0 \in K$, la suite (u^n) définie par (4.84) converge vers la solution u de (4.80).*

Démonstration. La démonstration reprend celle du Théorème 4.4.4 où l'on a montré que l'application $v \mapsto v - \mu J'(v)$ est strictement contractante lorsque $0 < \mu < 2\alpha/C^2$, c'est-à-dire que

$$\exists\, \gamma \in]0,1[\quad , \quad \big\| \big(v - \mu J'(v)\big) - \big(w - \mu J'(w)\big) \big\| \leq \gamma \|v - w\| \; .$$

Puisque la projection P_K est faiblement contractante ($\|P_k v - P_k w\| \leq \|v - w\|$), l'application $v \mapsto P_K\big(v - \mu J'(v)\big)$ est strictement contractante, ce qui prouve la convergence de la suite (u^n) définie par (4.84) vers la solution u de (4.80). □

Exercice 4.4.5 Soit $V = \mathbb{R}^N$ et $K = \{x \in \mathbb{R}^N \text{ tel que } \sum_{i=1}^N x_i = 1\}$. Expliciter l'opérateur de projection orthogonale P_K et interpréter dans ce cas la formule (4.83) en terme de multiplicateur de Lagrange.

Algorithme d'Uzawa

Le résultat précédent montre que la méthode de gradient à pas fixe avec projection est applicable à une large classe de problèmes d'optimisation convexe avec contraintes. Mais cette conclusion est largement un leurre du point de vue pratique, car l'opérateur de projection P_K n'est pas connu explicitement en général : la projection d'un élément $v \in V$ sur un convexe fermé quelconque de V peut être très difficile à déterminer !

Une exception importante concerne, pour $V = \mathbb{R}^M$, les sous-ensembles K de la forme

$$K = \prod_{i=1}^{M} [a_i, b_i] \tag{4.85}$$

(avec éventuellement $a_i = -\infty$ ou $b_i = +\infty$ pour certains indices i). En effet, il est alors facile de voir que, si $x = (x_1, x_2, \ldots, x_M) \in \mathbb{R}^M$, $y = P_K(x)$ a pour composantes

$$y_i = \min\big(\max(a_i, x_i), b_i\big) \quad \text{pour} \quad 1 \leq i \leq M \; , \tag{4.86}$$

autrement dit, il suffit juste de "tronquer" les composantes de x. Cette propriété simple, jointes aux remarques sur la dualité énoncée dans la Section 4.3, va nous conduire à un nouvel algorithme. En effet, même si le problème primal fait intervenir un ensemble K des solutions admissibles sur lequel la projection P_K ne peut être déterminée explicitement, le problème dual sera fréquemment posé sur un ensemble de la forme (4.85), typiquement sur $(\mathbb{R}_+)^M$. Dans ce cas, le problème dual peut être résolu par la méthode du gradient à pas fixe avec projection, et la solution du problème primal pourra ensuite être obtenue en résolvant un problème de minimisation **sans contrainte**. Ces remarques sont à la base de l'algorithme d'Uzawa, qui est en fait une méthode de recherche de point-selle.

Considérons le problème de minimisation convexe

$$\min_{F(v) \leq 0} J(v) \; , \tag{4.87}$$

où J est une fonctionnelle α-convexe définie sur $V = \mathbb{R}^N$ et F une fonction convexe de $V = \mathbb{R}^N$ dans $\mathbb{R}^M$. Comme pour le problème (4.80) on sait qu'il existe une unique

solution de (4.87). Sous les hypothèses du Théorème de Kuhn et Tucker 4.3.3, la résolution de (4.87) revient à trouver un point-selle (u, p) du Lagrangien

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \ , \tag{4.88}$$

sur $V \times (\mathbb{R}_+)^M$. A partir de la Définition 4.3.1 du point-selle

$$\forall\, q \in (\mathbb{R}_+)^M \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in V \ , \tag{4.89}$$

on déduit que $(p - q) \cdot F(u) \geq 0$ pour tout $q \in (\mathbb{R}_+)^M$, d'où on tire, pour tout réel $\mu > 0$,

$$(p - q) \cdot \big(p - \big(p + \mu F(u)\big)\big) \leq 0 \quad \forall\, q \in (\mathbb{R}_+)^M \ ,$$

ce qui, d'après la caractérisation de la projection orthogonale sur un convexe fermé, montre que

$$p = P_{\mathbb{R}_+^M}(p + \mu F(u)) \quad \forall\, \mu > 0 \ , \tag{4.90}$$

$P_{\mathbb{R}_+^M}$ désignant la projection de $\mathbb{R}^M$ sur $(\mathbb{R}_+)^M$.

Au vu de cette propriété et de la seconde inégalité dans (4.89), nous pouvons introduire **l'algorithme d'Uzawa** : à partir d'un élément quelconque $p^0 \in (\mathbb{R}_+)^M$, on construit les suites (u^n) et (p^n) déterminées par les itérations

$$\begin{aligned} &\mathcal{L}(u^n, p^n) = \inf_{v \in V} \mathcal{L}(v, p^n) \ , \\ &p^{n+1} = P_{\mathbb{R}_+^M}(p^n + \mu F(u^n)) \ , \end{aligned} \tag{4.91}$$

μ étant un paramètre positif fixé. On peut interpréter l'algorithme d'Uzawa en disant qu'alternativement il minimise le Lagrangien par rapport à v avec q fixé et il maximise (par un seul pas de l'algorithme du gradient projeté) ce même Lagrangien par rapport à q avec v fixé. Une autre manière de voir l'algorithme d'Uzawa est la suivante : il prédit une valeur du multiplicateur de Lagrange q et effectue une minimisation sans contrainte du Lagrangien par rapport à v, puis il corrige la prédiction de q en l'augmentant si la contrainte est violée et en le diminuant sinon. Nous verrons une troisième interprétation de l'algorithme d'Uzawa dans le cadre de la théorie de la dualité ci-dessous.

Théorème 4.4.7 *On suppose que J est α-convexe différentiable, que F est convexe et Lipschitzienne de V dans $\mathbb{R}^M$, c'est-à-dire qu'il existe une constante C telle que*

$$\|F(v) - F(w)\| \leq C\|v - w\| \quad \forall\, v, w \in V \ , \tag{4.92}$$

et qu'il existe un point-selle (u, p) du Lagrangien (4.88) sur $V \times (\mathbb{R}_+)^M$. Alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme d'Uzawa converge : quel que soit l'élément initial p^0, la suite (u^n) définie par (4.91) converge vers la solution u du problème (4.87).

Démonstration. Rappelons d'abord que l'existence d'une solution u de (4.87) découle de celle du point-selle (u, p) (voir la Proposition 4.3.2), alors que son unicité est une conséquence de l'α-convexité de J. De même, p^n étant fixé, le problème de

minimisation dans (4.91) a bien une solution unique u^n. D'après l'Exercice 4.2.11, les inéquations d'Euler satisfaites par u et u^n s'écrivent

$$\langle J'(u), v-u\rangle + p\cdot\big(F(v)-F(u)\big)\geq 0 \quad \forall\, v\in V\ , \tag{4.93}$$

$$\langle J'(u^n), v-u^n\rangle + p^n\cdot\big(F(v)-F(u^n)\big)\geq 0 \quad \forall\, v\in V\ . \tag{4.94}$$

Prenant successivement $v=u^n$ dans (4.93) et $v=u$ dans (4.94) et additionnant, on obtient

$$\langle J'(u)-J'(u^n), u^n-u\rangle + (p-p^n)\cdot\big(F(u^n)-F(u)\big)\geq 0\ ,$$

d'où en utilisant l'α-convexité de J et en posant $r^n=p^n-p$

$$r^n\cdot\big(F(u^n)-F(u)\big)\leq -\alpha\|u^n-u\|^2\ . \tag{4.95}$$

D'autre part, la projection $P_{\mathbb{R}_+^M}$ étant faiblement contractante, en soustrayant (4.90) à (4.91) on obtient

$$\|r^{n+1}\|\leq \|r^n+\mu(F(u^n)-F(u))\|\ ,$$

soit

$$\|r^{n+1}\|^2\leq \|r^n\|^2+2\mu r^n\cdot\big(F(u^n)-F(u)\big)+\mu^2\|F(u^n)-F(u)\|^2\ .$$

Utilisant (4.92) et (4.95), il vient

$$\|r^{n+1}\|^2\leq \|r^n\|^2+\big(C^2\mu^2-2\mu\alpha\big)\|u^n-u\|^2\ .$$

Si $0<\mu<2\alpha/C^2$, on peut trouver $\beta>0$ tel que $C^2\mu^2-2\mu\alpha<-\beta$, d'où

$$\beta\|u^n-u\|^2\leq \|r^n\|^2-\|r^{n+1}\|^2\ . \tag{4.96}$$

Ceci montre alors que la suite $\|r^n\|^2$ est décroissante : le membre de droite de (4.96) tend donc vers 0, ce qui entraîne que u^n tend vers u. □

Ainsi, l'algorithme d'Uzawa permet d'approcher la solution de (4.87) en remplaçant ce problème avec contraintes par une suite de problèmes de minimisation sans contraintes (4.91). A chaque itération, la détermination de p^n est élémentaire, puisque d'après (4.86) l'opérateur de projection $P_{\mathbb{R}_+^M}$ est une simple troncature à zéro des composantes négatives. Il faut aussi noter que le Théorème 4.4.7 ne dit rien de la convergence de la suite (p^n). En fait, cette convergence n'est pas assurée sous les hypothèses du théorème, qui n'assurent d'ailleurs pas l'unicité de l'élément $p\in(\mathbb{R}_+)^M$ tel que (u,p) soit point-selle (voir la Remarque 4.3.11 et l'Exercice 4.4.6 ci-dessous).

Il reste à faire le lien entre l'algorithme d'Uzawa et la théorie de la dualité, comme nous l'avons déjà annoncé. Rappelons d'abord que le problème dual de (4.87) s'écrit

$$\sup_{q\geq 0}\mathcal{G}(q)\ , \tag{4.97}$$

où, par définition

$$\mathcal{G}(q)=\inf_{v\in V}\mathcal{L}(v,q)\ , \tag{4.98}$$

et que le multiplicateur de Lagrange p est une solution du problème dual (4.97). En fait, sous des hypothèses assez générales, on peut montrer que $\mathcal{G}$ est différentiable et que le gradient $\mathcal{G}'(q)$ est précisément égal à $F(u_q)$, où u_q est l'unique solution du problème de minimisation (4.98). En effet, on a

$$\mathcal{G}(q) = J(u_q) + q \cdot F(u_q),$$

et en dérivant formellement par rapport à q

$$\mathcal{G}'(q) = F(u_q) + \langle J'(u_q) + q \cdot F'(u_q), u_q' \rangle = F(u_q),$$

à cause de la condition d'optimalité pour u_q. On voit alors que **l'algorithme d'Uzawa n'est autre que la méthode du gradient à pas fixe avec projection appliquée au problème dual** puisque la deuxième équation de (4.91) peut s'écrire $p^{n+1} = P_{\mathbb{R}_+^M}\big(p^n + \mu \mathcal{G}'(p^n)\big)$ (le changement de signe par rapport à (4.84) vient du fait que le problème dual (4.97) est un problème de maximisation et non de minimisation). Le lecteur vérifiera très facilement cette assertion dans le cas particulier étudié à l'exercice suivant.

Exercice 4.4.6 Appliquer l'algorithme d'Uzawa au problème de la Remarque 4.3.11 (fonctionnelle quadratique et contraintes affines en dimension finie). Si la matrice B est de rang M, ce qui assure l'unicité de p d'après la Remarque 4.3.11, montrer que la suite p^n converge vers p.

Remarque 4.4.8 Une variante, plus simple, de l'algorithme d'Uzawa est **l'algorithme d'Arrow-Hurwicz** qui s'interprète lui aussi comme un algorithme de point selle. Simplement, au lieu de minimiser exactement en v à chaque itération de (4.91), l'algorithme d'Arrow-Hurwicz effectue un seul pas d'une méthode gradient à pas fixe $\nu > 0$. Concrètement, à partir d'éléments quelconques $p^0 \in (\mathbb{R}_+)^M$ et $u^0 \in V$, on construit les suites (u^n) et (p^n) déterminées par les itérations

$$\begin{aligned} u^{n+1} &= u^n - \nu \left(J'(u^n) + p^n \cdot F'(u^n)\right) , \\ p^{n+1} &= P_{\mathbb{R}_+^M}(p^n + \mu F(u^{n+1})) , \end{aligned} \tag{4.99}$$

$\mu > 0, \nu > 0$ étant deux paramètres positif fixés. Autrement dit, (4.99) recherche un point selle en alternant un pas de minimisation en v et un pas de maximisation en q. •

Pénalisation des contraintes

Nous concluons cette sous-section en décrivant brièvement un autre moyen d'approcher un problème de minimisation avec contraintes par une suite de problèmes de minimisation sans contraintes ; c'est la procédure de **pénalisation** des contraintes. Nous évitons de parler ici "d'algorithme" car la pénalisation des contraintes n'est pas un algorithme mais une approche qui consiste à remplacer le problème avec contraintes par un problème "pénalisé" ou "approché" sans contraintes, dépendant

d'un petit paramètre $\varepsilon > 0$. La résolution effective du problème pénalisé sans contraintes doit être réalisée à l'aide de l'un des algorithmes de la Sous-section 4.4.2. Cette résolution peut d'ailleurs soulever des difficultés, car le problème pénalisé est souvent "mal conditionné" (voir la Sous-section 3.5.2). C'est seulement dans la limite ε tendant vers 0 que l'on retrouve le problème d'origine avec contraintes.

Comme d'habitude nous nous plaçons dans le cas où $V = \mathbb{R}^N$, et nous considérons de nouveau le problème de minimisation convexe

$$\inf_{F(v)\leq 0} J(v) \ , \tag{4.100}$$

où J est une fonction convexe continue de $\mathbb{R}^N$ dans $\mathbb{R}$ et F une fonction convexe continue de $\mathbb{R}^N$ dans $\mathbb{R}^M$.

On commence par proposer une méthode de **pénalisation extérieure** au sens où les contraintes ne sont pas respectées. Pour $\varepsilon > 0$, nous introduisons le problème sans contraintes

$$\inf_{v\in\mathbb{R}^N} \left(J(v) + \frac{1}{\varepsilon} \sum_{i=1}^{M} \left[\max\left(F_i(v), 0\right)\right]^2 \right) \ , \tag{4.101}$$

dans lequel ont dit que les contraintes $F_i(v) \leq 0$ sont "pénalisées". On peut alors énoncer le résultat suivant, qui montre que, pour ε petit, le problème (4.101) "approche bien" le problème (4.100).

Proposition 4.4.9 *On suppose que J est continue, strictement convexe, et infinie à l'infini, que les fonctions F_i sont convexes et continues pour $1 \leq i \leq M$, et que l'ensemble*

$$K = \left\{ v \in \mathbb{R}^N \, , \quad F_i(v) \leq 0 \quad \forall\, i \in \{1, \ldots, M\} \right\}$$

est non vide. En notant u l'unique solution de (4.100) et, pour $\varepsilon > 0$, u_ε l'unique solution de (4.101), on a alors

$$\lim_{\varepsilon\to 0} u_\varepsilon = u \ .$$

Démonstration. L'ensemble K étant convexe fermé, l'existence et l'unicité de u découlent du Théorème 4.1.3 et de la stricte convexité de J. De plus, la fonction $G(v) = \sum_{i=1}^{M} \left[\max\left(F_i(v), 0\right)\right]^2$ est continue et convexe puisque la fonction de $\mathbb{R}$ dans $\mathbb{R}$ qui à x associé $\max(x, 0)^2$ est convexe et croissante. On en déduit que la fonctionnelle $J_\varepsilon(v) = J(v) + \varepsilon^{-1} G(v)$ est strictement convexe, continue, et infinie à l'infini puisque $G(v) \geq 0$, ce qui implique l'existence et l'unicité de u_ε. Comme $G(u) = 0$, on peut écrire

$$J_\varepsilon(u_\varepsilon) = J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J_\varepsilon(u) = J(u) \ . \tag{4.102}$$

Ceci montre que

$$J(u_\varepsilon) \leq J_\varepsilon(u_\varepsilon) \leq J(u) \ , \tag{4.103}$$

et donc que u_ε est borné d'après la condition "infinie à l'infini". On peut donc extraire de la famille (u_ε) une suite (u_{ε_k}) qui converge vers une limite u_* lorsque ε_k tend vers 0. On a alors $0 \leq G(u_{\varepsilon_k}) \leq \varepsilon_k (J(u) - J(u_{\varepsilon_k}))$ d'après (4.102). Passant à la

limite, on obtient $G(u_*) = 0$, qui montre que $u_* \in K$. Comme (4.103) implique que $J(u_*) \leq J(u)$, on a alors $u_* = u$, ce qui conclut la démonstration, toutes les suites extraites (u_{ε_k}) convergeant vers la même limite u. □

Exercice 4.4.7 En plus des hypothèses de la Proposition 4.4.9, on suppose que les fonctions J et $F_1, \ldots, F_M$ sont continûment différentiables. On note de nouveau $I(u)$ l'ensemble des contraintes actives en u, et on suppose que les contraintes sont qualifiées en u au sens de la Définition 4.2.17. Enfin, on suppose que les vecteurs $\big(F_i'(u)\big)_{i\in I(u)}$ sont linéairement indépendants, ce qui assure l'unicité des multiplicateurs de Lagrange $\lambda_1, \ldots, \lambda_M$ tels que $J'(u) + \sum_{i=1}^M \lambda_i F_i'(u) = 0$, avec $\lambda_i = 0$ si $i \notin I(u)$. Montrer alors que, pour tout indice $i \in \{1, \ldots, M\}$

$$\lim_{\varepsilon\to 0} \left[\frac{2}{\varepsilon} \max\left(F_i(u_\varepsilon), 0\right)\right] = \lambda_i \, .$$

Remarque 4.4.10 Au lieu d'utiliser une pénalisation quadratique comme dans (4.101), on peut utiliser une pénalisation de type L^1, c'est-à-dire considérer le nouveau problème pénalisé

$$\inf_{v\in\mathbb{R}^N} \left(J_\varepsilon(v) = J(v) + \frac{1}{\varepsilon} \sum_{i=1}^M \Big| \max\left(F_i(v), 0\right)\Big| \right) \, . \tag{4.104}$$

On peut facilement généraliser la Proposition 4.4.9 à (4.104) qui est un problème convexe, admet une unique solution u_ε, convergeante vers la solution u de (4.100). L'aspect remarquable de cette pénalisation L^1 est qu'elle est exacte pour des petites valeurs de ε ! Autrement dit, pour ε suffisamment petit, on a $u_\varepsilon = u$. Le revers de la médaille pour cette propriété remarquable est que la fonction objectif dans (4.104) n'est pas dérivable car la valeur absolue et le maximum ne sont pas dérivables en 0. Il faut donc des algorithmes particuliers (par exemple de sous-gradient, voir [8]) que nous ne décrivons pas ici. Expliquons brièvement, dans le cas d'une seule contrainte $M = 1$, ce "miracle" de la pénalisation exacte. Supposons que la contrainte soit violée au minimum, $F(u_\varepsilon) > 0$. Alors, la valeur absolue et le maximum disparaissent et on peut écrire une condition d'optimalité très simple

$$J'(u_\varepsilon) + \frac{1}{\varepsilon} F'(u_\varepsilon) = 0 \, . \tag{4.105}$$

Mais comme u_ε converge vers u qui vérifie la condition d'optimalité $J'(u) + \lambda F'(u) = 0$ pour un multiplicateur de Lagrange $\lambda \geq 0$, on a aussi

$$\lim_{\varepsilon\to 0} \Big(J'(u_\varepsilon) + \lambda F'(u_\varepsilon) \Big) = 0 \, ,$$

ce qui est une contradiction avec (4.105) pour $1/\varepsilon$ plus grand que λ. Par conséquent, pour ε petit, le point de minimum u_ε vérifie la contrainte $F(u_\varepsilon) \leq 0$ et comme $J(u_\varepsilon) = J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(u) = J(u)$, par unicité de la solution de (4.100), on en déduit que $u_\varepsilon = u$. •

Remarque 4.4.11 Une variante de la méthode de pénalisation extérieure consiste à combiner cette dernière avec l'introduction du Lagrangien usuel : on parle alors de la **méthode du Lagrangien augmenté**. Concrêtement, on remplace le problème original (4.100) par la recherche du point selle du Lagrangien "augmenté"

$$\inf_{v\in\mathbb{R}^N}\sup_{\mu\in(\mathbb{R}_+)^M}\left(J(v)+\mu\cdot F(v)+\frac{1}{\varepsilon}\sum_{i=1}^{M}\left[\max\left(F_i(v),0\right)\right]^2\right)\ . \tag{4.106}$$

La maximisation en μ conduit à satisfaire les contraintes, mais la minimisation en v (à μ fixé) aussi. •

On propose maintenant une méthode de **pénalisation intérieure** au sens où les contraintes sont strictement respectées et jamais actives. Pour $\varepsilon > 0$, nous introduisons le problème sans contraintes

$$\inf_{v\in\mathbb{R}^N}\left(J(v)-\varepsilon\sum_{i=1}^{M}\frac{1}{\min(F_i(v),0)}\right)\ , \tag{4.107}$$

dans lequel ont dit que les contraintes $F_i(v) \leq 0$ sont "pénalisées". Notons que la fonction objectif pénalisée n'est finie que si $F_i(v) < 0$ et que, si la valeur d'une contrainte $F_i(v)$ s'approche de 0 par valeurs négatives, alors la fonction objectif tend vers $+\infty$. Ainsi donc, les éventuels points de minimum de (4.107) restent toujours à distance du "bord" des contraintes.

4.4.4 Méthode de Newton

On se place encore en dimension finie $V = \mathbb{R}^N$. Expliquons le principe de la méthode de Newton. Soit F une fonction de classe C^2 de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Soit u un zéro régulier de F c'est-à-dire que

$$F(u) = 0 \quad \text{et} \quad F'(u) \text{ matrice inversible.}$$

Une formule de Taylor au voisinage de v nous donne

$$F(u) = F(v) + F'(v)(u-v) + \mathcal{O}\left(\|u-v\|^2\right),$$

c'est-à-dire

$$u = v - \left(F'(v)\right)^{-1} F(v) + \mathcal{O}\left(\|v-u\|^2\right).$$

La méthode de Newton consiste à résoudre de façon itérative cette équation en négligeant le reste. Pour un choix initial $u^0 \in \mathbb{R}^N$, on calcule

$$u^{n+1} = u^n - \left(F'(u^n)\right)^{-1} F(u^n) \quad \text{pour} \quad n \geq 0. \tag{4.108}$$

Rappelons que l'on ne calcule pas l'inverse de la matrice $F'(u^n)$ dans (4.108) mais que l'on résout un système linéaire par l'une des méthodes exposées à la Section 3.5. Du point de vue de l'optimisation, la méthode de Newton s'interprète de la

manière suivante. Soit J une fonction de classe C^3 de $\mathbb{R}^N$ dans $\mathbb{R}$, et soit u un minimum local de J. Si on pose $F = J'$, on peut appliquer la méthode précédente pour résoudre la condition nécessaire d'optimalité $J'(u) = 0$. Cependant, on peut aussi envisager la méthode de Newton comme une méthode de minimisation. A cause du développement de Taylor

$$J(w) = J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) + \mathcal{O}\left(\|w - v\|^3\right), \quad (4.109)$$

on peut approcher $J(w)$ au voisinage de v par une fonction quadratique. La méthode de Newton consiste alors à minimiser cette approximation quadratique et à itérer. Le minimum de la partie quadratique du terme de droite de (4.109) est donné par $w = v - (J''(v))^{-1} J'(v)$ si la matrice $J''(v)$ est définie positive. On retrouve alors la formule itérative (4.108).

L'avantage principal de la méthode de Newton est sa convergence bien plus rapide que les méthodes précédentes.

Proposition 4.4.12 *Soit F une fonction de classe C^2 de $\mathbb{R}^N$ dans $\mathbb{R}^N$, et u un zéro régulier de F (i.e. $F(u) = 0$ et $F'(u)$ inversible). Il existe un réel $\epsilon > 0$ tel que, si u^0 est assez proche de u au sens où $\|u - u^0\| \leq \epsilon$, la méthode de Newton définie par (4.108) converge, c'est-à-dire que la suite (u^n) converge vers u, et il existe une constante $C > 0$ telle que*

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2 . \quad (4.110)$$

Démonstration. Par continuité de F' il existe $\epsilon > 0$ tel que F' est inversible en tout point de la boule de centre u et de rayon ϵ. Supposons que u^n soit resté proche de u, au sens où $\|u - u^n\| \leq \epsilon$, donc $F'(u^n)$ est inversible. Comme $F(u) = 0$, on déduit de (4.108)

$$u^{n+1} - u = u^n - u - (F'(u^n))^{-1} (F(u^n) - F(u))$$

qui, par développement de Taylor autour de u^n, devient

$$u^{n+1} - u = (F'(u^n))^{-1} \mathcal{O}\left(\|u^n - u\|^2\right).$$

Comme $\|u - u^n\| \leq \epsilon$, on en déduit qu'il existe une constante $C > 0$ (indépendante de n et liée au module de continuité de F' et de F'' sur la boule de centre u et de rayon ϵ) telle que

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2. \quad (4.111)$$

Si ϵ est suffisamment petit de manière à ce que $C\epsilon \leq 1$, on déduit de (4.111) que u^{n+1} reste dans la boule de centre u et de rayon ϵ. Cela permet de vérifier par récurrence l'hypothèse que $\|u - u^n\| \leq \epsilon$ pour tout $n \geq 0$, et (4.111) est bien la conclusion désirée. □

Remarque 4.4.13 Bien sûr, il faut conserver à l'esprit que chaque itération de la méthode de Newton (4.108) nécessite la résolution d'un système linéaire, ce qui est

coûteux. De plus, la convergence rapide (dite "quadratique") donnée par (4.110) n'a lieu que si F est de classe C^2, et si u^0 est assez proche de u, hypothèses bien plus restrictives que celles que nous avions utilisées jusqu'à présent. Effectivement, même dans des cas très simples dans $\mathbb{R}$, la méthode de Newton peut diverger pour certaines données initiales u^0 ; il faut noter aussi que la convergence quadratique (4.110) ne se produit qu'au voisinage d'un zéro régulier, comme le montre l'application de la méthode de Newton à la fonction $F(x) = \|x\|^2$ dans $\mathbb{R}^N$, pour laquelle la convergence n'est que géométrique. Par ailleurs, si on applique la méthode de Newton pour la minimisation d'une fonction J comme expliqué ci-dessus, il se peut que la méthode converge vers un maximum ou un col de J, et non pas vers un minimum, car elle ne fait que rechercher les zéros de J'. La méthode de Newton n'est donc pas supérieure en tout point aux algorithmes précédents, mais la propriété de convergence locale quadratique (4.110) la rend cependant particulièrement intéressante. •

Remarque 4.4.14 Un inconvénient majeur de la méthode de Newton est la nécessité de connaître le Hessien $J''(v)$ (ou la matrice dérivée $F'(v)$). Lorsque le problème est de grande taille ou bien si J n'est pas facilement deux fois dérivable, on peut modifier la méthode de Newton pour éviter de calculer cette matrice $J''(v) = F'(v)$. Les méthodes, dites de quasi-Newton, proposent de calculer de façon itérative aussi une approximation S^n de $(F'(u^n))^{-1}$. On remplace alors la formule (4.108) par

$$u^{n+1} = u^n - S^n F(u^n) \quad \text{pour} \quad n \geq 0.$$

En général on calcule S^n par une formule de récurrence du type

$$S^{n+1} = S^n + C^n$$

où C^n est une matrice de rang 1 qui dépend de $u^n, u^{n+1}, F(u^n), F(u^{n+1})$, choisie de manière à ce que $S^n - (F'(u^n))^{-1}$ converge vers 0. Pour plus de détails sur ces méthodes de quasi-Newton nous renvoyons à [5], [8], [15]. •

4.4.5 Méthodes d'approximations successives

Considérons un problème général d'optimisation sous contraintes d'égalité

$$\inf_{F(v)=0} J(v), \tag{4.112}$$

où $J(v)$ et $(F_1(v), ..., F_M(v)) = F(v)$ sont des fonctions régulières de $\mathbb{R}^N$ dans $\mathbb{R}$. Les remarques qui suivent s'appliquent de la même manière aux problèmes avec contraintes d'inégalité, moyennant des modifications évidentes.

Si l'application directe des algorithmes d'optimisation ci-dessus est trop compliquée ou coûteuse pour (4.112), une stratégie courante consiste à remplacer ce dernier par un problème approché obtenu par développement de Taylor des fonctions J et F. L'idée sous-jacente est qu'il est plus facile de résoudre le problème approché que le problème exact. Ces approximations n'ayant qu'un caractère local, il faut itérer cette stratégie en faisant un nouveau développement de Taylor au point de minimum obtenu sur le précédent problème approché.

La première méthode, dite de **programmation linéaire successive** (ou séquentielle), consiste à remplacer les fonctions J et F par des approximations affines (cette méthode est connue aussi sous l'acronyme SLP pour l'anglais "sequential linear programming"). Etant donné une initialisation $v^0 \in V$ (ne vérifiant pas nécessairement la contrainte $F(v) = 0$), on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme les solutions de

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \Big\{ J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) \Big\}, \tag{4.113}$$

qui n'est rien d'autre qu'un programme linéaire, comme étudié dans la Section 4.3.4 et pour lequel on dispose d'algorithmes extrêmement efficaces. Une difficulté immédiate dans la résolution de (4.113) est que sa valeur minimum puisse être $-\infty$ et qu'il n'y ait pas de solution optimale. (Notons que, sous une condition de qualification standard, $(F_1'(v^{n-1}), ..., F_M'(v^{n-1}))$ famille libre de $\mathbb{R}^N$, l'ensemble admissible de (4.113) n'est pas vide.) C'est pourquoi, en pratique, cette méthode s'accompagne d'une contrainte supplémentaire, dite de **région de confiance**, qui prend la forme

$$\|v - v^{n-1}\| \leq \delta, \tag{4.114}$$

où $\delta > 0$ est un paramètre qui définit la taille du voisinage de v^{n-1} dans lequel (4.113) est une bonne approximation de (4.112). La norme dans (4.114) peut-être soit la norme $\|v\|_\infty = \max_{1\leq i\leq N} |v_i|$, soit la norme $\|v\|_1 = \sum_{i=1}^N |v_i|$, ce qui dans les deux cas préserve le fait que le problème approché est un programme linéaire. Ce dernier a alors nécessairement au moins une solution optimale puisque l'ensemble admissible est désormais borné.

Une deuxième méthode, dite de **programmation quadratique séquentielle**, consiste à remplacer la fonction J par une approximation quadratique et F par une approximation affine (cette méthode est connue aussi sous l'acronyme SQP pour l'anglais "sequential quadratic programming"). Etant donné une initialisation $v^0 \in V$ (ne vérifiant pas nécessairement la contrainte $F(v) = 0$), on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme les solutions de

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \Big\{ J(v^{n-1}) + J'(v^{n-1})\cdot(v-v^{n-1}) + \frac{1}{2} Q^{n-1}(v-v^{n-1})\cdot(v-v^{n-1}) \Big\}, \tag{4.115}$$

où Q^{n-1} est une matrice symétrique de taille N. Si Q^{n-1} est définie positive, alors on sait résoudre explicitement le problème (4.115) (voir l'Exercice 4.2.13). Le point crucial dans cette méthode SQP est que Q^{n-1} **n'est pas** la Hessienne de la fonction objectif $J''(v^{n-1})$ mais est la Hessienne du Lagrangien

$$Q^{n-1} = J''(v^{n-1}) + \lambda^{n-1} \cdot F''(v^{n-1}),$$

où λ^{n-1} est le multiplicateur de Lagrange dans la condition d'optimalité pour v^{n-1} (solution à l'itération précédente). En effet, ce qui importe n'est pas l'approximation de J par son développement de Taylor à l'ordre 2 dans tout $\mathbb{R}^N$ mais seulement sur la variété définit par la contrainte $F(v) = 0$. Concrêtement, on écrit

$$J(v) \approx J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) + \frac{1}{2} J''(v^{n-1})(v - v^{n-1}) \cdot (v - v^{n-1}) \tag{4.116}$$

et

$$0 = F(v) \approx F(v^{n-1}) + F'(v^{n-1}) \cdot (v - v^{n-1}) + \frac{1}{2} F''(v^{n-1})(v - v^{n-1}) \cdot (v - v^{n-1}), \quad (4.117)$$

puis on multiplie (4.117) par le multiplicateur de Lagrange λ^{n-1} et on somme le résultat à (4.116), ce qui donne exactement la fonction objectif de (4.115) (à une constante près, en utilisant la contrainte linéaire). L'exercice suivant montre que la matrice Q^{n-1} est positive si v^{n-1} est le point de minimum de (4.112), ce qui entraine que (4.115) admet au moins une solution optimale. Par contre, la matrice $J''(v^{n-1})$ n'a aucune raison d'être positive en général. Néanmoins, si v^{n-1} n'est pas un point de minimum, il peut être nécessaire de recourir à nouveau à une contrainte de région de confiance du type de (4.114). Pour plus de détails nous renvoyons à [15].

Exercice 4.4.8 On considère le problème (4.112) où l'on suppose que les fonctions J et $F_1, ..., F_M$ sont deux fois continûment dérivables et que les vecteurs $\left(F_i'(u)\right)_{1 \le i \le M}$ sont linéairement indépendants pour un minimum local u de J sur $K = \{v \in \mathbb{R}^N,\ F(v) = 0\}$. Soit $\lambda \in \mathbb{R}^M$ le multiplicateur de Lagrange défini par le Théorème 4.2.13. Montrer que la condition nécessaire d'optimalité du second ordre pour u est

$$\left(J''(u) + \sum_{i=1}^{M} \lambda_i F_i''(u) \right) (w, w) \ge 0 \quad \forall\, w \in K(u) = \bigcap_{i=1}^{M} [F_i'(u)]^{\perp}.$$

Bibliographie

[1] ALLAIRE G., *Analyse numérique et optimisation,* Editions de l'Ecole Polytechnique, Palaiseau (2005).

[2] ALLAIRE G., KABER S. M., *Algèbre linéaire numérique. Cours et exercices,* Éditions Ellipses, Paris (2002).

[3] ALOUGES F., *Analyse variationnelle des équations aux dérivées partielles,* Cours de l'Ecole Polytechnique, Palaiseau (2015).

[4] BONNANS J., *Optimisation continue,* Mathématiques appliquées pour le Master / SMAI, Dunod, Paris (2006).

[5] BONNANS J., GILBERT J.-C., LEMARECHAL C., SAGASTIZABAL C., *Optimisation numérique,* Mathématiques et Applications 27, Springer, Paris (1997).

[6] CHVÁTAL V., *Linear programming*, Freeman and Co., New York (1983).

[7] CIARLET P.G., LIONS J.-L., *Handbook of numerical analysis,* North-Holland, Amsterdam (1990).

[8] CULIOLI J.-C., *Introduction à l'optimisation,* 2ème édition, Éditions Ellipses, Paris (2012).

[9] DANAILA I., JOLY P., KABER S. M., POSTEL M., *Introduction au calcul scientifique par la pratique,* Dunod, Paris (2005).

[10] DAUTRAY R., LIONS J.-L., *Analyse mathématique et calcul numérique pour les sciences et les techniques,* Masson, Paris (1988).

[11] EKELAND I., TEMAM R., *Convex analysis and variational problems,* Studies in Mathematics and its Applications, Vol. 1, North-Holland, Amsterdam (1976).

[12] GAUBERT S., *Recherche opérationnelle : aspects mathématiques et applications,* Cours de l'Ecole Polytechnique, Palaiseau (2014).

[13] GOLSE F., *Distribution, analyse de Fourier, équations aux dérivées partielles,* Cours de l'Ecole Polytechnique, Palaiseau (2013).

[14] GOLSE F., LASZLO Y., PACARD F., VITERBO C., *Analyse réelle et complexe,* Cours de l'Ecole Polytechnique, Palaiseau (2014).

[15] NOCEDAL J., WRIGHT S., *Numerical optimization,* Springer Series in Operations Research and Financial Engineering, New York (2006).

[16] SALENÇON J., *Mécanique des milieux continus,* Éditions de l'École Polytechnique, Palaiseau (2002).

Index

algorithme d'Uzawa, 141

base, 132
bien posé, 24

centrée, 14
coercive, 56, 57
condition aux limites de Dirichlet, 4
condition aux limites de Fourier, 4
condition aux limites de Neumann, 4
condition CFL, 17, 21, 33, 39, 76, 78
condition de qualification, 117
condition de stabilité de Von Neumann, 78
condition initiale, 3
conditionnement, 85
conditions aux limites périodiques, 32
consistance, 28
contrainte, 101
contrainte active, 117
contrainte qualifiée, 117
contrôle, 100
convexe, 104
α-convexité, 136
convexité forte, 136
convexité stricte, 104

diffusif, 45
diffusion numérique, 45
différences finies, 13
différentiabilité au sens de Fréchet, 106
différentiabilité au sens de Gateaux, 107
directions admissibles, 112
directions alternées, 40
dispersif, 46
divergence, 3
domaine de dependance, 9
dual, 134

décentré, 14
dérivée seconde, 108

égalité d'énergie, 54
élasticité, 11
élimination de Gauss, 86
énergie complémentaire, 100
équation aux dérivées partielles, 1
équation de la chaleur, 2
équation de la diffusion, 2
équation de Schrödinger, 10
équation des ondes, 8
équation des plaques, 12
équation équivalente, 44
erreur de troncature, 28
explicite, 15

factorisation de Cholesky, 89
factorisation LU, 87
fonction coût ou objectif, 101
fonction propre, 78
fonction test, 54
formulation variationnelle, 53, 71, 72
formule de Green, 52

gradient, 3, 92
gradient conjugué, 140
gradient projeté, 140
gradient à pas fixe, 139
gradient à pas optimal, 138

implicite, 15
infimum, 102
infini à l'infini, 102
instable, 16
interpolation, 65
intégration numérique, 63
inéquation d'Euler, 109

irréversible, 7

Lagrangien, 116, 120, 122
Lamé, 11
Laplacien, 3
lemme de Céa, 59
lemme de Farkas, 119

maillage, 13, 60
maillage uniforme, 60
matrice bande, 90
matrice d'itération, 31
matrice de masse, 74
matrice de rigidité, 58, 59, 62, 74, 82
mesure surfacique, 52
minimum global, 102
minimum local, 102
modélisation, 2, 4, 6
multiplicateurs de Lagrange, 120
méthode de Gauss-Seidel, 92
méthode de Jacobi, 91
méthode de la puissance, 94
méthode de Newton, 147
méthode de relaxation, 92
méthode directe, 83
méthode du gradient, 92
méthode du gradient conjugué, 92
méthode itérative, 83

nombre de Péclet, 5
normale extérieure, 52
norme subordonnée, 83

ordre d'un schéma, 28

pas d'espace, pas de temps, 13
point-selle, 122
polyèdre, 131
primal, 134
principe des travaux virtuels, 54
principe du maximum, 7, 18, 31
problème aux limites, 23
problème bien posé, 23
problème de Cauchy, 23
problème dual, 125, 134
problème primal, 125
programme linéaire, 129
propagation à vitesse finie, 7, 9
précision, 28
pénalisation, 144

quadrature, 63

réversible, 7, 9, 49

schéma linéaire, 31
schéma multiniveau, 27, 36
solution admissible, 131
solution basique, 132
solution optimale, 131
sommets, 61
splitting, 40
stable, 16, 30, 76
stationnaire, 10
stencil, 27
Stokes, 11
suite minimisante, 102
système, 11
système linéaire, 83

théorème de Kuhn et Tucker, 123
transport, 98

variable d'écart, 130
Von Neumann, 34, 38